Investigation of automated sleep staging from cardiorespiratory signals regarding clinical applicability and robustness

Miriam Goldammer [a,*], Sebastian Zaunseder [b], Moritz D. Brandt [c,d], Hagen Malberg [a], Felix Gräßer [a]

[a] Institute of Biomedical Engineering, TU Dresden, Dresden, Germany
[b] Department of Information Technology, FH Dortmund, Dortmund, Germany
[c] Department of Neurology, University Hospital Carl Gustav Carus, TU Dresden, Germany
[d] German Centre for Neurodegenerative Diseases (DZNE), Dresden, Germany

1. Introduction

Sleep stage classification based on polysomnography (PSG) data is an essential step in clinical sleep evaluation. PSG based information about sleep structure and sleep-related pathological events is mandatory for the diagnosis of several sleep disorders. But even though PSG is the gold standard for sleep evaluation, it is expensive and uncomfortable to acquire. Innovative approaches focusing on the evaluation of sleep structure based on frequently and simply collected signals will give us the opportunity to facilitate diagnostic and treatment of sleep disorders. Moreover, in most countries sleep laboratories are too rare to cover the existing and rapidly growing medical need in sleep medicine. These disadvantages and the general growth in mobile and unobtrusive technologies for biosignal acquisition have led to research into sleep staging from fewer signals (e.g. Electrocardiogram (ECG) only [1]) and even non-contact or unobtrusive sources [2] (e.g. radar [3] or 3D-cameras [4]). Using such signals, many works address sleep staging from heart rate, mainly by feature extraction from the ECG. Fonseca et al. use 132 HRV features [5], Li et al. use a combination of deep learning for spectrograms and ECG features [6], and Geng et al. compare three methods of feature extraction [7], yielding a kappa of 0.60, 0.54 and 0.65, respectively, for classifying into three to four sleep stages. However, there is no consensus concerning what aspects of the ECG identify specific sleep stages. Some recent approaches skip feature extraction and use a signal or time series as input for machine learning models directly. Thereby, feature extraction and classification are merged into one model, e.g. by using Convolutional Neural Networks (CNNs). Sun et al. [8] used a binary sequence from the ECG and the downsampled

* Corresponding author.
E-mail address: Miriam.Goldammer@tu-dresden.de (M. Goldammer).

respiratory effort, with a model consisting of convolutional and gated recurrent unit (GRU) layers. Their approach resulted in a Cohen's kappa of 0.65 for classifying into Wakefulness, NREM and REM (W/N/R). Korkalainen et al. [9] used the downsampled photoplethysmogram (PPG) as input for a model, again consisting of convolutional and GRU layers. Their model yielded a Cohen's kappa of 0.65 for classifying into W/N/R. Casal et al. [10] also classified the PPG into Wakefulness and Sleep with a GRU-based model, yielding a Cohen's kappa of 0.74. Sridhar et al. [11] classified sleep stages from the instantaneous heart rate derived from the ECG with a model consisting of different types of convolutional layers. They yielded a Cohen's kappa of 0.66 for distinguishing Wakefulness, light sleep (i.e. NREM 1 and NREM 2), deep sleep (i.e. NREM 3) and REM.

These works underline the potential of an end-to-end learning approach. However, towards real world usage, we see further requirements, such as a deeper characterization of factors that affect the model accuracy like training extent and data quality (particularly important with respect to unobtrusive but less reliable sensing techniques [2]), further improvements of classification accuracy, and investigation on the clinical suitability of such methods.

In this work, we propose a two-channel Convolutional Recurrent Neural Network (2cCRNN). The network was inspired by Malik et al. [1] and is based on our preliminary research [12] but extended in complexity. R-peak-to-R-peak intervals (RRIs) and breath-to-breath intervals (BBIs) derived from the ECG and respiratory effort serve as inputs. Our aims are threefold: we aim to optimize sleep staging, investigate model robustness of cardiorespiratory sleep assessment, and assess practical clinical applicability.

