Improving sleep/wake classification with recurrence quantification analysis features

Citation for published version (APA):

Rolink, J., Fonseca, P., Long, X., & Leonhardt, S. (2019). Improving sleep/wake classification with recurrence quantification analysis features. *Biomedical Signal Processing and Control*, 49, 78-86.
<https://doi.org/10.1016/j.bspc.2018.11.005>

DOI:

[10.1016/j.bspc.2018.11.005](https://doi.org/10.1016/j.bspc.2018.11.005)

Document status and date:

Published: 01/03/2019

Document Version:

Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Improving Sleep/Wake Classification with Recurrence Quantification Analysis Features

Authors: Jérôme Rolink, Pedro Fonseca, Xi Long, Steffen Leonhardt

J. Rolink (corresponding author) and S. Leonhardt are with

Philips Chair for Medical Information Technology (MedIT), RWTH Aachen University,

Pauwelsstrasse 20, 52074 Aachen, Germany

Tel.: +49-241-8023211

Fax: +49-241-8082442

E-mail: rolink@hia.rwth-aachen.de

P. Fonseca and X. Long are with

Philips Research, HTC, Prof. Holstlaan 4, 5656

AE Eindhoven, The Netherlands

Abstract

In this work the method of Recurrence Quantification Analysis (RQA), often used for the analysis of complex dynamic systems, is employed to extract novel features for sleep/wake classification only using cardio-respiratory signals like electrocardiogram (ECG), heart rate (HR) and respiratory effort (RE). A polysomnography data set consisting of 313 full-night recordings is used to evaluate the features. The sleep/wake classification is performed with a classifier based on Linear Discriminant Analysis (LDA) and it is validated in a Leave One Subject Out cross-validation (LOSOVC) scheme. More than 1300 features are extracted from five different cardio-respiratory modalities (ECG, RE, HR, and combinations of ECG+RE and RE+HR). Each modality is processed to obtain 13 basic RQA features, five post-processed versions and, furthermore, three normalizations are applied, leading to a total count of 975 RQA features. From literature, 126 known cardio-respiratory and actigraphy features, normalized with the same procedures thus resulting in 378 distinct features, are used for performance comparison. A feature selection method based on the Mahalanobis distance and the inter-feature correlation is used to determine the most relevant features of both sets, resulting in a set of 158 from the existing features (set A) and 232 from the RQA based features (set B). The pooled Cohen's kappa coefficient for set A and set B is 0.586 and 0.522, respectively. The combination of both feature sets (set C) improves the kappa value to 0.625.

In addition, ROC and PR curves with their corresponding Area Under Curve (AUC) values are computed. It is also shown that derived sleep statistics (sleep efficiency, sleep onset, total sleep time, etc.) deviate less with respect to the ground truth annotations using the additional RQA features compared to the use of literature-based features only.

Besides the overall performance, the data set is split into six different but almost equally sized age and sex groups to allow unbiased comparisons among each other. It is shown, that the classification performance decreases with increasing age but is almost independent from sex. Hence, for future work it is suggested to implement the RQA features and also to focus on the improvement of features especially for elderly persons.

Keywords: Recurrence Quantification Analysis, feature extraction, sleep/wake classification

1 INTRODUCTION

Determining sleep stages is originally performed via the acquisition of the electrical activity of the brain using electroencephalography (EEG) during the night, which is a complex measurement task. To perform a polysomnography (PSG), a trained sleep technician has to apply several electrodes onto the scalp of the subject before going to sleep, according to the acquisition guidelines defined by the American Academy of Sleep Medicine (AASM) [1]. Besides possible faulty application of the electrodes, the EEG acquisition is also very sensitive to environmental electrical noise, body movements and even would disturb regular sleep since the wearing of the electrodes and the cables is not very comfortable. In addition to the EEG, nowadays, the electrocardiogram (ECG), respiratory activity (e.g. by measuring flow or respiratory effort, RE), electrooculogram (EOG) and limb muscle activity are recorded during a regular PSG. Normally, the acquisition of ECG is done with two standard self-adhesive electrodes. Same for EOG, which measures the eye movement electrically with electrodes attached next to and/or above the eyes. Respiration effort is recorded with an induction plethysmography or a piezoelectric belt attached to the chest and abdomen, and respiration flow is measured with a thermistor under the nostrils, an oronasal cannula or, in case of positive airway pressure titration sessions, a facial mask. Muscle activity is quantified with electromyography (EMG) measurements at the limbs. Even though the additional measures of EOG and EMG can be captured much more easily than EEG it is still necessary to attach electrodes to the skin. Respiration and cardiac activity may be derived in a contactless or non-obtrusive way, enabling long-term measurements, which is nearly impossible for EEG. New possibilities, e.g. data acquisition in a familiar environment for unobtrusive sleep/wake detection to identify lack of sleep or to monitor quality of sleep over a longer time than a single night are made available.

Researchers have been trying to use these simpler cardio-respiratory modalities to detect sleep stages. The major aim is to obtain a good alternative to the traditional cost-intensive EEG method without deteriorating detection performance. However, this is not a trivial task since some of the sleep stages, e.g. the first stage of non-rapid eye movement (NREM), are very difficult to be distinguished without EEG. A first step is hence to detect the difference between the wake and any sleeping stage, because not only the quantity of sleep - described by the total sleep time (TST) - is important for restorative sleep but also the sleep quality. Increased counts of wake stages or arousals are indicators for a sleep or breathing disorder, causing bad sleep quality. For obstructive sleep apnea (OSA) patients, sleep fragmentation – an indicator of restlessness during sleep – can be estimated by respiratory events, such as the airway narrowing or collapse which can cause the patient to be aroused from sleep or to wake up [2].

