Single-channel EEG sleep stage classification based on a streamlined set of statistical features in wavelet domain

3 authors:

Thiago Lopes Trugillo da Silveira
Universidade Federal do Rio Grande do Sul
42 PUBLICATIONS   **358** CITATIONS

Alice Kozakevicius
Universidade Federal de Santa Maria
70 PUBLICATIONS   **576** CITATIONS

Cesar Rodrigues
Federal University of Santa Catarina
55 PUBLICATIONS   **399** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project   Methodologies and systems for brain-machine interfaces View project

Project   Spherical image processing View project

CrossMark

ORIGINAL ARTICLE

Single-channel EEG sleep stage classification based on a streamlined set of statistical features in wavelet domain

Thiago L. T. da Silveira[1] · Alice J. Kozakevicius[2] · Cesar R. Rodrigues[3]

Abstract The main objective of this study was to enhance the performance of sleep stage classification using single-channel electroencephalograms (EEGs), which are highly desirable for many emerging technologies, such as telemedicine and home care. The proposed method consists of decomposing EEGs by a discrete wavelet transform and computing the kurtosis, skewness and variance of its coefficients at selected levels. A random forest predictor is trained to classify each epoch into one of the Rechtschaffen and Kales' stages. By performing a comprehensive set of tests on 106,376 epochs available from the Physionet public database, it is demonstrated that the use of these three statistical moments has enhanced performance when compared to their application in the time domain. Furthermore, the chosen set of features has the advantage of exhibiting a stable classification performance for all scoring systems, i.e., from 2- to 6-state sleep stages. The stability of the feature set is confirmed with ReliefF tests which show a performance reduction when any individual feature is removed, suggesting that this group of feature cannot be further reduced. The accuracies and kappa coefficients yield higher than 90 % and 0.8, respectively, for all of the 2- to 6-state sleep stage classification cases.

1 Introduction

Sleep analysis has been classically performed by specialists in sleep medicine via the visual interpretation of electroencephalogram (EEG) signals. Nevertheless, automatic scoring of sleep stages is generally performed in three steps: signal preprocessing, features extraction and machine learning classification [32]. Recently, two comprehensive studies using single-channel EEG [7, 28] compared the classification results of the most used features and classifiers on 5- and/or 6-state sleep staging. Şen et al. [7] and Radha et al. [28] analyzed several nonlinear, time and frequency domain-based features, ranking the features according to their computational cost and their relevance to the classifiers' performances. Both studies, in which several EEG properties in many representation forms were analyzed, can be used as a starting point for investigating or designing EEG classification systems. Nonetheless, statistical parameters like variance, kurtosis and skewness, which are prone to sense variations [2], were intriguingly among the worst time-domain feature choices for EEG classification [7, 19], which instigates further investigation of their behavior in other domains.

The current work contributes to improving comparative analysis by assessing the classification performance of statistical moments computed in the wavelet domain. Because the different scales of the wavelet decomposition are associated with frequency bands, the variance, kurtosis and skewness are obtained for each approximate sleep-related sub-band after applying a 5-level Daubechies discrete

✉ Thiago L. T. da Silveira
thiago@inf.ufsm.br

1 Graduate Program in Informatics, Federal University of Santa Maria, Santa Maria, RS, Brazil

2 Department of Mathematics, Federal University of Santa Maria, Santa Maria, RS, Brazil

3 Department of Electronics and Computing, Federal University of Santa Maria, Santa Maria, RS, Brazil

🙋 Springer

wavelet transform (DWT) to single-channel EEG signals. Both the DWT and the statistical moments are tools for recognizing intrinsic variations within signals [2, 11]. Their association is therefore explored to reinforce subtle variations capable of distinguishing different patterns.

According to the results reported in [7], wavelet-based features are much more relevant to classification compared to their counterparts in the time domain. In [19], statistical moments of EEGs both in the time domain and for the entire frequency spectrum were investigated. Furthermore, the DWT has been shown to be an important tool for analyzing and processing non-stationary signals such as EEG [9, 34] and is preferred in terms of resource consumption compared to the continuous wavelet transform (CWT) [24].

Regarding classifier selection, a random forest is chosen because it presents the best trade-off between accuracy and computation time [7] compared to other classification algorithms, such as support vector machines, feed-forward neural networks and radial basis networks. The superiority of random forest algorithms has also been reported in [10] and [28]. Furthermore, computational cost tests conducted in this work demonstrated that its performance is suitable to real-time applications.

The current study addresses the classification of all five possible sleep stage arrangements according to the Rechtschaffen and Kales' (R&K) recommendations [30]: the 6-state (awake, sleep stages 1, 2, 3 and 4 [S1–S4] and rapid eye movement [REM]), 5-state (awake, S1, S2, slow wave sleep [SWS], REM), 4-state (awake, merge of sleep sates 1 and 2 [S1/S2], SWS and REM stages), 3-state (awake, non-rapid eyes movement [NREM] and REM) and 2-state (awake and sleep stages) sleep stages. One contribution of the proposed methodology is the assumption of the same streamlined set of features for classifying these five different sleep arrangements. Therefore, the method presented here offers an advantage with respect to previously reported methodologies, for which different feature sets were extracted depending on the classification scenario [3, 32, 36].

To assess the proposed methodology, the remaining sections of this study are organized as follows. In Sect. 2, the main steps involved are presented: data description, preprocessing, DWT, feature extraction and the random forest classifier. Section 3 reports the obtained results, and Sect. 4 discusses comparisons of the proposed methodology with other related studies. Finally, conclusions and further remarks are made in Sect. 5.

