An evaluation of cardio-respiratory and movement features with respect to sleep stage classification

T. Willemen, D. Van Deun, V. Verhaert, M. Vandekerckhove, V. Exadaktylos, J. Verbraecken, S. Van Huffel, B. Haex, and J. Vander Sloten

***Abstract*— Polysomnography is considered the gold standard to assess sleep accurately, but it can be expensive, time-consuming and uncomfortable, specifically in long-term sleep studies. Actigraphy, on the other hand, is both cheap and user-friendly, but depending on the application lacks detail and accuracy. Our aim was to evaluate cardio-respiratory and movement signals in discriminating between Wake, REM, light (N1N2) and deep (N3) sleep. The dataset comprised 85 nights of polysomnography from a healthy population. Starting from a total of 750 characteristic variables (features), problem-specific subsets of 40 features were forwardly selected using the combination of a wrapper method (Cohen's Kappa statistic on RBF-kernel Support Vector Machine (SVM) classifier) and filter method (minimum Redundancy Maximum Relevance criterion on Mutual Information). Final classification was performed using an RBF-kernel SVM. Non subject-specific Wake versus Sleep classification resulted in a Cohen's kappa value of 0.695, while REM versus NREM resulted in 0.558 and N3 versus N1N2 in 0.553. The broad pool of initial features gave insight in which features discriminated best between the different classes. The classification results demonstrate the possibility of making long-term sleep monitoring more widely available.**

***Index Terms*— biomedical signal processing, data analysis, medical information systems, sleep research, supervised learning**

I. Introduction

SLEEP can roughly be divided into two main states, labeled Rapid-Eye-Movement (REM) and NREM (N1-N2-N3) sleep, which alternate in cycles of about 90 minutes [1]. NREM sleep, especially deep sleep (N3), is more prominent during the first hours of sleep and is essential towards physical recovery [2],[3]. REM sleep, linked to dreaming and more prominent during the last hours of sleep, acts towards the recovery of our mental state [4].

Demographics show that up to 24 % of the population is faced with regular sleep problems [5], due to e.g. insomnia (the inability to initiate and maintain sleep), obstructive sleep apnea syndrome (OSAS; upper airway collapse during sleep), or a mere lack of sleep hygiene.

Polysomnography (PSG) [6] is considered the gold standard in sleep research and allows to assess most aspects of sleep accurately. Unfortunately, it requires at least one night in a specialized sleep lab. Expert technicians apply an extensive amount of sensors to the patient (possibly affecting sleep) and evaluate the collected data manually in 30 second intervals. Due to the complexity of sleep, dependent on physiological as well as psychological factors, great inter-night variability can be present, requiring multiple recording nights to prevent high incidence of false positives and negatives [7]. These factors make the method costly and time-consuming, limiting its application on a large scale, especially when long-term monitoring is considered. The addition of automatic sleep classifiers on PSG signals can lessen the burden of manual evaluation for a small drop in accuracy [8].

Actigraphy (ACT) [9] on the other hand, which makes use of movement information to differentiate between Wake and Sleep, is both cheap and user-friendly. Unfortunately, for many cases, it lacks detail (such as the REM-NREM distinction) and accuracy [10],[11]. No other methods reached the point beyond prototype or have been sufficiently validated to fill the gap between PSG and ACT [12].

Over the years however, extensive research has been performed on changes in heart rate (HR) and breathing rate (BR) across sleep stages and other related events. In 1923, MacWilliams et al. [13] were an early pioneer in noting the influence of sleeping and dreaming on blood pressure and heart activity. In 1956, Brooks et al. [14] published the first

Manuscript received October 8, 2012. This work was supported in part by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT).

T. Willemen, D. Van Deun, V. Verhaert and J. Vander Sloten are with the Biomechanics Section, Mechanical Engineering Department, KU Leuven, Heverlee 3001, Belgium (email: tim.willemen@mech.kuleuven.be; dorien.vandeun@mech.kuleuven.be; vincent.verhaert@mech.kuleuven.be; bart.haex@kuleuven.be; jos.vandersloten@mech.kuleuven.be).

M. Vandekerckhove is with the Research Group of Biological Psychology, Vrije Universiteit Brussel, Brussels 1050, Belgium (email: marie.vandekerckhove@vub.ac.be).

V. Exadaktylos is with the Division of Measure, Model and Manage Bioresponses, KU Leuven, Heverlee 3001, Belgium (email: vasileios.exadaktylos@biw.kuleuven.be).

J. Verbraeken is with the Multidisciplinary Sleep Disorders Centre, Antwerp University Hospital, University of Antwerp, Edegem 2650, Belgium (email: johan.verbraecken@uza.be).

S. Van Huffel is with the Department of Electrical Engineering-ESAT SCD (SISTA), and iMinds Future Health Department, KU Leuven, Heverlee 3001, Belgium (email: sabine.vanhuffel@esat.kuleuven.be).

B. Haex is with the Biomechanics Section, Mechanical Engineering Department, KU Leuven, and Imec, Heverlee 3001, Belgium (email: bart.haex@kuleuven.be).