In literature, several linear and non-linear features extracted from ECG, respiration and actigraphy are known for the task of sleep classification [3]–[20]. They are either

derived from time or frequency domain or based on statistical methods. This work will investigate the use of the so called Recurrence Quantification Analysis (RQA) for sleep/wake classification. It has been shown in previous investigations that RQA principally gives new possibilities for the extraction of more sophisticated features to differentiate between sleep and wake just using cardio-respiratory modalities [21]. Hence, the next step, and at the same time the main goal of this work, is to integrate those features into a classification system to improve the sleep/wake classification performance. The direct comparison with other studies is very difficult up to impossible, because other modalities, recording settings and evaluation techniques have been employed. Therefore, the most promising features known from literature are also included in this work and will be used as baseline to evaluate the performance. Hence, the difference in classification results with the newly developed RQA features is very precisely distinguishable as we investigate the impact on classification performance with and without using those well-known features.

2 MATERIALS AND METHODS

2.1 DATA SET

This study comprises 313 full-night polysomnography data sets (PSG), including simultaneously recorded actigraphy (Actiwatch, Philips Respironics), that originate from three different databases. About 2467 hours of data (294697 epochs representing approx. 8 hours per night) are totally considered. Around 2014 hours were spent sleeping which results in an average sleep efficiency of 81.6%. The Pittsburgh Sleep Quality Index (PSQI) of all included subjects was less than 6 and no evidence nor earlier diagnosis of respiratory or sleep disorders was present. The trials were approved by the ethics committee of the local organizations and all subjects signed a consent form.

- ① Nine subjects monitored in the Sleep Health Center, Boston, USA (Alice 5 PSG, Philips Respironics). Average age of 31.0 (± 10.4) years.
- ② Six subjects monitored in the Philips Experience Lab, Eindhoven, the Netherlands (Vitaport 3 PSG, TEMEC). Average age of 29.5 (± 10.0) years.
- ③ Two-night recordings of 149 subjects (298 nightly recordings in total) acquired in the context of the EU SIESTA project between 1997 and 2000 in seven different sleep laboratories all following the same measurement protocol (see <http://www.ofai.at/siesta/> for further information)[22]. Average age of 52.0 (± 36.3) years.

For ① and ②, sleep staging was performed by a sleep technician according to the guidelines of the American Academy of Sleep Medicine (AASM) [1], distinguishing between wake, REM¹, N1, N2 and N3. For ③, sleep stages were scored by trained sleep technicians in six classes (wake, REM, S1, S2, S3, S4) according to the Rechtschaffen & Kales (R&K) guidelines [23]. Since only sleep/wake classification will be performed in this study, the stages REM, N1/N2/N3 and S1/S2/S3/S4 are merged into a common

¹ REM – Rapid Eye Movement

SLEEP class. Additionally, the whole data set was divided into almost equally large age and sex groups (M=male, F=female) for further analysis (see Figure 1). Groups 1 to 3 represent subjects younger than 38 years, between 38 and 61 years old and older than 61 years, respectively.

Figure 1: Age distribution of subjects by sex, M=male, F=female, groups 1-3

2.2 KNOWN FEATURES

In literature, many features have been found to work well for sleep/wake detection using cardio-respiratory and activity-based features. Features with major impact have been chosen and implemented in the computational and classification framework:

- 6x DFA² features (including WDFA) from ECG [3]–[5]
- 2x actigraphy features [6], [7]
- 70x HRV³ related features extracted from time and frequency domain [8]–[14]
- 21x entropy related features (respiration and ECG) [15], [16]
- 5x features based on cardio-respiratory coordination [13], [17]
- 22x features from respiratory effort [18]–[20]

In total, 126 employed features from literature were used to build a performance baseline (state of the art), since classification measures from other studies are hardly comparable with each other. In addition, the circumstances and study protocols are always slightly different, because no standard procedures are defined. Time-variant a priori probabilities are known to improve classification performance [6], [9]. The method consists in computing the probability of the sleep and wake states for each epoch of the night based on the training data. To achieve maximum classification performance, the procedure is also part of this work, but will not be explained further.

2.3 RECURRENCE QUANTIFICATION ANALYSIS

With the RQA method, it is possible to investigate complex behaviors of dynamic systems, e.g. in meteorology or physiology [24]–[28]. Recurring states in the phase space are describable with the use of RQA. In the last decades, computational power has

² (W)DFA – (windowed) detrended fluctuation analysis

³ HRV – heart rate variability

significantly increased so that large amount of data can be processed and computationally complex methods as RQA can be employed. A reconstruction of recurring states can be found with a time delay embedding technique proposed by Takens [29]. Then, Eckmann et al. presented a way to visualize recurrences in multi-dimensional phase space with the so called (Cross-)Recurrence Plots (C-RP) [30]. Out of those visualizations, several indicators have been derived (see [27], [31] and [32]) and are available in a Cross Recurrence Plot Toolbox for Matlab® [33]. In this work, we used 13 different indicators, listed in Table 1 with their physical interpretation. The mathematical computations are the same as in preliminary work and hence is not focus of this paper (see [21]).