2 Methods

The proposed method for the classification of sleep stages begins with a preprocessing step that is performed on the entire dataset. Next, feature extraction is performed based on a wavelet multi-resolution analysis. The extracted features feed a random forest algorithm that performs the classification. The following subsections explain in detail each of these steps.

2.1 Data description and preprocessing

The experimental data used in this study were obtained from the Sleep-EDF [Expanded] database [18], which is part of the Physionet Bank [12]. All available recordings[1] of healthy subjects who claimed not to use any sleep-related medication were included in our study. These recordings correspond to a two-night sleep analysis of 10 male and 10 female subjects, 25–34 years of age. The second night recording for one of the volunteers, labeled 13, was not available in the database. The recordings contain two EEG channels (Fpz-Cz and Pz-Oz) sampled at 100 Hz in addition to electrooculogram (EOG) and electromyogram (EMG) signals.

This work utilizes a single EEG channel to perform the sleep stage classifications, as done in [3, 32, 36]. The Pz-Oz channel is selected for being more accurate than the Fpz-Cz channel for sleep stage identification [36]. Other polysomnographic (PSG) signals are discarded in the current study. Hypnogram annotations written by sleep experts for every 30 s are jointly provided in the recordings. The annotations are accepted as a correct reference and are usually considered for evaluating the classification methods, including the proposed method. Nevertheless, these annotations may introduce intrinsic errors, as the inter-expert agreement is not always 100 % independent of the chosen dataset [5].

Epochs from Physionet Bank classified by the sleep experts as movement time or not scored are removed from the analysis [36], as their patterns were already classified as not being related to the events of interest. This removal is done only for fairly training and testing the proposed method. After this preliminary selection, a total of 106,376 remaining valid epochs are utilized for training and evaluation through cross-validation technique, as in [32, 36]. This dataset characterizes a real scenario, of which approximately 68, 2.6, 16.7, 3.2, 2.2 and 7.2 % were scored as awake, S1, S2, S3, S4 and REM, respectively.

2.2 Discrete wavelet transform

Wavelet transforms (WTs) allow for the discrimination of non-stationary signals with different frequency properties [8], such as the EEGs. Furthermore, in contrast to Fourier transforms, WTs provide a multi-resolution analysis in

---

[1] Available data in [27] until April 19, 2016.

both time and scaling domains; nevertheless, frequency and scale are related [8]. Various WTs have already been used in sleep analysis, such as DWT in [34], wavelet packet in [9, 33] and CWT in [10].

In the current study, the normalized DWT of the Daubechies family with two vanishing moments and four filters (Db2) [8] is employed to analyze and to extract relevant features from the signals. The Db2 is adopted due to its efficacy in capturing data variations with only two null moments, as reported in [34]. Furthermore, increasing the number of null moments is avoided, which requires the extrapolation of a greater number of points to perform the wavelet transform on the boundaries [20].

In the Daubechies classical approach for the discrete wavelet algorithm [8], the length of the input vector C_0 must be L, with $L = 2^J$ and $J \in \mathbb{N}$. Considering that the length of C_0 is $L = 2^K \times R$, with R and $K \in \mathbb{N}$ and R being odd, it is possible to apply the direct transform K times on C_0. The final set after K decompositions contains R points. Because the epochs have a length of 3000 samples, each epoch can be decomposed just three times via this heuristic. Therefore, to increase the number of decomposition levels, another slight adaptation is made. Instead of considering each epoch with 3000 points, each one is increased by eight points. As EEG signals are subdivided only for the sake of convention, the eight missing points in each epoch are exactly obtained from the next one, avoiding any type of extrapolation. The last epoch of each EEG signal is discarded because, by the annotations, all subjects were awake at this moment. Thus, the proposed adjustment easily solves the issue of obtaining extra points on the boundary in all levels of the wavelet transform.

In fact, there is no shift in the starting point of the next epoch, but rather an overlap of eight points that are analyzed twice by consecutive epochs. Thus, the new extended set containing 3008 samples per epoch allows C_0 to be decomposed six times. However, just five wavelet decompositions are identified as necessary (see Sect. 2.3) and are in fact considered. Figure 1 presents the adjusted algorithm, which assumes input data with an even length.

Figure 2 illustrates the application of the five-level DWT algorithm with the boundary treatment over an epoch of the subject labeled 11, in accordance with the Physionet Bank notation. The upper plot shows the original signal, followed by the wavelet coefficients for each wavelet decomposition level.

2.3 Feature extraction

The sleep stage classification is classically performed by identifying the characteristics extracted from cerebral rhythms. The delta rhythm (1–4 Hz) arises intensely during the S3 and S4 stages, containing 20–50 % and more

```
1:  procedure DDWT(C_j, h, g, E, D)
2:      for l ← 0 to l < length(C_j)/2 do
3:          C_{j+1,l} ← 0
4:          D_{j+1,l} ← 0
5:          for k ← 0 to k < D do
6:              aux ← 0
7:              i ← (2 × l + k)
8:              if i < length(C_j) then
9:                  aux ← C_{j,i}
10:             else                       ▷ boundary treatment
11:                 m ← i mod length(C_j)
12:                 aux ← E_m
13:             end if
14:             C_{j-1,l} ← C_{j-1,l} + h_k × aux
15:             D_{j-1,l} ← D_{j-1,l} + g_k × aux
16:         end for
17:     end for
18:     return(C_{j+1}, D_{j+1})
19: end procedure
```