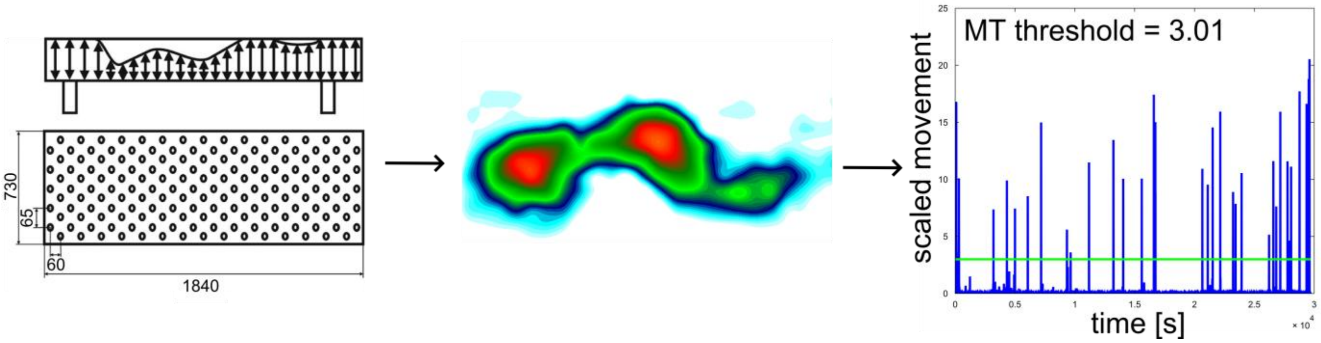

Fig. 1 Visualisation of the 2D measuring grid of the DynaSleep system (Custom8, Leuven, Belgium) on the left. In the center, an example of the measured indentation pattern from a person lying in the right lateral position. On the right, the extracted movement signal for one night, with the MT threshold indicated in green.

definitive measurements of HR variation during sleep in six individuals, each observed for 50 nights. In 1964, Snyder et al. [15] studied 12 subjects across 30 nights, reporting significant differences in blood pressure, HR and BR with varying sleep depth. Already in 1973 and 1976, attempts were made to distinguish active REM-N1 sleep from quiet N2-N3 sleep using HR data [16],[17]. HR and BR are currently known to be linked to the autonomic nervous system (ANS), from which sympathetic and parasympathetic variations correlate well with changes in sleep architecture [18]-[20]. HR and BR can be measured reliably using Electrocardiography (ECG) and Respiratory Inductance Plethysmography (RIP), limiting the amount of sensors attached.

This paper evaluates the use of heart rate, breathing rate and movement information in discriminating between between Wake, REM, light (N1N2) and deep (N3) sleep. It presents methods and results on data preprocessing, the extraction and selection of relevant features and the final classification step.

II. Materials and Methods

A. Data collection

The dataset comprises a total of 85 nights from a population of 36 healthy volunteers (age 22.1 ± 3.2 years), with one to four nights per subject. Participants were recruited through advertisement. Inclusion criteria were a good general health condition, regular sleep-wake schedule and an average of 6 to 9 hours of sleep per night. Exclusion criteria were sleeping disturbances, intake of medication or drugs influencing sleep, smoking habits or intake of more than three beverages containing caffeine or alcohol. The study was approved by the Ethics Committee at the ‘Vrije Universiteit Brussel’.

For all nights, complete polysomnographic recordings were performed (Dream system, Medatec NV, Brussels, Belgium), after which experts classified sleep stages (Wake, REM, N1, N2, N3) in 30 second epochs according to AASM rules (American Academy of Sleep Medicine) [6]. Additionally, HR was registered through Electrocardiogram (ECG), BR using Respiratory Inductance Plethysmography (RIP) and movement (MOV) by the DynaSleep system (Custom8, Leuven, Belgium). The latter continuously measures the perpendicular indentation of the mattress surface in a 2D-grid of 170 points. This allows for a sensitive and accurate registration of body movements, twitches and sleep postures, as described by Verhaert et al. [21].

B. Data pre-processing

HR was extracted from 200 Hz ECG measurements using the Pan-Tompkins algorithm [22]. A search-back post-processing algorithm was applied to identify and correct false positive and false negative R-peak detections. BR was extracted by first applying a cubic spline interpolation to the raw 200 Hz RIP data, smoothing the breathing signal. This allowed for an easy extraction of the valleys and peaks by differentiation, identifying inspiration, expiration and total breathing length intervals. Again, a search-back post-processing algorithm was applied to correct false positive and false negative peak and valley detections. The 1 Hz movement signal was evaluated by 2D integration of the derivative of the continuously measured indentation over the mattress surface. Normalization was performed by scaling this movement signal with the inverse of the subject’s Body Mass Index (BMI). Significant body movement events occur when the signal raises above the threshold value 3.01, as heuristically determined in [21]. Figure 1 visualizes the measurement system and shows an example of the movement signal for one night, including the MT threshold.

The interval size of the feature vectors to be classified was chosen to be 60 seconds, which is a trade-off between feature quality and time resolution. Especially for the breathing rate, smaller intervals can lead to unreliable ventilatory features due to the slow rhythm of the breathing signal. Also, since the lower boundary of the Low Frequency (LF) interval in Heart Rate Variability (HRV) analysis is put at 0.04 Hz, and the HRV Task Force [23] recommends an interval length of at least 10 times the wavelength of the lowest frequency bound (which would require an interval length of 250 seconds), the chosen 60 second interval is already quite short for HRV analysis. An artificial increase in time resolution can always be accomplished afterwards using a 60 second moving window, with a step size of e.g. 15 seconds. To transform the 30 second interval PSG-scored hypnogram values to 60 second interval values, the following decisions were made (in order of priority):

- intervals containing a significant body movement event were stored as MT
- intervals containing at least one PSG-scored Wake epoch were stored as Wake
- intervals containing at least one PSG-scored REM epoch were stored as REM
- the remaining intervals were stored as NREM

For example, a 60 second interval containing 30 seconds of Wake and 30 seconds of NREM was stored as Wake; a 60 second interval containing 30 seconds of MT and 30 seconds of REM was stored as MT.

