The addition of entropy-based regularity parameters improves sleep stage classification based on heart rate variability

M. Aktaruzzaman · M. Migliorini · M. Tenhunen · S. L. Himanen · A. M. Bianchi · R. Sassi

Received: 24 July 2014 / Accepted: 2 February 2015 / Published online: 18 February 2015

1 Introduction

Sleep has a significant effect on many physiological functions and plays a fundamental role in the genesis and insurgence of different pathologies (cardiologic, neurological, and metabolic). Its quality is one of the aspects that mostly influence our everyday life. It has a strong impact on natural processes like memorization, learning, and concentration [29]. Poor sleep quality or too short sleep time have been identified among the main causes of car or work accidents [24]. Also, sleep disturbances (i.e., the ones related to breathing) have a strong association with cardiovascular pathologies. A bad quality of sleep has an impact on blood pressure, decreases the immunity defenses, and may increase the insurgence probability of metabolic disturbances such as obesity and diabetes [7, 9, 17, 34].

Sleep quality is generally evaluated through polysom-nography (PSG), which consists of many physiological signals recorded during one or more nights of sleep: elec-troencephalogram (EEG), electromyogram (EMG), and electro-oculogram (EOG), besides respiration activity and electrocardiogram (ECG). Rules and guidelines provided by the American Academy of Sleep Medicine (AASM) allow the evaluation of wakefulness, sleep macrostructures, built through the alternation of different sleep stages as rapid eye movement (or REM), non-REM light sleep (stage 1 and 2), non-REM (or NREM) deep sleep (stage 3, also called slow-wave sleep, SWS), and microstructures, such as the cyclic alternating pattern (CAP sleep), K-complexes, microarousals [13, 16, 30].

M. Aktaruzzaman · R. Sassi
Dipartimento di Informatica, Università degli Studi di Milano, Crema, Italy

M. Migliorini (✉) · A. M. Bianchi
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Golgi 39, 20133 Milano, Italy
e-mail: matteo.migliorini@polimi.it

M. Tenhunen · S. L. Himanen
Department of Clinical Neurophysiology, Pirkanmaa Hospital District, Tampere, Finland

The standard practice is to perform sleep evaluation by the visual or semi-automatic scoring of polysomnographic traces [3]. This technique requires specific instrumentation and signals which are recorded and scored by trained personnel. In addition, their acquisition may be so comfortless to affect the sleep quality itself. On the other hand, many different studies have demonstrated that sleep strongly affects the peripheral system, particularly the autonomic nervous system, so that the heart rate variability (HRV) sig-nal presents different patterns during different sleep stages [14, 20, 23, 27] and during sleep phasic events [12, 28]. For these reasons, many recent studies have focused on the effects of sleep stage transitions on peripheral systems. Most of the works found in the literature have given empha-sis on the correspondence of different pattern of heart rate with different sleep stages [35], and more recently, a few works [8, 33] described methods to perform sleep stag-ing through HRV analysis. One of the advantages of using HRV for sleep evaluation is the possibility of employing less intrusive devices, such as a sensorized T-shirt or mat-tress [4, 15, 21].

Different features have shown to be promising in terms of differentiation between different sleep stages, using HRV analysis, e.g., features in both time and frequency domains (mean and standard deviation of HRV; total power (TP); very low frequency power (VLF); low-frequency power (LF), and high-frequency power (HF) [20, 25, 31]. In addition, approximate entropy (ApEn) and sam-ple entropy (SampEn), two commonly used tools for non-linear dynamic analysis of HRV, were considered. Estrada et al. [10] employed ApEn, as well as time and frequency domain features, of EEG and, then separately, of HRV. They found that features extracted during REM sleep, both from EEG and HRV, always overlapped with the features of any other stage. The use of SampEn for char-acterizing sleep stages from HRV was first demonstrated by Vigo et al. [32]. They analyzed 5-min epochs of wake state, stage 3 of NREM, and REM sleep. They found that SampEn in REM and deep sleep was significantly different ($p < 0.005$).

The goal of this study is to improve the automatic dis-tinction of wakefulness (WAKE) from sleep (SLEEP) and also of NREM from REM sleep, using HRV-based fea-tures only. During NREM stage 1, people drips in and out of sleep. Inspired by Kortelainen [15], who merged stage 1 with wake, and to reduce the ambiguity between wake and NREM, stage 1 was not considered in this study.

We first introduce new features that reflect the changes in regularity of the RR series (i.e., the series of intervals between successive R peaks of an ECG signal), among the different sleep stages. These nonlinear features enhance the characterization of sleep autonomic regulations and have been evaluated along with more established ones (classical time and frequency metrics as well as other nonlinear features). The classification performances depend on the features as well as on the proper selection of a classifier. However, the focus of this study is not to check the best classification strategy, but on the contrary, to verify the possible merit of new parameters in sleep classification. The actual classification performances of the entire feature set, and of selected features only, were evaluated using a feed forward neural network (FFNN), a special kind of artificial neural network (ANN) commonly used in pattern recognition and classification problems. FFNN is a very established tool for pattern recognition and classification. Thus, it was selected because, in our experience, it is also very effective in describing separation manifolds in the parameter space. Existing methods, reporting high accuracy in classifying sleep stages [8, 33], used a large set of features, making the system computationally expensive. In here, an improved accuracy has been sought using a smaller number of features, which would be helpful in implementing the system on low-power computers or smartphones.

2 Methods

The block diagram in Fig. 1 summarizes the proposed method. First, RR series of the considered epochs have been preprocessed to remove artifacts or ectopic beats. After preprocessing, a set of features has been extracted from RR series, and the performance of FFNN for WAKE versus SLEEP and NREM versus REM classifications has been tested using this set of features. Finally, a reduced set of best relevant features has been selected, from the extracted feature set, and the performance of the classifiers has been tested using these reduced set as well.

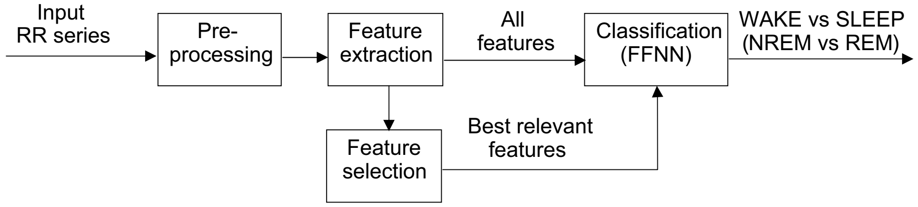

Fig. 1 Block diagram of the proposed method

2.1 Dataset

Full PSG of 20 patients with suspected sleep-disordered breathing was recorded for one night, each at the Sleep Center of Tampere University Hospital, Finland. The Ethical Committee of the Pirkanmaa Hospital District approved the study, and all the subjects gave an informed consent to be included into the study. The age of the subjects was between 49 and 68 years; the BMI varied between 21.8 and 40.6; 13 patients were females. The patients suffered from a variety of sleep disorders, including either different degrees of nocturnal apnea/hypopnea and/or insomnia (34 % of subjects: no-apnea, 32 %: mild/moderate apnea, 34 %: severe apnea).