Fig. 1 Algorithm for computing the direct Daubechies DWT for input vectors with even lengths. The procedure parameters C_j, h, g, E and D are, respectively, the scaling coefficients at the jth level, the scaling filter coefficients, the wavelet filter coefficients, the extrapolated samples at the jth level and twice the number of vanishing moments ($D = 4$). The procedure returns the C_{j+1} and D_{j+1} which are, respectively, the scaling and wavelet coefficient sets at coarser level $j + 1$
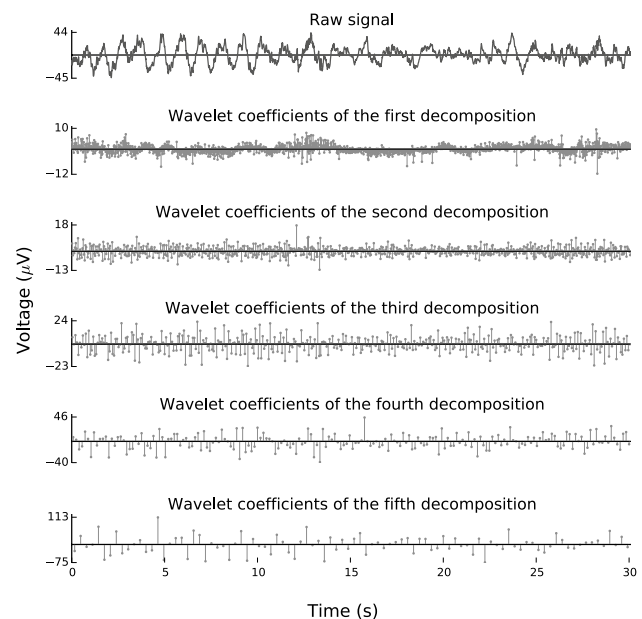

Fig. 2 Five-level wavelet decomposition of the 1800th epoch from the second night recording of the subject labeled as 11. At this moment, the subject's sleep is scored as S4 state according to the Physionet. The raw EEG signal is shown in the first plot, while the details in each wavelet decomposition are presented in the remaining ones

than 50 %, respectively, of the total energy of the spectrum [32]. At the transition from wakefulness to a resting condition, the theta rhythm (4–8 Hz) gradually increases, while the alpha rhythm (8–13 Hz) decreases [34]. According to Corsi-Cabrera et al. [6], the beta rhythm (13–30 Hz) is

Table 1 Frequency ranges in five levels of decomposition for Db2 with sampling frequency at 100 Hz

Coefficient set	Frequency range (Hz)	Associated brain rhythm
D_1	25–49	Low-gamma
D_2	12.5–25	Beta
D_3	6.25–12.5	Alpha
D_4	3.125–6.25	Theta
D_5	1.5625–3.125	Delta
C_5	0–1.5625	Delta

Fig. 3 Normalized variance, kurtosis and skewness of the wavelet coefficients in each decomposition level (D_1–D_5 and C_5, respectively, from top to bottom) of several epochs from the second night recording of subject 02. *White*, *light gray* and *dark gray* backgrounds correspond to the awake, S1 and S2 stages, respectively, according to the Physionet Bank annotations

primarily associated with the awake stage. The low-gamma rhythm (30–50 Hz) is related to the working memory and attention [17] and, hence, arises primarily during the awake stage. The high-gamma waves (higher than 50 Hz) are not associated with the sleep, but with motor tasks [29]. For more information on sleep and brain rhythms, see [21].

To distinguish all of the mentioned sleep-related cerebral rhythms, relevant features are extracted from the wavelet and scaling DWT coefficients for each epoch. Feature extraction drastically reduces the data dimensionality, while keeping only the relevant properties. For the considered sampling rate, obeying the Nyquist theorem [8], five levels of the DWT are necessary to accomplish the brain rhythm separation from the wavelet coefficients (D_1–D_5) (in [1.5625,49] Hz), each corresponding to one level. The set of scale coefficients (C_5) (in [0,1.5625] Hz) is related to the coarsest level, as shown in Table 1.

The proposed feature set is composed as follows: The second (variance), third (skewness) and fourth (kurtosis) moments of the statistical analysis are computed for all selected DWT coefficient sets (D_1–D_5 and C_5). In previous experiments, it was also considered the first statistical moment, i.e., the mean. However, the removal of this quantity from the feature set increased the classification rates, pointing out the mean was not sensitive enough to the sleep stages. Taking the selected DWT coefficient sets as *X*, the variance, skewness and kurtosis can be computed as

$$\mathrm{Var}(X) = \frac{\sum_{i=1}^{T}\left(X_i - \overline{X}\right)^2}{T}, \quad \mathrm{Skew}(X) = \frac{1}{T}\frac{\sum_{i=1}^{T}\left(X_i - \overline{X}\right)^3}{\mathrm{Var}(X)^{3/2}},$$

and $$\mathrm{Kurt}(X) = \frac{1}{T}\frac{\sum_{i=1}^{T}\left(X_i - \overline{X}\right)^4}{\mathrm{Var}(X)^2} - 3,$$

where $\overline{X}$ is the mean of *X* and *T* is its size.