Table 1: 13 different Recurrence Plot (RP) indicators with their physical interpretation

RP indicator		Interpretation
<i>REC</i>	Recurrence Rate	Correlation of the signal
<i>DET</i>	Determinism	Ratio of number of diagonal line points to total number of points. Larger ratio means that a more deterministic signal
<i>ENT</i>	Entropy	Entropy of RP (small value for periodic or uncorrelated signals)
<i>L</i>	Average length of diagonal lines	Average length of repetitive trajectories
<i>L_{max}</i>	Longest diagonal line	Measure for trajectory convergence
<i>LAM</i>	Laminarity	Stationarity of signals/system
<i>TT</i>	Trapping Time	Duration of signal/system laminarity
<i>T(1)</i>	Recurrence time of 1 st type	Average recurrence time of a Poincaré point (similar state)
<i>T(2)</i>	Recurrence time of 2 nd type	Like <i>T(1)</i> , but only the first recurrent point is considered
<i>V_{max}</i>	Longest vertical line	No physical interpretation possible
<i>T_{rec} / RPDE</i>	Recurrence period density entropy	Measure for aperiodicity of signals/system
<i>clust</i>	Cluster coefficient	Measure for interconnectedness of recurrence matrices
<i>trans</i>	Transitivity	Transitions in signals/system

Following our preliminary investigations on RQA features [21], five different RQA modalities were extracted out of the ECG, respiratory effort (RE) and the heart rate (HR) using the same parameter configurations for $m, \tau, \varepsilon, w_{size}$ and w_{step} and producing 65 basic RQA features:

1. R_{ECG} : RQA of the ECG
2. R_{RE} : RQA of thoracic RE
3. R_{HR} : RQA of HR, derived from ECG
4. $R_{ECG,RE}$: cross-RQA of ECG and RE
5. $R_{RE,HR}$: cross-RQA of RE and HR

2.4 PRE-PROCESSING

The same normalization procedures were employed as in our preliminary work [21]. This comprises newly developed RE signal normalization and the resampling of ECG data.

2.4.1 RESPIRATORY EFFORT

For RQA features, special pre-processing methods were developed to have optimal parameter definitions for the extraction (see [21]). Due to large motion artifacts or changes in the body position, the respiration amplitude is not constant over the whole night. RE, using a Respiratory Induction Plethysmography (RIP) belt positioned around the thorax was acquired with a sample rate of 10 Hz and filtered with an adaptive comb filter that normalizes the amplitude (see Figure 2).

Figure 2: Representation of adaptive comb filter for Respiratory Effort

The method was originally developed by Cyrill *et al.* for filtering ECG signals [34]. The present filter bank consists of $N = 188$ individual band pass filters with a bandwidth of 0.2 Hz and a center frequency spacing of 0.002 Hz resulting in a 99% overlap between two adjacent filters. The fundamental center frequencies are equally distributed within the regular breathing frequency range of 0.125-0.5 Hz. In addition, a second bandpass filter bank includes the first harmonic of the signal for less signal distortion. Knowing the best filter $F_{max}(k)$ means knowing the fundamental frequency $f_{fund}(k)$ of the signal. It is possible to select the correct filter of the bandpass filter bank (see $u_{bp,1}, u_{bp,2} \dots u_{bp,N}$ in Figure 2) and to extract the signal $u_{sel}(k)$ containing the fundamental frequency, only. In order to obtain the optimal bandpass filter, the fundamental frequency needs to be estimated. In theory, assuming no disturbances, the signal power is maximized when the filter frequencies match the desired signal frequencies, i.e. respiration. The “fundamental frequency estimation” and “filter selection” blocks in Figure 2 determine the best filter $F_{max}(k)$ by choosing the filtered signal with the maximum power $P_{max}(k)$ of the bandpass filtered signal $d_{BP}(k, x)$ at each discrete time step k :

$$P_{max}(k) = d_{BP}(k, F_{max}(k)) \quad (1)$$

$$F_{max}(k) = \underset{x=1 \dots N}{\operatorname{argmax}} \{d_{BP}(k, x)\} \quad (2)$$

$$u_{sel}(k) = u_{BP, F_{max}(k)}(k) \quad (3)$$

The switching between filter outputs causes (high frequency) discontinuities and offset errors which were removed by a following bandpass filter with 0.125 Hz and 1 Hz as left and right edge frequencies, respectively. Finally, the signal envelope for normalization was extracted applying Hilbert transform.

For the known literature features, a method described by Long *et al.* was applied instead of the described comb filter [35]. The signal was lowpass-filtered (10th order Butterworth filter with 0.6 Hz cut-off frequency) and then normalized by dividing with the average peak-to-peak amplitude. A large noise reduction is possible at the expense of good signal representation due to the low cut-off frequency.

2.4.2 ECG

The ECG data was re-sampled to 150 Hz to reduce the needed computational effort for RQA features. In addition, the baseline drift was removed and the data was high-pass filtered and normalized to cover a range of ± 1 .

2.4.3 HEART RATE

The instantaneous heart rate (HR) was determined by localizing R-peaks in the raw ECG. For detection we used an improved QRS detector based on the well-known Hamilton-Tompkins algorithm [36]. The resulting differences in time intervals were used to determine the instantaneous HR. To obtain a signal with equidistant samples, the HR was up-sampled by linear interpolation to a sample rate of 10 Hz thus having the same sampling rate as for the RE signal.