We considered the inter-beats (RR) series obtained from the ECG recordings, as well as the sleep scoring automatically derived from the complete PSG recordings (mainly using the EEG traces) through the Somnologica® software; the scoring was based on 30-s epochs. ECG R peak positions were detected automatically, also using the Somnologica® software.

2.2 Preprocessing

RR intervals may contain artifacts due to ectopic beats, movements, or detachments of the leads. These artifacts have been removed using a two-step processing. In the first step, RR intervals that lay outside the interval $I = [Q_1 - 3 \times \text{IR}, \ Q_3 + 3 \times \text{IR}]$ were marked as artifacts. Q_1 and Q_3 are, respectively, the first and the third quartiles of the RR series, while the interquartile range, IR, is defined as $Q_3 - Q_1$. In the second step, we labeled as normal those RR values, which varied <20 % of the previously accepted RR interval [18] (considering that the very first accepted RR interval was within the IR of the entire series).

RR segments of different lengths (corresponding to 2, 6, and 10 epochs, 30 s each, classified as pertaining to a common sleep stage) were considered for the study only if at least 20 % of the beats were finally marked as normal. Transitions were not considered, and each segment belonged to one sleep stage only.

2.3 Feature extraction

Feature extraction is a critical step for classification. The performance of the classifier depends on how robust the feature set is in distinguishing the entities. In this study, we first considered features that were shown as effective in previous studies; further, we added features describing the regularity of the series. They can be classified as (1) time-domain; (2) frequency domain; (3) detrended fluctuation analysis (DFA); and (4) regularity features.

1. *Time-domain features*: Standard deviation (SDNN) and the mean value (Mean_{NN}) [19] of the normal RR intervals were selected. SDNN and Mean_{NN} of each segment were, respectively, normalized by SDNN and mean of the entire RR series.
2. *Frequency domain features*: To estimate spectral features of each window, segments of normal beats were fit to an autoregressive model (AR) of fixed order. A previous work [20] showed that a model order of about 8 was sufficient for sleep classification purposes. In particular, it permitted to have at least one pole in each relevant spectral band. In this study, the order of the model was further increased to 9, such that the Anderson's test [5], which checks the whiteness of the prediction error, failed for <5 % of the examined cases. From the estimated model, three frequency band powers were extracted using the spectral decomposition technique described in [2]. The bands were: VLF from 0.003 to 0.04 Hz, LF from 0.04 to 0.15 Hz, and HF from 0.15 to 0.4 Hz. Figure 2 shows the RR signal considered for about 1.5 h in a single recording and the spectral components of the different sleep stages considered. Each of these powers was normalized by TP, and the LF/HF ratio (ratio of the power in the two bands) was also considered. As shown in the example of Fig. 2, the spectral components of HR during NREM and REM appear to be different. Finally, the modulus of the pole with the largest residual in the HF band (Pole_{HF}) was also included into the feature set. Pole_{HF} is strongly related to the respiratory frequency and periodicity [20].
3. *Detrended fluctuation analysis* (DFA): DFA is a scaling analysis method that provides a simple quantitative parameter to estimate the autocorrelation properties of a non-stationary signal. It has proven useful in characterizing correlations in apparently irregular time series [22]. In DFA, an integrated time series is constructed from the original one. Then, this integrated time series is divided into non-overlapping "time-windows" of increasing size n, and local trends are subtracted. The fluctuation of the remaining signal is determined while increasing the window size. The slope of the variance of the fluctuations versus the window size defines the scaling exponent. For many biological signals, among which most RR series, the DFA plot in logarithmic scale consists of two distinct linearly scaling regions of different slopes, separated at a break point. The two slopes are termed as "short-range scaling exponent" (α_1) and "long-range scaling exponent" (α_2). In this paper, only the short-range scaling exponent has been considered. For estimating α_1, n was varied from 4 to 11. The short-range DFA scaling exponent needs