Generally, the variance expresses the data variation around the mean value. The skewness shows the extent of signal distribution asymmetry. For signals with symmetric probability density function such as Normal, Laplace and Cauchy, this quantity is equal to zero. A large value of skewness indicates a sharp transition from low to high or from high to low value in the excerpted segment.
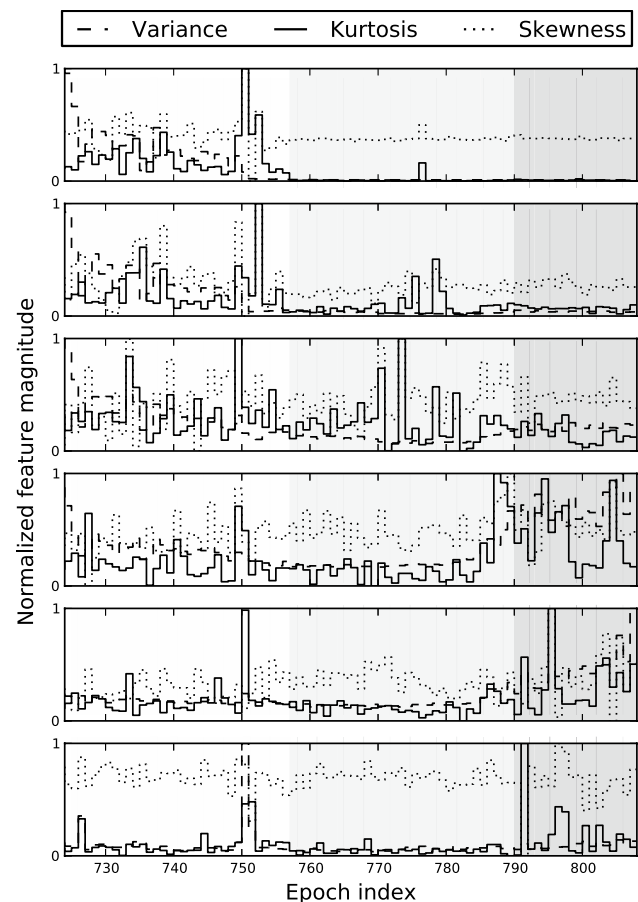
Consequently, this measure detects the probable edges of the signal in the analysis window. On the other hand, the kurtosis indicates the extent of flatness (smoothness or impulsive peaks) of samples in the analysis window. This measure allows to detect sharp ascending or descending regimes occurred in the excerpted segment [11].

Therefore, the already level-by-level relevant variations initially exposed by the DWT through the behavior of its wavelet coefficients are further noticed and captured by the chosen features. Figure 3 exemplifies the features' sensitivities at the sleep stage transitions (S1 to S2 and awake to S1) for each decomposition level of the second night recording from subject 11. This sensitivity, shown with normalized values in Fig. 3, is similar for all of the analyzed recordings of the database.

To ensure the relevance of all the selected features, the ReliefF filter [7, 13] is applied to rank them by their contribution to the classes' identification. Even when removing

the worst ranked quantity from the feature set, a reduction in the overall classification performances is observed, ranging from 0.01 to 0.1 %, depending on the number of classes, i.e., 2- to 6-state sleep stages. On the other hand, the removal of the best ranked feature causes a diminishing from 0.1 to 0.4 % in the overall performance. This actually indicates that the proposed feature set cannot be reduced in any of the five analyzed scenarios, and moreover, all features *together* are necessary to allow the good results achieved in this study.

The use of statistical measures in sleep scoring was already emphasized by Ronzhina et al. [32]. According to them, the kurtosis, skewness, mean, median and standard deviation have already been applied for analyzing physiological signals in the time domain. Furthermore, Koley and Dey [19] considered the first four statistical moments in both time and frequency domains as part of a 39-feature set for classifying sleep stages, in contrast to the streamlined 18-feature set adopted here. In their analysis, the computation of the features was performed for the complete frequency spectrum. In the current study, the statistical measures computation is performed only for most relevant sleep-related frequency ranges (see Table 1). A preliminary test, assuming the considered dataset and the selected classification algorithm (see next section for more detailed explanation about the classifier), showed that considering only the variance, kurtosis and skewness both in the time domain and in the whole frequency domain, it is possible to achieve accuracies just around 65 % for the 6-state problem.

According to Zhu et al. [36], the good classification performance is highly dependent on an efficient feature selection. And for accomplishing this performance in Zhu et al. [36], specific sets of features had been chosen according to the sleep stage arrangement. An important improvement of the present study is that now only one set of features is considered for all classification scenarios, providing more robustness for the proposed methodology. Aiming to perform the construction and evaluation of the classifier model, as suggested in [10], all $106,376 \times 18$ extracted features are firstly grouped into a single set.

2.4 Random forest classifier

The current study uses WEKA [13] as a classification tool, similarly to [10, 28]. WEKA (Waikato Environment for Knowledge Analysis) is an open-source software developed at University of Waikato, New Zealand, originally designed for performing several standard data mining tasks, as detailed in [35]. Here, all available classifiers were evaluated to explore the potential of WEKA in the context of signal analysis. The tested classifiers include multilayer perceptron, Naïve Bayes network, AdaBoost M1 and 1-nearest

neighbors [35], which achieved 85.3, 78.2, 75.2 and 73.3 % accuracies, respectively, for the 6-state sleep stages. These classifiers' performances presented a large range of variation according to the tested scenarios (number of sleep stage stages). However, the random forest algorithm performed better in all tested cases, confirming its superiority, as pointed out by other studies [7, 10, 28].