Parts of the data had to be excluded due to bad signal registration (e.g. a loose ECG electrode or bad tension of an RIP breathing belt) in order to ensure proper training and validation data. Given the size of the dataset, two automatic signal quality evaluation methods were implemented for automatic exclusion analysis. For the ECG, the robust Mahalanobis distance between the low frequency energy content of every 60 seconds of ECG data was calculated. Energy content was estimated by integration of the squared wavelet transform coefficients of 10 levels of Daubechies-6 wavelets (db6) [24], representing the logarithmically spaced frequency spectrum between 0.3 and 10 Hz. The robust Mahalanobis distance is calculated as

$$D_M(x)=\sqrt{(x-\mu_r)^T S_r^{-1}(x-\mu_r)}, \qquad (1)$$

where $\boldsymbol{x}$ is the multivariate vector containing the 10 energy variables for that interval, $\boldsymbol{\mu_r}$ contains the robust average values of these 10 energy variables over all intervals of that night and $\boldsymbol{S_r}$ is the robust covariance matrix of the data. The robust Mahalanobis distance allows for an estimation of the similarity of the data in that interval to the complete set of intervals. A suitable rejection threshold for ECG intervals having a too large Mahalanobis distance was heuristically determined at a distance value of 30. Calculations were performed using Libra, a Matlab Library for Robust Analysis [25], more specifically using the mcdcov function.

For the RIP, the signal to noise ratio (SNR) of every interval was estimated by taking the ratio of the signal and noise variances,

$$SNR = 10 * \log_{10}\left(\frac{\mathrm{var}(originalSignal)}{\mathrm{var}(originalSignal - splinedSignal)}\right), \qquad (2)$$

were the ***originalSignal*** is the RIP signal and the ***splinedSignal*** is the cubic spline interpolated RIP signal. A suitable rejection threshold for RIP intervals having a too low SNR was heuristically determined at a dB value of 5. At lower SNR values, irregularities start to arise in the breathing waveforms due to insufficient tension in the breathing belt.

In the following sections, 57 nights were used for feature selection and training of the classification model parameters, while 28 nights were used as validation set. No nights from the same subject were present in both training and test set to ensure that there was no person-specific training possible. Overall, 12.00% of the training set 60 second intervals and 12.09% of the test set 60 second intervals were not included in the analysis based on the above described signal quality evaluation methods. Table I gives an overview of the amount and percentage of accepted intervals for the different sleep states. For the MT intervals, never rejected based on the signal quality evaluation methods, an equal percentage (respectively 12.00% and 12.09% for training and test set) was excluded from analysis to maintain a balanced data set. Since from every interval type a relatively equal amount of data was rejected, a realistic distribution of sleep data was ensured.

TABLE I
OVERVIEW OF THE AMOUNT AND PERCENTAGE OF ACCEPTED INTERVALS IN THE FINAL DATASET

Sleep state	Training set		Validation set	
	#	%	#	%
MT	2107	88.00	987	87.91
WAKE	1297	86.35	808	88.99
REM	3470	83.19	1595	85.48
N1N2	10396	90.76	5113	90.98
N3	6166	86.74	3045	84.23
TOTAL	23550	88.00	11574	87.91

C. *Feature extraction*

After the data preprocessing step, the following signals were obtained, divided in 60 second intervals and labeled with their respective sleep states (as defined in section II.B):

- Start time and length of each beat-to-beat RR-interval (HR)
- Start time and length of each breathing cycle (BR)
- Start time and length of the inspiratory phase of each breathing cycle, expiratory phase of each breathing cycle and the ratio between these lengths (BRin, BRout and BRinoutratio)
- The 1 Hz movement signal derived from the DynaSleep system (MOV)

For each of these intervals an extensive number of features was defined. The choice of evaluating a broad set of features instead of a limited one, will allow to find a more optimal set of features, with respect to the specific classification problem and population. An overview of the defined features is given in table II. For the HR signal, feature types 1 through 8 were extracted, giving a total of 81 HR features. The features within feature type 4 were calculated by first detrending the HR signal by substracting the mean HR value of the previous x seconds, with x equal to 150, 600, 1800 and 7200. The features within feature type 7 were calculated by integration of the squared wavelet transform coefficients of 40 levels of db6 wavelets representing the linearly spaced frequency spectrum between 0.01 Hz and 0.40 Hz. The wavelet transforms were executed on the 2 Hz cubic spline interpolated HR signal. The features within feature type 8 were calculated by taking the natural logarithm of ‘1 + the original feature value’.

For the BR, BRin, BRout and BRinoutratio signals, the same feature types were extracted, except for feature type 7 and its logarithmic transform within feature type 8, giving a total of 65 features each. For the MOV signal, feature types 9 through 13 were extracted, giving a total of 34 MOV features. Movement intensity within feature type 9 through 12 was calculated by integrating the MOV signal over each movement

TABLE II
FEATURES DEFINED ON THE SIGNALS HR (1-8), BR, BRIN, BROUT, BRINOUTRATIO (1-6, 8) AND MOV (9-13)

#	Feature name	Short description	Amount per signal
1	mean	mean length of RR intervals or breath cycles	1
2	percentiles	2.5, 10, 25, 50 (=median), 75, 90 and 97.5 percentile	7
3	inter percentile ranges	range between percentile 2.5-97.5, 10-90, and 25-75	3
4	detrended of 1-2-3	by subtracting the mean value of the previous x seconds, x ϵ [150 600 1800 7200].	44
5	variance	variance of RR intervals/breath cycles	1
6	median absolute deviation	median(abs(x_i-median)), with x_i = data points of interval; variants by switching median with mean	4
7	frequency HRV	VLF (0.01 – 0.03), LF (0.04 – 0.15), HF (0.16-0.40), LF/HF (absolute and relative)	8
8	log of 5-6-7	natural logarithm of '1 + original feature value'	13
9	time to next movement	time in seconds until movement signal ≥ intensity threshold x ϵ [0 0.75 2.5 7.5 25]	5
10	time to previous movement	cfr. feature 9	5
11	time in between movements	cfr. feature 9	5
12	log of 9-10-11	natural logarithm of '1 + original feature value'	15
13	amount of movements	amount of movements within centered window ϵ [60 150 300 600]	4

interval with all its values above a minimal threshold of 0.15 (to cancel out noise). The values of the different movement intensity thresholds were 0, 0.75, 2.5, 7.5 and 25. The window sizes used in feature type 13 were 60, 150, 300 and 600.