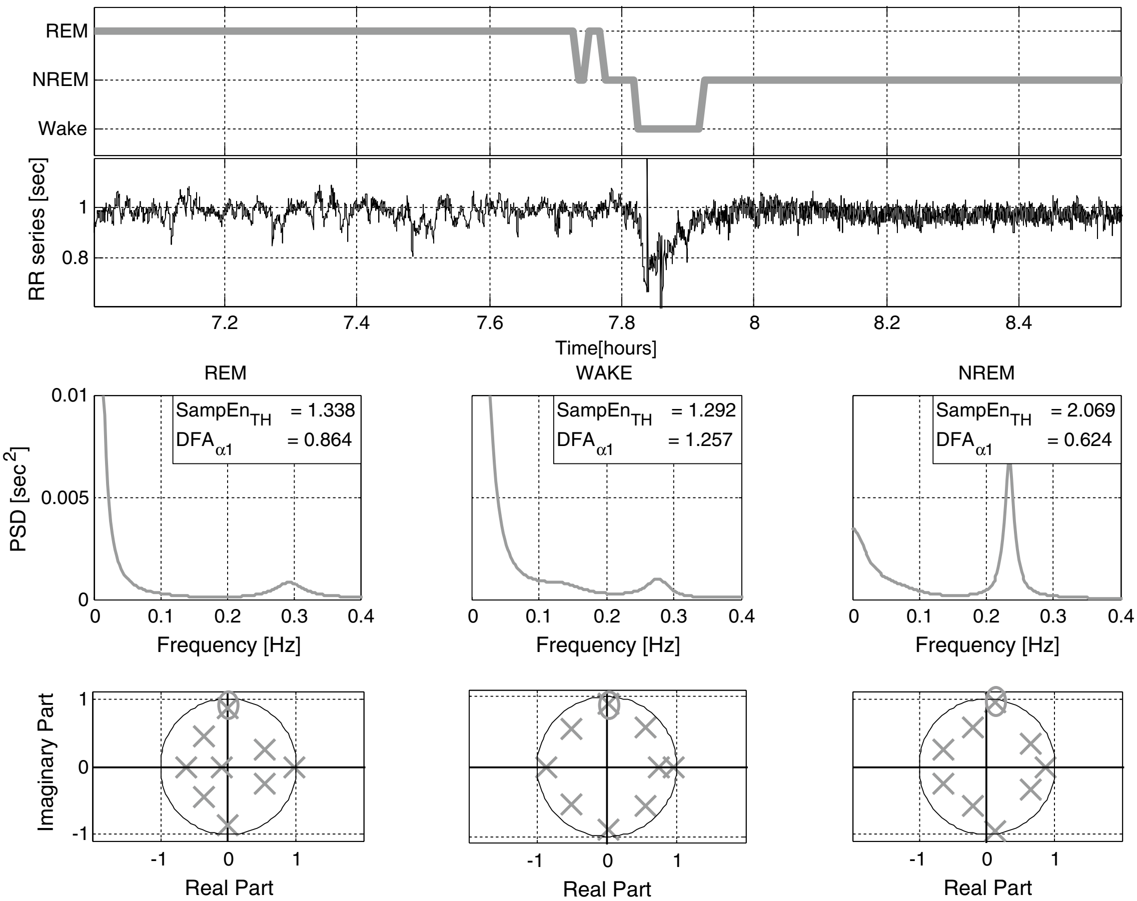

Fig. 2 *Top panel* one hour and a half of a PSG recording: a 3-stage hypnogram *top* and the corresponding RR series *bottom*. *Middle panel* power spectra of the RR series during the different sleep stages. *Bottom* positions of the poles of an AR model fitted to an RR series during each of the three stages considered (Pole$_{\text{HF}}$ was marked with a *circle*)

at least $4n$ samples to be computed reliably [22]. When considering windows of 2 epochs, the number of beats is around 60. Thus, the requirements are satisfied for computing short-term DFA and not for long-term DFA (which was not considered).

4. *Regularity features:* Sample entropy is a statistic commonly used to measure the regularity and complexity of physiological and clinical time series. SampEn [26] quantifies the regularity of a time series by matching a pattern of length m with any other pattern of the same length within a tolerance r; then, the comparison is repeated at an extended length $m + 1$. In this study, $m = 1$ and $r = 0.2 \times \text{SDNN}$ were used, given the short length of the series. SampEn was computed for both RR series and for series produced by AR models fitted on the RR series as in [1]. The numerical estimation of SampEn (SampEn$_{\text{NN}}$) for a time series RR(i), for $1 \leq j \leq N$ starts with constructing the templates $U_m(j) = \{\text{RR}(j), \text{RR}(j+1), \ldots, \text{RR}(j+m-1)\}$ of size m for $1 \leq j \leq N - m$ and defining the distance between $U_m(i)$ and $U_m(j)$ by $d\left[U_m(i),\ U_m(j)\right] = \max\limits_{0 \leq k \leq m-1} |U(i+k) - U(j+k)|$.

Now, let A_j^m be the number of templates $U_m(i)$ such that d$[U_m(i),\ U_m(j)] \leq r$, for $1 \leq i \neq j \leq N - m$ and $C_j^m(r) = A_j^m/(N - m - 1)$ and let A_j^{m+1} be the number of templates $U_{m+1}(i)$ such that $d\left[U_{m+1}(i), U_{m+1}(j)\right] \leq r$ for $1 \leq i \neq j \leq N - m$ and $C_j^{m+1}(r) = A_j^{m+1}/(N - m - 1)$.

If we define $A(m,\ r) = \sum_{j=1}^{N-m} C_j^m(r)/(N - m)$ and $A(m + 1,\ r) = \sum_{j=1}^{N-m} (r)/(N - m)$, then SampEn$_{\text{NN}}$ $(m,\ r,\ N) = \log A(m,\ r) - \log A(m + 1,\ r)$.

SampEn$_{\text{NN}}$ may be affected by nonlinearity, non-Gaussianity, or non-stationarity present in the series. Thus, starting from the same AR model employed for computing frequency domain features, the value of SampEn (SampEn$_{\text{TH}}$) was instead derived analytically using the formula pro-posed by Aktaruzzaman and Sassi [1]. In fact, for a station-ary stochastic process (thus, for an AR process $x[n]$), the probability of matching two templates of size m, within the error tolerance r, can be represented by

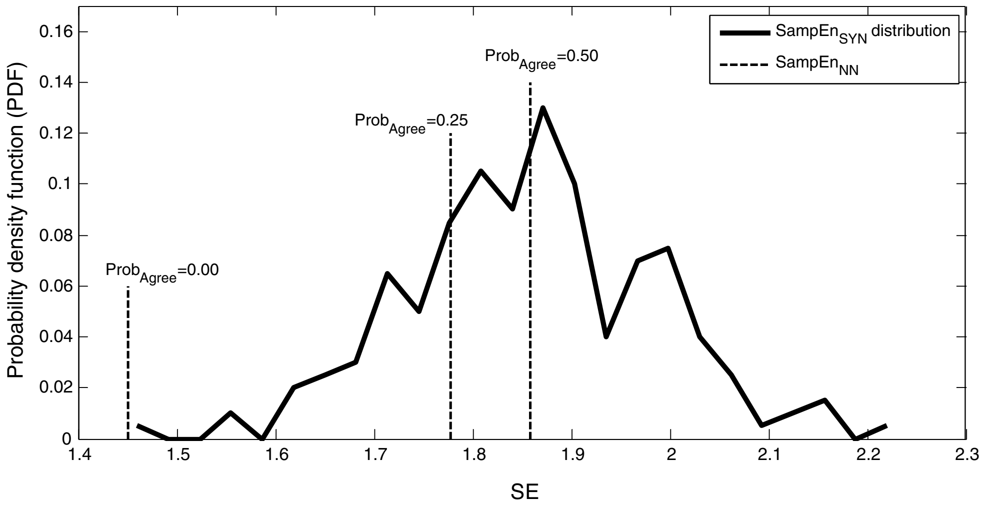

Fig. 3 Probability density of the values of SampEn computed on 200 synthetic series (*thick black line*), of which SampEn$_{SYN}$ is the average value, and the probability of agreement (Prob$_{Agree}$) for three distinct values of SampEn$_{NN}$ (*vertical bars*). The probability of agreement is indicated for each SampEn$_{NN}$

$$P_m = \int_{x(m)-r}^{x(m)+r} \cdots \int_{x(1)-r}^{x(1)+r} \frac{\mathrm{e}^{-\Sigma_m^T \Sigma_m^{-1} \Sigma_m}}{(2\pi)^{m/2} \det{(2\Sigma_m)}^{1/2}} d\xi_1 \ldots d\xi_{m\prime}$$

where Σ_m is the Toeplitz covariance matrix of the AR process. Then, SampEn can be expressed analytically as

$$\mathrm{SampEn}_{\mathrm{TH}} = \log P_m - \log P_{m+1}.$$

Alternatively, an expected value of SampEn (SampEn$_{SYN}$) of the AR model was estimated computing the mean value of SampEn obtained for 200 synthetic series (of the same length of the original ones), generated through the AR model itself (a Monte Carlo approach). These parametric estimations of SampEn (SampEn$_{TH}$ and SampEn$_{SYN}$) are truly affected only by the linear behavior of the model. Finally, the capability of the AR model to well approximate the series in terms of SampEn$_{NN}$ was tested using the Monte Carlo simulations result. To this aim, the distribution of SampEn values, estimated from the synthetic series, was compared with SampEn$_{NN}$. The value of SampEn$_{NN}$ may fall within or outside this distribution (Fig. 3). The probability of agreement (Prob$_{Agree}$) between SampEn$_{NN}$ and the distribution increases from 0 (SampEn$_{NN}$ lies out-side the distribution) to 0.5 (SampEn$_{NN}$ corresponds to the median of SampEn$_{SYN}$). Prob$_{Agree}$ was calculated nonpara-metrically using the ranks of SampEn$_{SYN}$.