The random forest—a fast, robust to noise and overfitting free [14, 31] ensemble method introduced by Breiman [4]—combines the classification outputs of N tree predictors, each one grown with F features. The parameter F obeys $F = \lfloor \log_2(M) + 1 \rfloor$, in which M is the total number of features of the dataset. In this study, $M = 18$, and thus, each tree predictor uses $F = 5$ features. The F features that are used to grow each of the N tree predictors are chosen randomly [4], and their precedence in the trees' bodies is determined according to the information gain internal estimator [10].

In Oshiro et al. [26], dataset density measures were proposed to identify how many trees a random forest must have to provide accurate results. Those authors performed experiments showing that using 64, 128, 256, 512 or 1024 trees resulted in nearly steady accuracies for low-density datasets. According to their metrics, the dataset used in the current study can be classified as a low-density dataset, and therefore, $N = 64$ was chosen here. Breiman [4] endorsed the conclusions stated above, demonstrating that incrementing the number of trees increases the computational cost, while the classifier's accuracy converges to a limit.

The random forest construction is performed as follows. Bootstrapping (sampling with replacement) technique is applied over the entire training set, generating $N = 64$ called in-bag sets that have the same length as the original training set. Each tree predictor takes into account its in-bag set for training. Afterward, F features are randomly selected to grow each tree predictor. Breiman [4] claimed that, considering the generated in-bag sets and the random features, it is possible to enhance the classification accuracy. According to Fraiwan et al. [10], many tree-building methods can be used in a random forest, such as random trees, classification and regression trees (CART), reduced error trees (REPTree), C4.5 and CHAID. WEKA uses random trees as individual predictors for a random forest classifier. Each tree predictor grows as much as possible, and no pruning technique is applied. The random forest classifier's output is the mode of the N trees' votes [10]. Figure 4, adapted from [10], illustrates the random forest construction procedure.

The proposed method's performance is estimated via a tenfold cross-validation. The k-fold cross-validation approach yields a robust evaluation of a classification method, and $k = 10$, according to [1], is nearly the right number of folds to get the best error estimate. The dataset
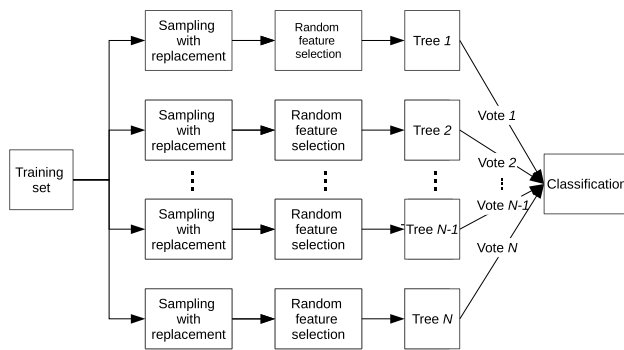
Fig. 4 Random forest structure illustrating each step of classification procedure. Firstly, bootstrapping sampling is applied to generate the in-bag sets for all tree predictors. Random feature selection is performed and the decision trees are grown according to them. The output classification is the mode of all trees' votes

is divided into k approximately equal size subsets. Of these, $k-1$ subsets are used to train the classifier, and the remaining subset is used in the test step. This means that training set is different from the testing one. This process is repeated k times. Each subset is tested only once. The final evaluation measure is given by the mean of the k executions of this process [32]. Method performance assessments by considering new subjects and splitting the complete dataset into random testing and training sets are given in the Appendix.

3 Results

Performance indicators such as precision, recall, overall accuracy and the Cohen's kappa coefficient [35] are considered here. These indicators can be computed through the analysis of the confusion matrix **M** [35]. It is a square matrix of order Q that relates the classifier output to a known correct score (sleep experts score), presenting an overview of the number of hits and misses for each class (sleep stage). The precision and recall of the i-th class,

$i = 1, 2, \ldots, Q$, as well as the overall accuracy and kappa coefficient are computed, respectively, through

$$P_i = \frac{\mathbf{M}_{ii}}{\sum_{j=1}^{Q} \mathbf{M}_{ji}}, \quad R_i = \frac{\mathbf{M}_{ii}}{\sum_{j=1}^{Q} \mathbf{M}_{ij}},$$

$$\text{Acc} = \frac{\sum_{i=1}^{Q} \mathbf{M}_{ii}}{\sum_{i=1}^{Q} \sum_{j=1}^{Q} \mathbf{M}_{ij}} \quad \text{and} \quad \kappa = \frac{\text{Acc} - \text{EA}}{1 - \text{EA}},$$

where $\quad \text{EA} = \dfrac{\sum_{i=1}^{Q} \left(\sum_{j=1}^{Q} \mathbf{M}_{ij} \sum_{j=1}^{Q} \mathbf{M}_{ji} \right)}{\left(\sum_{i=1}^{Q} \sum_{j=1}^{Q} \mathbf{M}_{ij} \right)^2},$

is the expected agreement [23] and $\mathbf{M}_{ij}$ is the element at the ith row and jth column of the confusion matrix.

Table 2 presents the confusion matrix for the 6-state sleep stages classification. The quantities of correctly classified epochs for each class, presented in bold, are disposed on the main diagonal of the confusion matrix. Of note, the S1 stage is completely discriminated from the S3 and S4 stages using the proposed method. As shown in [36] and already reported in [6], the truth S1 instances are prone to be mistakenly scored as awake, S2 or REM stages. In addition, the artificial neural networks (ANN) architectures presented in [32] did not achieve good results for the S1 and S3 classification. In terms of recall, the awake stage identification with the proposed method exhibited the best performance, yielding 99.3 %, and S1 the worst.

