# Cardiorespiratory sleep stage detection using conditional random fields

# Cardiorespiratory Sleep Stage Detection Using Conditional Random Fields

Pedro Fonseca, Niek den Teuling, Xi Long, *Member, IEEE*, and Ronald M. Aarts, *Fellow, IEEE*

***Abstract*—This paper explores the probabilistic properties of sleep stage sequences and transitions to improve the performance of sleep stage detection using cardiorespiratory features. A new classifier, based on conditional random fields, is used in different sleep stage detection tasks (N3, NREM, REM, and wake) in night-time recordings of electrocardiogram and respiratory inductance plethysmography of healthy subjects. Using a dataset of 342 polysomnographic recordings of healthy subjects, among which 135 with regular sleep architecture, it outperforms hidden Markov models and Bayesian linear discriminants in all tasks, achieving an average accuracy of 87.38% and kappa of 0.41 (87.27% and 0.49 for regular subjects) for N3 detection, 78.71% and 0.55 (80.34% and 0.56 for regular subjects) for NREM detection, 88.49% and 0.51 (87.35% and 0.57 for regular subjects) for REM, and 85.69% and 0.51 (90.42% and 0.52 for regular subjects) for wake. In comparison with the state of the art, and having been tested on a much larger dataset, the classifier was found to outperform most of the work reported in the literature for some of the tasks, in particular for subjects with regular sleep architecture. It achieves a comparable accuracy for N3, higher accuracy and kappa for REM, and higher accuracy and comparable kappa for NREM than the best performing classifiers described in the literature.**

***Index Terms*—Cardiac, conditional random fields (CRFs), respiratory, sleep staging.**

## I. Introduction

MANUAL sleep staging is a time-consuming task that requires the help of a sleep technician. Automatic sleep stage classification has been an active area of research for the past decades. It removes the human element from sleep staging, allowing for new applications such as real-time sleep staging (useful for intervention studies) and remote monitoring (in-home sleep studies) and eliminates the problem of inter-scorer variability resulting from manual scoring [1]. Sleep stages are ordinarily measured from electroencephalogram (EEG), but these sensors can disrupt sleep and often require care to apply correctly (i.e., requiring the help of an expert). Cardiorespiratory information provides a promising alternative to EEG for the purpose of sleep staging, with the benefit that it can be measured by unobtrusive methods. Cardiorespiratory-based sleep stage classification has been increasingly studied in recent years. Many studies have reported results on the classification of wake, REM, light sleep, and deep sleep stages [2]–[4], detecting REM sleep [3], [5], [6], or differentiating light and deep sleep stages with an ambulatory device [3], [7]. The results in the literature are promising, but not yet at the level required for reliable sleep staging. Classification is generally done using an extensive set of features. Many different classifiers have been tested over the years, such as linear discriminants (LDs) [6], [8], [9], hidden Markov models (HMMs) [10], and support vector machines [3]. For many classification tasks, the LD classifier was found to be among the best performing. The strength of LD lies in its underlying simplicity, providing a robust model of the features over the different sleep stages. However, LD classification is independent of time, whereas sleep is a structured process where the state and characteristics of each epoch are not independent from each other. Temporal classifiers can make use of this structure, improving classification. It has been shown that the process of sleep can be modeled using a Markov process [11]. The HMM classifier incorporates this Markov assumption, but the model does not handle correlations between features well. Furthermore, the model assumes that features are discriminative during the entire sleep stage. Some features could be indicative of a stage transition instead, but this information is not used by either HMM or LD.

This paper investigates the use of a temporal classifier based on Markov networks, named conditional random fields (CRF) [12] as a new approach to automatic cardiorespiratory sleep stage classification. To the best of the authors' knowledge, this method has not been used in the domain of sleep staging before, other than for a single paper on EEG sleep staging [13] which, given the difference in the sensing modalities, does not allow for a fair comparison. CRF is a generalization of HMM that conditions the model on the given observations. This allows for a more expressive model that can model feature dependences.

## II. Material and Methods

### A. Datasets

The dataset comprises full polysomnographic (PSG) data of 180 subjects from three different databases. The first database, with 327 recordings of 165 subjects (most subjects were monitored for two consecutive nights), was part of the database

Manuscript received November 13, 2015; revised January 15, 2016 and March 7, 2016; accepted March 24, 2016. Date of publication April 4, 2016; date of current version June 29, 2017.

P. Fonseca, X. Long, and R. M. Aarts are with the Department of Electrical Engineering, Eindhoven University of Technology, 5612, AZ, Eindhoven, The Netherlands, and also with Philips Research, 5656, AE, Eindhoven, The Netherlands (e-mail: pedro.fonseca@philips.com; xi.long@philips.com; ronald.m.aarts@philips.com).

N. den Teuling is with Philips Research, 5656, AE, Eindhoven, The Netherlands, and also with the Department of Mathematics and Computer Science, Eindhoven University of Technology, 5612, AZ, Eindhoven, The Netherlands (e-mail: niek.den.teuling@philips.com).

Digital Object Identifier 10.1109/JBHI.2016.2550104

created during the EU Siesta project [14] between 1997 and 2000 in seven different sleep laboratories. The database was restricted to subjects considered healthy (no sleep disorders, no shift work, no depression, usual bedtime before midnight), with a Pittsburgh Sleep Quality Index [15] of at most 5. Sleep stages were scored by consensus agreement between two sleep technicians blind to the condition of the participants in six classes (wake, REM, S1, S2, S3, S4) according to the R&K guidelines [16]. The second database, comprising single-night recordings of six healthy subjects, was collected at the Philips Experience Lab of the High Tech Campus, Eindhoven, The Netherlands, during 2010 (Vitaport 3 PSG, TEMEC). The third database, comprising nine healthy subjects, was collected at the Sleep Health Center, Boston, USA, during 2009 (Alice 5 PSG, Philips Respironics). Electrocardiogram (ECG) was recorded with a Modified Lead II derivation, sampled at 500 Hz (Boston data), 256 Hz (Eindhoven data), and 200 Hz (Siesta data). Respiratory effort was recorded with thoracic respiratory inductance plethysmography (RIP) sampled at 10 Hz (Boston data), 256 Hz (Eindhoven data), and 200 Hz (Siesta data). Subjects with diagnosed sleep disorders (sleep apnea, insomnia, restless legs syndrome, etc.) were excluded from the three data collections. Sleep stages for subjects in the second and third databases were scored by individual sleep technicians blind to the condition of the participants in five classes (wake, REM, N1, N2, N3) according to the AASM guidelines [17]. The protocols for the three data collection studies were reviewed and accepted by the local ethics committees where the experiments took place, and all participants signed informed consent forms. Although all subjects in the three databases were considered healthy, it is reasonable to expect an impact of monitoring on the sleep quality of the subjects. This effect, also called "first night effect"[18] actually often lasts more than the first night and is variable from subject to subject. However, since it is likely to have an impact on the performance of sleep stage classifiers, two sets were created and analyzed separately: the first comprises only recordings which have a minimum percentage of each sleep stage, representative of minimum sleep statistics of healthy adult sleep: at least 5% of deep sleep, 15% of REM sleep, a sleep efficiency of at least 75%, and a minimum of 7 h in bed [19] (set "regular"). This resulted in a total of 135 recordings (101 subjects). The second set comprises all 342 recordings from the three datasets (set "all"). Table I summarizes the subject demographics and sleep statistics of both sets.

TABLE I
DEMOGRAPHICS AND SLEEP STATISTICS OF SUBJECTS IN THE TWO SETS USED IN THE STUDY

| | "regular" | | "all" | |
|---|---|---|---|---|
| Parameter | Mean ± std | Range | Mean ± std | Range |
| N | 101 subjects, 135 recordings | | 180 subjects, 342 recordings | |
| Sex | 57 female subj. (56.44 %) 76 female recs. (56.30 %) | | 98 female subj. (54.44 %) 185 female recs. (54.09 %) | |
| Age (year) | 42.8 ± 16.8 | 20 – 83 | 51.0 ± 19.6 | 20 – 95 |
| BMI (kg/m$^2$) | 23.8 ± 3.1 | 17.2 – 31.3 | 24.6 ± 3.5 | 17.0 – 35.3 |
| TIB (hour) | 8.0 ± 0.4 | 7.2 – 9.6 | 7.9 ± 0.5 | 5.3 – 9.6 |
| SE (%) | 88.9 ± 5.5 | 75.5 – 99.0 | 81.3 ± 11.9 | 21.7 – 99.0 |
| REM (%) | 21.7 ± 3.3 | 16.3 – 31.3 | 18.3 ± 5.7 | 0.0 – 34.8 |
| N3 (%) | 17.2 ± 5.7 | 6.0 – 36.2 | 15.2 ± 8.3 | 0.0 – 42.7 |

"regular" is a subset of the "all" dataset where only subjects with a minimum of 5% N3, 15% REM sleep, and 75% sleep efficiency were considered.
REM and N3 percentages were calculated over the total sleep time for each recording.
BMI: body mass index, TIB: time in bed, SE: sleep efficiency.