2.4 Classification

The goals of this study were to distinguish different sleep stages. A FFNN was first trained using a set of features extracted from a training set of data. Then, the trained FFNN was used for classification of the test dataset.

The data from the 20 recorded patients were divided according to two different cross-validation techniques: leave one out (LOO) among recordings and tenfold on the total amount of data. In the dataset considered, the distribution of the classes was unbalanced. This may neg-atively influence the training of the FFNN [36]. For this reason, the entire study was repeated using both unbal-anced and balanced proportions of classes for training. To balance the populations, samples were selected ran-domly. Also, the different initializations of the randomly selected weights might lead to (slightly) different clas-sification results. To minimize this problem, the training and testing of the FFNN was repeated five times, and the average performances were taken into account. Finally, the classification performances provided by the network were evaluated by means of accuracy (ACC), sensitivity (SENS), specificity (SPEC), and Cohen's Kappa (K) reli-ability [6].

2.5 Feature selection

The possibility of reducing the feature set dimension was investigated. To this aim, a feature selection strategy was applied, using the following two approaches:

- Greedy backward elimination: The classification reliability (K) was checked by leaving one feature out at a time. The feature discarded at each round was the one leading to the highest K value of the remaining set. This procedure was repeated until only the most significant feature was retained.
- Greedy forward selection: The procedure was started with the single best feature estimated using the previous approach. In here, at each round, another feature, from the remaining ones, was added to the set. The additional feature was chosen so that the new set of features was leading to the highest K value. The procedure was repeated until all the features were included.

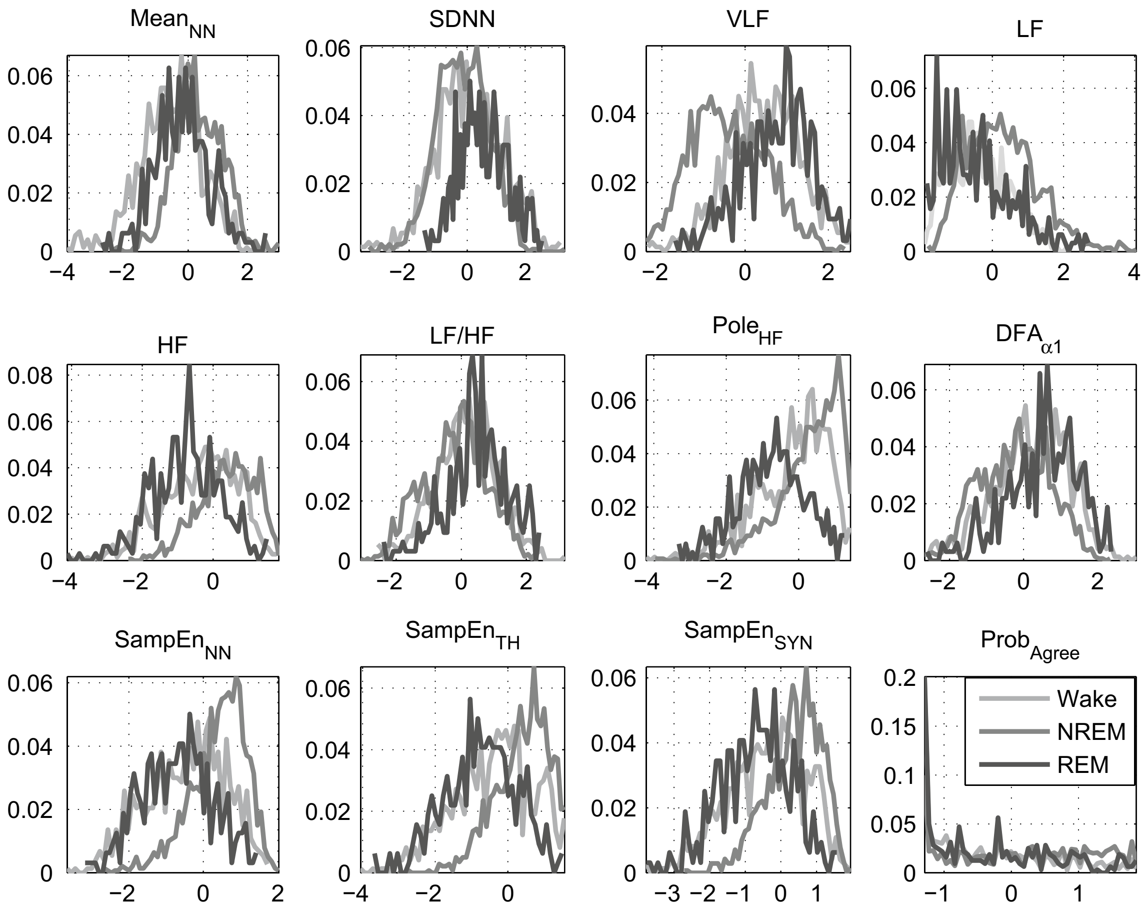

Fig. 4 Probability distributions of the features in the full set, for different sleep stages (after logarithmic transformation when necessary, see text for details). They were obtained from 6 epoch long RR series

The feature selection strategy was performed on windows of 6 epochs, with a tenfold cross-validation (CV) procedure. We limited the feature selection procedure to 6 epochs, because shorter time periods may be not sufficient for a reliable estimation of certain parameters (spectral estimation, entropy evaluation), while longer periods HRV series may be affected by non-stationarity and may be too long when compared with the 30-s epochs clinically used for sleep classification.