Table 3 summarizes the precision and recall measures computed for the ns-state sleep stage classification, with $ns \in \{2, 3, 4, 5, 6\}$. Notably, the awake stage has high and nearly constant values for both metrics that are independent of the ns value. Similar behavior is observed for the REM stage when $ns \neq 2$. In addition, the SWS stage is clearly distinguishable from the other stages when $ns = 4$ and $ns = 5$. The proposed method exhibits 3.6 and 9.1 % increased recall and precision for the awake compared to the results obtained in [3] for the 5-state sleep stages. Furthermore, the recall measure for awake and precision for S1

Table 2 Confusion matrix for 6-state sleep stages

	Proposed method's scoring					
	Awake	S1	S2	S3	S4	REM
Expert's scoring						
Awake	**71836**	40	239	4	2	232
S1	1176	**164**	746	0	0	718
S2	690	20	**15605**	492	37	955
S3	74	0	1102	**1734**	457	3
S4	35	0	94	618	**1586**	0
REM	605	47	1755	2	1	**5307**
P (%)	96.5	60.5	79.9	60.8	76.1	73.6
R (%)	99.3	5.8	87.7	51.5	68.0	68.8

Table 3 Precision (in italic) and recall for *ns*-state sleep stages classification

Stage	*ns* = 2 (%)	*ns* = 3 (%)	*ns* = 4 (%)	*ns* = 5 (%)	*ns* = 6 (%)
Awake	*97.4*	*96.9*	*96.9*	*96.6*	*96.5*
	98.7	99.1	99.1	99.3	99.3
S1	*97.2*	*89.4*	*82.0*	*59.9*	*60.5*
	94.3	88.5	82.4	6.1	5.8
S2				*80.3*	*79.9*
				87.0	87.7
S3			*87.1*	*87.4*	*60.8*
			79.8	79.8	51.5
S4					*76.1*
					68.0
REM		*77.1*	*77.0*	*73.6*	*73.6*
		63.5	68.3	69.1	68.8

Table 4 Accuracy and kappa coefficient comparisons

ns value	Method in [3]	Method in [32]	Method in [36]	Proposed method
ns = 2	95.4 %/0.83	96.9 %/–	97.9 %/0.96	97.3 %/0.94
ns = 3	88.3 %/0.76	90.3 %/–	92.6 %/0.87	93.9 %/0.87
ns = 4	74.5 %/0.63	81.4 %/–	89.3 %/0.83	92.3 %/0.84
ns = 5	71.2 %/0.61	–/–	88.9 %/0.83	91.5 %/0.83
ns = 6	–/–	76.7 %/–	87.5 %/0.81	90.5 %/0.80

is 9.7 and 6.9 % higher, respectively, than those obtained in [10].

It was not possible to establish further comparisons with the considered methods in terms of class-specific metrics, as they are not clearly included in their reports. The results presented in Table 3 highlight the ability of the proposed methodology to distinguish different sleep stages for all values of *ns*. The Cohen's kappa coefficient is also used to assess the proposed method. The 6- to 2-stage sleep states achieved 0.80, 0.83, 0.84, 0.87 and 0.94 kappa values, respectively. According to Landis and Koch [23], classifiers for which the kappa value is greater than or equal to 0.80 are considered *excellent*.

In terms of computational resources consumption, according to the numerical experiments performed here, the random forest model, once built, spends on average 0.6 or 1.3 ms to classify an epoch having $M = 1$ (practical minimum) or $M = 18$ features, respectively. As the acquisition time of each epoch is 30 s, the method is sufficiently fast for real-time EEG processing. Additionally, the time spent to build the classifier model for $M = 18$, using a consumer-grade computer (Core i7, 2.0 GHz, 8 Gb RAM), is approximately 3 min for the 106,376 epochs.

4 Discussion

The performance results obtained in the current study are compared to those found in the state-of-art literature, in which the classification of sleep stages with 2, 3, 4, 5 and 6 states is addressed. Table 4 compares the overall accuracies and kappa coefficients of the proposed method and three other previously reported methods [3, 32, 36] that also used recordings from the Physionet Bank, the same EEG channel and the same guideline (R&K) for classifying the 2 to 6 states of sleep stages. Since the methodologies are not completely comparable to the one presented here, we clearly state the most relevant aspects considered in each study.

Zhu et al. [36] employed a fifty percent random splitting for the training and testing sets used to build and assess their method. They reported to have processed 14,963 epochs. Ronzhina et al. [32] presented several ANN architectures, and the ANN that they found as performing best is used in our comparison here. The authors employed recordings from eight subjects from the Physionet Bank, although they did not state how many epochs were used in their computations. Their results are reported after a ten-fold cross-validation assessment. Cohen's kappa coefficient and the 5-state sleep stage classification were not assessed in their study. Berthomier et al. [3] performed an epoch-by-epoch comparison between the output of their method and the score given by sleep experts. They used approximately 8,500 epochs. Classification of the 6 states was not assessed in their study.

Regarding the kappa coefficient, for *ns* = 4, the proposed methodology exhibited the best performance compared to [3] and [36], as shown in Table 4. The proposed method's kappa values for *ns* = 3 and *ns* = 5 are equal to those presented in [36] and greater than those published in [3]. For *ns* = 2 and *ns* = 6, the proposed method performed slightly poorer than the method reported in [36]. Regarding overall accuracy, the proposed method exhibited the best performance for the 3- to 6-state sleep stage classifications compared to the three previous studies. The overall accuracy of the proposed method for *ns* = 2 is 0.4 and 1.9 % greater than the accuracies obtained in [32] and [3], respectively. For the group with six classes (*ns* = 6), the proposed method's accuracy is 3 and 13.8 % greater than those reported in [36] and [32], respectively.