TABLE II
RESPIRATORY FEATURES USED IN THE STUDY

| Number | Feature name |
|---|---|
| 1–3 | Respiratory frequency calculated in time and frequency domain and its spectral power [6], [25] |
| 4–7 | PSD VLF, LF, and HF power and LF-to-HF ratio [25] |
| 8–10 | SD of respiratory frequency over 150, 210, and 270 s [6] |
| 11–13 | Mean and SD of breath-by-breath correlations, SD of breath lengths [25] |
| 14-15 | Sample entropy and variance of respiratory effort over 180 s [6], [26] |
| 16-17 | Respiratory dynamic time and frequency warping self-(dis)similarity [27] |
| 18–21 | Standardized mean and median of respiratory peaks and troughs [24] |
| 22-23 | Sample entropy of respiratory peaks and troughs [24] |
| 24-25 | Median peak-to-trough difference and dynamic time warping similarity [24] |
| 26–31 | Median volume and flow rate of breaths, inhalations, and exhalations [24] |
| 32-33 | Ratio of inhalation-to-exhalation volume and flow rate [24] |

PSD, power spectral density; VLF, very low frequency; LF, low frequency; HF, high frequency; SD, standard deviation.
Features 1–7 and 11–17 were computed with windows of 30 s; features 18–33 were computed with windows of 150 s; for the remaining features the window length is indicated.

### *B. Feature Extraction*

The respiratory effort and ECG signals of all subjects were preprocessed before feature extraction. The respiratory effort signal was first resampled to a common sampling rate of 10 Hz. It was then filtered with a tenth-order Butterworth low-pass filter with a cutoff frequency of 0.6 Hz. Baseline was removed by subtracting the median amplitude calculated in a sliding window of 120 s. The baseline wander of the ECG signal was filtered with a linear phase high-pass filter using a Kaiser window of 1.016 s, with a cutoff frequency of 0.8 Hz and a side-lobe attenuation of 30 dB [20]. QRS complexes were detected and localized from the ECG signals using a Hamilton–Tompkins QRS detector [21], [22] followed by a post-processing localization algorithm [23]. The resulting R–R interval time series was resampled using linear interpolation at a sampling rate of 4 Hz.

Since the goal is to classify each 30-s epoch in each recording, all features were automatically extracted using sliding windows centered on each epoch. Tables II–IV give a list of all respiratory, cardiac, and cardiorespiratory features used in this study, together with the window lengths used to compute them and a reference to the original work(s), where these features were first used for different sleep stage classification tasks.

During feature extraction, an automatic process was used to determine whether or not portions of the ECG and RIP signals were adversely affected by noise or motion artifacts. Regarding the ECG signal, each 30-s epoch was separately evaluated in regard to the coverage of R–R intervals. The corresponding cardiac and cardiorespiratory features were only extracted for

TABLE III
CARDIAC FEATURES USED IN THE STUDY

| Number | Feature name |
|---|---|
| 1–3 | Mean HR, mean RR, and detrended mean RR [25], [30] |
| 4–8 | SDNN, RR range, pNN50, RMSSD, and SDSD [30] |
| 9–12 | RR logarithmic VLF, LF, and HF power and LF-to-HF ratio [30], [31] |
| 13–16 | RR mean respiratory frequency and power, max phase and module in HF pole [5] |
| 17–36 | Multiscale sample entropy of RR intervals at length 1 and 2, scales 1–10 over 510 $s^1$ [32] |
| 37–42 | RR DFA, its short, long exponents and all scales, and WDFA over 330 s and PDFA over nonoverlapping segments of 64 heartbeats [33]–[35] |
| 43–46 | Mean absolute difference in HR and RR and in detrended HR and RR [4] |
| 47–56 | RR and HR percentiles (10%, 25%, 50%, 75%, and 90%) [4] |
| 57–66 | Detrended RR and HR percentiles (10%, 25%, 50%, 75%, and 90%) [4] |
| 67 | Sample entropy of symbolic binary changes in RR intervals [36] |
| 68–70 | Power, fourth power and curve length of the ECG [37] |
| 71–73 | Nonlinear energy, Hjorth mobility, and complexity of the ECG [37] |
| 74–77 | Peak power and corresponding frequency, mean, and median ECG PSD [37] |
| 78 | Spectral entropy of the ECG [37] |
| 79 | Hurst exponent of the ECG [38] |
| 80–81 | Short- and long-range phase coordination of R–R intervals in patterns of up to eight consecutive heartbeats [39], [40] |

HR heart rate; RR R–R interval; SDNN standard deviation of RR; pNN50 percentage of successive RR differences $>50$ ms; RMSSD, root mean square of successive RR differences; SDSD, standard deviation of successive RR differences; VLF, very low frequency; LF, low frequency; HF, high frequency; DFA, detrended fluctuation analysis; PDFA, progressive DFA; WDFA, windowed DFA; PSD, power spectral density.
Features 1–16 and 43–67 were computed with windows of 270 s; features 68–79 were computed with windows of 30 s; for the remaining features, the window length is indicated.
[1] The estimation accuracy of sample entropy is lower in series shorter than $10^m$ (where $m$ is the pattern length, in samples) [26], [41]. In practice, this means that this feature will be accurate for all scales with $m = 1$ and for scales below 6 with $m = 2$. The choice of window size was discussed in our earlier work [42].

TABLE IV
CARDIORESPIRATORY FEATURES USED IN THE STUDY

| Number | Feature name |
|---|---|
| 1 | Co-power between RR and respiratory effort over nine epochs [43] |
| 2-3 | Short- and long-range phase coordination between respiration and RR in patterns of up to eight consecutive heartbeats [39], [40] |

RR R–R interval.

a given epoch if the sum of the length of all detected R–R intervals in the window used to extract each feature was equal or larger than 50% of the length of that window. Regarding the RIP signal, peaks and troughs were first detected based on the sign change of the respiratory effort signal slope, and then marked as false detections if 1) the sum of two successive peak-to-trough intervals was less than the median of all intervals or if 2) the peak-to-trough distance was less than 15% of the median of all intervals [24]. Corresponding respiratory features were only extracted for epochs where the window used to extract a feature did not have false peak/trough detections. After automatically rejecting epochs with artifacts, and after extracting each feature from the remaining epochs on each recording, the values of rejected epochs were estimated with linear interpolation between the values of neighboring (valid) epochs. Finally, in order to reduce physiological and equipment-related variations between subjects, all features were normalized in terms of mean and standard deviation (Z-score normalization). This allowed common decision thresholds to be used for all subjects.

### C. Bayesian Linear Discriminant

The Bayesian LD is based on the Bayes decision rule for minimizing the probability of error, i.e., to choose the class that maximizes its posterior probability given an observation (feature vector) $\mathbf{x}$, or, in the case of two classes $a$ and $b$, to choose class $a$ if

$$g_a(\mathbf{x}) - g_b(\mathbf{x}) \geq D \tag{1}$$

where $D$ is a decision threshold and $g_a(\mathbf{x})$ is a so-called discriminant function, $g_a(\mathbf{x}) = \ln P(a|\mathbf{x})$. Using the Bayes rule, and under the assumption that the observations of each class are drawn from multivariate normal distributions and that the covariance matrices of all classes are identical, the discriminant function is given [28] by

$$g_a(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_a)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu_a) + \ln P(a) \tag{2}$$

where $\mu_a$ is the mean feature vector for class $a$, $\Sigma$ is the pooled covariance matrix for all classes, and $P(a)$ is the prior probability of class $a$. Considering the prior probabilities of all classes as constant, the decision boundary $D$ in (1) is given by

$$(\mathbf{x} - \mu_b)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu_b) - (\mathbf{x} - \mu_a)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu_a). \tag{3}$$

It is clear that the further apart the mean vectors $\mu_a$ and $\mu_b$ are in the feature space, i.e., the larger the factor $\mathbf{\Sigma}^{-1}(\mu_a - \mu_b)$ is, the more separable the classes are.