3 Results

The twelve features {Mean$_{\mathrm{NN}}$, SDNN, VLF, LF, HF, LF/HF, Pole$_{\mathrm{HF}}$, DFA$_{\alpha1}$, SampEn$_{\mathrm{NN}}$, SampEn$_{\mathrm{TH}}$, SampEn$_{\mathrm{SYN}}$, and Prob$_{\mathrm{Agree}}$} were extracted from each RR segment of 2, 6, and 10 epochs. A logarithmic transformation was applied to SDNN, HF, and LF/HF in order to get statistical distribu-tions closer in shape to a Gaussian function. Figure 4 shows the distributions of the extracted features with respect to the sleep stages considered (after average value subtraction and normalization of the standard deviation). For some of them, like VLF, Pole$_{\mathrm{HF}}$, and the various SampEn, the distributions of REM and NREM stages differ, even visually. The power of each single feature, in separating sleep stages, has been statistically tested through a Kruskal–Wallis nonparametric analysis of variance. All the features significantly ($p < 0.01$) discriminate SLEEP from WAKE and NREM from REM.

The number of neurons in the hidden layer of the FFNN affects its capabilities. To determine it, the performances of the classifier were preliminary observed using different number of hidden neurons (8, 12, 15, 20, 25) in a smaller case study (only a subset of the subjects and epochs was used). The classification performances, in terms of accuracy, did not improve using more than 12 neurons, which is what was employed in the following of the study.

The results for discriminating the different sleep stages using the full feature set have been summarized in Table 1, which reports ACC along with *K*, for balanced and unbalanced number of samples and using both tenfold and LOO validation techniques. The accuracy of WAKE versus SLEEP classification was 77.16 and 71.65 %, for tenfold

Table 1 Sleep stage classification using the full feature set. (a) Results for WAKE versus SLEEP classification. (b) NREM versus REM classification results

Number of epochs	Distribution type	Tenfold				LOO			
		ACC (%)	SENS (%)	SPEC (%)	K	ACC (%)	SENS (%)	SPEC (%)	K
(a)									
2	Unbalanced	77.16 ± 0.14	40.40 ± 0.71	92.62 ± 0.25	0.38 ± 0.01	71.65 ± 13.77	41.39 ± 20.78	89.61 ± 8.75	0.28 ± 0.18
	Balanced	69.93 ± 0.69	66.22 ± 0.71	73.64 ± 1.10	0.40 ± 0.01	68.37 ± 11.79	65.56 ± 21.27	70.93 ± 17.00	0.27 ± 0.21
6	Unbalanced	77.22 ± 0.68	47.20 ± 2.80	89.99 ± 0.62	0.41 ± 0.02	69.66 ± 16.69	42.35 ± 25.58	88.28 ± 10.11	0.26 ± 0.22
	Balanced	72.40 ± 0.95	73.51 ± 1.36	71.29 ± 1.14	0.45 ± 0.02	67.59 ± 12.09	71.97 ± 22.28	70.26 ± 17.12	0.31 ± 0.24
10	Unbalanced	77.91 ± 0.27	52.17 ± 1.42	88.90 ± 0.76	0.44 ± 0.01	71.92 ± 18.24	43.68 ± 27.34	88.96 ± 7.78	0.29 ± 0.24
	Balanced	75.09 ± 1.10	76.17 ± 2.18	74.00 ± 2.15	0.50 ± 0.02	70.76 ± 10.68	73.28 ± 23.74	72.10 ± 15.31	0.32 ± 0.24
(b)									
2	Unbalanced	83.17 ± 0.14	96.02 ± 0.22	29.33 ± 1.19	0.32 ± 0.01	82.07 ± 5.14	94.42 ± 6.33	30.30 ± 22.68	0.27 ± 0.16
	Balanced	71.96 ± 0.86	72.48 ± 2.02	71.44 ± 1.39	0.44 ± 0.02	68.69 ± 16.55	69.69 ± 22.88	67.59 ± 23.85	0.29 ± 0.19
6	Unbalanced	86.74 ± 0.46	94.78 ± 0.36	51.38 ± 3.35	0.51 ± 0.02	84.63 ± 6.51	93.13 ± 9.33	46.39 ± 26.96	0.41 ± 0.21
	Balanced	80.50 ± 1.02	80.58 ± 2.04	80.42 ± 1.75	0.61 ± 0.02	75.83 ± 16.92	75.57 ± 22.85	78.20 ± 23.00	0.44 ± 0.24
10	Unbalanced	88.21 ± 0.63	94.91 ± 0.54	57.88 ± 2.72	0.57 ± 0.02	84.62 ± 8.12	91.47 ± 11.41	52.27 ± 33.69	0.42 ± 0.25
	Balanced	82.79 ± 2.17	84.00 ± 2.18	81.59 ± 2.92	0.66 ± 0.04	79.39 ± 15.65	79.40 ± 19.98	80.17 ± 22.38	0.49 ± 0.25

Table 2 Results of the features selection procedure for WAKE versus SLEEP classification, using windows of 6 epochs with balanced datasets. (a) Classification performances after removing one feature at a time (the feature removed is indicated in each row). (b) Classification performances after adding one feature at a time

	ACC (%)	SENS (%)	SPEC (%)	K		ACC (%)	SENS (%)	SPEC (%)	K
(a)					(b)				
All	74.40	73.80	74.90	0.49	Mean$_{\text{NN}}$	67.10	58.90	75.30	0.34
VLF	76.10	76.60	75.60	0.52	VLF	70.10	69.90	70.30	0.40
DFA$_{\alpha1}$	74.90	75.90	74.00	0.50	DFA$_{\alpha1}$	72.10	71.20	72.90	0.44
SampEn$_{\text{TH}}$	75.00	75.30	74.70	0.50	Prob$_{\text{Agree}}$	72.50	73.20	71.80	0.50
SampEn$_{\text{SYN}}$	74.70	74.80	74.50	0.49	SDNN	72.90	74.20	71.50	0.46
LF/HF	74.80	74.40	75.20	0.50	LF/HF	73.30	72.90	73.70	0.47
Pole$_{\text{HF}}$	75.80	77.40	74.20	0.52	SampEn$_{\text{SYN}}$	74.80	75.90	73.70	0.50
SampEn$_{\text{NN}}$	74.90	74.40	75.50	0.50	SampEn$_{\text{TH}}$	74.30	74.50	74.10	0.49
SDNN	73.30	74.20	72.30	0.47	HF	75.00	74.50	75.50	0.50
Prob$_{\text{Agree}}$	71.60	71.20	72.10	0.43	SampEn$_{\text{NN}}$	74.60	75.30	73.80	0.49
HF	68.90	62.50	75.30	0.38	LF	74.20	74.00	74.40	0.48
LF	67.10	58.90	75.30	0.34	Pole$_{\text{HF}}$	74.40	73.80	74.90	0.49
Mean$_{\text{NN}}$ (Mean$_{\text{NN}}$ is the only features left after removal of LF)					All (All features are included after the addition of Pole$_{\text{HF}}$)				

and LOO techniques, respectively, with unbalanced number of samples and 2 epochs. ACC and *K* did not change considerably neither increasing the number of epochs nor changing the proportion of samples (balanced or unbalanced). It is worth noting that SENS and SPEC represent here the true recognition of WAKE and SLEEP stages, respectively. There was an incremental trend in SENS (from 40.40 %, 2 epochs to 52.17 %, 10 epochs) and K (from 0.38 to 0.44) with increasing the RR segment length, when tenfold validation was used. The same happened in NREM versus REM classification, where ACC increased (from 83.17 to 88.21 %), as well as SPEC (29.33 to 57.88 %) and K (0.32 to 0.57). The classifier showed a slightly smaller recognition accuracy (84.62 % instead of 88.21 %) when LOO was used.

