A comparison of probabilistic classifiers for sleep stage classification

To cite this article before publication: Pedro Fonseca et al 2018 Physiol. Meas. in press https://doi.org/10.1088/1361-6579/aabbc2

Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is © 2018 Institute of Physics and Engineering in Medicine.

As the Version of Record of this article is going to be / has been published on a gold open access basis under a CC BY 3.0 licence, this Accepted Manuscript is available for reuse under a CC BY 3.0 licence immediately.

Everyone is permitted to use all or part of the original content in this article, provided that they adhere to all the terms of the licence https://creativecommons.org/licenses/by/3.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected and is not published on a gold open access basis under a CC BY licence, unless that is specifically stated in the figure caption in the Version of Record.

View the article online for updates and enhancements.
A comparison of probabilistic classifiers for sleep stage classification

Pedro Fonseca\textsuperscript{1,2}, Niek den Teuling\textsuperscript{1,3}, Xi Long\textsuperscript{1,2}, Ronald M. Aarts\textsuperscript{1,2}

\textsuperscript{1} Philips Research, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands
\textsuperscript{2} Department of Electrical Engineering, Eindhoven University of Technology, Postbus 513, 5600MB Eindhoven, The Netherlands
\textsuperscript{3} Department of Mathematics and Computer Science, Eindhoven University of Technology, Postbus 513, 5600MB Eindhoven, The Netherlands.

E-mail: pedro.fonseca@philips.com

Abstract. Objective: To compare conditional random fields (CRF), hidden Markov models (HMMs) and Bayesian linear discriminants (LDs) for cardiorespiratory sleep stage classification on a five-class sleep staging task (Wake/N1/N2/N3/REM), to explore the benefit of incorporating time information in the classification and to evaluate the feasibility of sleep staging on obstructive sleep apnea (OSA) patients.

Approach: The classifiers with and without time information were evaluated with 10-fold cross-validation on five-, four- (Wake/N1+N2/N3/REM) and three-class (Wake/NREM/REM) classification tasks using a data set comprising 443 night-time polysomnography (PSG) recordings of 231 participants (180 healthy participants, 100 of which with a ‘regular’ sleep architecture, and 51 participants previously diagnosed with OSA).

Main results: CRF with time information (CRF\textsuperscript{t}) outperforms all other classifiers on all tasks, achieving a median accuracy and Cohen’s $\kappa$ for all participants of 62.8% and 0.44 for five classes, 68.8% and 0.47 for four classes, and 77.6% and 0.55 for three classes. An advantage was found in training classifiers specifically for ‘regular’ and ‘OSA’ participants, achieving an improvement in classification performance for these groups. For ‘regular’ participants, CRF\textsuperscript{t} achieved a median accuracy and Cohen’s $\kappa$ of 67.0% and 0.51, 70.8% and 0.53 and 81.3% and 0.62 for five-, four- and three-classes respectively, and for ‘OSA’ patients, of 60.0% and 0.40, 70.0% and 0.45, and 75.8% and 0.51 for five-, four- and three-classes respectively.

Significance: The results suggest that CRF\textsuperscript{t} is not only better at learning and predicting more complex and irregular sleep architectures, but that it also performs reasonably well in five-class classification, the standard for sleep scoring used in clinical PSG. Additionally, and albeit with a decrease in performance when compared with healthy participants, sleep stage classification in OSA patients using cardiorespiratory features and CRF\textsuperscript{t} seems feasible with reasonable accuracy.
A comparison of probabilistic classifiers for sleep stage classification

1. Introduction

Despite substantial advances in the last decades in simpler and less obtrusive techniques for sleep monitoring, overnight polysomnography (PSG) remains the gold standard for sleep assessment and diagnosis of sleep disorders. However valuable, it is a costly, highly specialized procedure, with substantial burden to the participants under investigation and often limited to one or two consecutive nights. Because PSG is not practical for long-term sleep monitoring, significant research effort has been put in the development of sensors and techniques which allow sleep to be measured with less expensive sensors that are easier to set up and which can be used at home. (Redmond et al. 2007, Hedner et al. 2011, Van De Water et al. 2011, Kortelainen et al. 2010, Willemen et al. 2014, Fonseca et al. 2015). Given the limitations of home sleep monitoring, however, many sensors commonly used in PSG such as electroencephalography (EEG), electrooculography (EOG) or electromiography (EMG), are essentially excluded. The only physiological characteristics that can conveniently and reliably monitored are body movements, for example by means of actigraphy (Ancoli-Israel et al. 2003), cardiac activity, for instance with electrocardiography (ECG), photoplethysmography (PPG) (Yu et al. 2006, Beattie et al. 2017), or ballistocardiography (BCG) (Mack et al. 2009), and surrogates of respiratory effort based on respiratory movements measured with respiratory inductance plethysmography (RIP) belts around the thorax, BCG, or even Doppler radars (de Chazal et al. 2011). Although these modalities can be comfortably used in a home setting, and even set up by the subjects themselves, the relation between these physiological characteristics and sleep stages cannot be trivially annotated using visual rules as is done with standard PSG. Exploiting the varying autonomic characteristics during different stages of sleep, algorithms have been described which automate this process, typically using cardiac characteristics computed from heart rate variability and respiratory features from respiratory effort to classify epochs in states of Wake, ‘light sleep’ (typically combining N1 and N2), ‘deep sleep’ or N3, and rapid eye movement (REM) (Willemen et al. 2014, Fonseca et al. 2015, Beattie et al. 2017).

In earlier work (Fonseca et al. 2017) we explained how transitional features and non-separable features are better handled by conditional random fields (CRF) than by probabilistic classifiers traditionally used in sleep stage classification literature, such as hidden Markov models (HMMs) or Bayesian linear discriminants (LDs). However, that work only explored binary classification problems and did not address the more complex problem of multiple-class classification. Here we investigate the use of CRF for the simultaneous classification of multiple sleep stages using the same cardiorespiratory features as in our previous work, and compare it with HMM and LD for the same task. In addition, we will evaluate the extension of the multiple-class sleep stage classification problem to the five classes defined by the American Academy of Sleep Medicine (AASM) (Iber et al. 2007) to separately classify N1, N2, N3, REM sleep and Wake. This is in contrast with most work done in the area of cardiorespiratory sleep staging where classification is usually limited to either three classes (Wake, REM and non-REM...
A comparison of probabilistic classifiers for sleep stage classification

(NREM) sleep) (Redmond et al. 2007, Domingues et al. 2014), or four classes (N1 and N2 combined, N3, REM sleep and Wake) (Fonseca et al. 2015, Willemen et al. 2014). Finally, we will evaluate the feasibility of sleep stage classification of obstructive sleep apnea (OSA) patients using CRF and the same set of cardiorespiratory features. Earlier work reported in literature has shown that the presence of this condition negatively impacts the performance of cardiorespiratory classifiers, achieving modest performance on three-class classification of Wake, REM and NREM with accuracy and Cohen’s $\kappa$ of 67% and 0.32 (Redmond & Heneghan 2006) and 70% and 0.41 (Willemen et al. 2015).

2. Material and Methods

2.1. Data sets

The data sets used in this project were the same as used in our previous work (Fonseca et al. 2017), further extended with participants previously diagnosed with OSA. They comprise a total of 443 recordings of 231 participants from three databases recorded in different locations during different time periods. The first data set consists of 428 recordings of 216 participants, and was collected as part of the EU SIESTA project (Klosch et al. 2001) involving clinical partners and research groups from Austria, Germany, United Kingdom, Finland, Spain, Denmark and The Netherlands, between 1997 and 2000.

The study was approved by the local ethical committee of each research group and all participants signed informed consent. None of the participants had a history of drug or alcohol use, were working at night, or had been diagnosed with a medical disorder interfering with the goal of the study. All participants had a Mini-Mental State Examination score (Folstein et al. 1975) larger than 25 and a Pittsburgh Sleep Quality Index (PSQI) score (Buysse et al. 1989) of at most 5. The apnea patients in this data set had been diagnosed according to the International Classification of Diseases, Tenth Edition (ICD-10) (G47.3) and at the date of investigation were free of psychopharmacotherapeutic drugs for at least five times the half-life of the medication, and in a steady state for on-going non-psychopharmacotherapeutic medication at least two months.