2.4.4 INTERPOLATION

Disturbances, e.g. caused by body movements, on ECG or respiratory signals led to erroneous feature data. First, segments with artifacts were detected and invalidated. Some features that do not rely on filtering windows may work with those invalid epochs, other need a continuous course of the raw signal. Hence, before performing feature extraction, invalid data had to be interpolated. In this work, a non-smoothing cubic spline interpolation method⁴ was used.

2.4.5 FEATURE NORMALIZATION

The resulting cardio-respiratory activity during sleep is very different from subject to subject. A subject-based normalization of features before classification generally improves classification performance, because it reduces subject-to-subject variance [37]. For the raw RQA features (see 2.3), four normalization procedures were separately applied to create additional sets of features resulting in 325 RQA features:

- ZSCORE, i.e. removed mean and scaled by the standard deviation

⁴ see Matlab® function `interp1(x, y, xi, 'spline')`

- Windowed standard deviation (window size of 30 s)
- High-pass filter by removing moving average (window size of 15 s)
- High-pass filter by removing moving average (window size of 30 s)

Furthermore, all features, including those known from literature, were normalized with ZSCORE and a Rosenblatt transformation was applied [37]. The latter transforms the arbitrary distribution of a feature into a normally distributed version.

2.5 CLASSIFICATION

The main five steps of the whole classification process are illustrated in Figure 3 and described in the following subsections.

1. ECG with derived HR, respiratory effort and actigraphy data of the polysomnography recordings
2. Feature extraction, after raw data pre-processing
3. Feature selection
4. Cross-validation (subject-by-subject)
5. Overall evaluation

Figure 3: Processing steps for classification

2.5.1 DATA ACQUISITION STEP

The raw ECG including the derived instantaneous heart rate signal, respiratory effort and actigraphy signals were used to generate the features after being pre-processed (see section 2.4). ECG was recorded using a modified Einthoven II electrode configuration and respiratory using a Respiratory Induction Plethysmography (RIP) sensor around the chest. Activity was captured with a wrist-watch (Actiwatch, Philips Respironics). All

signals were finally synchronized to a common time base. Manual annotations by trained sleep technicians were converted to the binary sleep/wake class (see section 2.1).

2.5.2 FEATURE EXTRACTION STEP

In total, a feature set comprising 1353 features was created, based on 126 basic features from literature (see 2.2) and 65 RQA features (see 2.3) with their additional normalized versions (see 2.4.5).

2.5.3 FEATURE SELECTION STEP

For the sleep/wake classification task, it is very likely that the available feature set is over-fitting the problem causing classification performance loss, e.g. due to feature noise. Therefore, not only the non-relevant features had to be removed from the set, but also the correlated ones. This is done with feature selection (step 3 in Figure 3), using an algorithm based on allowed minimal Mahalanobis distance and maximal inter-feature correlation (mMMC). First, all features are sorted by descending Mahalanobis distance. Those with distances lower than a defined threshold are excluded. For the remaining ones, the maximum Pearson correlation to the other features is determined and removed if the value is above a certain threshold and another feature with a higher Mahalanobis distance is already present in the selection. In this work, the Mahalanobis distance threshold was set to 0.15 and the correlation threshold to 0.9. Compared to exhaustive feature selection methods, e.g. Sequential Forward/Backward Search (SFS/SBS), the algorithm works very efficiently and is suited for classification with linear discriminant analysis (LDA) since both are employing the Mahalanobis distance as class separation metric. The mMMC algorithm was validated, even for small data sets, in a previous work [38].

2.5.4 CROSS-VALIDATION STEP

The classification (step 4a in Figure 3) was performed using LDA due to its mathematical simplicity and intuitive use without lacking classification performance. LDA assumes a set of normally distributed features, i.e. sampled at a 30 s epoch interval. The classifier specific data, in the case of LDA the covariance matrices and feature mean vectors, were generated with a specific training set (supervised learning) which should be excluded from the validation set to prevent biasing the results. LDA assumes homoscedasticity, i.e. equal class covariances. In this work, this was achieved by pooling the covariance matrices of the sleep and the wake class to a common covariance matrix:

$$\Sigma = \sum_{i=1}^C \frac{n_i - 1}{n - C} \cdot \Sigma_i, \quad (4)$$

with n_i as the number of samples in class i of C classes. n represents the total number of samples. A discriminant function of an unknown feature vector $\vec{x}$ for each class i was then determined:

$$g_i(\vec{x}) = -\frac{1}{2} \cdot (\vec{x} - \vec{\mu}_i)^T \cdot \Sigma^{-1} \cdot (\vec{x} - \vec{\mu}_i) + \ln(P(\omega_i)) = -\frac{1}{2} \cdot r_{M,i}^2 + \ln(P(\omega_i)), \quad (5)$$

with $\vec{\mu}_i$ as the mean vector determined in the training procedure. $P(\omega_i)$ represents the a-priori priority of the given class ω_i . $r_{M,i}$ is called the Mahalanobis distance, a metric which takes the covariance of the data into account. Note that for each sample of the feature vector, the classifier determines the largest value of the discriminant functions between each class. A sample was assigned to class ω_i otherwise to ω_j if

$$g_i(\vec{x}) - g_j(\vec{x}) > T \quad \forall i \neq j \in 1 \dots C, \quad (6)$$

for a defined threshold T .