Table 2 shows the average results of the feature selection procedure for WAKE versus SLEEP classification. In Table 2a, the first row corresponds to the results for the full set of features, as well as the last row of Table 2b. The value of *K* increased from 0.34 (when only Mean$_{\text{NN}}$

Table 3 Results of the features selection procedure for NREM versus REM classification, using windows of 6 epochs with balanced datasets. (a) Classification performances after removing one feature at a time (the feature removed is indicated in each row). (b) Classification performances after adding one feature at a time

	ACC (%)	SENS (%)	SENS (%)	K		ACC (%)	SENS (%)	SENS (%)	K
(a)					(b)				
All	84.10	83.70	67.70	0.68	Pole_{HF}	75.90	74.20	77.60	0.52
VLF	84.40	85.40	83.40	0.69	LF	80.50	79.50	81.60	0.61
HF	84.20	84.60	83.90	0.69	Mean_{NN}	83.30	81.80	84.80	0.67
$\text{SampEn}_{\text{TH}}$	84.50	85.50	83.40	0.70	$\text{SampEn}_{\text{SYN}}$	84.60	84.60	84.50	0.70
$\text{DFA}_{\alpha 1}$	84.90	85.70	84.20	0.70	$\text{SampEn}_{\text{NN}}$	85.10	85.10	85.20	0.70
LF/HF	84.50	84.80	84.10	0.69	SDNN	84.70	85.20	84.30	0.70
SDNN	84.80	85.60	84.10	0.67	$\text{DFA}_{\alpha 1}$	85.00	85.60	84.40	0.70
$\text{SampEn}_{\text{SYN}}$	84.40	83.90	84.80	0.69	VLF	85.00	85.50	84.50	0.70
$\text{Prob}_{\text{Agree}}$	84.70	84.80	84.50	0.69	LF/HF	84.60	84.80	84.30	0.69
$\text{SampEn}_{\text{NN}}$	83.60	82.60	84.60	0.67	HF	84.60	84.70	84.50	0.69
Mean_{NN}	80.90	79.10	82.70	0.62	$\text{Prob}_{\text{Agree}}$	83.50	83.80	83.30	0.67
LF	75.90	74.20	77.60	0.52	$\text{SampEn}_{\text{TH}}$	75.90	74.20	77.60	0.52
Pole_{HF} (Pole_{HF} is the only features left after removal of LF)					All (All features are included after the addition of $\text{SampEn}_{\text{TH}}$)				

Table 4 Sleep stages classification using 4 relevant features only. (a) Results (mean ± std) for WAKE versus SLEEP classification. (b) NREM versus REM classification

Number of epochs	Distribution type	Tenfold				LOO			
		ACC	SENS	SPEC	K	ACC	SENS	SPEC	K
(a)									
2	Balanced	67.69 ± 0.44	62.89 ± 0.91	72.48 ± 0.77	0.35 ± 0.01	67.55 ± 9.93	65.29 ± 17.72	70.37 ± 13.81	0.26 ± 0.22
6	Balanced	70.79 ± 0.81	72.99 ± 1.49	68.60 ± 2.20	0.42 ± 0.02	69.02 ± 11.34	75.13 ± 18.99	68.57 ± 15.92	0.31 ± 0.24
10	Balanced	73.30 ± 0.94	76.39 ± 2.21	70.21 ± 1.20	0.47 ± 0.02	71.34 ± 11.73	77.15 ± 17.00	69.27 ± 14.02	0.33 ± 0.23
(b)									
2	Balanced	71.73 ± 0.66	74.08 ± 0.78	69.38 ± 0.83	0.43 ± 0.01	68.35 ± 15.32	68.34 ± 21.57	71.11 ± 22.58	0.29 ± 0.18
6	Balanced	80.28 ± 1.33	79.40 ± 2.29	81.16 ± 0.77	0.61 ± 0.03	74.54 ± 19.72	73.74 ± 25.54	80.24 ± 20.34	0.44 ± 0.24
10	Balanced	83.78 ± 2.06	82.41 ± 2.60	85.15 ± 2.73	0.68 ± 0.04	79.77 ± 15.72	79.68 ± 20.07	81.13 ± 22.88	0.51 ± 0.25

was considered) to 0.45 (when VLF, $\text{DFA}_{\alpha 1}$, and $\text{Prob}_{\text{Agree}}$ were also added). The addition of the remaining features just increased the feature set dimension without any major significant contribution to the value of K. A similar behavior was observed when reducing the size of the feature set. Thus, the set of {Mean_{NN}, VLF, $\text{DFA}_{\alpha 1}$, and $\text{Prob}_{\text{Agree}}$} was further considered for WAKE versus SLEEP classification. Similarly, the results of the feature selection strategy for NREM versus REM classification are instead illustrated in Table 3. Also, in this case, the value of K did not improve considerably beyond the addition of four features. The four best relevant features were: {Mean_{NN}, LF, Pole_{HF}, and $\text{SampEn}_{\text{SYN}}$}.

The classification results obtained using only the four most relevant features are reported in Table 4. The accuracy obtained for training with unbalanced samples of classes might prove unsatisfactory (K value was very small). Also, K values for training with balanced number of samples are always higher than those with unbalanced sets. Thus, in Table 4, we have summarized only the results for training with balanced samples. Overall, there was no considerable difference between ACC (88.22 %) and K (0.56), obtained using only four relevant features, and ACC (88.21 %) and K (0.57) obtained with the full set of features. The mean results using tenfold and the LOO cross-validation were comparable for every cases, while the standard deviation was higher using the LOO technique, suggesting that the inter-subject variability is large.

As a final confirmation that four features are sufficient for describing the variability of the data, we further verified using principal component analysis (PCA), a linear technique which is often used to guide feature selection with traditional classification methods [11], that four transformed variables captured 99 % of the variance of the data in both classification problems. When considering the WAKE versus SLEEP problem, the features with the largest normalized weight with respect to each of the first four principal components were: $\text{SampEn}_{\text{SYN}}$, LF, $\text{Prob}_{\text{Agree}}$, and Mean_{NN}, while for the NREM versus REM classification, they were: HF, LF, $\text{Prob}_{\text{Agree}}$, and Mean_{NN}. Interestingly also with this simplified approach, the new features we introduced were significantly relevant in the classification problems.