The second data set comprises six healthy participants, recorded at the Philips Experience Lab of the High Tech Campus, Eindhoven, The Netherlands, during 2010 and the third data set comprises nine healthy participants, recorded at the Sleep Health Center, Boston, USA, during 2009. Both studies included participants with no primary history of neurological, cardiovascular, psychiatric, pulmonary, or sleep disorders. In addition, none of the participants were using sleep, antidepressant or cardiovascular medication, recreational drugs, and had a Pittsburgh Sleep Quality Index (Buysse et al. 1989) of at most 5. The protocols were reviewed by the Internal Committee of Biomedical Experiments of Philips Research and all participants signed an informed consent.
A comparison of probabilistic classifiers for sleep stage classification

All sleep technicians were blind to the condition of the participants and scored sleep stages on the second and third data sets in five classes (Wake, REM, N1, N2, N3) according to the AASM guidelines (Iber et al. 2007), and on the first data set, in six classes (Wake, REM, S1, S2, S3, S4) according to the R&K guidelines (Rechtschaffen & Kales 1968). To use a common list of ground-truth classes, S1 was mapped to N1, S2 to N2, and S3 and S4 were merged in a single N3 class. These sleep stage annotations were used as ground-truth for training and validation in our study.

Besides all modalities needed for offline visual scoring of sleep stages according to the corresponding guidelines, all recordings comprise ECG recorded with a Modified Lead II derivation, and respiratory effort recorded with thoracic RIP. Although all participants in the data set were healthy, we expect an effect of monitoring in the actual characteristics of sleep. The well-known ‘first night effect’ has an impact on sleep architecture, namely with a reduction in sleep efficiency, a delay in the onset of N3 and REM, and increased number of stage transitions (Agnew et al. 1966). In order to analyze the impact of these changes, and similar to our previous study, a subset of the complete data set was analyzed separately, comprising 134 recordings (100 participants) with a minimum percentage of each sleep stage considered representative of sleep statistics of healthy adults: at least 5% of N3, 15% of REM sleep, and a minimum of 7 hours in bed (Ohayon et al. 2004) (set ‘regular’). Regarding sleep efficiency, we used a minimum threshold of 75% which, albeit low to describe healthy sleep, accounted for the increased sleep onset latencies recorded in many participants, but which otherwise had N3 and REM durations characteristic of healthy adult sleep. Although all participants in the data set were healthy, we expect an effect of monitoring in the actual characteristics of sleep. The well-known ‘first night effect’ has an impact on sleep architecture, namely with a reduction in sleep efficiency, a delay in the onset of N3 and REM, and increased number of stage transitions (Agnew et al. 1966). In order to analyze the impact of these changes, and similar to our previous study, a subset of the complete data set was analyzed separately, comprising 134 recordings (100 participants) with a minimum percentage of each sleep stage considered representative of sleep statistics of healthy adults: at least 5% of N3, 15% of REM sleep, and a minimum of 7 hours in bed (Ohayon et al. 2004) (set ‘regular’). Regarding sleep efficiency, we used a minimum threshold of 75% which, albeit low to describe healthy sleep, accounted for the increased sleep onset latencies recorded in many participants, but which otherwise had N3 and REM durations characteristic of healthy adult sleep. The second set, also analyzed separately, comprised all 102 recordings of the 51 OSA patients (set ‘OSA’). The remaining 80 participants - not analyzed separately - did not have a sleep disorder and were as such considered healthy, but had a sleep architecture which, likely due to the first night effect, did not fit the normative values chosen for the ‘regular’ set described above. The third set comprised all 443 recordings from all three data sets (set ‘all’). Table 1 summarizes the participant demographics and sleep statistics of the three sets.
A comparison of probabilistic classifiers for sleep stage classification

Table 1. Demographics and sleep statistics of participants in the two subsets used in the study.

| Parameter     | ‘regular’       | ‘OSA’           | ‘all’            |
|---------------|-----------------|-----------------|------------------|
|               | Mean (SD)       | Range           | Mean (SD)        | Range           | Mean (SD)       | Range           |
| N             | 100 subjs., 134 recs. | 51 subjs., 102 recs. | 231 subjs., 443 recs. |
| Sex           | 56 female subj. (56.0 %) | 7 female subj. (13.7 %) | 105 female subj. (45.5 %) |
| Age (year)    | 44.3 (17.2)     | 20 – 83         | 51.4 (9.8)       | 29 – 73         | 50.4 (17.95)    | 20 – 95         |
| BMI (kg/m²)   | 24.0 (3.0)      | 19.0 – 31.3     | 30.7 (5.1)       | 22.9 – 43.3     | 25.9 (4.67)     | 17.0 – 43.3     |
| TIB (hour)    | 8.0 (0.4)       | 7.2 – 9.7       | 8.3 (0.3)        | 7.8 – 9.6       | 8.0 (0.50)      | 5.8 – 9.7       |
| SE (%)        | 88.8 (5.6)      | 75.5 – 99.0     | 82.9 (11.1)      | 25.3 – 97.7     | 81.6 (11.73)    | 24.7 – 99.0     |
| N1 (%)        | 8.5 (3.9),      | 1.6 – 21.8      | 18.3 (10.66)     | 4.0 – 63.9      | 13.0 (8.09)     | 1.6 – 63.9      |
| N2 (%)        | 51.8 (6.4)      | 30.6 – 67.1     | 53.0 (10.8)      | 14.0 – 74.2     | 53.8 (9.10)     | 14.0 – 76.2     |
| N3 (%)        | 17.1 (5.7)      | 6.0 – 36.2      | 9.4 (7.7)        | 0.0 – 44.2      | 13.8 (8.50)     | 0.0 – 44.2      |
| REM (%)       | 21.7 (3.3)      | 16.3 – 31.3     | 18.1 (5.6)       | 0.0 – 30.0      | 18.3 (5.63)     | 0.0 – 34.8      |
| AHI (event/hour) | n.a.            | 38.3 (23.7)     | 1.7 – 97.1       | n.a.            |

‘regular’ is a subset of the healthy participants in ‘all’ data set where only participants with a minimum of 5% N3, 15% REM sleep and 75% sleep efficiency were considered. ‘OSA’ is a subset of the participants in ‘all’ data set that had a previous diagnosis of sleep apnea. N1, N2, N3, and REM percentages were calculated over the total sleep time for each recording. BMI: body mass index, TIB: time in bed, SE: sleep efficiency, AHI: apnea/hypopnea index, calculated as the sum of all apnea and hypopnea events divided by the total sleep time.

2.2. Feature extraction

The classifiers used in this study use the same set of cardiac and respiratory features described in our previous work (Fonseca et al. 2017) and will only be summarized here. The baseline wander of the ECG signal was first removed with a high-pass filter using a Kaiser window of 1.016 sec, a cut-off frequency of 0.8 Hz and a side-lobe attenuation of 30 dB. QRS complexes were detected with a Hamilton-Tompkins QRS detector (Hamilton 2002, Hamilton & Tompkins 1986) and further localized with a post-processing algorithm (Fonseca et al. 2014). RIP signals were first resampled to a common sampling rate of 10 Hz, and then low-pass filtered with a 10th order Butterworth filter with a cut-off frequency of 0.6 Hz. Baseline was removed by subtracting the median value calculated in a sliding window of 120 s. After the signals were processed, a total of 81 features from ECG, 33 from RIP and 3 from the simultaneous analysis of both modalities, were extracted using windows centered on non-overlapping epochs of 30 s, aligned with the ground-truth annotations from PSG. The length of the extraction windows varied from feature to feature as described in earlier work (Fonseca et al. 2017). An automatic artifact rejection procedure was employed to determine whether segments of the ECG and RIP signals were too affected by noise or motion artifacts. For each feature extracted from ECG, a given epoch was marked as invalid if the sum of the length of the R-R intervals in the window used to compute that feature amounted to less than 50% of the length of that window. For all features extracted from RIP,
A comparison of probabilistic classifiers for sleep stage classification

respiratory peaks and troughs on each epoch were first detected, and a criteria was used to determine whether they had a plausible length (Long et al. 2014). Epochs with one or more implausible peaks or troughs were marked as invalid. After feature extraction and artifact rejection, the values of features for epochs marked as invalid were estimated with linear interpolation between the values of neighboring (valid) epochs. Finally, to reduce between-subject physiological and equipment-related variations (Long et al. 2015), all features were normalized to have zero mean and unit standard deviation.