According to such aims, we first determine how good our model's sleep staging can become by different inputs and target labels. Second, we survey whether there are significant differences between patient groups by comparing the classification quality between subgroups by sex, age, apnea hypopnea index (AHI) and body mass index (BMI). Third, we examine the clinical applicability by calculating sleep metrics from the hypnograms and evaluating their overall reliability. Fourth, we study model robustness on variations and errors in the input data. And fifth, we investigate how the amount of training data affects the classification quality to specify data requirements in terms of robustness and transferability of our architecture.

2. Methods

2.1. Data

We used data from the first part of the Sleep Heart Health Study (SHHS1) [13,14]. The database contains one full-night PSG for each participant together with sleep stage annotations. From the PSG, we only used the ECG, the thoracic respiratory effort and the sleep stage annotations. RRIs were extracted from the raw ECG with the filter band algorithm proposed by Afonso et al. [15] in its implementation from [16]. RRIs were additionally filtered for implausible values according to [17], BBIs were extracted by algorithm *respdetect* implemented in [16]. Both time series were linearly interpolated, resampled at 4 Hz, and normalized to z-score for each signal and recording. This resulted in an interpolated RRI time series (iRRI) and an interpolated BBI times series (iBBI). The first and last five minutes from each signal were truncated, due to generally poor signal quality in this time windows. All gaps in the original time series (e.g. no R-peaks detected for several minutes) were linearly interpolated from the previous RRI (resp. BBI) to the next RRI (resp. BBI). From the 5804 participants in SHHS1, we were able to extract RRIs and BBIs for 5036 participants (the remaining 786 participants were sorted out automatically owing to conspicuous data).

SHHS1 contains sleep stage annotations according to Rechtschaffen and Kales. To convert those annotations to AASM sleep stages, we combined stages S3 and S4 into one stage analogue to NREM 3 and replaced Movement by whatever sleep stage followed in the next epoch. Except for these AASM alike combinations of the sleep stage labels, we used some more combinations during our experiments: (a) we combined NREM 1 and NREM 2 to light sleep (L), in contrast to NREM 3 as deep sleep (D) and (b) we combined all NREM stages into one group (N). We will further refer to these ground truth groupings as AASM, W/L/D/R and W/N/R. This approach of combining sleep stages is similarly used in many related publications, e.g., [8–11].

For further analysis, we grouped the participants by the criteria sex, age, BMI and AHI. The number of participants in each subgroup for both training and hold-out test data are listed with the results. We configured the bins for the subgroups to avoid bias by underrepresentation in the training data and to be comparable to the closest related literature [8,11].

The data underlying this article was accessed from the National Sleep Research Resource, Sleep Heart Health Study Part One, https://sleepdata.org/datasets/shhs. Details on derived data and models will be shared on request.

2.2. Validation strategy

PSGs from 998 participants served as hold-out test data. They were selected to match the AHI distribution of the whole dataset but were selected randomly otherwise. This resulted in 4038 participants for training and validation.

We optimized our architecture with an extensive grid search and three-fold cross validation, using the ground truth labels W/L/D/R. For the final evaluation on the hold-out test data, we performed a ten-fold cross validation for all different ground truth groupings and inputs. We then classified the hold-out test data by predicting with each of the ten cross validation models independently, and afterwards using their mode classification for each epoch. In other words, we combined the ten models from the cross validation to a simple ensemble classifier with majority vote.

When comparing subgroups of patients, we tested all results for significance with Student's *t*-test (resp. Welch's *t*-test, if necessary). We only consider differences with $p < 0.05$ to be significant.