For each subject, an individual classification was performed (step 4b in Figure 3), leading to a Leave One Subject Out Cross Validation (LOSOVC). Hence, data from all subjects except the subject itself are used during the training step. After this, all single results are evaluated separately or merged (step 4c and 5 in Figure 3) [18]–[20].

2.5.5 EVALUATION STEP

In general, comparing the estimated classes with the annotations performed by a trained physician yields the classification performance. The evaluation of all positively and negatively classified epochs returns the True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) measures which are used to compute amongst others accuracy (ACC), precision (PPV), sensitivity (TPR), False Positive Rate (FPR) and specificity (TNR) (see Table 2). For unequally distributed classes, as for example in the sleep/wake-classification task, the Cohen's kappa coefficient (κ) is a good metric to measure agreement. Simply said, it measures the normalized distance from random classification. Kappa ranges from 0-0.2 (slight agreement) to 0.81-1 (almost perfect agreement) [39].

Table 2: Confusion matrix with derived ratios and kappa value

test/ reality	true	false	Σ		
positive	TP	FP	N_{pos}	$ACC = \frac{TP + TN}{N_{total}}$	$TPR = \frac{TP}{N_{true}}$
negative	FN	TN	N_{neg}	$PPV = \frac{TP}{N_{pos}}$	$TNR = \frac{TN}{N_{false}}$
Σ	N_{true}	N_{false}	N_{total}	$FPR = \frac{FP}{N_{false}}$	$\kappa = \frac{ACC - p_c}{1 - p_c}$

All these measures require the definition of threshold T (see eq. (6)). Small inaccuracies could lead to large differences in the above mentioned measures and it is impossible to draw conclusions about the overall classification performances. For this, often the so called Receiver Operating Characteristic (ROC) or the Precision Recall (PR) curves are used. ROC curves map FPR against TPR by varying the threshold T so that FPR is between 0 and 1. Similarly, PR curves show FPR (also called recall) against PPV (synonym for precision). It is possible to superimpose different classifier results or even to optimize the threshold T for a specific application, e.g. for finding the desired

compromise between sensitivity and specificity or the optimal κ -coefficient. The optimal performance for ROC curves is located in the top left corner, for PR curves in the top right corner. Additionally, the Area Under the Curve (AUC) helps comparing classification performances with a single metric.

3 RESULTS AND DISCUSSION

At first, the obtained overall classification performance of three different feature sets, using literature known and RQA features, is presented with ROC and PR curves. Then, sex and age dependencies for the feature selection process and for classification results are examined. The sections have slightly different focuses, so that we discuss the results paragraph-wise, instead of summarizing the discussion in a separate section. Final conclusions are stated in section 4.

3.1 OVERALL CLASSIFICATION PERFORMANCE

In this section we present the classification performance, when pooling all individual subject results into a single result set using three different feature sets (see Table 3). The employed LDA based classification procedure is described in section 2.5 using pre-computed a-priori probabilities $P(\omega_i)$ as employed in previous work [6] [9].

Table 3: Feature sets definition

Feature set	Description	Feature count
A: Lit	Best literature features only	158 (42% of literature features)
B: RQA	Best RQA features only	232 (24% of RQA features)
C: Lit+RQA	Combination of literature and RQA features	390 (29% of all features)

The direct comparison of results in published literature is very difficult, since the measurement setup, the data acquisition, the feature sets, target epoch lengths, and other modalities are widely varied and prevent a fair comparison. The biggest obstacle for a valid comparison between different research work is the absence of publicly available large scale and representative data sets. To obtain a fair comparison in this work we therefore examine amongst others a feature set comprised of already well-known features (set A, Table 3). The kappa coefficient, beside the AUCs of the ROC and PR curves, is used to describe the overall performance and for comparison. The best features of each set were determined with the mMMC algorithm as described in section 2.5.3 (minimum Mahalanobis distance of 0.15 and maximum correlation of 0.9). The total feature count of the best literature features is 158 (set A) and 232 for the RQA features (set B). Set C was extracted using all features in the same way, which gave exactly the combination of both sets A and B meaning independency of both sets.

Generally, to have strictly independent training and testing data sets so that no biasing of the results occurs, the feature selection should only be applied on the training set of each LOSOCV step. Since the number of subjects is large enough and the impact of each subject on the selection is negligible, we only did the feature selection step (see step 3 in section 2.5) on the whole data set once. This considerably facilitates the comparison of

different feature sets. Note that in a real world scenario, this is not applicable, because the ground truth is generally not known and the feature selection therefore has to be performed on the training data set, only.

In Figure 4, the pooled ROC and PR curves are shown for set A, B and C. The coefficients κ_o and κ_+ represent the average classification performance values using a detection threshold $T = 0$ (see eq. (6)) and the maximum kappa value after having computed the optimal T , respectively. Here, κ_o and κ_+ only differ by 0.001-0.002, hence the default zero threshold gives a very good performance in terms of kappa. For simplicity κ_o will be used for further comparisons, e.g. for the computation of sleep statistics.