2.3. Multiple-sleep stage classification

A detailed explanation of Bayesian LDs, Bayesian networks (BNs), HMMs and CRFs, and how they behave in the presence of transitional and non-separable features was given in our previous work (Fonseca et al. 2017). This section will give only a brief summary of the three classifiers.

The Bayesian LD classifier is a single-epoch classifier which uses the Bayes’ rule for minimizing the probability of error by selecting, based on the observations (features), the most likely class (sleep stage). In contrast with LDs, BNs can make use of directed acyclic graphs to better exploit the structured nature of sleep, where a given sleep stage depends also on the previous stage. HMMs are a special type of generative BNs which model the probability of an observation given an unobserved (hidden) state, and the probability of a state given the previous state. CRF classifiers are a type of discriminative models which, in contrast with generative models, avoid the need to learn the emission probabilities from the training data and model directly the probability of a class given an observation and given the class of the previous state. This has the benefit of requiring substantial less training data, or less simplifications about the feature space, and can make better use of correlated features.

In the following, the three classifiers are described in terms of their defining equations.

Applied in previous work on multiple-sleep stage classification (Fonseca et al. 2015), the LD classifier uses Bayes’ rule for minimizing the probability of error by selecting, separately for each epoch, the most likely class, i.e., the class that maximizes the discriminant function:

$$g_i(x) = -\frac{1}{2} (x - \mu_i)^T \Sigma^{-1} (x - \mu_i) + \ln P(\omega_i)$$

(1)

where \(\mu_i\) is the average feature vector for class \(i\), \(\Sigma\) is the pooled covariance matrix for all classes, and \(P(\omega_i)\) is usually a fixed term with the prior probability of each class. It has been observed that the probability of different sleep stages changes with sleep time (Redmond et al. 2007). In order to exploit this characteristic, and in addition to the standard LD of (1), a variation of this classifier with time-varying prior probabilities (LDtvp) was considered (Redmond et al. 2007, Fonseca et al. 2015). In this case, the last term of the discriminant equation corresponds to \(P(\omega_i, t)\), representing the prior probability of class \(i\) at time (since lights off) \(t\), with \(t = 1, ..., T\) where \(T\) corresponds
A comparison of probabilistic classifiers for sleep stage classification

to the duration of the recording. All parameters $\mu_i$, $\Sigma$, and $P(\omega_1; t)$ were determined during training.

Unlike LD classifiers, where the most likely class is considered separately for each epoch, HMM classifiers make use of BNs where the dependency between consecutive states can be modeled by means of directed edges. HMM are a special case of BNs where the class of each state, $w_t$, remains hidden, and the only observable variables are the feature vectors, $x_t$, which depend on them (Rabiner 1989). Allowing the current state to depend on the previous state and on the current observations, the joint probability of the variables in this network is given by

$$P(W, X) = P(w_1) P(x_1 | w_1) \prod_{t=2}^{T} P(w_t | w_{t-1}) P(x_t | w_t).$$

(2)

Using the Viterbi algorithm (Viterbi 1967), the most likely sequence of states is progressively determined as the sequence that maximizes the joint probability up to each state,

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

(3)

with the tail function for $t = 1,$

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

(4)

Emission probabilities expressing the likelihood of an observation given a class, $P(X_t | w_t)$, and transition probabilities expressing the likelihood of transitioning to a state $w_t$ given the previous state $w_{t-1}$, i.e., $P(w_t | w_{t-1})$ are learned during training. In our study we used a discrete HMM classifier (Rabiner 1989) whereby the estimation of emission probabilities is obtained by mapping the feature space to a discrete set of symbols, or clusters. The number of clusters $k$ was automatically determined during training as the value of $k$ that maximized the normalized mutual information between the labels (class) of each feature vector and the corresponding closest clusters obtained with $k$-means clustering (Strehl & Ghosh 2002). During classification, the cluster index of each feature vector was determined by finding the closest cluster centroid. The index of this cluster was used as input to the Viterbi algorithm to obtain the posterior probability of each class and the most likely class for each state.

HMM, like other generative models, require the distribution of the observations to be modeled, or somehow learned from the training data. When the features are not independent this requires large amounts of training data or strong assumptions about the distribution and the relation between features. In the case of our study, this was achieved by mapping the feature space to a limited, discrete, number of clusters. Discriminative models such as CRF (Lafferty et al. 2001) do not have this requirement, because they compute the probability of a class based on an observation, $P(w | X)$, directly from data. Because discriminative models do not require any assumptions about the independence of the observed variables, they make better use of interdependent, correlated features, very common in the case of cardiorespiratory sleep stage classification. Parameter
A comparison of probabilistic classifiers for sleep stage classification

learning with CRF is usually done with factor graphs (Kschischang et al. 2001) and
the conditional probability \( P(W|X) \) can be obtained from

\[
P(W|X) = \frac{1}{Z(X)} \prod_t^T \exp \left( \sum_k^K \lambda_k f_k(w_t, w_{t-1}, x_t) \right),
\]

with

\[
Z(X) = \sum_w \prod_t^T \exp \left( \sum_k^K \lambda_k f_k(w_t, w_{t-1}, x_t) \right),
\]

where \( \lambda_k \) are the model parameters and \( f_k \) the feature functions, in this study specified
as the set of state-transition functions \( f_t(w_t, w_{t-1}, x_t) \) defined for all possible state-
transition pairs \( (s, r) \), and state-observation functions \( f_e(w_t, w_{t-1}, x_t) \) defined for all
state-observation pairs \( (s, x) \),

\[
f_t(w_t, w_{t-1}, x_t) = 1_{\{w_s = s\}} 1_{\{w_{s-1} = r\}}, \text{ for each } (s, r),
\]

and

\[
f_e(w_t, w_{t-1}, x_t) = 1_{\{w_s = s\}} x_t, \text{ for each } (s, x),
\]

where the notation \( 1_{\{\text{condition}\}} \) has a value of 1 if the condition is true, and 0 otherwise.
A total of \( K = S^2 + S \) parameters \( \theta = \{\lambda_{s,r}, \lambda_{s,x}\} \), with \( S \) as the number of classes,
are estimated during training using \( N \) input sequences (corresponding to state and
observation sequences from \( N \) different recordings), by maximizing the conditional log-
likelihood (Sutton & McCallum 2007),

\[
L(\theta) = \sum_{i=1}^N \log P(W|X^i) = \sum_{i=1}^N \sum_t^T \sum_{k=1}^K \lambda_k f_k(w_t^i, w_{t-1}^i, x_t^i) - \sum_{i=1}^N \log Z(X^i),
\]

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 algorithm (Dennis, Jr & More 1997).
Decoding of a linear-chain CRF is done in a similar way as HMM, using a modified
version of the Viterbi algorithm (Sutton & McCallum 2007) where the most likely
sequence of states is progressively determined as the sequence of classes that maximizes
the recursion

\[
V_t(w_t) = \max_{w_{t-1}} \exp \left( \sum_k^K \lambda_k f_k(w_t, w_{t-1}, x_t) \right) V_{t-1}(w_{t-1}).
\]

In a manner comparable to the use of time-varying prior probabilities for LD, a
trivial way of encoding time information in Bayesian chains is to simply add a time
feature to the feature set. By defining the lights-off and lights-on instants as time
\( t = 0 \) and \( t = 1 \) respectively, and by linearly interpolating the values for all epochs,
a comparable (and normalized) time feature can be used for each recording. This
 technique was used for both the HMM and CRF classifiers, and will be further specified
as HMM\( t \) and CRF\( t \) to explicitly indicate the use of time information.
A comparison of probabilistic classifiers for sleep stage classification

2.4. Evaluation

To compare the performance of the three classifiers (and their corresponding time-variations) in the multiple-stage classification tasks, a 10-fold cross-validation procedure was used. Furthermore, to eliminate the possibility of subject-dependent bias between training and validation, when a participant had more than one recording they were both made part of the same fold for either training or validation. The same folds were used for all classifiers to allow for paired comparisons.