2.3. Model architecture and configuration

As stated previously, we enhanced our CNN architecture [12] as following. Firstly, we added the iBBI to the input, yielding two input channels with input sequences of 1200 samples (300 s) for each epoch. The epoch to be classified is in the middle of that 300 s window. Secondly, we appended a bidirectional LSTM layer with 40 units to the CNN architecture. The model now takes data from 240 epochs as input and accordingly generates a series of (240, n_s) labels as output, with n_s being the number of labels in the sleep stage grouping, i.e. (240, 4) for W/L/D/R as in Table 1. See the model summary with additional comments in Table 1 for more detailed information. We implemented the model with TensorFlow [18] and Keras [19], and selected Adam [20] optimizer with learning rate 0.001 and categorical cross-entropy loss function for training. Additionally, we applied early stopping by validation loss with patience of ten epochs.

2.4. Model evaluation

We evaluated the performance of our model by Cohen's kappa (κ) [21] and confusion matrices. Cohen's kappa is a metric to measure interrater agreement, which takes random agreement i.a. due to class distribution into account. It is therefore less distorted by non-uniform class distributions than e.g. accuracy. Since sleep stages are not uniformly distributed, κ became a popular metric for evaluating sleep staging performance and is probably the most commonly reported metric except for accuracy in sleep staging publications, e.g. [1,5–11]. When describing κ, we use the nomenclature by Landis et al.[22], which considers κ values greater 0.6 substantial agreement, and values greater

Table 1
Model summary. The model has 1,806,356 trainable parameters. Model input: iRRI and iBBI, model output: W/L/D/R. Sleep stage groupings: W: Wakefulness, R: REM, L: Light Sleep (NREM 1 + NREM 2), D: Deep Sleep (NREM 3), iRRI: interpolated time series of intervals between subsequent R-peaks in the electrocardiogram, iBBI: interpolated time series of intervals between subsequent breathes in the respiratory effort.

Layer type	Output shape	Number of parameters	Comment
Input Layer	[(None, 240, 1200, 2)]	0	
Conv1D	(None, 240, 1185, 64)	2,112	64 filters, kernel size 16, stride 1, activation ReLu
Conv1D	(None, 240, 585, 64)	65,600	64 filters, kernel size 16, stride 2, activation ReLu
Conv1D	(None, 240, 570, 64)	65,600	64 filters, kernel size 16, stride 1, activation ReLu
Conv1D	(None, 240, 278, 64)	65,600	64 filters, kernel size 16, stride 2, activation ReLu
Conv1D	(None, 240, 263, 64)	65,600	64 filters, kernel size 16, stride 1, activation ReLu
Conv1D	(None, 240, 124, 64)	65,600	64 filters, kernel size 16, stride 2, activation ReLu
Conv1D	(None, 240, 109, 64)	65,600	64 filters, kernel size 16, stride 1, activation ReLu
Conv1D	(None, 240, 47, 64)	65,600	64 filters, kernel size 16, stride 2, activation ReLu
Flatten	(None, 240, 3008)	0	
Dropout	(None, 240, 3008)	0	Dropout rate 0.3
Dense	(None, 240, 400)	1,203,600	Activation ReLu
Dropout	(None, 240, 400)	0	Dropout rate 0.3
Bidirectional LSTM	(None, 240, 80)	141,120	40 units
Dense	(None, 240, 4)	324	Activation Softmax

0.8 almost perfect agreement. To look into potential biases, we also calculated the mean κ for some subgroups, according to age, sex, BMI and AHI.

To evaluate the predicted hypnograms concerning further processing, we calculated sleep metrics according to the Sleep Scoring Data in the AASM manual (Section 2, Table B: Sleep Scoring Data) [23]. The metrics we use are total sleep time, sleep latency, REM latency, wake after sleep onset (WASO), sleep efficiency and percentage of time in each stage. The values calculated from the predicted hypnograms are interpreted as regression outputs and therefore evaluated by Spearman's rho (ρ_s) correlation coefficient against the values computed from the annotations.

2.5. Variations on the RRI data

Since RRIs are harder to acquire with high quality compared to respiration, we assume iRRI will be affected more by changes in the setting than iBBI. Therefore, we used a version of our model that only takes iRRI as input, and investigated the robustness towards changes in the data like smoothing, gaps, different filters, and missing values.