Although Fraiwan et al. [10] also used a random forest to classify sleep stages, a different database and EEG channel were employed, along with another guideline for sleep classification: the one developed by the American Academy of Sleep Medicine (AASM) [16]. Nevertheless, according to Moser et al. [25], the 5-state sleep stages following both the AASM and R&K standards are equivalent.

Table 5 Methodology comparison with Fraiwan et al. [10]

Resource	Method in [10]	Proposed method
Standard for classification	AASM	R&K
Database	thoracic clinic at the University of Heidelberg	Sleep-EDF database
EEG channel	C3A1 at 100 Hz	Pz-Oz at 100 Hz
Feature extractor	CWT	DWT
Feature(s)	Renyi's entropy	Variance, kurtosis and skewness
Classifier	Random forest	Random forest
Number of epochs	20,269	106,376
5-state sleep stages accuracy	83 %	91.5 %

Another difference between the current work and [10] is the adoption of the DWT, which contrasts with their choice of a CWT. In general, the DWT spends less computational resources than the CWT [24]. In addition, the DWT does not produce redundant coefficients [8], which is of interest for increasing the classifier performance. Fraiwan et al. [10] considered three feature extractor algorithms. However, the current study analyzes all of the 2- to 6-state sleep stages using the same algorithm, representing a step toward online classification procedures. Table 5 summarizes the similarities and differences between the current work and [10].

As previously mentioned by related studies [32, 36], methods for classifying sleep stages that are based on single-channel EEG signals can produce very competitive results compared to methods based on several PSG channels. Our results corroborate this trend as the proposed methodology yields similar or even better overall performances when compared to recently published studies that utilize more than one EEG channel [15] or combinations of EEG, EOG and EMG channels [22] as input signals.

5 Conclusion

This study presents a robust method for sleep stage classification according to the R&K standard using a single EEG channel. The methodology was broadly validated using a dataset composed of more than 100,000 EEG epochs of 30 s. The open-access signals are available on the Sleep-EDF [Expanded] database. The primary contributions are the use of the discrete wavelet transform as a feature extractor in the frequency ranges related to the cerebral rhythms and the adoption of a single set of 18 features to classify all of the 2- to 6-state sleep stages, which is formed by variance, skewness and kurtosis computed per each one of the 5 decomposition levels of the DWT, including the coarsest scaling coefficient block.

The use of these features in wavelet domain increases the classifier's performance nearly 25 % when compared to their use in time and whole frequency domains. The efficacy of the feature set was tested with the ReliefF filter ranker, which pointed out no possible subset selections without degrading the classification performance. Individually removing each of the features caused accuracy decrements from 0.01 to 0.4 %, demonstrating that the good performance achieved here is due to the union of all features.

Precision, recall, accuracy and Cohen's kappa coefficient are assumed as figures of merit of the proposed method, enabling a wide range of comparisons with well-established and state-of-art methodologies. Quantitative comparisons with state-of-art studies show that the current method yields comparable or higher accuracies when classifying any of the sleep stages. The best results are observed for 3- to 6-state sleep stages, with overall accuracies equal to 93.9, 92.3, 91.5 and 90.5 %, respectively. Kappa values are maintained higher than 0.8 for all cases.

Appendix

To reaffirm the results obtained via the proposed methodology, three additional random forest models were tested. In the first model, a random split of the complete feature dataset was considered. Approximately two-thirds of data were used to construct the model, and the remaining one-third

was used for testing. The accuracies reached 90.2, 90.8, 92.0, 93.7 and 97.3 % for the 6- to 2-state sleep stages, respectively. The difference in the accuracy of this model compared to the model reported by Fraiwan et al. [10], who used the same data distribution for training and testing, is 7.8 %.

In the second model, a fifty percent random splitting of data for training and testing sets was considered. The accuracies achieved for the 6- to 2-state sleep stages were 90.1, 91.2, 92.0, 93.7 and 97.2 %, respectively. A direct comparison with Zhu et al. [36], considering their model for training and testing, indicated that the proposed method's accuracies are 2.6, 2.3, 2.7 and 1.1 % higher than those obtained in the previous study for the 6- to 3-state sleep stages.

In the third model, 35 EEG signals from the 39 available were considered. The remaining 4 EEG signals (recordings from the second night of subjects 00, 01, 02 and 03) were individually used to test the model. The average accuracies reached 88.8, 90.0, 91.7, 93.2 and 97.7 % for the 6- to 2-state sleep stages, respectively. These three models were built with 64 random trees. Among the compared studies, the best results in terms of accuracies for the 3- to 6-state classifications, independent of the chosen approach for assessment, were achieved using the proposed methodology.