Figure 4: ROC and PR curves of feature sets A, B and C

With the literature-based features (set A) a kappa value of 0.558 with ROC/PR AUC values of 0.898/0.724 and a PPV of 0.65 was obtained. The RQA features alone (set B) only reached a kappa value of 0.522 with just as smaller ROC/PR AUC values of 0.866/0.656 and a PPV of 0.63 although the number of features is remarkably higher (see Table 3). This means that the RQA features on their own do not perform as well as the known literature features although more of them have been selected. Nevertheless, combining the features of set A and B shifts the kappa value to 0.625 with ROC/PR AUC values of 0.918/0.764 and a PPV of 0.70 (set C) which shows the benefit of additionally using the RQA features. Note that this also demonstrates that the literature features are independent of and complementary to the RQA features, hence adding information that is relevant for the sleep/wake classification task.

3.2 AGE AND SEX DIFFERENCES

It is well-known that both the sleep structure and the cardio-respiratory physiology alter with age and probably also with sex [40]. To analyze age and sex relationships, the

subject database was divided into six almost equally sized subsets (M1-M3 and F1-F3, see section 2.1).

3.2.1 FEATURE SELECTION

In this section, the age and sex dependency is examined on feature selection level. For all subjects, as aforesaid, 390 features were selected in set C. For the following analysis, the features selection is performed in the same way for each of the six almost equally sized age/sex groups (see Figure 1). The bar graphs in Figure 5 are divided into black and white sections, differentiating the amount of RQA features and those from literature.

Figure 5: Feature count of literature and RQA features in set C for each sex and age group; M1-M3=male, F1-F3=female, group {1, 2, 3}={<38 yrs, 38-61 yrs, >61 yrs} (see Figure 1)

The total number of selected features varies from 366 to 471. Generally, the number of relevant and non-correlating features decreases with age, but differences between male and female subjects are only marginal. The amount of literature features across the subsets shows less variation than the RQA features hence having less dependency on age or sex. Therefore, we conclude that the new RQA features are especially suited for the analysis of sleep of younger persons.

3.2.2 CLASSIFICATION PERFORMANCE

The kappa values for individual subjects in each subject subset were separately investigated after the overall classification with feature set C. The range of kappa values is depicted in Figure 6 with boxplots (quartiles as box around median and whiskers with 1.5 times the quartiles, outliers as circles). The kappa distribution skews are slightly negative (tending to higher values), with some outliers in the lower kappa region. The overall inter-quartile range is about 0.2, which means that 50% of the subjects are approximately within the median kappa value of 0.68 ± 0.1 . The inter-quartile ranges

within the groups are slightly smaller. Some subjects have very poor classification performance (0.1-0.2, especially in F3, M2 and M3) where other were nearly perfectly classified (>0.9 , especially in F1, F2 and M1). It is confirmed that the performance is altered with higher age and sex when using the overall selected features. From an autonomic point of view a smaller cardio-respiratory coupling is observed with increasing age, linked with lower respiratory sinus arrhythmia, often reflected in a lower power in the high-frequency band of the HRV spectrum. The reduced cardio-respiratory activity during sleep hence gives less information about sleep stages [40]. It might also happen, that the alignment of the annotation and the sensor data is too inaccurate, since the start of annotation is arbitrarily set and leading to a random delay of ± 30 s. In addition, studies have shown the inter-rater performance when visually annotating sleep only achieve kappa values of 0.68-0.82, even with the same annotators consecutively [41]. This is one of the main drawbacks of epoch-based classification which cannot be avoided. At the end, the classification is only as good as the annotated data and those worsening effects are used within the training of a numerical classifier. Another fundamental limitation of using cardio-respiratory and actigraphy measurements might be the generally missing possibility of differentiation between laying awake in bed in a relaxed way and a regular sleep stage since the autonomic response of the body may be very similar and only visible in the EEG signals.

Figure 6: Boxplot of kappa values for all age and sex groups of feature set C;
A=overall, M1-M3=male, F1-F3=female, group {1, 2, 3}={<38 yrs, 38-61 yrs, >61 yrs} (see Figure 1)

3.2.3 SLEEP STATISTICS

The last paragraph handles the sleep statistics between feature set A to C of each individual subject, which can be derived from sleep and wake stages. It is more important for this analysis to examine the error between the real annotation and the classifier results. Hence, the differences/deltas in Sleep Efficiency (SE), Sleep Onset (SO), Wake After Sleep Onset (WASO) and the Total Sleep Time (TST) are shown as boxplots in Figure 7. The average SE error is only 0.9% for set A and 0.2% for set C with a variation of a few percent.

In general, the age and sex group dependency is more visible in set A than in set C for the parameters SE, WASO and TST. In set C, the variation of these parameters is clearly less and also the average detection error is smaller. For the groups F1 and M1 the parameters SE and TST are more likely to be overrated, tending to be underrated with higher age in both sex groups. The inverse effect is visible for WASO.

For SO no prominent differences are visible between the two sets. The average error range is within a few minutes.

Figure 7: Sleep statistics of classification with feature sets A and C compared to annotations

4 SUMMARY AND CONCLUSIONS

In this work, we demonstrate the use of novel RQA features derived from cardio-respiratory modalities for the sleep/wake classification task in combination with known literature features in a classification system based on LDA. The classification was performed with a straight-forward Leave One Subject Out Cross-Validation (LOSOVCV) on recorded data of 313 individual nights from 164 different subjects. The dataset was split into six almost equally sized subject groups based on the subject age and sex to have unbiased results when comparing the different groups.