Thus, to investigate robustness and transferability of our approach to different data sources, we applied the following modifications to our RRI input data:

- No filtering of RRIs after detection (i.e. different preprocessing)
- Only using 90%, 80% and 70% of RRIs; in other words, randomly dropping 10–30% of RRIs (i.e. different R-peak detector or noisy signal)
- Adding up to 5% of noise to each RRI; to be precise, shifting each RRI by a random offset of up to +/- 5% of the RRI's magnitude (i.e. different R-peak detector)
- Using a 5-, 20-, and 40-sample mean for each sample of the interpolated 4 Hz input, resulting in a 1 s-, 5 s and 10 s-mean iRRI (i.e. using a device that only gives a mean HR calculated from the last few seconds)
- Random gaps in the iRRI of 5–10 s that make up 30% of the input (i.e. noisy signal or device with transmission problems)

Fig. 1 shows segments of 1200 samples (300 s) of the same data with different modifications.

2.6. Variation of train set size

To investigate whether there is a saturation in classification quality dependent on the train set size, we retrained our model with randomly sampled, growing subsets from the complete training data. To reduce computing time, we mostly used a three-fold cross validation and only increased to five and ten folds to yield larger train sets.

3. Results

3.1. General classification quality

Our model shows high accuracy (κ 0.80, accuracy 88%) for classifying into three sleep stage groupings (W/N/R) and substantial performance (κ 0.66, accuracy 76%) for classifying into five stages corresponding to AASM staging. The complete results in terms of κ for different inputs and stage groupings are summarized in Table 2 and detailed results of two models are displayed in Table 3. A hypnogram of an average PSG with AASM sleep stages is shown in Fig. 2. Note that even some of the short changes are detected correctly, e.g. single epochs of wakefulness during REM and frequent changes between wakefulness and light sleep. However, especially rapidly repeating stage changes between light sleep and deep sleep are often predicted as stable phases of either light sleep or deep sleep.

Concerning classification performance by mean κ for participant subgroups, we only detected a sex difference in classification quality when classifying into W/L/D/R. We generally saw a stable performance according to mean κ for participants with any BMI and all sleep stage groupings. Only for participants with BMI between 30 and 35, mean κ is significantly larger compared to other BMI intervals when classifying into W/L/D/R and AASM. Comparing age groups, the results show that classification performance according to mean κ significantly decreased for participants older than 60 years. Concerning AHI, we find that our model generally performs well for all participants with AHI < 30, with peak performance by κ for the subgroup with AHI 5–15 and a decrease in accuracy for AHI > 30. For a detailed comparison for all these subgroups including statistical hypothesis test results, see Table 4. Fig. 3 displays boxplots of the results in the subgroups when classifying into AASM stages, with significant differences highlighted in the plot.

3.2. Clinical sleep metrics

Calculating typical sleep metrics from our predicted hypnograms and comparing them to the annotations, we found that total sleep time, sleep latency, REM latency and sleep efficiency are predicted with a very strong confidence ($\rho_s \geq 0.8$, $p < 0.001$). Furthermore, WASO, percentage of NREM and percentage of REM are predicted with moderate confidence ($0.8 > \rho_s \geq 0.6$, $p < 0.001$). However, percentage of light sleep and deep sleep are predicted with only fair confidence ($0.6 > \rho_s \geq 0.3$, $p < 0.001$). See Fig. 4 for graphical illustration by scatter plots and detailed results of Spearman's rho and its significance.