References

1. Alpaydin E (2010) Introduction to machine learning, 2nd edn. The MIT Press, Cambridge
2. Asadzadeh M, Hashemi E, Kozakevicius A (2013) Efficiency of combined Daubechies and statistical parameters applied in mammography. Appl Comput Math 12(3):289–305
3. Berthomier C, Drouot X, Herman-Stoïca M et al (2007) Automatic analysis of single-channel sleep EEG: validation in healthy individuals. Sleep 30:1587–1595
4. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
5. Chapotot F, Becq G (2010) Automated sleep–wake staging combining robust feature extraction, artificial neural network classification, and flexible decision rules. Int J Adapt Control 24(5):409–423
6. Corsi-Cabrera M, Muoz-Torres Z, del Ro-Portilla Y, Guevara MA (2006) Power and coherent oscillations distinguish REM sleep, stage 1 and wakefulness. Int J Psychophysiol 60(1):59–66
7. Şen B, Peker M, Çavuşoğlu A, Çelebi FV (2014) A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. J Med Syst 38(3):1–21
8. Daubechies I (1992) Ten lectures on wavelets. Society for Industrial and Applied Mathematics, Philadelphia
9. Ebrahimi F, Mikaeili M, Estrada E, Nazaren H (2008) Automated sleep stage classification based on EEG signals by using neural networks and wavelet packet. In: 30th annual international IEEE EMBS conference, pp 1151–1154
10. Fraiwan L, Lweesy K, Khasawneh N et al (2012) Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier. Comput Methods Programs Biomed 108(1):10–19
11. Ghaffari A, Homaeinezhad MR, Khazraee M, Daevaeiha MM (2010) Segmentation of holter ECG waves via analysis of a discrete wavelet-derived multiple skewness–kurtosis based metric. Ann Biomed Eng 38(4):1497–1510
12. Goldberger AL, Amaral LAN, Glass L et al (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation 101(23):e215–e220
13. Hall M, Frank E, Holmes G et al (2009) The WEKA data mining software: an update. SIGKDD Explor 11(1):10–18
14. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York
15. Huang CS, Lin CL, Ko LW et al (2014) Knowledge-based identification of sleep stages based on two forehead electroencephalogram channels. Front Neurosci 8:263
16. Iber C, Ancoli-Israel S Jr, ALC, Quan SF (2007) The AASM manual for the scoring of sleep and associated events: rules. Terminology and Technical Specifications, American Academy of Sleep Medicine
17. Jia X, Kohn A (2011) Gamma rhythms in the brain. PLoS Biol 9(4):1–4
18. Kemp B, Zwinderman AH, Tuk B et al (2000) Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. IEEE Trans BioMed Eng 47:1185–1194
19. Koley B, Dey D (2012) An ensemble system for automatic sleep stage classification using single channel EEG signal. Comput Biol Med 42(12):1186–1195
20. Kozakevicius A, Schmidt AA (2013) Wavelet transform with special boundary treatment for 1D data. Comput Appl Math 32(3):447–457
21. Kryger MH (2009) Atlas of clinical sleep medicine: expert consult, 2nd edn. Saunders, Philadelphia
22. Lajnef T, Chaibi S, Ruby P et al (2015) Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods 250:94–105
23. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
24. Mallat S (2008) A wavelet tour of signal processing: the sparse way, 3rd edn. Academic Press, Cambridge
25. Moser D, Anderer P, Gruber G et al (2009) Sleep classification according to AASM and Rechtschaffen and Kales: effects on sleep scoring parameters. Sleep 32(2):139–149
26. Oshiro TM, Perez PS, Baranauskas JA (2012) How many trees in a random forest? In: Perner P (ed) Machine learning and data mining in pattern recognition, vol 7376. Lecture notes in computer science. Springer, Berlin Heidelberg, pp 154–168
27. Physionet (2013) The sleep-EDF-X database. http://www.physionet.org/physiobank/database/sleep-edfx. Accessed 12 May 2015
28. Radha M, Garcia-Molina G, Poel M, Tononi G (2014) Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal. In: Engineering in Medicine and Biology Society, 2014 36th Annual International Conference of the IEEE
29. Rao RPN (2013) Brain computer interface: an introduction, 1st edn. Cambridge University Press, Cambridge
30. Rechtschaffen A, Kales A (1969) A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Electroencephalogr Clin Neurophysiol 26(6):644
31. Robnik-ikonja M (2004) Improving random forests. In: Boulicaut JF, Esposito F, Giannotti F, Pedreschi D (eds) Machine

learning: ECML 2004, vol 3201. Lecture notes in computer science. Springer, Berlin Heidelberg, pp 359–370

32. Ronzhina M, Janousek O, Kolarova J et al (2012) Sleep scoring using artificial neural networks. Sleep Med Rev 16:251–263

33. Silveira TLT, Kozakevicius AJ, Rodrigues CR (2016) Automated drowsiness detection through wavelet packet analysis of a single EEG channel. Expert Syst Appl 55:559–565

34. Subasi A (2005) Automatic recognition of alertness level from EEG by using neural network and wavelet coefficients. Expert Syst Appl 28(4):701–711

35. Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques, 3rd edn. Morgan Kaufmann/Elsevier, Burlington

36. Zhu G, Li Y, Wen P (2014) Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal. IEEE J Biomed Health Inform 18(6):1813–1821

Alice J. Kozakevicius holds a PhD in Applied Mathematics (2002) and is a member of the Graduate Program in Informatics (PPGI) from UFSM. Her interests include wavelet methods for solving PDEs and signal processing.

Thiago L. T. da Silveira is a MD student in Informatics at the Universidade Federal de Santa Maria (UFSM), Brazil. His current research interests include pattern recognition and digital signal processing.

Cesar R. Rodrigues holds a PhD in Electrical Engineering (2002) and is with the Department of Electronics and Computing, UFSM. His interests include the design of integrated circuits and their applications in biomedical engineering.