In total, this sums up to 375 features per 60 second interval. Finally, for each of the previously defined feature types, a night-specific normalized version of the feature type was defined. For every night in the dataset and for every feature, the robust mean and standard deviation were calculated (again using Libra [25], more specifically the median and madc functions). These parameters were then used to normalize the original feature values of every 60 second interval of that night, using the formula

$$feature_{normalized} = \left(\frac{feature_{original} - median}{madc} \right), \qquad (3)$$

This normalization should allow for a better interchangeability between features from different subjects and from different nights. The final amount of features thus sums up to 750.

D. *Feature selection*

In order to reduce computational cost, complexity and noise on the classification result, a subset of features was selected on the training set as being most descriptive for the targeted classification problem and population. For this purpose, a combination of a filter method and wrapper method was used. Filter methods are low in computational costs since they work independent from a classification algorithm, while wrapper methods require parameter training and effective classification. As classification algorithm, the libSVM C-implementation [26] of an RBF-kernel (Radial Basis Function) Support Vector Machine (SVM) was used, which requires two parameters to train, namely a cost parameter and regularization parameter. To save on computation time, the wrapper method was thus only used for the evaluation of single feature classification performance by calculating a five-fold cross-validated value of Cohen's kappa [27].

For evaluating synergies between features (constructing an optimized set of features that work well together), a filter method based on mutual information and a minimum Redundancy Maximum Relevance-criterion was used [28]. The criterion is defined as,

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j;c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j;x_i) \right], \qquad (4)$$

a greedy forward selection algorithm with ***X*** representing the complete set of features x_j, ***S*** the subset of already selected features x_i (of size *m*), *c* representing the class identifier and *I* the mutual information, defined as

$$I(x;y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy, \qquad (5)$$

with ***p(x)***, ***p(y)*** and ***p(x,y)*** probability density functions. These functions were approximated by an adaptive diffusion-based kernel density estimator [29]. The left side of figure 2 shows an example of the probability density functions **$p(x_j,c_1)$** and **$p(x_j,c_2)$**, used for the calculation of $I(x_j;c)$. The more overlap, the lower the mutual information will be. The right side of figure 2 shows **$p(x_j,x_i)$**, used for calculating $I(x_j,x_i)$. The more the distribution follows the diagonal, the higher the mutual information between the two features.

The feature selection process was thus performed in two steps. In step one, the previously described wrapper method (RBF-kernel SVM) was used in order to obtain a fast reduction in feature amount, and this for each of the four signal types separately (HR, BR, BRin/BRout/BRinoutratio and MOV) in order to ensure the presence of features across all signals. As a result, 75% of features having the lowest values of Cohen's kappa were discarded from the feature set, as to only keep those features already reaching a decent classification accuracy on their own. In step two, a subset of 10 synergetic features per signal type was selected out of the remaining feature set using the previously described filter method (mRMR-criterion), leading to a final subset of 40 features. Further optimization of this set (e.g. by further reducing its size, while evaluating the impact on classification accuracy) was no part of this study.

E. *Feature classification*

In order to distinguish Wake from Sleep, both at sleep onset (the first hour of every night) and during the night (the remaining hours of the night), and to distinguish NREM sleep from REM sleep and light sleep (N1N2) from deep sleep (N3),

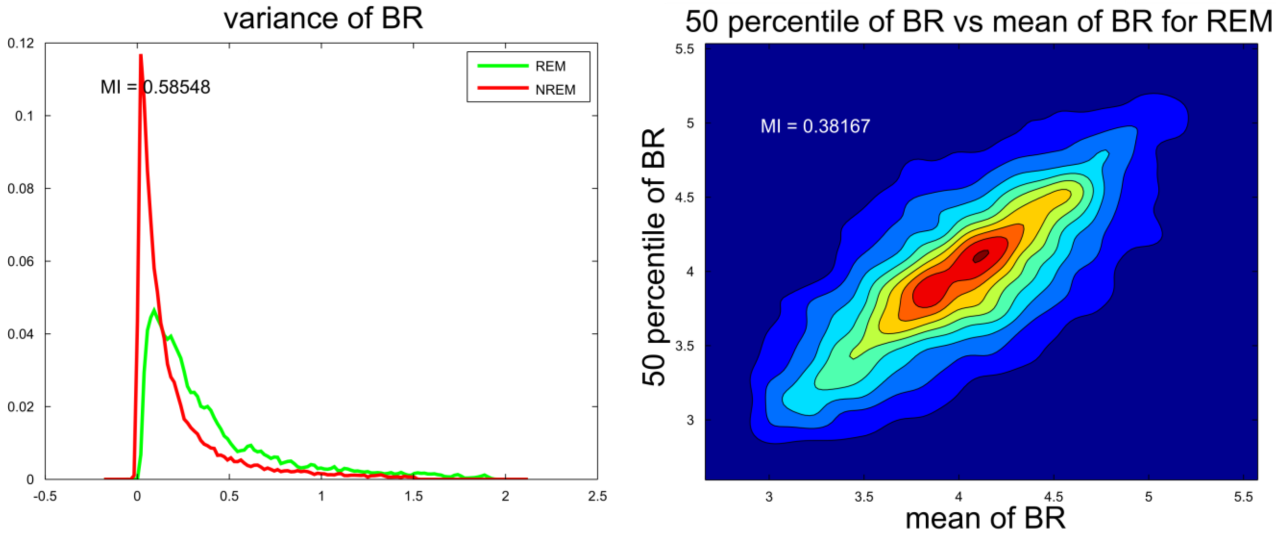

Fig. 2 Examples of feature-class (on the left) and inter-feature (on the right) probability density functions used to calculate the mutual information values of the mRMR-criterion. More overlap in the feature-class distributions will result in smaller MI between the feature and the classes. The more the inter-feature distributions follow the diagonal, the higher their MI will be.

four separate binary classification problems were assessed. For this purpose, four separate feature sets were selected, one for each classification problem. The results of the different binary classification problems were also combined to evaluate the complete Wake-Sleep, Wake-REM-NREM and WAKE-REM-N1N2-N3 classification problems. In all classification problems, MT intervals were scored as Wake.