The performance of each of the six classifiers (LD, LDtvp, HMM, HMMt, CRF and CRFt) was evaluated on epoch-per-epoch agreement between the predicted classes on a five-class classification task (Wake/N1/N2/N3/REM) in terms of accuracy (percentage of epochs where the predicted and ground-truth sleep stage agreed) and Cohen’s $\kappa$ coefficient of agreement, an agreement measure that compensates for the chance of random agreement (Cohen 1960). $\kappa$ is usually interpreted with the following terms: below 0.20 indicates ‘slight agreement’, 0.21 to 0.4 ‘fair agreement’, 0.41 to 0.60 ‘moderate agreement’, 0.61 to 0.8 ‘substantial agreement’, and above 0.81 ‘almost perfect agreement’. In addition, the performance of all classifiers was also evaluated on four-class (Wake/N1+N2/N3/REM) and three-class (Wake/NREM/REM) tasks. These results were obtained by merging both the predicted and ground-truth classes in the following way: for the four-class task, N1 and N2 were merged into a single N1+N2 class. For three-class classification, N1, N2 and N3 were merged into a single NREM class. A two-tailed Wilcoxon signed-rank test (Wilcoxon 1945) was used to test for significant differences between the best and the remaining classifiers for each task. This evaluation procedure was performed separately for the ‘regular’, the ‘OSA’ and the ‘all’ data sets.

In order to evaluate whether the class merging process is an adequate procedure when a higher performance is desired at the cost of simplifying the classification problem, the performance of each classifier was compared with that obtained with dedicated classifiers, trained specifically for the four- and three-class problem. This was done by merging the ground-truth annotation classes before training. Two-tailed Wilcoxon signed-rank tests were used to test for significant differences between the class-merging and the dedicated-training approaches.

Finally, and given the difference in complexity between all classifiers and to heuristically check for the presence of over-fitting or generalization problems, the training and validation performance of all classifiers was computed using cross-validation for a varying number of participants and so-called learning curves were plotted.

3. Results

Tables 2, 3 and 4 compare the classification performance of each classifier for the ‘regular’, ‘OSA’ and ‘all’ data sets, for the four- and five-class classification tasks, in terms of $\kappa$ and accuracy. For the three-class classification task (results omitted in the tables, for brevity), CRFt performed significantly better than the other classifiers,
A comparison of probabilistic classifiers for sleep stage classification

Table 2. Mean (SD), and median $\kappa$ and accuracy after sleep staging of ‘regular’ participants (N=134).

| Classifier | Wake/N1+N2/N3/REM | Wake/N1/N2/N3/REM |
|------------|--------------------|--------------------|
|            | $\kappa$ (-)       | accuracy (%)       | $\kappa$ (-)       | accuracy (%)       |
| LD         | 0.44 (0.11), 0.45  | 62.66 (7.25), 63.25| 0.39 (0.09), 0.40  | 53.33 (7.49), 54.54|
| LDtvp      | 0.49 (0.12), 0.50  | 68.41 (7.14), 69.99| 0.45 (0.11), 0.47  | 63.61 (7.79), 64.95|
| HMM        | 0.31 (0.11), 0.32  | 50.75 (8.76), 51.47| 0.29 (0.11), 0.31  | 49.86 (8.56), 47.95|
| HMMt       | 0.33 (0.11), 0.34  | 52.17 (8.05), 53.08| 0.31 (0.10), 0.32  | 48.38 (8.12), 48.55|
| CRF        | 0.49 (0.14), 0.51  | 68.76 (8.69), 69.46| 0.47 (0.13), 0.49  | 64.66 (8.86), 65.66|
| CRFt       | **0.50 (0.15)**, **0.53**$^*$| **69.59 (9.03)**, **70.79**$^*$| **0.48 (0.13)**, **0.51**$^*$| **65.66 (9.20)**, **67.02**$^*$|

$^*$CRFt performs significantly better than all others ($p < .001$).

Table 3. Mean (SD), and median $\kappa$ and accuracy after sleep staging of ‘OSA’ participants (N=102).

| Classifier | Wake/N1+N2/N3/REM | Wake/N1/N2/N3/REM |
|------------|--------------------|--------------------|
|            | $\kappa$ (-)       | accuracy (%)       | $\kappa$ (-)       | accuracy (%)       |
| LD         | 0.38 (0.10), 0.38  | 60.72 (7.30), 61.71| 0.32 (0.09), 0.33  | 47.91 (6.88), 48.15|
| LDtvp      | 0.42 (0.11), 0.43  | 67.98 (8.05), 69.90| 0.36 (0.10), 0.37  | 57.27 (9.30), 57.62|
| HMM        | 0.19 (0.12), 0.20  | 46.28 (11.07), 46.63| 0.17 (0.11), 0.17  | 38.27 (10.26), 38.53|
| HMMt       | 0.19 (0.12), 0.20  | 46.71 (10.69), 47.12| 0.17 (0.11), 0.17  | 38.98 (10.38), 39.99|
| CRF        | 0.44 (0.13), 0.44  | **68.85 (8.38)**, **69.66**$^*$| 0.39 (0.12), 0.40  | 59.26 (10.27), 60.19|
| CRFt       | **0.44 (0.13)**, **0.45**$^{1,2}$| 68.67 (8.31), 69.67| **0.40 (0.12)**, **0.40**$^{1,2}$| **59.30 (10.20)**, **59.92**$^{1,2}$|

$^{1,2}$CRF performs significantly better than LD, HMM and HMMt ($p < .001$), but not LDtvp ($p = .17$) nor CRFt ($p = .64$).

achieving an average $\kappa$ of 0.59 ± 0.15 (with a median of 0.62) and accuracy of 80.67 ± 6.93% (median 81.29%) in the ‘regular’ data set, a $\kappa$ of 0.50 ± 0.13 (median 0.51) and accuracy 75.40 ± 6.83% (median 75.77%) for ‘OSA’, and a $\kappa$ of 0.54 ± 0.14 (median 0.55) and accuracy of 76.70 ± 7.43% (median 77.62%) for ‘all’. For three- and four-class classification tasks, the results were obtained by merging the classes resulting from five-class classification. The subset comparisons of Table 4 were obtained with cross-validation on the ‘all’ data set but analyzed separately for the recordings that were part of the ‘regular’, ‘OSA’ and ‘irregular’ (healthy participants not part of the ‘regular’ set) data sets.

An unpaired Wilcoxon rank sum test (Wilcoxon 1945) was used to compare, for the CRFt classifier, the performance obtained between the subsets when performing five-class classification. The performance was not significantly different between the ‘regular’ and the ‘irregular’ sets, neither for accuracy ($p = .64$) nor for $\kappa$ ($p = .46$). It was, however, significantly higher for ‘regular’ than for ‘OSA’, ($p < .001$ and $p < .001$), and for ‘irregular’ than for ‘OSA’ ($p < .01$ and $p < .001$).
A comparison of probabilistic classifiers for sleep stage classification

Table 4. Mean (SD), and median $\kappa$ and accuracy after sleep staging of ‘all’ participants (N=443).