### D. Probabilistic Graphical Models

As explained, the decision rule for Bayesian LDs is based on the estimation of the posterior probabilities $P(a|\mathbf{x})$ of each class, which depends, for classification purposes, solely on the likelihood $P(\mathbf{x}|a)$ and on the prior probability $P(a)$. Sleep is a structured process, and when the body enters a certain sleep stage, it will stay in this stage for a while before transitioning to the next stage [29]. Provided that the body is in a certain stage, the chance of the body being in that stage in the next epoch is generally larger than switching to another stage. This suggests that a classifier can benefit from taking past information into account. Although the LD classifier is adequate for many problems, in the case of sleep, which is a structured process, it might not exploit the full potential of all information available. Consider, for simplicity, a process, illustrated in Fig. 1, comprised of a sequence of states which can have one of two classes $a$ and $b$. Now, consider that the posterior probability of a class $w_i$ for a given state also depends on the class $w_j$ of the previous state and on the likelihood $P(\mathbf{x}_i|w_i)$ of observing $\mathbf{x}_i$ given the class $w_i$:

$$P(w_i|\mathbf{x}_i, w_j) = \frac{P(w_i|w_j)P(\mathbf{x}_i|w_i)}{P(\mathbf{x}_i)}. \tag{4}$$

The class chosen for the current state, $w_i$, should correspond to the class that yields the largest posterior probability. Since the

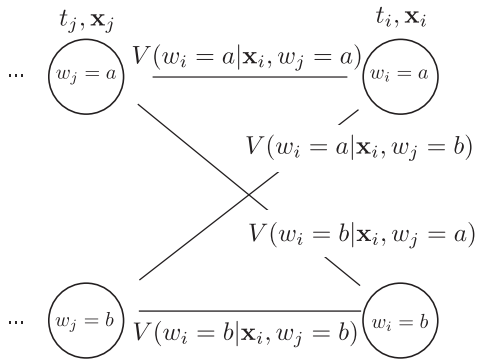


Fig. 1. Trellis diagram for process with two classes.

probability $P(\mathbf{x}_i)$ is irrelevant for the choice of the most likely class $w_i$, we can write

$$P(w_i|\mathbf{x}_i, w_j) \propto P(w_i|w_j)P(\mathbf{x}_i|w_i). \tag{5}$$

In the previous example, it was assumed that $w_j$ is known. However, when presented with a sequence of observations $X$, the previous class is not known, since the likelihood of the model for a given sequence of states might not be optimal on the next observation. The choice of $w_i$, therefore, depends on the previous time point only, but the optimal $w_j$ can only be chosen after the optimal path up to time point $t_i$ is determined, which depends on the observations $\mathbf{x}_i$. Consider a state $w_k$ that happens at time $t_k$, just before $t_j$. In this case, $t_k$ is conditionally independent from $w_i$ and can be fixed depending on the choice of $w_j$. Writing out the dependences, we obtain the recursive relation

$$V(w_i|\mathbf{x}_i, w_j) = P(w_i|w_j)P(\mathbf{x}_i|w_i)\, V(w_j|\mathbf{x}_j, w_k). \tag{6}$$

In the following, for brevity, the choice of a class $a$ for a state $w_i$ will be represented by $i_a$ and the conditions for choosing class $a$ will be derived. The conditions for choosing class $b$ can be derived in an analogous way.

Since the state $w_j$ corresponding to the optimal path until time $t_i$ is not fixed until $\mathbf{x}_i$ is observed, determining the class $w_i$ requires an evaluation of the likelihood of four possible combinations, $V(i_a|\mathbf{x}_i, j_a)$, $V(i_b|\mathbf{x}_i, j_a)$, $V(i_a|\mathbf{x}_i, j_b)$, and $V(i_b|\mathbf{x}_i, j_b)$. The class at time $t_i$ will be given by

$$\arg\max_{w_i} V(w_i|\mathbf{x}_i, w_j) \quad \forall w_i, w_j = a, b. \tag{7}$$

Expanding the previous equation, class $a$ is selected if

$$\begin{aligned} &V(i_a|\mathbf{x}_i, j_a) \geq V(i_b|\mathbf{x}_i, j_a) \\ &\quad \wedge\ V(i_a|\mathbf{x}_i, j_a) \geq V(i_b|\mathbf{x}_i, j_b) \end{aligned} \tag{8}$$

which corresponds to the case where in the optimal path $j = a$, or

$$\begin{aligned} &V(i_a|\mathbf{x}_i, j_b) \geq V(i_b|\mathbf{x}_i, j_a) \\ &\quad \wedge\ V(i_a|\mathbf{x}_i, j_b) \geq V(i_b|\mathbf{x}_i, j_b) \end{aligned} \tag{9}$$

where in the optimal path $j = b$. Using (6) in the previous equations, the conditions become

$$\begin{aligned} &\frac{P(\mathbf{x}_i|i_a)}{P(\mathbf{x}_i|i_b)} \geq \frac{P(i_b|j_a)}{P(i_a|j_a)} \\ &\quad \wedge\ \frac{P(\mathbf{x}_i|i_a)}{P(\mathbf{x}_i|i_b)} \geq \frac{P(i_b|j_b)\, V(j_b|\mathbf{x}_j, w_k)}{P(i_a|j_a)\, V(j_a|\mathbf{x}_j, w_k)} \end{aligned} \tag{10}$$

or

$$\begin{aligned} &\frac{P(\mathbf{x}_i|i_a)}{P(\mathbf{x}_i|i_b)} \geq \frac{P(i_b|j_a)\, V(j_a|\mathbf{x}_j, w_k)}{P(i_a|j_b)\, V(j_b|\mathbf{x}_j, w_k)} \\ &\quad \wedge\ \frac{P(\mathbf{x}_i|i_a)}{P(\mathbf{x}_i|i_b)} \geq \frac{P(i_b|j_b)}{P(i_a|j_b)}. \end{aligned} \tag{11}$$

*1) Transitional Features:* One of the important consequences of using state-dependent posterior probabilities for classification is that the estimation will benefit from features that are only discriminative at state transitions. Consider the (univariate) toy example of Fig. 2(a), where $P(\mathbf{x}_i|a) \approx P(\mathbf{x}_i|b)$ during the majority of intervals. However, after each class transition, the feature is extremely discriminative, with $P(\mathbf{x}_i|a) \gg P(\mathbf{x}_i|b)$ or vice versa. At time point $t_1$ in the figure, and supposing that at that point $V(j_b|\mathbf{x}_j, w_k) > V(j_a|\mathbf{x}_j, w_k)$, the rules (10) and (11) can be applied to determine whether class $a$ is chosen. Assuming that the probability of changing states is smaller than the probability of remaining in the same state and that the probability of staying in the same state is approximately the same for both classes, the first conditions of (10) and of (11) are met and the conditions can be simplified as

$$\frac{P(\mathbf{x}_i|i_a)}{P(\mathbf{x}_i|i_b)} \geq \frac{V(j_b|\mathbf{x}_j, w_k)}{V(j_a|\mathbf{x}_j, w_k)} \vee \frac{P(\mathbf{x}_i|i_a)}{P(\mathbf{x}_i|i_b)} \geq \frac{P(i_b|j_b)}{P(i_a|j_b)}. \tag{12}$$

As long as one of these conditions is satisfied, time point $t_1$ will be classified with $a$, and consequently, $V(i_a|\mathbf{x}_i, w_j) > V(i_b|\mathbf{x}_i, w_j)$. This condition will persist after the transition until the time point $t_2$ in the figure, i.e., as long as $P(\mathbf{x}_i|a) \approx P(\mathbf{x}_i|b)$, (8) will favor the choice of class $a$.

*2) Nonseparable Features:* Another situation where state-dependent posterior probabilities will help classification is in the case of features which are not easy to (linearly) separate. Consider the toy example of Fig. 2(b), and the corresponding histogram for the values of each class in Fig. 2(c). Since there is a large overlap between the histograms of both classes, the classification error with the LD of (3) will be correspondingly large. However, during the intervals of class $a$, for example, between $t_1$ and $t_2$, the feature occasionally has a higher value, uncharacteristic of $b$. These points would be correctly classified using an LD since, here, $P(\mathbf{x}|a) > P(\mathbf{x}|b)$. However, while an LD would make a classification error as soon as the feature value would be lower again, these points will have an important effect on state-dependent factors $V(w_i|\mathbf{x}_i, w_j)$. With the same assumptions as before, class $a$ would be chosen if one of the conditions in (12) is satisfied. However, unlike the previous case, the likelihood $P(\mathbf{x}_i|i_a)$ will never be substantially higher than $P(\mathbf{x}_i|i_b)$, meaning that the right condition of (12) will likely not be satisfied. Each time $P(\mathbf{x}_i|i_a)$ is slightly larger than $P(\mathbf{x}_i|i_b)$, the ratio $V(j_b|\mathbf{x}_j, w_k)/\, V(j_a|\mathbf{x}_j, w_k)$ will be-

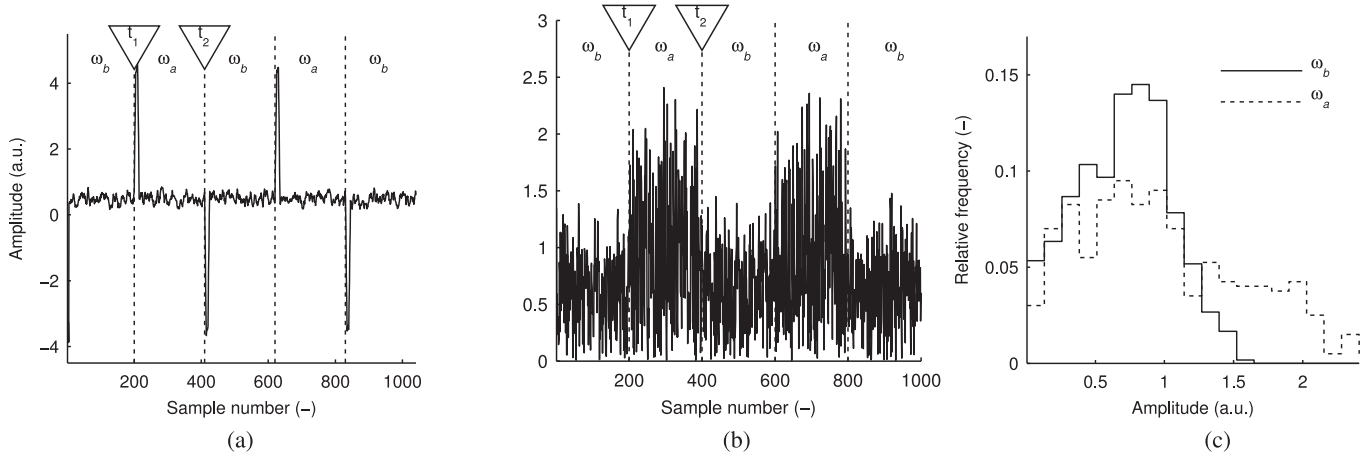


Fig. 2. Toy example of a feature which is (a) only discriminative at the transition between states and (b) difficult to separate linearly. (c) Histogram of the feature in (b).

come smaller, eventually allowing the left condition of (12) to be satisfied and class $a$ to be chosen.