In total, more than 1300 different features were extracted. The selection of relevant features was done with a very efficient features selection method (mMMC) to give high classification performance. 390 (158 literature-based and 232 RQA features) relevant and non-correlating features were identified in the whole feature set. It has been shown that the new RQA features improve the overall classification performance in combination with the literature features, although solely used they cannot outperform the latter. These improvements include the better ROC/PR curve (larger Cohen's kappa and AUC values) as well as the derived sleep statistics that give hints on the sleep quality.

On feature level, the analyses show a dependency on age. The classification performances are clearly better for younger subjects. More relevant RQA features were selected for younger subjects, regardless of the sex. For literature features, this trend is not clearly visible, but generally fewer features were selected during the proposed feature selection method. This means – in addition to the better classification performance – that more information about sleep and wake stages is stored in less features giving more relevance to each feature.

In summary, the additional use of RQA features should be considered in future work. However, the general focus should be set on the improvement of the sleep/wake discrimination for elderly, because for younger subjects the human rating performance is already almost attained. With this work, the hypothesis that sleep/wake classification for elderly is much more difficult and less performing than for younger persons is supported.

5 ACKNOWLEDGEMENTS

This work was funded by Philips Research. The authors would like to thank all reviewers for their insightful comments.

6 REFERENCES

- [1] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan, *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. Westchester, IL: American Academy of Sleep Medicine (AASM), 2007.
- [2] D. J. Pitson and J. R. Stradling, "Autonomic markers of arousal during sleep in patients undergoing investigation for obstructive sleep apnoea, their relationship to EEG arousals, respiratory events and subjective sleepiness," *J. Sleep Res.*, vol. 7, no. 1, pp. 53–59, Mar. 1998.
- [3] J. Kantelhardt, Y. Ashkenazy, P. Ivanov, A. Bunde, S. Havlin, T. Penzel, J.-H. Peter, and H. Stanley, "Characterization of sleep stages by correlations in the magnitude and sign of heartbeat increments," *Phys. Rev. E*, vol. 65, no. 5, pp. 1–6, May 2002.
- [4] M. Staudacher, S. Telser, A. Amann, H. Hinterhuber, and M. Ritsch-Marte, "A new method for change-point detection developed for on-line analysis of the heart beat variability during sleep," *Phys. A Stat. Mech. its Appl.*, vol. 349, no. 3–4, pp. 582–596, Apr. 2005.
- [5] M. Adnane and Z. Jiang, "Automatic sleep-wake stages classifier based on ECG," in *ICCAS-SICE*, 2009, pp. 493–498.
- [6] S. Devot, R. Dratwa, and E. Naujokat, "Sleep/wake detection based on cardiorespiratory signals and actigraphy," in *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2010, pp. 5089–5092.
- [7] P. Fonseca, R. M. Aarts, X. Long, J. Rolink, and S. Leonhardt, "Estimating actigraphy from motion artifacts in ECG and respiratory effort signals," *Physiol. Meas.*, vol. 37, no. 1, pp. 67–82, 2016.
- [8] M. Malik, "Heart Rate Variability," *Ann. Noninvasive Electrocardiol.*, vol. 1, no. 2, pp. 151–181, Apr. 1996.
- [9] S. J. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 485–496, Mar. 2006.
- [10] 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, no. 4, pp. 245–256, Oct. 2007.
- [11] M. Basner, B. Griefahn, U. Müller, G. Plath, and A. Samel, "An ECG-based algorithm for the automatic identification of autonomic activations associated with cortical arousal," *Sleep*, vol. 30, no. 10, pp. 1349–1361, Oct. 2007.
- [12] M. Mendez and M. Matteucci, "Sleep staging from Heart Rate Variability: time-varying spectral features and Hidden Markov Models," *Int. J. Biomed. Eng. Technol.*,

vol. 3, pp. 246–263, 2010.