4 Discussion

This study supports the possibility of building a fully automatic classification of sleep stages into WAKE, NREM, and REM, using RR series analysis. In fact, the features considered were not only statistically different in different sleep stages, but also relevant when used in a classifier.

A set of 12 features, including new ones based on the regularity of the series, was considered for discriminating NREM versus REM and WAKE versus SLEEP. The feature selection strategy reduced the feature set dimension from 12 to 4, by removing redundant parameters which did not carry additional information. The overall classification performances (as measured by ACC, K) did not change significantly when a subset of four features instead of the full set was considered.

Two nearly distinct sets of four features (with the exception of Mean_{NN} contained in both) were selected for the two classification problems. In addition to time and frequency domain parameters (already reported in the literature), three additional features were included in these sets.

Mean_{NN} and Pole_{HF} were previously reported [20] as significant features for sleep staging into NREM versus REM and also here proved so. In particular, Mean_{NN} increased from WAKE to SLEEP and from REM to NREM indicating an augmented vagal control. Similarly, VLF and LF proved here valuable in discriminating between WAKE versus SLEEP and NREM versus REM, confirming previous studies [20, 25]. VLF has been normalized as a percentage of the total power: its increase during WAKE indicates that the total variance is influenced also by different factors in addition to the sympatho-vagal system, while during SLEEP, the main source of variability is the sympatho-vagal balance. Moreover, LF was expected to increase during REM. However, in this study, LF increased during NREM. This can be explained with the fact that LF is normalized with respect to the total power. Thus, the higher values of LF during NREM are also determined by the lower values of VLF during NREM. Finally, the module of the strongest pole in the HF band (Pole_{HF}) was previously employed for discriminating between NREM and REM [20] and here proved highly informative. This feature captures the periodicity of the respiration rhythm which is high during NREM sleep and decreases significantly during REM [20].

The use of three measures of SampEn has been inspired by the fact that their estimates may vary differently according to series characteristics, such as the presence of nonlinearity, non-stationarity, and non-Gaussianity. However, in here, the three methods showed really similar behaviors (Fig. 4). $\text{SampEn}_{\text{NN}}$, $\text{SampEn}_{\text{SYN}}$, and $\text{SampEn}_{\text{TH}}$ in practice carried the same information, even if the latter two are linear indexes while the former is a nonlinear metric. Overall, the entropy increased significantly during NREM, suggesting a higher regularity during REM sleep.

The relevance of $\text{DFA}_{\alpha1}$ and $\text{Prob}_{\text{Agree}}$ in the classification process suggests that there are evident changes in short-term correlations and nonlinear regularity of HR during different sleep stages. A part from increasing the overall classification performances, it gave also information about the physiology of sleep. In fact, during SLEEP, the decreased $\text{DFA}_{\alpha1}$ might reflect a reduced short-term (in the range 4–11 heart beats) persistence of HRV patterns in time, while, coherently, a larger $\text{Prob}_{\text{Agree}}$ suggests a possible lowering of nonlinearity or non-stationarity during such periods.

$\text{SampEn}_{\text{SYN}}$ and $\text{Prob}_{\text{Agree}}$ as well as the majority of the extracted features depend on autoregressive models estimation. Thus, at least 3 min of recording (6 epochs) was needed to get good classification performances (see Tables 1 and 4) using AR models. Also, ACC and K increased with the number of epochs considered. Further studies will focus on the selection of the best signal length for classification. In addition, the use of time-variant models will be introduced, in order to verify the possibility of reducing the number of samples needed to perform a similarly reliable classification. This would be a significant step toward a 30-s HRV-based hypnogram.

The recognition accuracy obtained using LOO is slightly smaller than what achieved using tenfold validation, and this is likely due to the fact that 20 subjects were not enough to capture the large variability of patterns across subjects. As a consequence, the features of the subject excluded from training were completely "new" for the FFNN and could not be predicted by the knowledge acquired from the rest of the population. LOO is closer to what happens in a practical application, where the FFNN is trained in a laboratory and then used on different subjects. However, to capture the possible variability in the features across subjects, a much larger population should be employed (and such a large population is not easy to obtain). Using only 20 subjects, the results are underestimating the possible accuracy of the method. Given, the relatively small number of subjects at disposal, LOO was considered as a sort of limiting bottom value for accuracy, while tenfold CV represented a correspondent limiting top value. In fact, with the latter samples in the test and training set, even if completely distinguished, might be correlated among them.

From a technical standpoint, the unbalance in the distribution of classes has become a crucial problem in machine learning algorithms, because the performances of the learning phase could be seriously compromised. So a balanced distribution of classes is often recommended [36] as we did consider in here. As expected, in comparing the results with balanced versus unbalanced number of samples for training and testing, the performances gave privilege to the most represented class (SLEEP for WAKE versus SLEEP and NREM for NREM versus REM classification) when unbalanced samples are used for training. This is typical when the sets are skewed toward one of the classes. This issue supports the statement that the classification is more reliable when training is performed with balanced number of samples. Also, the reliability factor K helped in checking if the classifier was biased to a specific class. Although such considerations are important in practice, only a few studies [15, 21, 25] considered this issue.

To the best of our knowledge, a few previous studies [8, 20, 25, 33] performed sleep stage classification from HRV analysis. It is difficult to compare the results obtained on different datasets, because the accuracy depends on the characteristics of the datasets themselves, and also on the specific type of classifier employed. However, the accuracy of NREM and REM classification obtained in this study was always larger than 82 %, which is an improvement over the results reported in Mendez et al. [20]. Although, Redmond et al. [25] reported a classification accuracy for WAKE versus SLEEP of 89 %, which is larger, they used a set of 30 features, collected not only from ECG but also from respiratory signals. The overall accuracy found in this work (Table 4) is equivalent to those reported in [33] (tables 5 and 6), even if here using only 4 features (which should be more computationally cost effective). Our results are also comparable with those reported in [8], even though they used 32 features, and in here only 4 or 12. The slightly better accuracy (84.4 %, instead of 79.8 %) reported in [8] for long (10 epochs) RR series may be due to the usage of bootstrapping to increase the number of samples (i.e., REM). In fact, due to the requirement of considering only homogenous consecutive sleep epochs, the number of long RR segments in our study is small, so it might lead to insufficient training of the neural network (which does not happen with short series). It is also worth noting that the population considered here included patients with different degree of apnea, but apnea-related events were not considered separately from the other signal segments and were just included blindly in the study (to mimic real-world situations). Thus, while our results were possibly negatively influenced by this methodological decision, the overall method proved robust to the presence of apnea.