*3) Hidden Markov Models:* The framework introduced by the Bayesian LD is not the most appropriate to model dependence on previous states. Bayesian networks, represented by directed acyclic graphs, however, are very well suited for this purpose. These graphs represent dependences between random variables using directed edges. The joint probability of the variables $V = \{v_1, \ldots, v_N\}$ in a Bayesian network is given by

$$P(V) = \prod_{i=1}^{N} P\left(v_i | \, pa(v_i)\right) \tag{13}$$

where $pa(v)$ denotes the parent nodes of $v$. In the problem of sleep stage detection, these dependences can be simplified by allowing the current state to depend on the previous state and on the observations on the current state. In this case, the network corresponds to an HMM where the class of each state, $w_t$, remains hidden, and the only observable variables are the feature vectors, $\mathbf{x}_t$, which depend on them [44]. The joint probability of the variables in this network is given by

$$P(W, X) = P(w_1)P(\mathbf{x}_1|w_1) \prod_{t=2}^{T} P(w_t|w_{t-1})P(\mathbf{x}_t|w_t) \tag{14}$$

where $P(w_t|w_{t-1})$ represents the probability of transitioning from one state to the following, $P(\mathbf{x}_t|w_t)$ the probability that observation $\mathbf{x}_t$ was generated by $w_t$, and $P(w_1)$ is the initial state probability, i.e., the initial belief about the starting state. For a given sequence of observations $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, the corresponding sequence of states $W = \{w_1, \ldots, w_T\}$ can be estimated as the sequence of states $\hat{w}_t$ that maximizes the conditional probability $P(w|X)$:

$$\hat{w}_t = \arg\max_{w_t} P(w_t|X). \tag{15}$$

Since $\max P(w_t|X) \propto \max P(w_t, X)$, the states can be estimated as the sequence which maximizes the joint probability of the whole chain. This can be efficiently computed with the Viterbi algorithm [45] where the most likely sequence of states is progressively determined as the sequence that maximizes the joint probability up to each state

$$V_t(w_t) = \max_{w_{1,\ldots,t-1}} P\left(w_{1,\ldots,t}, X_{1,\ldots,t}\right), \text{ for } 2 \leq t \leq T \tag{16}$$

which can be factorized and recursively computed as

$$V_t(w_t) = \max_{w_{t-1}} \left[P(X_t|w_t)P(w_t|w_{t-1})V_{t-1}(w_{t-1})\right] \tag{17}$$

with the tail function $V_1(w_t)$ given by

$$V_1(w_1) = \max_{w_1} \left[P(X_1|w_1)P(w_1)\right]. \tag{18}$$

The most probable class in state $t$, $\hat{w}_t$, will finally be given by

$$\hat{w}_t = \arg\max_{w_t} V_t(w_t). \tag{19}$$

Note that the classification rule in (19) is equivalent to the rule in (7). An HMM, decoded with the Viterbi algorithm, will therefore benefit from the properties described in the previous sections.

*4) Conditional Random Fields:* In generative models such as HMMs, the parameters are learned by maximizing the joint probability distribution $P(W, X)$, which in turn requires the distribution of the observations, $P(X)$, to be modeled or somehow learned from the data. When the features of the observed variable $X$ are not independent, the joint distribution may be extremely difficult to model, requiring either large amounts of training data, or strong assumptions about the variables. Discriminative models, such as CRF [12], avoid this problem by computing the probability $P(y|\mathbf{x})$ of a possible output $y = (y_1, y_2, \ldots, y_n)$ given an observation $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, avoiding the explicit modeling of the marginal $P(X)$. By simplifying the modeling problem and not requiring any assumption about the independence of the features (only about the states), discriminative models make better use of correlated interdependent features, which are common in the case of sleep stage detection using cardiorespiratory features. CRFs are a special case of undirected graphs, which are globally conditioned on the observation $X$. Parameter learning and inference is usually performed by means of factor graphs [46], a type of model which describes the probability distribution of

the network using non-negative factors to express interaction between random variables. The joint distribution can be factorized over all maximal cliques

$$P(V) = \frac{1}{Z} \prod_{c \in C} \Psi_c(v_c) \tag{20}$$

where the potential functions $\Psi_c$ are the factor potentials of a variable $v_c$ in a clique $c$, and $Z$ is a normalization function, needed since the potential functions can be any arbitrary function,

$$Z = \sum_{v} \prod_{c \in C} \Psi_c(v_c). \tag{21}$$

Because they are strictly positive, the joint distribution can be described by a log-linear model [47]

$$P(V) = \frac{1}{Z} \exp\left(\sum_{c \in C} \lambda_c f_c(v_c)\right) \tag{22}$$

$$Z = \sum_{v \in V} \exp\left(\sum_{c \in C} \lambda_c f_c(v_c)\right) \tag{23}$$

where the set of parameters $\lambda_c \in \mathbb{R}_0^+$ is selected to maximize the model fit. Defining a set of $K$ (with $K = S^2 + S$ and $S$ as the number of classes) parameters $\Theta = \{\lambda_{s,r}, \lambda_v\}$ and feature functions $f_k(y_t, y_{t-1}, \mathbf{x}_t)$ which incorporates transition functions and state-observation functions

$$f_k(y_t, y_{t-1}, \mathbf{x}_t) = \begin{cases} 1_{\{y_t = s\}} 1_{\{y_{t-1} = r\}}, & \text{for each } (s, r) \\ 1_{\{y_t = s\}} \mathbf{x}_t, & \text{for each } (s, x) \end{cases} \tag{24}$$

where the notation $1_{\{\text{condition}\}}$ has a value of 1 if the condition is true, and 0 otherwise; the joint probability in (22) can be simplified as

$$P(X, Y) = \frac{1}{Z} \exp\left(\sum_{t}^{T} \sum_{k}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right) \tag{25}$$

which corresponds to the general log-linear model of (22) and (23). Using Bayes' rule, the conditional probability $P(Y|X)$ can be obtained as

$$P(Y|X) = \frac{\frac{1}{Z} \exp\left(\sum_{t}^{T} \sum_{k}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right)}{\frac{1}{Z} \sum_{y} \exp\left(\sum_{t}^{T} \sum_{k}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right)} \tag{26}$$

which is equivalent to

$$P(Y|X) = \frac{1}{Z(X)} \prod_{t}^{T} \exp\left(\sum_{k}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right) \tag{27}$$

with

$$Z(X) = \sum_{y} \prod_{t}^{T} \exp\left(\sum_{k}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right). \tag{28}$$

These equations correspond to the general form of a CRF [see (20) and (21)] with the potentials $\Psi_t$ given by the function

$$\Psi_t = \exp\left(\sum_{k}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right). \tag{29}$$

It should be mentioned that the set of observations used at time point $t$ and denoted by $\mathbf{x}_t$ can actually contain observations from different points in time as well. In this study, and to allow a fair comparison with LD where the posterior probabilities of each class at a given time are determined by the feature vector (observations) at that time, only features observed at time point $t$ were used. The parameters $\theta = \lambda_{s,r}, \lambda_v$ need to be estimated such that the model has a good fit with the training dataset. Using $N$ input sequences (corresponding to state and observation sequences from $N$ different subjects), the model fit is expressed using a conditional log-likelihood [48]

$$L(\theta) = \sum_{i=1}^{N} \log P\left(Y^i|X^i\right) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k\left(y_t^i, y_{t-1}^i, \mathbf{x}_t^i\right) - \sum_{i=1}^{N} \log Z\left(X^i\right) \tag{30}$$

where the superscript $i$ indicates the $i$th state-observation sequence pair. The set of parameters $\theta$ that maximizes $L(\theta)$ can be found using numerical optimization techniques such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [49]. The decoding of a linear-chain CRF is done in a similar way as HMM, using a modified version of the Viterbi algorithm [48]. The goal is still to estimate the sequence of states that maximizes the conditional probability $P(w|X)$, as in (15). The relation $\max\ P(w|X) \propto \max\ P(w, X)$ still holds, and the estimation is equivalent to estimating the maximum $V_t\,(w_t)$:

$$\begin{aligned} V_t\,(w_t) &= \max_{w_{1,\ldots,t-1}} P\,(w_{1,\ldots,t}, \mathbf{x}_{1,\ldots,t}) \\ &= \max_{w_{1,\ldots,t-1}} \exp\left(\sum_{i}^{t} \sum_{k}^{K} \lambda_k f_k\,(w_i, w_{i-1}, \mathbf{x}_t)\right) \\ &= \max_{w_{1,\ldots,t-1}} \prod_{i}^{t} \Psi_t\,(w_i, w_{i-1}, \mathbf{x}_i). \end{aligned} \tag{31}$$