- [13] A. M. Bianchi, M. O. Mendez, and S. Cerutti, “Processing of signals recorded through smart devices: sleep-quality assessment,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 3, pp. 741–747, May 2010.
- [14] A. Noviyanto, S. Isa, and I. Wasito, “Selecting Features of Single Lead ECG Signal for Automatic Sleep Stages Classification using Correlation-based Feature Subset Selection,” *Int. J. Comput. Sci.*, vol. 8, no. 5, pp. 139–148, 2011.
- [15] N. Burioka, G. Cornélissen, F. Halberg, D. T. Kaplan, H. Suyama, T. Sako, and E. Shimizu, “Approximate Entropy of Human Respiratory Movement During Eye-Closed Waking and Different Sleep Stages,” *Chest*, vol. 123, no. 1, pp. 80–86, Jan. 2003.
- [16] M. Costa, A. Goldberger, and C.-K. Peng, “Multiscale entropy analysis of biological signals,” *Phys. Rev. E*, vol. 71, no. 2, pp. 1–18, Feb. 2005.
- [17] H. Bettermann, D. Cysarz, and P. Van Leeuwen, “Comparison of two different approaches in the detection of intermittent cardiorespiratory coordination during night sleep,” *BMC Physiol.*, vol. 2, pp. 1–18, Dec. 2002.
- [18] X. Long, J. Foussier, P. Fonseca, R. Haakma, and R. M. Aarts, “Respiration amplitude analysis for REM and NREM sleep classification,” in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2013, pp. 5017–5020.
- [19] X. Long, J. Foussier, P. Fonseca, R. Haakma, and R. M. Aarts, “Analyzing respiratory effort amplitude for automated sleep stage classification,” *Biomed. Signal Process. Control*, vol. 14, pp. 197–205, Nov. 2014.
- [20] X. Long, P. Fonseca, J. Foussier, R. Haakma, and R. M. Aarts, “Sleep and wake classification with actigraphy and respiratory effort using dynamic warping,” *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 4, pp. 1272–1284, 2014.
- [21] J. Rolink, M. Kutz, P. Fonseca, X. Long, B. Misgeld, and S. Leonhardt, “Recurrence quantification analysis across sleep stages,” *Biomed. Signal Process. Control*, vol. 20, pp. 107–116, 2015.
- [22] G. Klösch, B. Kemp, T. Penzel, A. Schlögl, P. Rappelsberger, E. Trenker, G. Gruber, J. Zeithofer, B. Saletu, W. Herrmann, S. L. Himanen, D. Kunz, M. J. Barbanoj, J. Röschke, A. Värri, and G. Dorffner, “The SIESTA project polygraphic and clinical database,” *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 51–57, 2001.
- [23] A. Kales and A. Rechtschaffen, *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Allan Rechtschaffen and Anthony Kales, editors, no. no 204. U. S. National Institute of Neurological Diseases and Blindness, Neurological Information Network, Bethesda, Md., 1968.
- [24] J. Zbilut, M. Koebbe, H. Loeb, and G. Mayer-Kress, “Use of recurrence plots in the analysis of heart beat intervals,” *Cardiology*, pp. 263–266, 1990.
- [25] P. I. Terrill, S. J. Wilson, S. Suresh, D. M. Cooper, and C. Dakin, “Application of recurrence quantification analysis to automatically estimate infant sleep states using a single channel of respiratory data,” *Med. Biol. Eng. Comput.*, vol. 50, no. 8, pp. 851–865, Aug. 2012.
- [26] H. Nguyen, B. Wilkins, Q. Cheng, and B. Benjamin, “An Online Sleep Apnea

- Detection Method Based on Recurrence Quantification Analysis," *IEEE J. Biomed. Heal. Informatics*, vol. 18, pp. 1285–1293, 2014.
- [27] N. Marwan, J. F. Donges, Y. Zou, R. V. Donner, and J. Kurths, "Complex network approach for recurrence analysis of time series," *Phys. Lett. A*, vol. 373, no. 46, pp. 4246–4254, Nov. 2009.
 - [28] Y. Peng and Z. Sun, "Characterization of QT and RR interval series during acute myocardial ischemia by means of recurrence quantification analysis," *Med. Biol. Eng. Comput.*, vol. 49, no. 1, pp. 25–31, Jan. 2011.
 - [29] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence, Warwick 1980*, Springer, 1981, pp. 366–381.
 - [30] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence Plots of Dynamical Systems," *Europhys. Lett.*, vol. 973, no. 4, pp. 973–977, 1987.
 - [31] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang, "Complex networks: Structure and dynamics," *Phys. Rep.*, vol. 424, no. 4, pp. 175–308, Feb. 2006.
 - [32] R. Donner, J. Heitzig, J. Donges, Y. Zou, N. Marwan, and J. Kurths, "The geometry of chaotic dynamics -- a complex network perspective," *Eur. Phys. J.*, vol. 84, no. 4, pp. 653–672, 2011.
 - [33] N. Marwan, "Cross Recurrence Plot Toolbox for Matlab, Version 5.15, Release 28.6." Transdisciplinary Concepts and Methods Potsdam Institute for Climate Impact Research (PIK), Potsdam (Germany), pp. 1–58, 2010.
 - [34] D. Cyrill, J. McNames, and M. Aboy, "Adaptive comb filters for quasiperiodic physiologic signals," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2003, vol. 3, pp. 2439–2442.
 - [35] X. Long, J. Yang, T. Weysen, R. Haakma, J. Foussier, P. Fonseca, and R. M. Aarts, "Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging," *Physiol. Meas.*, vol. 35, pp. 2529–2542, 2014.
 - [36] P. Fonseca, R. M. Aarts, J. Foussier, and X. Long, "A novel low-complexity post-processing algorithm for precise QRS localization," *Springerplus*, vol. 3, no. 376, pp. 1–13, 2014.
 - [37] S. J. van Albada and P. a Robinson, "Transformation of arbitrary distributions to the normal distribution with application to EEG test-retest reliability," *J. Neurosci. Methods*, vol. 161, no. 2, pp. 205–211, Apr. 2007.
 - [38] J. Foussier, P. Fonseca, X. Long, and S. Leonhardt, "Automatic Feature Selection for Sleep/Wake Classification with Small Data Sets," in *6th International Conference on Bioinformatics Models, Methods and Algorithms*, 2013, pp. 1–7.
 - [39] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
 - [40] X. Long, R. Haakma, R. M. Aarts, P. Fonseca, and J. Foussier, "Between-laboratory and Demographic Effects on Heart Rate and Its Variability During Sleep," in *Workshop on Smart Healthcare and Healing Environments in conjunction with the European Conference on Ambient Intelligence (Aml'14)*, 2014, pp. 1–4.
 - [41] M. Silber, S. Ancoli-Israel, M. H. Bonnet, S. Chokroverty, M. Grigg-Damberger, M. Hirshkowitz, S. Kapen, S. Keenan, M. Kryger, and T. Penzel, "The visual scoring of sleep in adults," *J. Clinical Sleep Med.*, vol. 3, no. 2, pp. 121–131, Mar. 2007.