As already stated in the previous section, classification was performed using the libSVM C-implementation of an RBF-kernel SVM. Its two parameters, the cost (penalizing value for misclassifications during training of the classifier) and the RBF-kernel parameter gamma, were optimized using five-fold cross-validation and gridsearch on the training set. The optimal parameters were selected based on a maximal averaged value of Cohen’s kappa over the five folds.

Besides agreements and Cohen’s kappa, also precision and recall were calculated. Recall quantifies the amount of class one instances (with class one being the smallest of the two classes) identified within the complete set of instances, while precision quantifies how precise the classifier was in this identification. Hence, there is normally a trade-off between precision and recall.

The cost penalizing value parameter of the SVM can have a different value for each of the two classes, in order to cope with unbalanced datasets in different ways. Increasing the cost value for class one will cause an increase in precision (less false positives), but a drop in recall (more false negatives). Tuning the difference between the cost value of class one and two allows to maximize Cohen’s kappa, although for some applications one could be more interested in maximizing precision or recall itself.

During the five-fold cross-validated training phase of our classifiers, we first optimized this ratio between the cost parameter of the two classes, using the default recommended values for cost (=1) and gamma (=1/feature amount = 1/40). In the next step, using this optimal ratio, the five-fold cross-validated gridsearch was performed over cost and gamma parameters.

III. Results

Table III gives an overview of the averaged cross-validated training results and the test set results for all four classification problems. The next subsections will only hold information about which features were selected, and thus deemed distinctive, for all classification problems.

A. *Wake vs Sleep at onset: <60 min*

As features, mean and percentile features (25th, median, 75th) were the most important for the HR and BR signal, with their average values being higher in Wake compared to Sleep. Normalization was important here, proven by the fact that 7/10 selected HR and 9/10 selected BR features were normalized. Detrending was less important (6/10 HR and 3/10 BR) mainly because no data was present to detrend with from before subjects went to bed, thus detrending was done with only the data from bed time on. Since no BRinoutratio features were selected, the perceived trend in increasing BRout interval lengths during sleep onset compared to BRin interval lengths was not significantly present enough for all subjects. The selected BRin and BRout features followed the same characteristics as the selected BR features. As for movement features, the logarithmic transform was important (7/10 selected MOV features) to make its distribution of values Gaussian. Selected features were time between higher intensity movements (2.5, 7.5, 25; no twitches) and time to the last higher intensity movement.

B. *Wake vs Sleep after onset: > 60 min*

As HR features, a raise in heart rhythm was found distinctive, characterized by mean and upper percentile (median, 75th, 90th, 97.5th) features being selected, coupled

TABLE III
RESULTS OF THE DIFFERENT CLASSIFICATION PROBLEMS (IN PERCENTAGES)

	Training set				Validation set			
	WAKE-SLEEP < 60 min[a]	WAKE-SLEEP > 60 min[b]	REM-NREM	N3-N1N2	WAKE-SLEEP < 60 min	WAKE-SLEEP > 60 min	REM-NREM	N3-N1N2
Agreement	90.21	91.14	84.21	78.59	89.55	91.96	86.24	79.48
Kappa	74.10	66.08	49.27	53.56	76.23	66.55	55.81	55.25
Precision	78.13	61.54	53.40	72.78	86.75	65.96	55.78	74.99
Recall	83.33	84.37	65.71	67.88	81.36	77.32	75.30	67.55

[a]Wake versus Sleep classification problem only using data from the first hour of every night. [b]Only using data after first hour.

to detrending over 150 seconds (selected for 7/10 features; 9/10 features were detrended). For BR features, it was primarily the higher variance (also its variants MAD and IQR) that was found distinctive, with again detrending found important (8/10; 4/10 was detrended over 150 seconds). Overall, normalization was less important here (5/10 HR and 1/10 BR), probably because longer periods of Wake (like with insomnia patients) were not present in the dataset, and periods of Wake were thus primarily coupled to a higher amount of movement activity, leading to a more similar overall higher heart and breathing rhythm for all subjects. There were 7 BRinoutratio features selected, distributed over variance (and its variants) and extreme percentiles at both ends (thus also reflecting variance). This probably reflects the difference between controlling our breathing more consciously (sympathetic) or it being controlled completely unconsciously during sleep (parasympathetic). For movement features, again the logarithmic transform was important (10/10), and selected features were the time to the last movement or time to the next movement with at least an intensity of 2.5 (so again no twitches).

C. *REM vs NREM*

As HR and BR features, mainly the presence of short periods of higher heart and breathing rhythm were found distinctive for REM sleep, reflected in the selection of upper percentile features (75th, 90th, 97.5th) but no mean. For BR, also a higher MAD and IQR in REM was characteristic. Normalization was again important (7/10 HR and 7/10 BR), together with detrending although this time over a longer timeframe of 7200 seconds (7/10 HR and 7/10 BR). There were also 3 BRinoutratio features selected (MAD and IQR), probably again reflecting the shifting balance between sympathetic (REM) and parasympathetic (NREM) activity. All other features were BRout features, and no BRin features, which seems to tell that the changes in breathing rhythm are mainly caused by changes in the expiration phase. Besides the necessary logarithmic transform (7/10), selected MOV features were all related to the time between movements of any intensity, which means also short twitches.