|                   | Wake/N1+N2/N3/REM | Wake/N1/N2/N3/REM |
|-------------------|-------------------|-------------------|
|                   | $\kappa$ (-)      | accuracy (%)      | $\kappa$ (-)      | accuracy (%)      |
| LD                | 0.39 (0.12), 0.40 | 59.16 (7.80), 59.98 | 0.34 (0.10), 0.35 | 48.76 (8.08), 49.59 |
| LDtvp             | 0.44 (0.13), 0.45 | 66.67 (8.46), 67.78 | 0.40 (0.12), 0.42 | 59.83 (9.62), 61.37 |
| HMM               | 0.26 (0.13), 0.26 | 46.05 (10.32), 46.74 | 0.25 (0.12), 0.26 | 42.92 (10.68), 43.46 |
| HMMt              | 0.25 (0.13), 0.26 | 45.18 (10.03), 45.73 | 0.25 (0.12), 0.25 | 42.10 (10.41), 42.62 |
| CRF               | 0.44 (0.14), 0.45 | 66.54 (8.97), 67.37 | 0.42 (0.13), 0.43 | 61.10 (9.77), 62.18 |
| CRFt              | **0.46 (0.14), 0.47** | **67.36 (8.93), 68.77** | **0.43 (0.13), 0.44** | **61.83 (9.86), 62.80** |

subset ‘regular’ (N=134)

|                   | Wake/N1+N2/N3/REM | Wake/N1/N2/N3/REM |
|-------------------|-------------------|-------------------|
|                   | $\kappa$ (-)      | accuracy (%)      | $\kappa$ (-)      | accuracy (%)      |
| LDtvp             | **0.46 (0.12), 0.49** | **67.11 (7.48), 67.76** | 0.44 (0.11), 0.46 | 62.65 (7.75), 63.47 |
| CRFt              | 0.46 (0.14), 0.47 | 66.51 (8.97), 68.26 | **0.45 (0.12), 0.46** | **63.11 (8.94), 63.82** |

subset ‘irregular’ (N=207)

|                   | Wake/N1+N2/N3/REM | Wake/N1/N2/N3/REM |
|-------------------|-------------------|-------------------|
|                   | $\kappa$ (-)      | accuracy (%)      | $\kappa$ (-)      | accuracy (%)      |
| LDtvp             | 0.43 (0.13), 0.44 | 66.47 (8.85), 67.78 | 0.38 (0.12), 0.39 | 58.61 (10.10), 59.45 |
| CRFt              | **0.45 (0.14), 0.47** | **67.73 (8.91), 69.04** | **0.42 (0.13), 0.43** | **61.27 (10.20), 62.04** |

subset ‘OSA’ (N=102)

|                   | Wake/N1+N2/N3/REM | Wake/N1/N2/N3/REM |
|-------------------|-------------------|-------------------|
|                   | $\kappa$ (-)      | accuracy (%)      | $\kappa$ (-)      | accuracy (%)      |
| LDtvp             | 0.41 (0.12), 0.43 | 66.51 (8.42), 68.41 | 0.35 (0.12), 0.35 | 56.43 (10.47), 57.53 |
| CRFt              | **0.43 (0.13), 0.44** | **67.18 (8.07), 68.96** | **0.38 (0.13), 0.39** | **58.52 (10.51), 59.51** |

*1: CRFt performs significantly better than LD, HMM, HMMt and CRF ($p < .001$) and LDtvp ($^{*1}p < .001, ^{*2}p < .01$).

For the ‘regular’ subset, there is no significant difference in performance between LDtvp and CRFt ($^{*1}p = .22, ^{*2}p = .06, ^{*3}p = .14, ^{*4}p = .40$).

For the ‘irregular’ subset, CRFt performs significantly better than LDtvp ($^{*3}p < .001$).

For the ‘OSA’ subset, CRFt performs significantly better than LDtvp for five classes ($^{*4}p < .001$) but there isn’t always a significant difference between CRFt and LDtvp for the remaining tasks ($^{*5}p < .05, ^{*6}p = .23$).

Table 5 indicates the results for LDtvp and CRFt obtained with a dedicated classifier trained specifically with three and four classes and compares them with the results obtained by merging the classes detected with the five-class classifier for each set.

Tables 6, 7, and 8 indicate the confusion matrix for five-class classification with the CRFt classifier for the three data sets. In addition, the table indicates the sensitivity and the positive predictive value (PPV) for each class and, for comparison, the performance obtained with LDtvp. For Tables 7 and 8 the performance obtained after evaluating the results of each subset after cross-validation with CRFt on ‘all’ data set are also indicated.

Finally, Figures 1a, 1b and 2 illustrate the so-called ‘learning curves’ for the LDtvp, HMM, and CRFt classifiers for the three data sets. For each classifier, the curves illustrate the training and validation performance, averaged after cross-validation, for a varying number of participants.
Table 5. Classification performance with classifiers trained with three and four classes.

|                          | kappa (-) | accuracy (%) | kappa (-) | accuracy (%) |
|-------------------------|-----------|--------------|-----------|--------------|
| **Wake/NREM/REM**       |           |              |           |              |
| ‘regular’ (N=102)       |           |              |           |              |
| LDtvp                   | 0.55 (0.15) | 80.4 (6.4)  | 0.46* (0.12) | 68.0 (7.0)  |
| CRFt                    | 0.59 (0.15) | **81.8** (6.1) | **0.51** (0.14) | **71.0** (8.4) |
| ‘OSA’ (N=102)           |           |              |           |              |
| LDtvp                   | 0.40*** (0.12) | 74.2 (7.7)  | 0.39 (0.11) | 68.2 (8.9)  |
| CRFt                    | **0.50** (0.14) | **77.0** (7.6) | **0.45** (0.13) | **70.6** (8.7) |
| ‘all’ (N=443)           |           |              |           |              |
| LDtvp                   | 0.47*** (0.15) | 76.2 (8.0)  | 0.41*** (0.13) | 66.1 (8.4)  |
| CRFt                    | **0.53** (0.15) | **78.0** (7.8) | **0.46** (0.14) | **68.8** (9.1) |

The table indicates for each classifier and classification task the mean (standard deviation) \( \kappa \) and accuracy. Indicated in **bold** are the results for which the dedicated classifiers obtained a higher performance than that achieved by merging the results of the five-class classifier. Significant differences are indicated (**\( p < .001 \), **\( p < .01 \), *\( p < .05 \)).

Table 6. Pooled confusion matrix for CRFt and LDtvp classification of ‘all’ participants.

| pred. \(-\) ref. \(\rightarrow\) | N1  | N2  | N3  | REM | Wake | PPV (%) |
|----------------------------------|-----|-----|-----|-----|------|---------|
| **CRFt**                         |     |     |     |     |      |         |
| N1                               | 1625| 2187| 51  | 450 | 1771 | **26.7**|
| N2                               | 21071| 152450| 30925| 17899| 16142| **63.9**|
| N3                               | 539 | 6340| 14968| 111 | 904 | **65.5**|
| REM                              | 6507| 8435| 242 | 37896| 6411 | **63.7**|
| Wake                             | 14053| 15875| 753 | 7136 | 49975| 56.9    |
| **Sens. (%)**                    | 3.7 |     | 82.3| 31.9| 59.7 | **66.5**|

| **LDtvp**                        |     |     |     |     |      |         |
| N1                               | 1607| 2032| 103 | 1170| 1523 | 25.0    |
| N2                               | 26648| 150789| 25877| 20638| 22609| 61.2    |
| N3                               | 567 | 13462| 18946| 328 | 1323 | 54.7    |
| REM                              | 7054| 10837| 721 | 35658| 8427 | 56.9    |
| Wake                             | 7919| 8167| 1292| 5698 | 41321| **64.2**|
| **Sens. (%)**                    | 3.7 | 81.4| **40.4**| 56.2 | 54.9  |         |

The confusion matrices compare the predicted and the reference sleep stages for all epochs of all participants in the data set after cross-validation with CRFt (above) and LDtvp (below). **Bold** indicates the classifier (between CRFt and LDtvp) which obtained the best sensitivity and PPV for each class.

4. Discussion

4.1. Performance comparison between classifiers

As indicated in Table 2, the CRF classifier with normalized time (CRFt) outperforms all other classifiers for the ‘regular’ data set for all classification tasks. The differences
A comparison of probabilistic classifiers for sleep stage classification

Table 7. Pooled confusion matrix for CRFt classification of ‘regular’ participants.

| pred. → ref. | N1     | N2     | N3     | REM    | Wake    | PPV (%) |
|-------------|--------|--------|--------|--------|---------|---------|
| N1          | 691    | 1200   | 20     | 155    | 520     | **26.7** | 24.7   | 22.4   |
| N2          | 3750   | 47869  | 10884  | 5939   | 2542    | **67.4** | 66.2   | 64.9   |
| N3          | 186    | 2473   | 8131   | 24     | 339     | 72.9    | **78.9** | 60.0   |
| REM         | 2006   | 3770   | 111    | 17256  | 1585    | 69.8    | **74.0** | 65.5   |
| Wake        | 2829   | 3076   | 230    | 1103   | 8696    | 54.6    | 44.1   | **65.2** |

| Sens. (%)   | CRFt   | 7.3    | **82.0** | 42.0   | **70.5** | 63.6    |
|             | CRFt ‘all’ | 6.7  | 80.4    | 31.1   | 64.1    | **74.4** |
|             | LDtvp  | 5.7    | 78.4    | **50.3** | 66.5    | 54.1    |

See notes to Table 6. For comparison, the sensitivity and PPV obtained with LDtvp for the same data set, and with CRFt evaluated for the ‘regular’ set after cross-validation on the ‘all’ data set are also given.