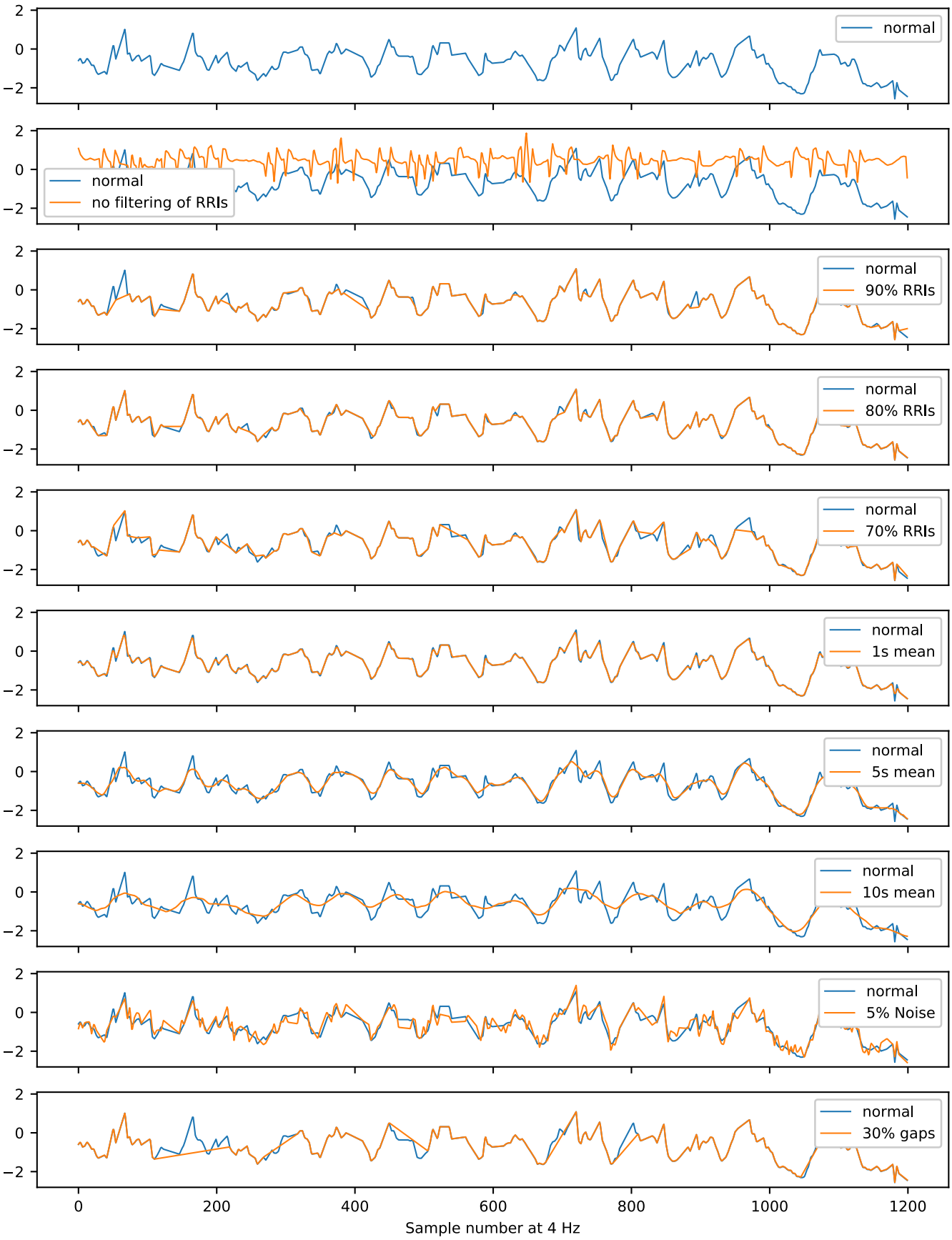

Fig. 1. Plot of each modification to iRRI time series on the same 300 s of iRRI. iRRI: interpolated time series of intervals between subsequent R-peaks in the electrocardiogram.

Table 2
Mean Cohen's kappa results on hold-out test data comparing results from the literature with our model for different groupings of sleep stages and different inputs. ECG: electrocardiogram, PPG: photoplethysmogram, RE: respiratory effort, sleep stage groupings: W: Wakefulness, N: all NREM, R: REM, L: Light Sleep (NREM 1 + NREM 2), D: Deep Sleep (NREM 3), S: all Sleep (NREM 1–3 + REM), AASM: W/N1/N2/N3/R.