D. *N3 vs N1N2*

As HR and BR features, the presence of a lower heart and breathing rhythm within N3 sleep was found distinctive, reflected in the selection of lower percentile features (2.5th, 10th, 25th) but this time also mean and median for HR. For BR, also variance, MAD and IQR was selected, them being significantly lower in N3 sleep. Normalization of features was again important (6/10 HR and 10/10 BR), as was detrending (9/10 HR detrended over 600 sec and 5/10 BR). The detrending length of 600 seconds is remarkable since most periods of N3 sleep lasted longer than 10 minutes, which would make a longer detrending period more logical. The next possible interval however (1800 seconds) could have been too large. Adding more possible interval lengths should be considered. No BRinoutratio features were selected; BRin and BRout features followed the same characteristics as the selected BR features. As for movement features, the logarithmic transform (7/10) and the time to the last movement of any intensity was important. This means that longterm absence of movement is a good predictor for N3 sleep.

IV. DISCUSSION

An extensive study of feature extraction, selection and classification was presented for Wake-Sleep, REM-NREM and N3-N1N2 classification. Table IV compares the obtained classification results to similar classification algorithms found in literature. The last row of the table lists the presented results of this study. It has by far the largest dataset, even with 10% of the data being excluded due to noise, mostly caused by the RIP-signal for the detection of breathing. The use of a nasal thermistor would have led to a better quality signal, but is less comfortable for the subject, with more interference of normal sleep. The large dataset allowed for a reliable feature selection process, algorithm training and classifier validation. The resemblance in Cohen's Kappa values on training and validation set (table III) proves the absence of overtraining, and thus the reliable interchangeability between different healthy subjects and nights. In the case of sleep-disordered patients however, separate training data sets will be required to classify for example breathing disorder related events as apneas or periodic leg movement disorder events. During these events, output from the currently presented classifier will not be meaningful and should be disregarded. For the case of insomnia, a separate training data set with a higher presence of wake after sleep onset would probably give a higher classification accuracy, since this information was barely present in the currently used data set.

The use of different cost parameter ratios within the classifier allowed to optimize the trade-off between false positives and false negatives, leading to the highest kappa

TABLE IV
OVERVIEW OF SIMILAR CLASSIFICATION ALGORITHMS FOUND IN LITERATURE

Author	Signals	Classification	Acc	Kappa	Nights	Subject specific	Interval length
Harper [30]	HR, BR	WAKE-REM-NREM	85%	-	25 healthy	No	60s
	HR		82%	-			
	BR		80%	-			
Redmond [33]	HR, BR	WAKE-REM-NREM	79%	0.56	37 OSAS	Yes[a]	30s
	HR, BR		67%	0.32		No	
	EEG		87%	0.75		Yes[a]	
	EEG		84%	0.68		No	
Redmond [34]	HR, BR	WAKE-SLEEP	89%	0.60	31 healthy	No	30s
		WAKE-REM-NREM	76%	0.46			
Canisius [35]	HR	WAKE-REM-NREM	76%	-	18 healthy	No	30s
Devot [31]	HR, BR, MOV	WAKE-SLEEP	96%	0.70	9 healthy	No	30s
	HR, BR, MOV		85%	0.61	27 insomniacs		
	Actigraphy		94%	0.51	9 healthy		
	Actigraphy		78%	0.39	27 insomniacs		
Kortelainen [36]	HR, MOV	WAKE-REM-NREM	79%	0.44	18 healthy	No[b]	30s
Mendez [32]	HR	REM-NREM	79%	-	25 healthy	No	30s
Migliorini [37]	HR, BR, MOV	WAKE-REM-NREM	77%	0.55	17 healthy	No[b]	30s
This paper	HR, BR, MOV	WAKE-SLEEP	92%	0.69	85 healthy	No	60s
		REM-NREM	86%	0.56			
		N3-N1N2	79%	0.55			
		WAKE-REM-NREM	81%	0.62			
		WAKE-REM-N1N2-N3	69%	0.56			

[a] Test and training set information derived from the same night → possible overestimation of subject-specific classification potential
[b] Test and training set contain nights from the same individuals → possible overestimation of non-subject-specific classification potential

values. In some applications, one can however prefer a higher sensitivity, or a higher precision, thus preferring a result with a lower kappa value.

Some authors [30]-[37] used time-dependent a priori probabilities to increase the accuracy of their classifiers, giving a higher chance to the classifier of scoring Wake, REM, N1N2 or N3 sleep in certain parts of the night. Although this leads in general to better results for normal healthy sleep, it can lead to distorted results when less than normal sleep characteristics occur. Others [31]-[32],[36]-[37] used smoothing algorithms such as median filters or averaging over e.g. 15 minute intervals, either on the classified hypnogram or on the feature values themselves. While this can improve overall accuracy, it gives rise to a drop in time resolution and will likely distort the time location of sleep stage transitions. In some applications however, when only a rough overview of the sleep stage distribution of a person is required, these methods could significantly improve the overall correlation between an EEG-based and a cardio-respiratory based sleep assessment.

Harper et al. [30] and Devot et al. [31] showed that a combination of signals (cfr. their results in table IV), can improve classification accuracy significantly. The use broad pool of initial features on this combination of signals is one of the main factors that contributed to the good classification accuracies in this study. A thorough description of the selected features is discussed in the results section and can be used to optimize future feature starting sets.