Factorizing the equation and rewriting it recursively, we get

$$V_t\,(w_t) = \max_{w_{t-1}} \Psi_t\,(w_t, w_{t-1}, \mathbf{x}_t)\, V_{t-1}\,(w_{t-1}) \tag{32}$$

with the tail function $V_1(w_1)$ given by

$$V_1\,(w_1) = \max_{w_1} \Psi_t\,(w_1, \mathbf{x}_1)\,. \tag{33}$$

CRF was used to classify the toy example of Fig. 2(a), and Fig. 3(a) illustrates the results. It is clear that the score obtained in the case of a feature with transitional properties matches almost perfectly the class annotations. This illustrates how such features can be exploited by CRF, but not by LD. Regarding the CRF classification of the toy example of Fig. 2(b), illustrated in Fig. 3(b), the score obtained in the case of a feature which is not easily separable also behaves as expected. Despite the large number epochs with ambiguous values, which would cause LD to erroneously classify many instances, as more and more unambiguous feature values are found, the classification converges to the correct class. Apart from a delay in the transition between detected classes, the classification remains correct until the end of that interval. Besides being able to respond to transitional

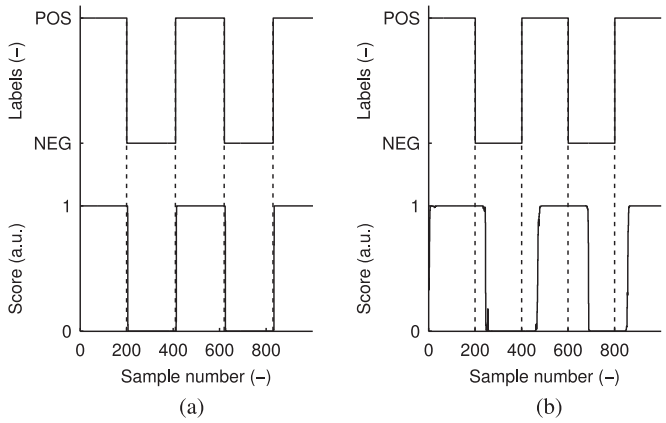


Fig. 3. CRF scores and corresponding classes for toy example of (a) Fig. 2(a) and (b) Fig. 2(b). In both cases, the positive label corresponds to class $w_b$ and the negative label to class $w_a$. A score closer to 1 indicates a higher posterior probability for class $w_b$. "a.u." stands for "arbitrary units."

features, CRF is thus also adequate to classify sequences where states remain stable for a certain amount of time, such as sleep.

### E. Training and Evaluation

In order to compare the performance of the LD, HMM, and CRF classifiers, four separate noncomplementary detection tasks were considered: N3, NREM, REM, and wake. For the N3 detection task, S3 and S4 from the Siesta data were merged into a single N3 class; for the NREM detection task, S1, S2, S3, and S4 from the Siesta data and N1, N2, and N3 from the Boston and Eindhoven sets were merged into a single NREM class; and for the REM and wake tasks, the corresponding classes in each data set were used. Each of these classes was considered in a "one versus rest" setting, and for each task, a tenfold cross-validation scheme was used. Furthermore, to guarantee that the validation gives, as much as possible, an unbiased estimate of subject-independent classification, care was taken to guarantee that two recordings of the same subject were part of the same fold. To allow a paired comparison, the same folds were used to validate each classifier.

*1) LD Classifier:* There is a large redundancy in the extensive set of cardiac and respiratory features used in this study. Since the LD classifier is particularly sensitive to the presence of redundant and, more importantly, nondiscriminative features, the correlation feature selection algorithm [50] was used to restrict the features to a set which maximizes their discriminative power, while minimizing redundancy between them. Feature selection was performed on each iteration of the cross-validation procedure to avoid biasing the validation performance. The classification score is obtained as the difference between the discriminant functions of the positive class and the remaining classes [the left-hand side of (1)].

*2) HMM Classifier:* A discrete HMM classifier [44] was used for sleep stage detection. In order to estimate the emission probabilities, a mapping between the feature space and a discrete set of symbols was first defined. This was achieved with $k$-means clustering, where the number of clusters and the centroid of each cluster were determined during the training step of each cross-validation iteration. The number of clusters was automatically determined for each classification task by computing, after $k$-means clustering for a varying value of $k$, the value that maximizes the normalized mutual information between the labels (class) of each feature vector and the corresponding closest clusters [51]

$$\arg\max_k \mathrm{NMI}\left(\lambda^w, \lambda_K^k\right) \tag{34}$$

with

$$\mathrm{NMI}\left(\lambda^w, \lambda_K^k\right) = \frac{\sum_{i=1}^{W}\sum_{j=1}^{K} n_{i,j}\log\left(\frac{n\cdot n_{i,j}}{n_i^w n_j^k}\right)}{\sqrt{\left(\sum_{i=1}^{W} n_i^w \log\left(\frac{n_i^w}{n}\right)\right)\left(\sum_{j=1}^{K} n_j^k \log\left(\frac{n_j^k}{n}\right)\right)}} \tag{35}$$

where $\lambda^w$ is the mapping between each feature vector and its class, and $\lambda_K^k$ is the mapping between each feature vector and the closest cluster after $k$-means clustering with $K$ clusters, $n_i^w$ is the number of feature vectors belonging to class $i$ according to $\lambda^w$, $n_j^k$ is the number of feature vectors assigned to cluster $j$ according to $\lambda_K^k$, and $n_{i,j}$ is the number of feature vectors with class $i$ and in cluster $j$. During classification, the cluster index of each feature vector is determined by finding the closest cluster centroid. The index of this cluster is used as input to the Viterbi algorithm to obtain a score expressing the posterior probability of the positive class according to (17).

*3) CRF Classifier:* The CRF classifier was trained with the BFGS algorithm, and classification was performed with the modified version of the Viterbi algorithm, yielding a score given by (32) for each epoch which can be interpreted as the posterior probability of the positive class in that epoch.

*4) Classification Performance:* In order to compare the classification performance of each classifier, the scores obtained for each test subject in each iteration of the cross-validation procedure were collected and aggregated (pooled). The precision–recall (PR) curve–which plots the positive predictive value (PPV) against the true positive rate (TPR)–and the receiver operating characteristic (ROC)–which plots the TPR versus the false positive rate (FPR)–were computed for a varying threshold on the scores output by each classifier. Although the PR curve depends on the priors of each class, it is most useful in the case of heavily imbalanced classes, such as when detecting wake (in average, 18.7% of all epochs in our "all" dataset), N3 (12.4% of all epochs), or REM (15.2% of all epochs). In these cases, this curve allows us to easily evaluate whether an increase in true positives comes at a disproportionate cost in false positives. The threshold leading to the maximum pooled Cohen's kappa coefficient of agreement [52] was then computed. Based on this threshold, the kappa coefficient for each subject was computed. Note that since this threshold was selected based on the pooled kappa, it will not necessarily correspond to the maximum kappa coefficient for each subject. Significance was tested with a two-tailed Wilcoxon signed-rank test [53] for each evaluation metric.

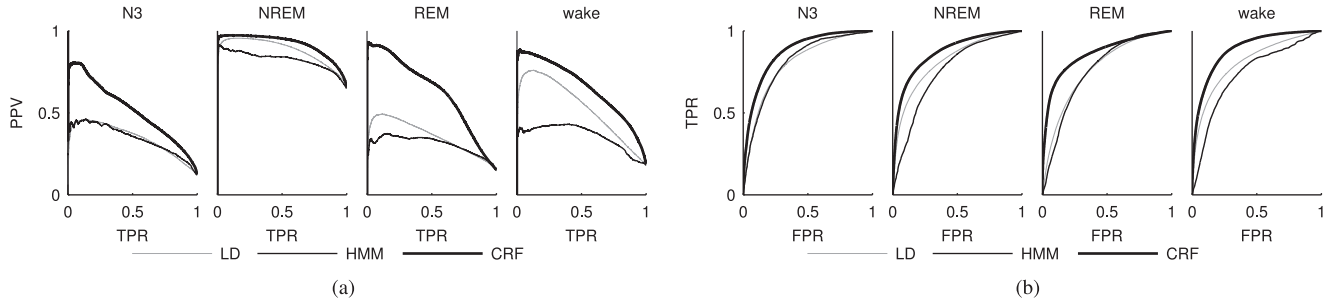


Fig. 4. Pooled (a) PR and (b) ROC curves per classifier and classification task, for dataset "all."