Table 8. Pooled confusion matrix for CRFt classification of ‘OSA’ participants.

| pred. → ref. | N1     | N2     | N3     | REM    | Wake    | PPV (%) |
|-------------|--------|--------|--------|--------|---------|---------|
| N1          | 1496   | 1405   | 16     | 486    | 1049    | 33.6    | 30.7   | **34.2** |
| N2          | 7996   | 36832  | 5210   | 3991   | 4020    | **63.4** | 62.5   | 60.4   |
| N3          | 159    | 1298   | 2050   | 50     | 272     | 53.5    | **54.8** | 49.6   |
| REM         | 2470   | 2198   | 93     | 9069   | 2154    | 56.7    | **58.9** | 53.0   |
| Wake        | 3079   | 2796   | 193    | 1641   | 9590    | 55.4    | 49.5   | **60.8** |

| Sens. (%)   | CRFt   | 9.8    | **82.7** | 27.1   | **59.5** | 56.1    |
|             | CRFt ‘all’ | 2.1  | 81.9    | **44.8** | 51.8    | 59.8    |
|             | LDtvp  | **13.1** | 81.8    | 37.6   | 49.8    | 47.8    |

See notes to Table 7.

are significant for all performance metrics, except for accuracy in the three-class task. The same can be observed for the ‘all’ data set in Table 4, where the differences are significant for both κ and accuracy, for all classification tasks. It is clear that the addition of normalized time information as a feature improves the classification performance for healthy participants in comparison with the standard CRF. Regarding the ‘OSA’ data set, as indicated in Table 3, CRFt achieves a significantly higher κ and for the five-class task, accuracy, than all other classifiers except CRF. Interestingly, CRF achieves a better accuracy for the three- and four-class task, albeit not significantly in comparison with CRFt and LDtvp. In contrast with the classification for healthy participants, it is clear that for the ‘OSA’ data set the addition of time information leads to smaller improvements over standard CRF. OSA is known to cause increased sleep fragmentation due to the arousals which often follow apnea and hypopnea events (Roehrs et al. 1989), and also a decrease of N3 (Ratnavadivel et al. 2009, Bianchi et al. 2010) and REM (Bianchi et al. 2010). Likewise, it is reasonable to expect that, at least for patients with
Figure 1. Average validation and training (suffix TR) performance $\kappa$ for a varying number of recordings in the (a) ‘regular’ and (b) ‘OSA’ data sets, for the LDtvp, HMM and CRFt classifiers.

Figure 2. Average validation and training (suffix TR) performance $\kappa$ for a varying number of recordings in the ‘all’ data set, for the LDtvp, HMM and CRFt classifiers.

A higher OSA severity, the occurrence of apnea events possibly dictates more changes in sleep architecture than the natural progression of sleep over time. While for healthy participants, for instance, there is a strong association between the occurrence of N3 and REM with sleeping time, the presence and progression of these sleep stages for an apnea patient may depend more on the disordered breathing events which can occur during the entire night.

4.2. Performance comparison between data sets

Comparing Tables 2, 3 and 4 it is obvious that the overall performance for ‘all’ participants is lower than for ‘regular’ participants but higher than for ‘OSA’
A comparison of probabilistic classifiers for sleep stage classification

participants. Accordingly, healthy sleep architectures seem easier to model and classify with the current set of cardiorespiratory features, in contrast with the more complex characteristics of healthy participants who do not have regular sleep architectures and even more so with OSA patients for which the architecture may be disrupted by their condition. Interestingly, as illustrated by the results for the subsets ‘regular’ and ‘OSA’ of the ‘all’ data set (also indicated in Table 4), the performance for those participants is overall lower than what was obtained when the classifier was trained separately with examples of only ‘regular’ and only ‘OSA’ participants (Tables 2 and 3). This suggests that by attempting to model the common characteristics of all participants, including those with regular, irregular and disrupted sleep architectures, the classifier compromises on the ability to learn characteristics that were specific of individual groups of participants to be able to better capture irregular sleep architectures. A prior stratification of the sleep architecture (at least between regular and irregular participants) and screening for a sleep disorder such as apnea would allow specific classifiers to be used instead of a globally trained one and improve overall classification performance.

Regarding the comparison between the performance of CRFt and LDtvp for the subsets of the ‘all’ data set, it can be observed that the differences depend on the characteristics of the subset. Although there are no significant differences in performance for the ‘regular’ subset, the differences are significant for all tasks for the ‘irregular’ subset and for the five-class task for the ‘OSA’ subset. This suggests that a CRFt classifier trained with examples of all participants is better able to learn complex sleep architectures than LDtvp, in particular for the most complex case of five sleep stages.

4.3. Learning curves

Regarding Figures 1a and 1b it is noticeable that the training and validation performances of the CRFt classifier converge at approximately the same number of recordings (around 70-80) as the LDtvp classifier and that both achieve a better training and classification performance than HMM, suggesting that the latter is simply not able to learn the characteristics of the sleep stages for either data set. The same can be seen in Fig. 2 although, because the complexity of the problem increases with the addition of participants with irregular sleep architectures, both classifiers require a larger number of recordings for the performances to converge (around 200). What is clear from the three figures is that for the ‘regular’, ‘OSA’ and ‘all’ sets, the number of recordings seems to be sufficient for an adequate training and validation. There is no evidence of over-fitting nor problems generalizing to new data, in which cases the training performance would seem substantially higher than the validation performance when using the complete data sets.
4.4. Comparison between training strategies

Table 5 shows there is, generally speaking, no clear advantage to training classifiers dedicated specifically to the three- and four-class tasks as compared with the results obtained by appropriately merging the classes resulting from five-class classification. For CRFt, the only significant improvement is in terms of accuracy for the three-class task in the ‘regular’ data set and also in terms of accuracy for both tasks in the ‘all’ set. $\kappa$ actually decreases (albeit not significantly) when compared with the results obtained by merging the five-class classification results. For LDtvp there are no significant improvements for any task, and the results are significantly worse in terms of $\kappa$ for the four-class task for ‘regular’ and ‘all’, and for the three-class task for the ‘regular’ and ‘all’ tasks. This is likely due to substantial differences between the classes: when merged together, the resulting class is likely very heterogeneous and a classifier trained with a fewer-class model still needs to learn - with a more complex model - the characteristics of the original classes. Consider for instance the CRF classifier, which uses the training data to model the probability $P(w|X)$ of a sleep stage given a set of observations or features. When this is done for a class (for example NREM) obtained by merging multiple stages (in this case, N1, N2 and N3) the merged feature space is highly heterogeneous since it comprises examples of substantially different stages and the model that needs to be learned is highly complex. Instead, if the classifier is trained to learn $P(w|X)$ separately for N1, N2 and N3, each individual model is simpler and even if classification errors are done when trying to distinguish between these three stages, the confusion between the predictions for the three classes disappears when the results are combined in a single NREM class.

4.5. Confusion matrices

Table 6 indicates the confusion matrices obtained after cross-validation with CRFt and LDtvp for ‘all’ participants. It is clear that the CRFt classifier achieves a superior performance both in terms of sensitivity and PPV for almost all classes, with the exception of a lower sensitivity to N3 and a lower PPV to Wake. Furthermore, it is interesting to note that the LDtvp classifier confuses more often classes which are, in principle, more disparate, than CRFt does. For instance, notice the much larger number of predictions of N1 in epochs of REM (1170 versus 450 with CRFt), or when predicting REM or Wake in epochs of N3 (721 and 1292 against 242 and 753 with CRFt respectively). This suggests that not only does the CRFt do an overall better job of sleep stage classification, but also that the output is overall more plausible than that with LDtvp. Regarding the confusion matrix for the ‘regular’ data set, Table 7 indicates once more that CRFt, either trained with ‘regular’ participants, or with ‘all’ participants, yields better sensitivity and PPV than LDtvp for the same classes as in Table 6, with all participants. Interestingly, the same is not observed for ‘OSA’ participants, as indicated in Table 8. In this case, LDtvp achieves the best PPV for Wake, but not the best sensitivity for N3. Where LDtvp outperforms CRFt seems to
be in the detection of N1, both in terms of sensitivity as well as PPV. A note should be made about the apparent superior performance of CRF trained with ‘all’ participants versus CRF trained with ‘OSA’ for some classes, for example achieving a much higher sensitivity to N3. Although Tables 3 and 4 show that a dedicated classifier trained with OSA participants outperforms a classifier trained with all participants and then applied to this subset, the results in the confusion matrix appear to contradict that observation. This apparent contradiction stems from the fact that the confusion matrices are based on all epochs aggregated over all participants in that data set. As we had seen in Table 1, these OSA patients have, in average, less REM and especially less N3 than the average for ‘regular’ participants and in the complete set. The confusion matrix thus hides the much lower performance obtained by CRF trained with ‘all’ participants, for example on participants with nearly no N3 nor REM sleep.