	This work			Sun et al. [8]		Korkalainen et al. [9]		Casal et al.[10]	Sridhar et al. [11]
Input from	W/N/R	W/L/D/R	AASM	W/N/R	AASM	W/N/R	AASM	W/S	W/L/D/R
ECG/PPG	0.76	0.65	0.63	0.65	0.49	0.65	0.51	0.74	0.66
RE	0.68	0.57	0.55		0.69	0.53			
ECG & RE	0.80	0.68	0.66		0.76	0.59			

Table 3
Detailed results for classification into AASM sleep stages (left) and W/N/R sleep stage grouping (right) on hold-out test data which consists of 955,346 epochs. The table states the absolute number of epochs for each combination of true and predicted stages, the corresponding sensitivity and precision for each sleep stage and the overall accuracy. Model input: iRRI and iBBI, model output: AASM (left) or W/N/R (right). Sleep stage groupings: W: Wakefulness, N: all NREM, R: REM, AASM: W/N1/N2/N3/R, iRRI: interpolated time series of intervals between subsequent R-peaks in the electrocardiogram, iBBI: interpolated time series of intervals between subsequent breathes in the respiratory effort.

Predicted Sleep Stage	True Sleep Stage					Precision
	W	R	N1	N2	N3	
W	201,832	3,840	10,182	16,863	1,138	86.31%
R	5,726	118,417	3,783	15,704	420	82.21%
N1	1,230	726	2,056	1,151	7	39.77%
N2	28,545	16,749	20,484	348,406	62,006	73.17%
N3	1,085	246	61	30,853	63,836	66.44%
Sensitivity	84.65%	84.60%	5.62%	84.36%	50.10%	
Accuracy						76.89%

Predicted Sleep Stage	True Sleep Stage			Precision
	W	R	N	
W	194,122	3,371	20,480	89.06%
R	5,234	116,124	17,515	83.62%
N	39,062	20,483	538,955	90.05%
Sensitivity	81.42%	82.96%	93.41%	
Accuracy				88.89%