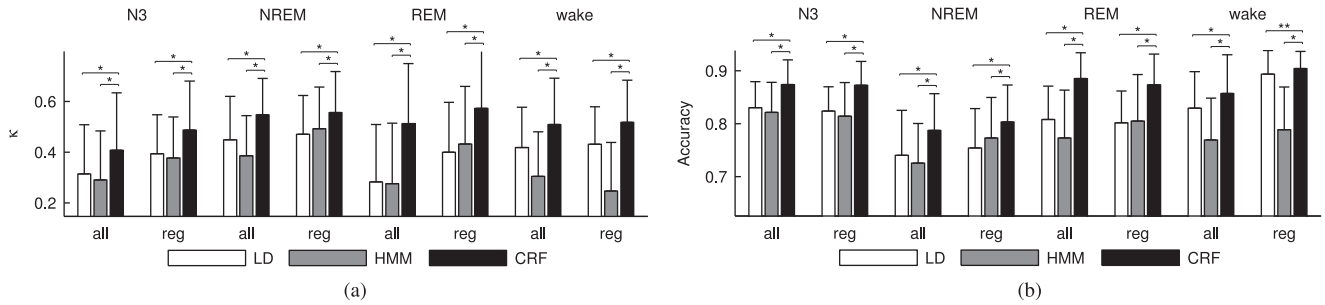


Fig. 5. Mean (a) Cohen's kappa coefficient of agreement and (b) accuracy, per classifier, per detection task for both datasets. * indicates significantly higher performance ($p < 0.001$), after a two-tailed Wilcoxon signed-rank test [53].

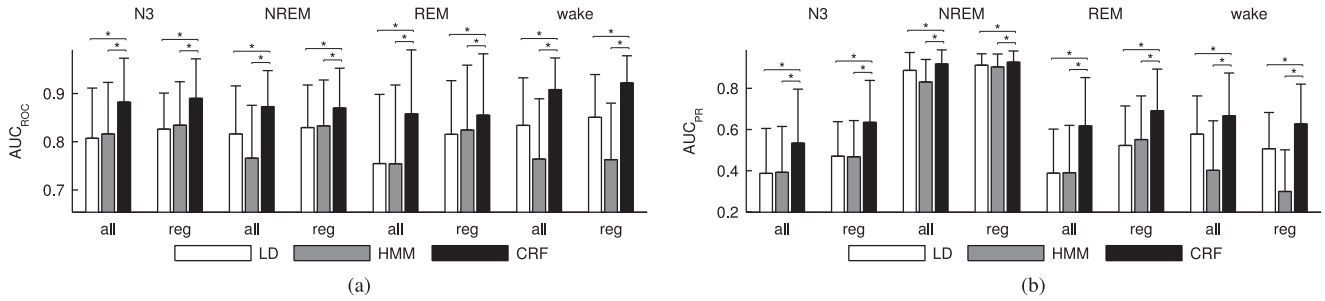


Fig. 6. Mean area under the (a) PR and (b) ROC curves, per classifier, per detection task for both datasets. * indicates significantly higher performance ($p < 0.001$), after a two-tailed Wilcoxon signed-rank test [53].

## III. Results and Discussion

Fig. 4 compares the pooled PR and ROC curves obtained with each classifier, for each detection task in the "all" dataset. In all detection tasks, the CRF classifier outperforms the other classifiers over the entire solution space.

Fig. 5 compares the average kappa coefficient and accuracy obtained with each classifier for the different classification tasks in both datasets. The performance of the CRF classifier is significantly higher than both the HMM and the LD classifiers in all tasks. The performance in the "regular" dataset is also higher than in the "all" dataset, reflecting the more regular sleep structure of those subjects. As illustrated in Fig. 6, the area under the curve (AUC) for the PR and ROC curves is also highest with the CRF classifier. Using a Wilcoxon rank sum test we found no significant differences ($p > 0.05$) in CRF performance for any task between the subjects of each of the three databases comprising the "all" dataset. Regarding the differences in performance improvement, particularly apparent in the PR curves, it is important to analyze both PR and ROC curves in context: while it seems that the improvement in PPV for NREM is much smaller than for N3 and REM, it is worth reminding that the prior probability of NREM is much higher than of the remaining classes, and that in this case, the PR curve will always seem optimistic. In this case, it is worth noticing that the improvement in AUC for the ROC curve of NREM is in line with the improvements obtained for the other classes.

To give some examples of CRF classification, Figs. 7 and 8 illustrate the reference hypnogram and corresponding results obtained for the subject closest to the average kappa performance (across all tasks) and for the subject with the best average performance, respectively. Regarding wake detection, we can see in Fig. 7 that some of the false positives occur during REM periods, most notably during the second half of the night. This is likely related to the increase in sympathetic activity during phasic REM [54], which shares some characteristics with a wake state. Regarding N3 detection, it is interesting to note, also in Fig. 7, that the first (shorter) periods of N3 are incorrectly classified. Paying closer attention to the posterior probabilities for this

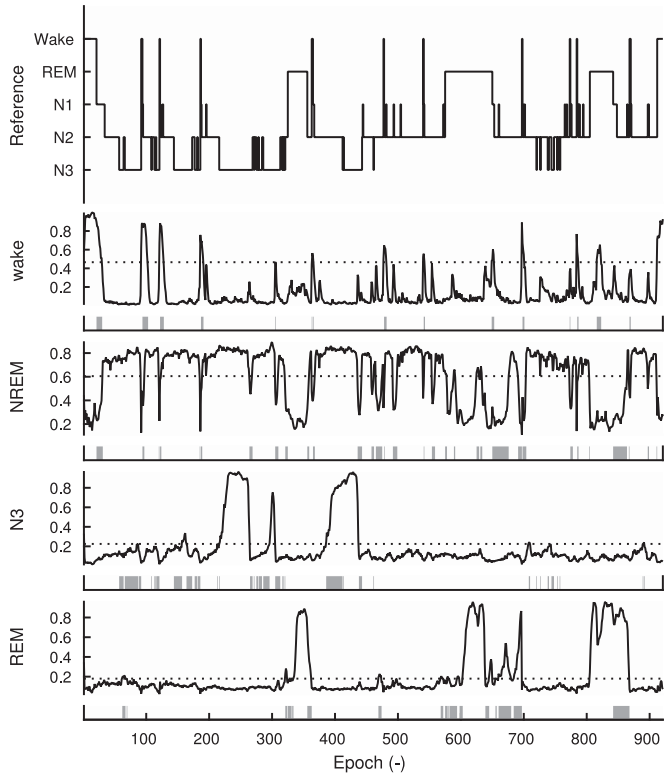


Fig. 7. Example reference hypnogram (top) and classification results (remaining plots) for the subject closest to the average performance (averaged over all classification tasks). For each classification task, each plot illustrates the posterior probability obtained with CRF (solid line), the classification threshold (dashed line), and epochs for which there was a classification error (gray bars). The kappa values obtained for this subject were 0.55 (wake), 0.58 (NREM), 0.45 (N3), and 0.56 (REM).

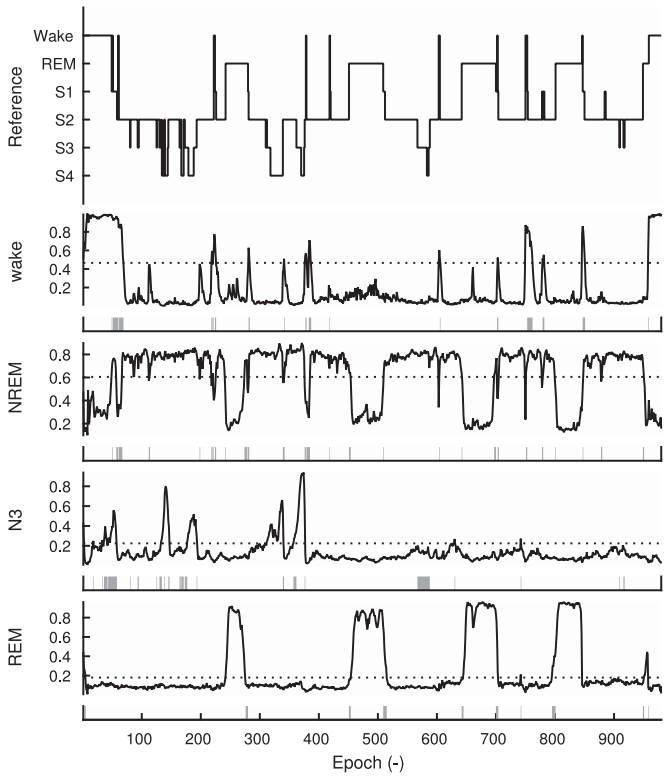


Fig. 8. Example reference hypnogram (top) and classification results (remaining plots) for the subject with the best average results (across all tasks). The kappa values for this subject were 0.72 (wake), 0.86 (NREM), 0.58 (N3), and 0.90 (REM).

class, we see a gradual increase which comes close to the detection threshold; although the periods were misclassified, this suggests that if the periods were longer, even for such a difficult detection of N3 the classification would probably eventually converge to the correct class, as happened for the correctly detected period after epoch 200. Finally, regarding REM detection in Fig. 7, it is interesting to note the missed REM detections between epochs 550 and 600: looking at the posterior probabilities of NREM in these epochs, it is clear that the classifier is confusing many epochs of this REM period with NREM. This is likely due to the presence of intervals of tonic REM, which are parasympathetically driven [54] and, therefore, more difficult to distinguish from NREM.