4.6. Examples of predicted hypnograms

To illustrate some examples of the classification output with the CRF classifier, Figures 3 and 4 illustrate a comparison between the ground-truth reference hypnogram and the classification output for the participant with the best performance and the participant with the performance closest to the median (both in terms of $\kappa$), in the ‘regular’ data set. As it can be clearly seen in Fig. 3, there is a substantial match between the reference and the prediction for the best participant, with most mistakes occurring in short transitions between classes, such as between epoch 50 and 100 where a sequence of transitions between N3 and N2 lead to most mistakes in the first half of that recording. Regarding the participant with the performance closest to the median, in Fig. 4, it is clear that many classification errors occur once more during periods with faster transitions between stages, such as between epochs 550 and 600 where there are successive transitions between REM and N2, or between epochs 700 and 800 and at the end of the recording where there are successive transitions between Wake, REM and N1. Similar observations are made for ‘OSA’ participants, where the best and the participant with performance closest to the median are illustrated in Figures 5 and 6. Although the cause for these successive transitions is more likely related to the occurrence of apnea events, the classifier is once more not being able to fully cope with fragmented sleep architectures.

Regarding the classification of participants of the ‘all’ data set, Figs. 7 and 8 illustrate a comparison between the ground-truth reference hypnogram and the classification output for the participant with the best performance and the performance closest to the median in the ‘all’ data set. The first remarkable aspect in Fig. 7 is the extremely low sleep efficiency of this participant, who spent a large part of the night awake. The classifier is correctly able to detect this period, and most mistakes occur once more during short periods of N3 and N1 which alternate with N2. Regarding Fig. 8 we again see that many mistakes occur during fast transitions to and from short periods of N3 although in this particular participant we see a rather low sensitivity to
A comparison of probabilistic classifiers for sleep stage classification

Figure 3. Classification example for the participant with the highest $\kappa$ of ‘regular’ participants ($\kappa$ of 0.74 and an accuracy of 84.2%). From top to bottom: reference hypnogram, result of classification with CRFt classifier, epochs for which there is a mismatch between reference and classification.

Figure 4. Classification example for the participant closest to the median $\kappa$ of ‘regular’ participants ($\kappa$ of 0.51 and an accuracy of 64.3%). See caption to Fig. 3.

N3, likely related to a weaker expression of some of the hallmark cardiac and respiratory characteristics of N3 seen in most participants in the training set.

4.7. Comparison with earlier work and state of the art

In earlier work we used a linear discriminant similar to the LDtvp classifier used in this manuscript and, with a $\kappa$ of 0.49 achieved a moderate agreement with PSG, on par with the best classifiers reported in literature at the time (Fonseca et al. 2015). In that work we used a smaller data set, comprised exclusively of healthy participants with regular sleep architecture. In comparison with the present work, we are now able to achieve a moderate agreement with nearly the same performance ($\kappa$ of 0.48) but for a five-class classification task using the CRFt classifier. In more recent work on four-class

ACCEPTED MANUSCRIPT
A comparison of probabilistic classifiers for sleep stage classification

sleep stage classification of healthy adults, Beattie et al. (2017) used an accelerometer and a PPG sensor to extract movement, breathing and heart rate variability features, achieving a \( \kappa \) of 0.52 and accuracy of 69% on healthy adults. Albeit on a different, larger data set, and when trained exclusively with ‘regular’ participants with cardiac and respiratory (but not movement) features, our CRFt classifier obtained comparable results with an average and median \( \kappa \) of 0.50 and 0.53 and accuracy of 69.6% and 70.8% respectively.

Additionally, CRFt seems better able to handle participants with irregular sleep architectures and with sleep disorders, achieving consistently better results than LDtvp both with a generally trained classifier evaluated on ‘irregular’ and ‘OSA’ subsets, as well as with a dedicated classifier trained only on ‘OSA’ patients. In this regard, it is interesting to compare the results reported in literature for sleep stage classification of OSA patients. Redmond and Heneghan reported an accuracy and \( \kappa \) on a three-class

Figure 5. Classification example for the participant with the highest \( \kappa \) of ‘OSA’ participants (\( \kappa \) of 0.71 and an accuracy of 79.9%). See caption to Fig. 3.

Figure 6. Classification example for the participant closest to the median \( \kappa \) of ‘OSA’ participants (\( \kappa \) of 0.40 and an accuracy of 61.1%). See caption to Fig. 3.
A comparison of probabilistic classifiers for sleep stage classification

Figure 7. Classification example for the participant with the highest $\kappa$ of ‘all’ participants ($\kappa$ of 0.74 and an accuracy of 83.8%). See caption to Fig. 3.

Figure 8. Classification example for the participant closest to the median $\kappa$ of ‘all’ participants ($\kappa$ of 0.44 and an accuracy of 63.4%). See caption to Fig. 3.

task of 67% and 0.32 (Redmond & Heneghan 2006), achieving a fair agreement for 37 recordings, and Willemsen et al. reported, also for a three-class task, accuracy and $\kappa$ of 70% and 0.41 (Willemsen et al. 2015), thus achieving a moderate agreement for 25 recordings. In comparison, the CRFt classifier in our work achieved an accuracy and $\kappa$ of 75.4% and 0.50 respectively, for the same three-class problem, representing a substantial improvement, albeit on a different data set.

Overall, these results suggest that the CRF classifier is better at modeling the transitional characteristics between sleep stages, while offering an improved separation between sleep stages with overlapping cardiorespiratory characteristics. These advantages allow it to not only better learn and predict more complex and irregular sleep architectures, but also to make the first steps towards five-class classification, the standard for sleep scoring used in clinical PSG and, importantly, to achieve first promising results in the sleep stage classification of patients with sleep
A comparison of probabilistic classifiers for sleep stage classification

disorders.

4.8. Limitations

Finally, we should highlight the limitations of this technology. As discussed, the classification performance on OSA patients is higher when using a dedicated classifier than when using a classifier trained with healthy and disordered participants. In a real-world application, this would imply knowing, a priori, whether a patient suffers from a sleep disorder (OSA in this case) or not, before the appropriate classifier can be used. This may of course limit the immediate applicability of this technology in a clinical context, where the analysis of sleep architecture might be relevant for the complete understanding of a given condition in the first place. Alternative methods for automatic detection of, in this case, respiratory events, would be necessary in order to obtain a first segmentation of the participants between those with OSA and those without. Furthermore, OSA is only one of the many sleep disorders, and the applicability of these methods to patients suffering from other pathologies remains to be demonstrated. Regarding the HMM classifier, it clearly underperforms when compared with the LD and the CRF classifiers. A possible explanation for this is related to the limitations imposed by the nature of the HMM classifier used, where the discretization into a limited number of symbols - needed to compute the emission and transition probabilities - possibly limits the learning capabilities in this complex task. Although this problem was not so evident in our earlier results, where HMM was superior to LD for certain binary sleep classification tasks, this simplification is clearly not appropriate in the more complex problem of multi-class classification. The applicability of continuous HMM models remains to be evaluated.

5. Conclusions

The performance of a CRF classifier was evaluated in a home sleep monitoring context with cardiorespiratory features, and compared to HMM and LD classifiers. In earlier work we have shown how CRF can exploit transitional features and better handle non-separable features, but the evaluation was limited to binary classification in ‘one class versus the rest’ tasks. In this work we show that CRF is superior to the other classifiers also in the more useful multiple-stage classification problem. Furthermore, we explored the advantage of incorporating time as a feature for classification.