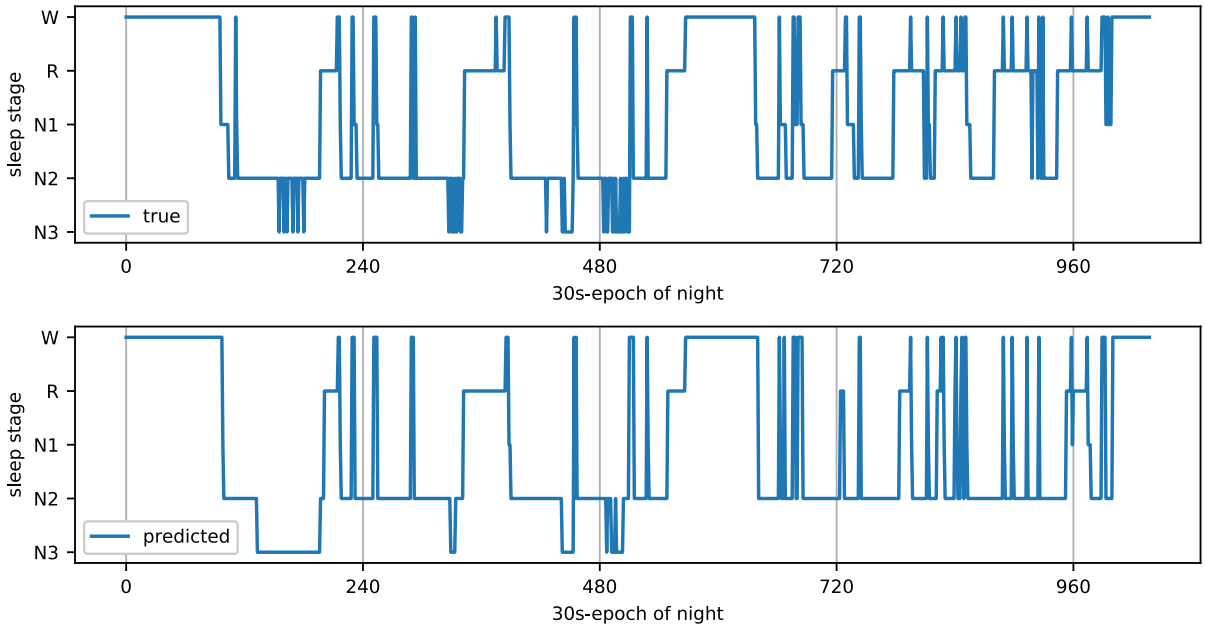

Fig. 2. Exemplary hypnogram from true and predicted sleep stages. For this participant and ground truth grouping, our model's performance by Cohen's kappa is 0.63 and the accuracy is 75%. Model input: iRRI and iBBI, model output: AASM. iRRI: interpolated time series of intervals between subsequent R-peaks in the electrocardiogram, iBBI: interpolated time series of intervals between subsequent breathes in the respiratory effort.

3.3. Robustness

As displayed in Table 5, we found that our trained model is robust to 70% missing RRIs or 1s mean. It is less robust for no filtering, inexact R detections, 5 s and 10 s mean, and 30% gaps (see first line of Table 5). However, training on data that was manipulated in the same way resulted in a performance comparable to that of the model trained and tested on the original data, according to the mean κ results (see diagonal of Table 5).

When training with increasing fold sizes with random training patients, we yield mean κ for classifying W/N/R on the test data as listed in Table 6. Note that the classification performance converges around 2500 PSGs, but still, we yield better results with more data.

4. Discussion

Results from most recent and most comparable state-of-the-art publications to complement our results are displayed in Table 2. By using a large and diverse dataset, optimizing the model architecture in several steps, and combining two promising input signals, we merged

Table 4
Mean Cohen's kappa results on hold-out test data by different ground truth groupings of sleep stages and different patient subgroups for both our model and one from the literature. Note, that the number of participants is only from our data. Mean values of Cohen's kappa that differed significantly from one subgroups to all others were marked with * (significant difference by *t*-test with $p < 0.05$). Model input: iRRI and iBBI, model output: see column heading. BMI: body mass index, AHI: apnoea hypopnoea index, Sleep stage groupings: W: Wakefulness, N: all NREM, R: REM, L: Light Sleep (NREM 1 + NREM 2), D: Deep Sleep (NREM 3), AASM: W/N1/N2/N3/R, iRRI: interpolated time series of intervals between subsequent R-peaks in the electrocardiogram, iBBI: interpolated time series of intervals between subsequent breathes in the respiratory effort.

		Our model					Sun et al. [8]	
		Number of participants in set		Sleep stage grouping			Sleep stage grouping	
		Train	Test	W/N/R	W/L/D/R	AASM	W/N/R	AASM
Sex	male	1,887	471	0.78	*0.68	0.66	0.76	0.59
	female	2,151	527	0.79	*0.66	0.65	0.77	0.58
Age (in years)	39–60	1,681	408	*0.80	*0.69	*0.67	0.77	0.59
	60–90	2,357	590	*0.77	*0.66	*0.64	0.74	0.55
BMI (in kg/cm^2)	18–25	1,001	241	0.78	0.66	0.64	0.76	0.59
	25–30	794	207	0.78	0.67	0.65	0.76	0.59
	30–35	1,154	304	0.80	*0.69	*0.67	0.77	0.59
	35–50	1,089	246	0.77	0.66	0.64	0.76	0.58
AHI	0–5	1,052	293	0.79	0.67	0.66	0.77	0.59
	5–15	1,726	395	0.80	0.68	0.66	0.76	0.58
	15–30	864	205	*0.77	0.67	0.65	0.75	0.58
	30–108	367	99	*0.72	0.64	*0.61	0.75	0.56