In the REM-NREM classification problem, the absence of features regarding the energy spectrum of Heart Rate Variability (HRV) is remarkable, since it was the main investigated feature in the study of Mendez et al. [32], and important in many others. While NREM sleep is associated with a decrease in both frequency and power of sympathetic bursts, no complete absence of this activity was observed. Somers et al. [18] and Bonnet et al. [19] described that sympathetic bursts will occur e.g. at locations of K-complexes and arousals. Furthermore, during REM sleep, sympathetic activity is mainly concentrated during periods of rapid-eye movement and complete muscle atonia. In between, there is a marked decrease in activity, which makes them looking more like NREM periods. The use of short intervals without smoothing or averaging could thus lead to false positive and negative classifications. While these sympathetic activity changes should also have an influence on heart and breathing rate rhythm and variance, their impact proves to be less. It could be hypothesized that the change in rhythm and variance is only a second order effect, thus inherently filtering out the more rapid changes in presence and absence of sympathetic activity. Further on, since the HRV Task Force [23] recommends interval lengths of at least 10 times the wavelength of the lowest frequency bound, the chosen 60 second intervals might just be too short and unreliable for frequency HRV analysis.

Bonnet et al. [19] also described that the shift in HRV (from low to high sympathetic activity) already starts a few minutes before PSG-scored REM onset, and also lasts a few minutes longer than the PSG-scored REM period. In PSG, the epochs after REM sleep are usually scored as light N1 sleep, and in over 50 percent of the cases the epochs before REM sleep too [38]. Interesting is the fact that, when only using EEG activity information, sleep stages N1 and REM are almost impossible to distinguish. This is one of the reasons why they used to be combined to describe a single sleep state. The above could be the source of difficult to avoid REM-NREM misclassifications. A possible solution would be to combine sleep stages N1 and REM again in cardio-respiratory based sleep scoring.

Redmond et al. [33] showed that an automatic classifier based on EEG signals has a much smaller drop in accuracy when going from subject-specific to non-subject-specific, compared to a classifier based on HR and BR. This can be related to the greater variability in heart and lung strength and capacity among different subjects, compared to a greater similarity in brain functioning. While this reveals a possible limitation in HR and BR classification accuracy, the current results show that sufficiently accurately results are feasible whilst using the right features.

Besides Harper et al. [30] and this study, all other authors made use of 30 second intervals instead of 60 second intervals, mainly because 30 second intervals are considered the gold standard in polysomnography, acting as a reference method. The predominant frequencies observed in our brain waves are however an order of two greater than the frequencies observed in our heart and breathing rhythm, making the choice for 30 second intervals in cardio-respiratory based sleep scoring less obvious. Future work must therefore investigate the influence of interval length on feature quality and classification accuracy.

It is important to mention that, due to inter-scorer variability present in expert polysomnography, the reference intervals are not always correct. Danker-Hopfe et al. [39] reported an inter-scorer Cohen's Kappa of 0.7626 in a database of 56 healthy and 16 sleep disordered patients, when considering a 5-class Wake-REM-N1-N2-N3 classification according to AASM rules. Automatic classifiers, even on EEG signals, are still not capable of attaining or surpassing this accuracy. They are however consistent in the decisions and mistakes they make, which is not the case in humans.

V. Conclusion

An extensive study of feature extraction, selection and classification was presented, for Wake-Sleep, REM-NREM and N3-N1N2 classification. The broad pool of initial features gave insight in which features discriminated best between the separate classes. Attributes such as normalization and detrending proved vital. The use of different cost ratio parameters within the classifier allowed for a trade-off between sensitivity and precision. Future work will investigate the influence of interval length on feature quality and classification accuracy, and will evaluate the classification method on sleep-disordered patients if the necessary datasets can be acquired.

References

[1] A. Rechtschaffen, and A. Kales, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," in *National Institutes of Health Publications No. 204*, Washington D.C.: U.S. Government Printing Office, 1968.

[2] B. Haex, *Back and Bed: Ergonomic Aspects of Sleeping*. Boca Raton, FL: CRC Press, 2004.

[3] K. Adam, and I. Oswald, "Sleep helps healing," *British Medical Journal*, vol. 289, pp. 1400-1401, Nov.1984.

[4] P. Meerlo, R. E. Mistlberger, B. L. Jacobs, H. C. Heller, and D. McGinty, "New neurons in the adult brain: the role of sleep and consequence of sleep loss," *Sleep Medicine Reviews*, vol. 13, pp. 187-194, Jun. 2009.

[5] The National Sleep Foundation. (2012). *What makes a good night's sleep* [online]. Available: http://www.sleepfoundation.org

[6] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan, "The AASM Manual for the Scoring of Sleep and Associated Events," Westchester, IL: American Academy of Sleep Medicine, 2007.

[7] J. Carlile, and N. Carlile, "Repeat study of 149 patients suspected of having sleep apnea but with an AHI < 5," *Sleep*, vol. 31, pp. A153, 2008.

[8] T P. Anderer, A. Moreau, M. Woertz, M. Ross, G. Gruber, S. Parapatics et al., "Computer-Assisted Sleep Classification according to the Standard of the American Academy of Sleep Medicine: Validation Study of the AASM Version of the Somnolyzer 24x7," *Neuropsychobiology*, vol. 62, pp. 250-264, Sept. 2010.

[9] T. Morgenthaler, C. Alessi, L. Friedman, J. Owens, V. Kapur, B. Boehlecke, T. Brown, A. C. Junior, J. Coleman, T. Lee-Chiong, J. Pancer, and T. J. Swick, "Practice Parameters for the Use of Actigraphy in the Assessment of Sleep and Sleep Disorders: An Update for 2007," *Sleep*, vol. 30, pp. 519-529, Apr. 2007.

[10] J. Paquet, A. Kwainska, and J. Carrier, "Wake detection capacity of actigraphy during sleep," *Sleep*, vol. 30, pp. 1362-1369, Oct. 2007.

[11] A. Bulckaert, V. Exadaktylos, G. De Bruyne, B. Haex, E. De Valck, J. Wuyts, J. Verbraecken, and D. Berckmans, "Heart rate based nighttime awakening detection," *European Journal of Applied Physiology*, vol. 109, pp. 317-322, May 2010.

[12] A. T. Van de Water, A. Holmes, and D. A. Hurley, "Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography: a systematic review," *Journal of Sleep Research*, vol. 20, pp. 183-200, Mar. 2011.