A direct comparison between the results obtained with the CRF classifier proposed in this paper and previous work in this area is not easy. With the exception of wake and deep sleep detection, almost no literature was found on binary classification. An additional factor which makes the comparison difficult is related to the large between-subject variations, where small datasets might not be representative enough of the problem at hand. Nevertheless, in order to compare the performance of the CRF classifier with the state-of-the-art in this area, the problem was restricted to binary classification using the same modalities described in the literature (ECG and/or respiratory effort). In cases where no "one versus rest" classifiers were found, binary classifiers were chosen which had the target class as one of the classes. For example when literature described NREM versus REM, we analyzed separately the performance of NREM and REM detections. Based on this selection, classifiers with the best reported kappa coefficient and with the best accuracy were chosen for comparison and are summarized in Table V.

Regarding deep sleep detection, the CRF classifier achieves a lower kappa coefficient but comparable accuracy than the LD classifier described in our recent work [55] for both datasets. It is relevant to note that in that work we used the same set of cardiorespiratory features and a similar LD classifier on a comparable dataset. However, that work proposed the use of spline fitting to reduce the effect of body motion artifacts and within-subject variability and more importantly, the use of a time delay of 5 min in the classification. It is interesting to note that even without these two critical elements, the CRF classifier still achieves a comparable accuracy in the same task. To provide a fair comparison, we also compared the CRF classifier with the results reported in [55] when spline fitting and time delay were not used. In that case, it is clear that CRF achieves a higher accuracy and kappa for the "regular" dataset, suggesting its superiority over LD for the same task, using the same base features. Regarding NREM detection, the CRF classifier achieves a comparable kappa and accuracy as the HMM classifier of Mendez *et al.* [5] although they use a much smaller dataset limited to a narrow age range between 40 and 50 years. Regarding REM detection, CRF achieves a comparable kappa coefficient for the "regular" dataset, and a lower coefficient for

TABLE V
COMPARISON WITH CLASSIFIERS REPORTED IN THE LITERATURE

| | Author | N | Modalities | Acc. (%) | kappa (-) |
|---|---|---|---|---|---|
| N3 | Long [55][1†] | 325 | ECG RE | 88.9 | 0.51 |
| | Long [55][*] | 257 | ECG RE | 86.1 | 0.45 |
| | CRF ("regular") | 135 | ECG RE | 87.27 (4.49) | 0.49 (0.19) |
| | CRF ("all") | 342 | ECG RE | 87.38 (4.64) | 0.41 (0.23) |
| NREM | Mendez [5][2†] | 12 | ECG | 79.3 | 0.56 |
| | CRF ("regular") | 135 | ECG RE | 80.34 (6.97) | 0.56 (0.16) |
| | CRF ("all") | 342 | ECG RE | 78.71 (6.93) | 0.55 (0.14) |
| REM | Mendez [5][2†] | 12 | ECG | 79.3 | 0.56 |
| | CRF ("regular") | 135 | ECG RE | 87.35 (5.79) | 0.57 (0.22) |
| | CRF ("all") | 342 | ECG RE | 88.49 (4.91) | 0.51 (0.24) |
| wake | Redmond [6] [3‡] | 31 | ECG RE | 89 | 0.6 |
| | Long [27] [3§] | 15 | RE | 94.2 | 0.58 |
| | CRF ("regular") | 135 | ECG RE | 90.42 (3.23) | 0.52 (0.17) |
| | CRF ("all") | 342 | ECG RE | 85.69 (7.32) | 0.51 (0.18) |

Comparison with best studies reported in the literature, restricted to classifiers which use the same modalities. The indicated accuracy and kappa correspond to the average, and between parenthesis (when available), the standard deviation.
Original classification task: [1] N3/other, [2] NREM/REM, [3] wake/sleep.
Classifier in the literature with the highest [§]kappa coefficient and accuracy, [‡]kappa coefficient, [§]accuracy.
[*] Same classifier as the best reported for N3 classification [55], but features are not smoothed with spline fitting and no time delay is used. Reported results restricted to subjects with more than 30 min of N3.

the "all" dataset. In both cases, it achieves a higher accuracy. Finally, regarding wake detection, the classifier achieves a lower kappa coefficient but higher accuracy (for the 'regular' dataset) than the classifier of Redmond *et al.* [6], and a lower kappa coefficient and accuracy than the classifier of Long *et al.* [27]. Regarding the latter, it is interesting to note that the set of features used was the same as in this work and that the dataset used in that work was also a subset of the "regular" dataset described here. Since the LD classifier used in this work is in all similar to the one used in that work, and that CRF outperformed LD for this data set, we speculate that the decrease in performance could come from the characteristics of the remaining subjects, arguably more difficult from a classification point of view. It should be noted that for wake/sleep classification there are studies which report higher performance [8]. However, they were excluded from this comparison since they use actigraphy, which remains the single most discriminative feature for this task.

Finally, it is important to highlight some of the limitations of this study. First, and despite the use of a large dataset with a wide range of ages, none of the subjects suffered from sleep disorders. This has obvious limitations in a clinical context for which disordered patients are of particular interest. Second, the classification tasks evaluated in this study were binary. Although it is clear from the examples in Figs. 7 and 8 that the output of each binary task is somewhat complementary, this technique will be most useful when the classifier is extended to a multiclass task. An additional note should be made in regard to the use of a discrete HMM classifier. Given the transitional nature of sleep stages (in particular during NREM sleep), it is possible that continuous HMM, modeled for example using Gaussian mixture models, could be better suited for the task at hand. Nevertheless, despite these limitations, this study demonstrates the superiority of CRF for sleep stage classification over the more popular LD and discrete HMM classifiers. Since the CRF framework presented is not limited to binary classification, the extension to multiple stage classification will be addressed in future work.

## IV. CONCLUSION

The probabilistic properties of sleep stage sequences and transitions are explored to improve the performance of sleep stage detection using cardiorespiratory features. A new classifier, based on CRFs, is compared with a classifier based on HMMs and a Bayesian LD for different sleep stage detection tasks: N3, REM, NREM, and wake. The CRF classifier was evaluated using cross-validation on a large dataset comprising 342 PSG recordings of healthy subjects. It was found to significantly outperform the other classifiers in all performance metrics, suggesting the adequacy of this classifier for the problem of sleep stage detection. In addition, it was tested on a larger dataset than most state-of-the-art work reported in the literature and, therefore, with a wider range of population characteristics. It achieves, in particular for subjects with regular sleep characteristics, comparable accuracy for N3, higher accuracy and kappa for REM, and higher accuracy and comparable kappa for NREM. Future work will explore the postprocessing of cardiorespiratory features used in classification, which has proven useful in our earlier work in N3 classification and the use of the CRF classifier in multistage classification problems and in recordings of subjects with sleep disorders such as insomnia and sleep-disordered breathing.

## ACKNOWLEDGMENT

The authors would like to thank Dr. J. Vanschoren and Dr. J. Karel for their critical review of this manuscript and for their valuable suggestions.

## REFERENCES

[1] N. Punjabi, N. Shifa, G. Doffner, S. Patil, G. Pien, and R. Aurora, "Computer-assisted automated scoring of polysomnograms using the somnolyser system," *Sleep*, vol. 38, no. 10, pp. 1555–1566, 2015.
[2] A. Noviyanto and A. M. Arymurthy, "Sleep stages classification based on temporal pattern recognition in neural network approach," in *Proc. Int. Joint Conf. Neural Netw.*, 2012, pp. 1–6.
[3] T. Willemen *et al.*, "An evaluation of cardio-respiratory and movement features with respect to sleep stage classification," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 2, pp. 661–669, Sep. 2014.
[4] B. Yılmaz, M. H. Asyalı, E. Arıkan, S. Yetkin, and F. Özgen, "Sleep stage and obstructive apneaic epoch classification using single-lead ECG," *Biomed. Eng. Online*, vol. 9, p. 39, Jan. 2010.
[5] M. O. Mendez, M. Matteucci, V. Castronovo, L. Ferini-Strambi, S. Cerutti, and A. M. Bianchi, "Sleep staging from Heart Rate Variability: Time-varying spectral features and hidden Markov models," *Int. J. Biomed. Eng. Technol.*, vol. 3, no. 3, pp. 246–63, 2010.
[6] S. J. Redmond, P. de Chazal, C. O'Brien, S. Ryan, W. T. McNicholas, and C. Heneghan, "Sleep staging using cardiorespiratory signals," *Somnologie-Schlafforschung und Schlafmedizin*, vol. 11, no. 4, pp. 245–256, Oct. 2007.
[7] M. Bresler, K. Sheffy, G. Pillar, M. Preiszler, and S. Herscovici, "Differentiating between light and deep sleep stages using an ambulatory device based on peripheral arterial tonometry," *Physiol. Meas.*, vol. 29, no. 5, pp. 571–84, May 2008.
[8] S. Devot, R. Dratwa, and E. Naujokat, "Sleep/wake detection based on cardiorespiratory signals and actigraphy," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 2010, Jan. 2010, pp. 5089–5092.
[9] M. Migliorini, A. M. Bianchi, D. Nisticò, J. Kortelainen, E. Arce-Santana, S. Cerutti, and M. O. Mendez, "Automatic sleep staging based on ballisto-