5 Conclusions

The proposed method appeared prospective for automatic sleep stage classification, based on HRV analysis only, with a focus on the distinction of wakefulness from sleep, and REM from NREM sleep. The regularity parameters were found as the most significant among the features considered, for both classification problems. Apart from increasing the overall classification performances, they also provided information about the physiology of sleep, in particular with respect to NREM stages. These findings paved the way to further investigations of the behavior of the autonomic nervous system during sleep. Time-variant autoregressive (TVAR) models, as well as other machine learning tools, less sensitive to the unbalanced proportion of samples, will be investigated in future works to improve the classification performances of the method.

References

1. Aktaruzzaman M, Sassi R (2014) Parametric estimation of sample entropy in heart rate variability analysis. Biomed Signal Process Control 14:141–147
2. Baselli G, Porta A, Rimoldi O et al (1997) Spectral decomposition in multichannel recordings based on multivariate parametric identification. IEEE Trans Biomed Eng 44:1092–1101
3. Berry RB, Budhiraja R, Gottlieb DJ et al (2012) Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events. J Clin Sleep Med 8:597–619
4. Bianchi AM, Mendez MO, Cerutti S (2010) Processing of signals recorded through smart devices: sleep-quality assessment. IEEE Trans Inf Technol Biomed 14:741–747
5. Box GEP, Jenkins GM (1976) Time series analysis: forecasting and control, Revised edition. Holden-Day, San Francisco
6. Cohen J (1968) Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70:213–220
7. Covassin N, de Zambotti M, Cellini N et al (2013) Cardiovascular down-regulation in essential hypotension: relationships with autonomic control and sleep. Psychophysiology 50:767–776
8. Ebrahimi F, Setarehdan S-K, Ayala-Moyeda J, Nazeran H (2013) Automatic sleep staging using empirical mode decomposition, discrete wavelet transform, time-domain, and nonlinear dynamics features of heart rate variability signals. Comput Methods Progr Biomed 112:47–57
9. Engeda J, Mezuk B, Ratliff S, Ning Y (2013) Association between duration and quality of sleep and the risk of pre-diabetes: evidence from NHANES. Diabet Med J Br Diabet Assoc 30:676–680
10. Estrada E, Nazeran H (2010) EEG and HRV signal features for automatic sleep staging and apnea detection. In: 20th International Conference on Electronics Communications and Computer 142–147

11. Everitt BS, Dunn G (2010) Applied multivariate data analysis, 2nd edn. Wiley, Chichester
12. Ferini-Strambi L, Bianchi A, Zucconi M et al (2000) The impact of cyclic alternating pattern on heart rate variability during sleep in healthy young adults. Clin Neurophysiol 111:99–101
13. Kales A, Rechtschaffen A, Los Angeles University of California, et al. (1968) A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. US National Institute of Neurological Diseases and Blindness, Neurological Information Network, Bethesda, Md
14. Kondo H, Ozone M, Ohki N et al (2014) Association between heart rate variability, blood pressure and autonomic activity in cyclic alternating pattern during sleep. Sleep 37:187–194
15. Kortelainen JM, Mendez MO, Bianchi AM et al (2010) Sleep staging based on signals acquired through bed sensor. IEEE Trans Inf Technol Biomed 14:776–785
16. Kuna ST, Badr MS, Kimoff RJ et al (2011) An official ATS/AASM/ACCP/ERS workshop report: research priorities in ambulatory management of adults with obstructive sleep apnea. Proc Am Thorac Soc 8:1–16
17. Logue EE, Scott ED, Palmieri PA, Dudley P (2014) Sleep duration, quality, or stability and obesity in an urban family medicine center. J Clin Sleep Med 10:177–182
18. Malik M, Cripps T, Farrell T, Camm AJ (1989) Prognostic value of heart rate variability after myocardial infarction. A comparison of different data-processing methods. Med Biol Eng Comput 27:603–611
19. Malik M, Bigger JT, Camm AJ et al (1996) Heart rate variability standards of measurement, physiological interpretation, and clinical use. Eur Heart J 17:354–381
20. Mendez MO, Matteucci M, Castronovo V et al (2010) Sleep staging from heart rate variability: time-varying spectral features and hidden Markov models. Int J Biomed Eng Technol 3:246–263
21. Migliorini M, Bianchi AM, Nisticò D, et al. (2010) Automatic sleep staging based on ballistocardiographic signals recorded through bed sensors. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2010:3273–3276
22. Peng C-K, Havlin S, Stanley HE, Goldberger AL (1995) Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 5:82–87
23. Penzel T, Bunde A, Grote L et al (2000) Heart rate variability during sleep stages in normals and in patients with sleep apnea. Stud Health Technol Inform 77:1256–1260
24. Pizza F, Contardi S, Antognini AB et al (2010) Sleep quality and motor vehicle crashes in adolescents. J Clin Sleep Med 6:41–45
25. Redmond DSJ, de Chazal P, O'Brien C et al (2007) Sleep staging using cardiorespiratory signals. Somnologie Schlafforschung Schlafmed 11:245–256
26. Richman JS, Moorman JR (2000) Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol 278:H2039–H2049
27. Scholz UJ, Bianchi AM, Cerutti S, Kubicki S (1997) Vegetative background of sleep: spectral analysis of the heart rate variability. Physiol Behav 62:1037–1043
28. Sforza E, Pichot V, Barthelemy JC et al (2005) Cardiovascular variability during periodic leg movements: a spectral analysis approach. Clin Neurophysiol 116:1096–1104
29. Stanley N (2005) The physiology of sleep and the impact of ageing. Eur Urol Suppl 3:17–23
30. Terzano MG, Parrino L, Smerieri A et al (2002) Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. Sleep Med 3:187–199
31. Vaughn BV, Quint SR, Messenheimer JA, Robertson KR (1995) Heart period variability in sleep. Electroencephalogr Clin Neurophysiol 94:155–162
32. Vigo DE, Dominguez J, Guinjoan SM et al (2010) Nonlinear analysis of heart rate variability within independent frequency components during the sleep–wake cycle. Auton Neurosci 154:84–88
33. Xiao M, Yan H, Song J et al (2013) Sleep stages classification based on heart rate variability and random forest. Biomed Signal Process Control 8:624–633
34. Zamarrón C, Valdés Cuadrado L, Alvarez-Sala R (2013) Pathophysiologic mechanisms of cardiovascular disease in obstructive sleep apnea syndrome. Pulm Med 2013:521087
35. Zemaitytė D, Varoneckas G, Sokolov E (1984) Heart rhythm control during sleep. Psychophysiology 21:279–289
36. Zhou Z-H, Liu X-Y (2006) Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans Knowl Data Eng 18:63–77