[13] J. A. MacWilliams, "Some applications of physiology to medicine: III. Blood pressure and heart action in sleep and dreams: their relation to haemorrhages angina and sudden death," *The British Medical Journal*, vol. 2, pp. 1196-1200, Dec. 1923.

[14] C. M. Brooks, K. S. Coleman, B. F. Hoffman, F. Kleyntjens, E. H. Koenig, E.E. Suckling, and H. J. Treumann, "Sleep and variations in certain functional activities accompanying cyclic changes in depth of sleep," *Journal of Applied Physiology*, vol. 9, pp. 97-104, Jul. 1956.

[15] F. Snyder, J. A. Hobson, D. F. Morrison, and F. Goldfrank, "Changes in respiration, heart rate and systolic blood pressure in human sleep," *Journal of Applied Physiology*, vol. 19, pp. 417-422, May 1964.

[16] A. J. Welch, and P. C. Richardson, "Computer sleep stage classification using heart rate data," *Electroencephalography and Clinical Neurophysiology*, vol. 34, pp. 145-152, Feb. 1973.

[17] M. J. Lisenby, P. C. Richardson, and A. J. Welch, "Detection of cyclic sleep phenomena using instantaneous heart rate," *Electroencephalography and Clinical Neurophysiology*, vol. 40, pp. 169-177, Feb. 1976.

[18] V. K. Somers, M. E. Dyken, A. L. Mark, and F. M. Abboud, "Sympathetic-nerve activity during sleep in normal subjects," *New England Journal of Medicine*, vol. 328, pp. 303-307, Feb. 1993.

[19] M. H. Bonnet, and D. L. Arand, "Heart rate variability: sleep stage, time of night and arousal influences," *Electroencephalography and Clinical Neurophysiology*, vol. 102, pp. 390-396, May 1997.

[20] H. J. Burgess, J. Trinder, and Y. Kim, "Cardiac autonomic nervous system activity during pre-sleep wakefulness and stage 2 NREM sleep," *Journal of Sleep Research*, vol. 8, pp. 1113-122, Jun. 1999.

[21] V. Verhaert, B. Haex, T. De Wilde, D. Berckmans, M. Vandekerckhove, J. Verbraecken, and J. Vander Sloten, "Unobtrusive assessment of motor patterns during sleep based on mattress indentation measurements," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, pp. 787-794, Mar. 2011.

[22] J. Pan, and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 32, pp. 230-236, Mar. 1985.

[23] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart Rate Variability: standards of measurement, physiological interpretation and clinical use," *Circulation*, vol. 93, pp. 1043-1065, Mar. 1996.

[24] I. Daubechies, Ten lectures on wavelets. Philadelphia, PA: CBMS-NSF, 1992.

[25] S. Verboven, and M. Hubert, “LIBRA: a Matlab library for robust analysis,” *Chemometrics and Intelligent Laboratory Systems*, vol. 75, pp. 127-136, Feb. 2005.

[26] C.-C. Chang, and C.-J. Lin, “LibSVM: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1-27, Dec. 2011.

[27] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol 33, pp. 159-174, Mar. 1977.

[28] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226-1238, Aug. 2005.

[29] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, “Kernel density estimation via diffusion,” *The Annals of Statistics*, vol. 38, pp. 2916-2957, 2010.

[30] R. M. Harper, V. L. Schechtman, and K. A. Kluge, Machine classification of infant sleep state using cardiorespiratory measures, Electroencephalography and Clinical Neurophysiology 67 (1987), 379-387.

[31] S. Devot, and R. N. E. Dratwa, “Sleep/wake detection based on cardiorespiratory signals and actigraphy,” in Proceedings *of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Buenos Aires, 2010, pp. 5089-5092.

[32] 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,” *International Journal of Biomedical Engineering and Technology*, vol. 3, pp. 246-263, Apr. 2010.

[33] S. J. Redmond, and C. Heneghan, “Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea,” *IEEE Transactions on Biomedical Engineering*, vol. 53, pp. 485-496, Mar. 2006.

[34] S. J. Redmond, P. de Chazal, C. O’Brien, S. Ryan, W. T. McNicholas, and C. Heneghan, “Sleep staging using cardiorespiratory signals,” *Somnologie*, vol. 11, pp. 245-256, Oct. 2007.

[35] S. Canisius, T. Ploch, V. Gross, A. Jerrentrup, T. Penzel, and K. Kesper, “Detection of sleep disordered breathing by automated ECG analysis,” in *Proceedings of the 30th Annual Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Vancouver, 2008, pp. 2602-2605.

[36] J. M. Kortelainen, M. O. Mendez, A. M. Bianchi, M. Matteucci, and S. Cerutti, “Sleep staging based on signals acquired through bed sensor,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, pp. 776-785, May 2010.

[37] M. Migliorini, A. M. Bianchi, N. Domenico, J. Kortelainen, E. Arce-Santana, S. Cerutti, and M. O. Mendez, “Automatic sleep staging based on ballistocardiographic signals recorded through bed sensors,” in *Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Buenos Aires, 2010, pp. 3273-3276.

[38] A. Kishi, H. Yasuda, T. Matsumoto, Y. Inami, J. Horiguchi, M. Tamaki, Z. R. Struzik, and Y. Yamamoto, “NREM sleep stage transitions control ultradian REM sleep rhythm,” *Sleep*, vol. 34, pp. 1423-1432, Oct. 2011.

[39] H. Danker-Hopfe, P. Anderer, J. Zeitlhofer, M. Boeck, H. Dorn, G. Gruber, E. Heller, E. Loretz, D. Moser, S. parapatics, B. Saletu, A. Schmidt, and G. Dorffner, “Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard,” *Journal of Sleep Research*, vol. 18, pp. 74-84, Mar. 2009.