cardiographic signals recorded through bed sensors," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Jan. 2010, vol. 2010, pp. 3273–3276.
[10] J. M. Kortelainen, M. O. Mendez, A. M. Bianchi, M. Matteucci, and S. Cerutti, "Sleep staging based on signals acquired through bed sensor," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 3, pp. 776–85, May 2010.
[11] J. W. Kim, J.-S. S. Lee, P. A. Robinson, and D.-U. U. Jeong, "Markov analysis of sleep dynamics," *Phys. Rev. Lett.*, vol. 102, no. 17, p. 178104, May 2009.
[12] J. Lafferty, A. Mccallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learning*, 2001, vol. 2001, pp. 282–289.
[13] G. Luo and W. Min, "Subject-adaptive real-time sleep stage classification based on conditional random field," in *Proc. AMIA Annu. Symp. Proc.*, Jan. 2007, vol. 2007, p. 488.
[14] G. Klosch *et al.*, "The SIESTA project polygraphic and clinical database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 51–57, May/Jun. 2001.
[15] D. J. Buysse, C. F. Reynolds, T. H. Monk, S. R. Berman, and D. J. Kupfer, "The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research," *Psychiatry Res.*, vol. 28, no. 2, pp. 193–213, 1989.
[16] A. Rechtschaffen and A. Kales, *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Washington, DC, USA: U.S. Government Printing Office, U.S. Public Health Service, 1968.
[17] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Darien, IL, USA: Amer. Acad. Sleep Med., 2007.
[18] H. W. Agnew, W. B. Webb, and R. L. Williams, "The first night effect: An EEG study of sleep," *Psychophysiology*, vol. 2, no. 3, pp. 263–266, Jan. 1966.
[19] M. M. Ohayon, M. A. Carskadon, C. Guilleminault, and M. V. Vitiello, "Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: Developing normative sleep values across the human lifespan," *Sleep*, vol. 27, no. 7, pp. 1255–1273, Nov. 2004.
[20] J. A. van Alsté, W. van Eck, and O. E. Herrmann, "ECG baseline wander reduction using linear phase," *Comput. Biomed. Res.*, vol. 19, no. 5, pp. 417–427, Oct. 1986.
[21] P. Hamilton, "Open source ECG analysis," in *Proc. Comput. Cardiol. Conf.*, Sep. 2002, pp. 101–4.
[22] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database," *IEEE Trans. Biomed. Eng.*, vol. BE-33, no. 12, pp. 1157–65, Dec. 1986.
[23] P. Fonseca, R. M. Aarts, J. Foussier, and X. Long, "A novel low-complexity post-processing algorithm for precise QRS localization," *SpringerPlus*, vol. 3, no. 1, p. 376, 2014.
[24] X. Long, J. Foussier, P. Fonseca, R. Haakma, and R. M. Aarts, "Analyzing respiratory effort amplitude for automated sleep stage classification," *Biomed. Signal Process. Control*, vol. 14, pp. 197–205, 2014.
[25] S. J. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 485–496, Mar. 2006.
[26] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Am. J. Physiol. Heart Circulatory Physiol.*, vol. 278, no. 6, pp. H2039–H2049, 2000.
[27] X. Long, P. Fonseca, J. Foussier, R. Haakma, and R. M. Aarts, "Sleep and wake classification with actigraphy and respiratory effort using dynamic warping," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 4, pp. 1272–1284, Oct. 2014.
[28] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley-Interscience, 2000.
[29] X. Long, J. B. Arends, R. M. Aarts, R. Haakma, P. Fonseca, and J. Rolink, "Time delay between cardiac and brain activity during sleep transitions," *Appl. Phys. Lett.*, vol. 106, no. 14, p. 143702, 2015.
[30] Task Force of the European Society ofCardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: Standards of measurement, physiologic interpretation, and clinical use," *Eur. Heart J.*, vol. 17, no. 3, pp. 354–381, 1996.
[31] P. Bušek, J. Vaňková, J. Opavský, J. Salinger, and S. Nevšímalová, "Spectral analysis of the heart rate variability in sleep," *Physiol. Res.*, vol. 54, no. 4, pp. 369–376, 2005.
[32] M. Costa, A. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Phys. Rev. Lett.*, vol. 89, no. 6, p. 068102, Jul. 2002.
[33] J. W. Kantelhardt, E. Koscielny-bunde, H. H. A. Rego, S. Havlin, and A. Bunde, "Detecting long-range correlations with detrended fluctuation analysis," *Phys. A, Statist. Mech. Appl.*, vol. 295, no. 3, pp. 441–454, 2001.
[34] T. Penzel, J. W. Kantelhardt, L. Grote, J.-H. H. Peter, and A. Bunde, "Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 10, pp. 1143–1151, Oct. 2003.
[35] S. Telser, M. Staudacher, Y. Ploner, A. Amann, H. Hinterhuber, and M. Ritsch-Marte, "Can one detect sleep Stage transitions for on-line sleep scoring by monitoring the heart rate variability?" *Somnologie*, vol. 8, no. 2, pp. 33–41, 2004.
[36] D. Cysarz, H. Bettermann, and P. van Leeuwen, "Entropies of short binary sequences in heart period dynamics," *Am. J. Physiol. Heart Circulatory Physiol.*, vol. 278, no. 6, pp. 2163–2172, 2000.
[37] A. Noviyanto, S. M. Isa, I. Wasito, and A. M. Arymurthy, "Selecting features of single lead ECG signal for automatic sleep stages classification using correlation-based feature subset selection," *Int. J. Comput. Sci. Issues*, vol. 8, no. 1, pp. 1178–1181, 2011.
[38] U. R. Acharya, E. C. Chua, O. Faust, T. C. Lim, and L. F. Lim, "Automated detection of sleep apnea from electrocardiogram signals using nonlinear parameters," *Physiol. Meas.*, vol. 32, no. 3, pp. 287–303, Mar. 2011.
[39] H. Bettermann, D. Cysarz, and P. Van Leeuwen, "Detecting cardiorespiratory coordination by respiratory pattern analysis of heart period dynamics–The musical rhythm approach," *Int. J. Bifurcation Chaos*, vol. 10, no. 10, pp. 2349–2360, Oct. 2000.
[40] D. Cysarz, H. Bettermann, S. Lange, D. Geue, and P. van Leeuwen, "A quantitative comparison of different methods to detect cardiorespiratory coordination during night-time sleep," *Biomed. Eng. Online*, vol. 3, no. 1, p. 44, Dec. 2004.
[41] J. M. Yentes, N. Hunt, K. K. Schmid, J. P. Kaipust, D. McGrath, and N. Stergiou, "The appropriate use of approximate entropy and sample entropy with short data sets," *Ann. Biomed. Eng.*, vol. 41, no. 2, pp. 349–365, 2013.
[42] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ECG and respiratory effort," *IOP Physiol. Meas.*, vol. 36, pp. 2027–2040, 2015.
[43] Y. Ichimaru, K. P. Clark, J. Ringler, and W. J. Weiss, "Effect of sleep stage on the relationship between respiration and heart rate variability," in *Proc. Comput. Cardiol. Conf.*, 1990, pp. 657–660.
[44] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[45] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 2, pp. 260–269, Apr. 1967.
[46] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.
[47] J. M. Hammersley and P. Clifford, "Markov fields on finite graphs and lattices," unpublished, 1971.
[48] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*, vol. 93. Cambridge, MA, USA: MIT Press, 2007, pp. 142–6.
[49] J. E. Dennis,Jr and J. J.More, "Quasi-Newton methods, motivation and theory," *SIAM Rev.*, vol. 19, no. 1, pp. 46–89, 1997.
[50] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Univ. Waikato, Hamilton, New Zealand, 1999.
[51] A. Strehl and J. Ghosh, "Cluster ensembles–A knowledge reuse framework for combining multiple partitions," *J. Mach. Learning Res.*, vol. 3, pp. 583–617, 2002.
[52] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.
[53] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.
[54] R. L. Verrier and R. M. Harper, *Cardiovascular Physiology: Central and Autonomic Regulation*, 5th ed., M. H. Kryger, T. Roth, and W. C. Dement, Eds. New York, NY, USA: Elsevier, 2011.
[55] X. Long, P. Fonseca, R. Aarts, R. Haakma, J. Rolink, and S. Leonhardt, "Detection of nocturnal slow wave Sleep based on cardiorespiratory activity in healthy adults," *IEEE J. Biomed. Health Informat.*, vol. PP, no. 99, pp. 1–1, 2015.

Authors' photographs and biographies not available at the time of publication.