The results also demonstrated that five-class classification is feasible, and that dedicated three- and four-class classifiers do not outperform the results obtained by merging classes with a five-class classifier. Finally, it was shown that CRF classifiers also achieve moderate performance in the sleep stage classification of OSA patients, outperforming earlier work reported in literature for this population.

More research is needed to improve the sensitivity to N1 stage, and to evaluate this technology in subjects suffering from other sleep disorders. Furthermore, and
A comparison of probabilistic classifiers for sleep stage classification

given the improvements obtained with CRF over the single-epoch LD, higher-order multivariate Markov chains could be explored to make use not only of the previous state, but of a longer sequence of past states. Although this will introduce additional training complexity, it should lead to further performance improvements.

Although the current performance of cardiorespiratory sleep stage classifiers is, overall, still modest, the technology has the potential to complement PSG for long-term monitoring at home, potentially replacing the role that actigraphy had over the last decades in the assessment of sleep-wake disorders, and extend it with more detailed information about sleep architecture.

References

Agnew H W, Webb W B and Williams R L 1966 The first night effect: An EEG study of sleep Physiologial Psychology 2 263–6

Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W and Pollak C P 2003 The role of actigraphy in the study of sleep and circadian rhythms Sleep 26 342–92

Beattie Z, Oyang Y, Statan A, Ghoreyshi A, Pantelopoulos A, Russell A and Heneghan C 2017 Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals Physiological Measurement 38 1968–79

Bianchi M T, Cash S S, Mietus J, Peng C K and Thomas R 2010 Obstructive sleep apnea alters sleep stage transition dynamics PLoS ONE 5 e11356

Buyse D J, Reynolds C F, Monk T H, Berman S R and Rupfer D J 1989 The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research Psychiatry Research 28 193–213

Cohen J 1960 A Coefficient of agreement for nominal scales Educational and Psychological Measurement 20 37–46

de Chazal P, Fox N, O’Hare E, Heneghan C, Zaffaroni A, Boyle P, Smith S, O’Connell C and McNicholas W T 2011 Sleep/wake measurement using a non-contact biomotion sensor Journal of Sleep Research 20 356–66

Dennis, Jr J E and More J J 1997 Quasi-Newton methods, motivation and theory SIAM Review 19 46–89

Domingues A, Paiva T and Sanches J M 2014 Hypnogram and sleep parameter computation from activity and cardiovascular data IEEE Transactions on Biomedical Engineering 61 1711–9

Folstein M F, Folstein S E and McHugh P R 1975 “Mini-mental state”: A practical method for grading the cognitive state of patients for the clinician Journal of Psychiatric Research 12 189–98

Fonseca P, Aarts R M, Fousier J and Long X 2014 A novel low-complexity post-processing algorithm for precise QRS localization SpringerPlus 3 376–1/13

Fonseca P, den Teuling N, Long X and Aarts R M 2017 Cardiorespiratory sleep stage detection using conditional random fields IEEE Journal of Biomedical and Health Informatics 21 956–66

Fonseca P, Long X, Rathia M, Haakma R, Aarts R M and Rolink J 2015 Sleep stage classification with ECG and respiratory effort Physiological Measurement 36 2027–40

Hamilton P 2002 Open source ECG analysis Computers in Cardiology pp 101–4

Hamilton P S and Tompkins W S 1986 Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database IEEE Transactions on Biomedical Engineering 33 1157–65

Hedner J, White D P, Malhotra A, Herscovici S, Pittman S D, Zou D, Grote L and Pillar G 2011 Sleep staging based on autonomic signals: A multi-center validation study Journal of Clinical Sleep Medicine 7 301–6

Iber C, Ancoli-Israel S, Cherson A and Quan S F 2007 The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications 1st edn American Academy of Sleep Medicine Westchester, IL.
A comparison of probabilistic classifiers for sleep stage classification

Kloch G, Kemp B, Penzel T, Schlogl A, Rappelsberger P, Trenker E, Gruber G, Zeitlhofer J, Himanen S L, Kunz D, Barbanoj M J, Roschke J, Varri A and Dorffner G 2001 The SIESTA project polygraphic and clinical database IEEE Engineering in Medicine and Biology 20 51–7

Kortelainen J M, Mendez M O, Bianchi A M, Matteucci M and Cerutti S 2010 Sleep staging based on signals acquired through bed sensor IEEE Transactions on Information Technology in Biomedicine 14 776–85

Kschischang F R, Frey B J and Loeliger H A 2001 Factor graphs and the sum-product algorithm IEEE Transactions on Information Theory 47 498–519

Lafferty J, McCallum A and Pereira F C N 2001 Conditional random fields: Probabilistic models for segmenting and labeling sequence data Proceedings of the 18th International Conference on Machine Learning (ICML) pp 282–9

Long X, Foussier J, Fonseca P, Haakma R and Aarts R M 2014 Analyzing respiratory effort amplitude for automated sleep stage classification Biomedical Signal Processing and Control 14 197–205

Long X, Haakma R, Leufkens T R M, Fonseca P and Aarts R M 2015 Effects of between- and within-subject variability on autonomic cardiorespiratory activity during sleep and their limitations on sleep staging: a multilevel analysis Computational Intelligence and Neuroscience 2015 583820–1/17

Mack D C, Patric J T, Suratt P M, Felder R A and Alwan M 2009 Development and preliminary validation of heart rate and breathing rate detection using a passive, ballistocardiography-based sleep monitoring system IEEE Transactions on Information Technology in Biomedicine 13 111–20

Ohayon M M, Carskadon M A, Guilleminault C and Vitiello M V 2004 Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: Developing normative sleep values across the human lifespan Sleep 27 1255–73

Rabiner L R 1989 A tutorial on hidden Markov models and selected applications in speech recognition Proceedings of the IEEE 77 257–86

Ratnavadivel R, Chau N, Studler D, Yeo A, McEvoy R D and Catcheside P G 2009 Marked reduction in obstructive sleep apnea severity in slow wave sleep Journal of Clinical Sleep Medicine 5 519–24

Rechtschaffen A and Kales A 1968 A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects Public Health Service, U.S. Government Printing Office Washington, DC

Redmond J and Henehan C 2006 Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea IEEE Transactions on Biomedical Engineering 53 485–96

Redmond S J, de Chazal P, O’Brien C, Ryan S, McNicholas W T and Henehan C 2007 Sleep staging using cardiorespiratory signals Somatologie-Schlaforschung und Schlafmedizin 11 245–56

Roeris T, Zorick F, Wittig R, Conway W and Roth T 1989 Predictors of objective level of daytime sleepiness in patients with sleep-related breathing disorders Chest 95 1202–6

Strehl A and Ghosh J 2002 Cluster ensembles - a knowledge reuse framework for combining multiple partitions The Journal of Machine Learning Research 3 583–617

Sutton C and McCallum A 2007 Introduction to Statistical Relational Learning Vol. 93 pp 142–6

Van De Water A T, Holmes A and Hurley D A 2011 Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography - a systematic review Journal of Sleep Research 20 183–200

Viterbi A J 1967 Error bounds for convolutional codes and an asymptotically optimum decoding algorithm IEEE Transactions on Information Theory 13 260–9

Wilcoxon F 1945 Individual comparisons by ranking methods Biometrics Bulletin 1 80–3

Willemsen T, Van Deun D, Verhaert V, Vandekerckhove M, Exadaktylos V, Verbraecken J, Huffel S V, Haex B and Vander Slooten J 2014 An evaluation of cardio-respiratory and movement features with respect to sleep stage classification IEEE Journal of Biomedical and Health Informatics 18 661–9

Willemsen T, Varon C, Caicedo Dorado A, Haex B, Vander Slooten J and Van Huffel S 2015 Probabilistic
cardiac and respiratory based classification of sleep and apneic events in subjects with sleep apnea *Physiological Measurement* **36** 2103–18

Yu C, Liu Z, McKenna T, Reisner A T and Reifman J 2006 A method for automatic identification of reliable heart rates calculated from ECG and PPG waveforms *Journal of the American Medical Informatics Association* **13** 309–20