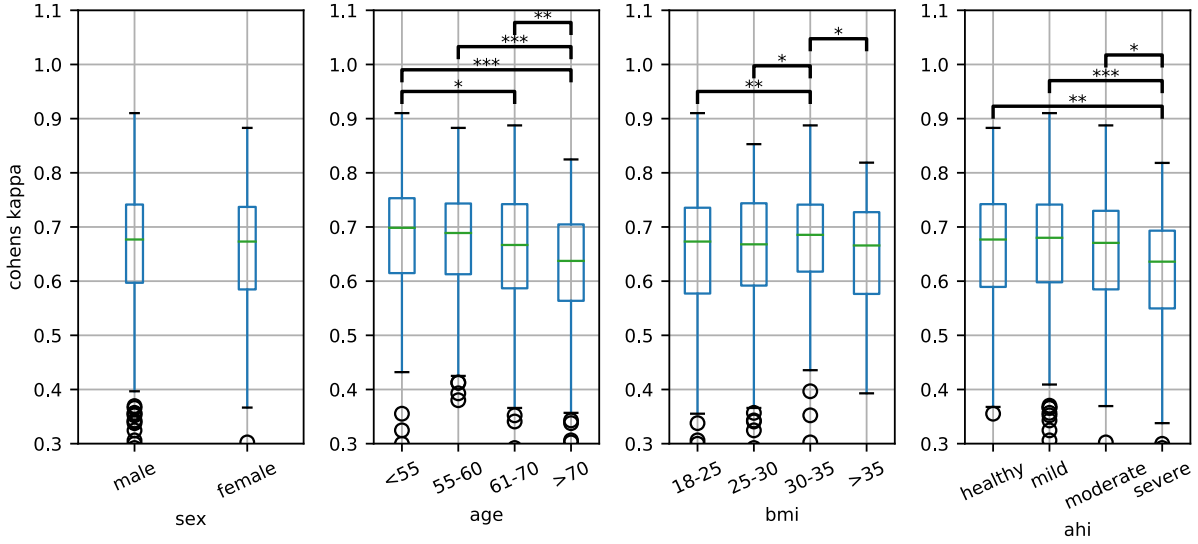

Fig. 3. Comparison of mean Cohen's kappa results for different participant subgroups from the hold-out test data. The green line shows the median, the box frames the upper and lower quartile, the whiskers extend up to one and a half times the interquartile range, any values outside this range are displayed as outliers (circles). All subgroups were tested for significant differences by *t*-test and the corresponding p-values are coded as: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Model input: iRRI and iBBI, model output: AASM. AHI: healthy: 0–5, mild: 5–15, moderate: 15–30, severe: >30, iRRI: interpolated time series of intervals between subsequent R-peaks in the electrocardiogram, iBBI: interpolated time series of intervals between subsequent breathes in the respiratory effort.

the advantages of several other approaches into one model. Thus, we optimized accuracy and κ compared to previous studies for diverse groupings of sleep stages. The results by κ, confusion matrices and hypnograms draw a coherent picture that our model is capable of distinguishing Wakefulness, NREM and REM very well by utilizing only information derived from tachogram and breathing. However, it is not able to reliably differentiate between light sleep and deep sleep or identify NREM 1.

There are only few approaches at time series based sleep stage classification that all show some distinct similarities, e.g. using CNNs and RNNs, and preprocessing the ECG to an RRI time series. In comparison, our model represents are very simple and straightforward architecture with a mere combinations of CNN and LSTM layers with fewer recurrent units and less filter kernels than others [8,9], even though it is more complex than just two layers of GRUs as in [10]. Also, we processed the two input signals parallel in one branch, rather than creating two parallel branches as in [8]. However, since the models process different input signals at different levels of preprocessing, it is not possible to compare model complexity with the available information. Furthermore, we deliberately chose to process rather long segments of 120 min (compared to 14 min [8] and 50 min [9]) but not the whole night (as in [10,11]) at a time. We assume that by this compromise, we supplied the model with the contextual information of at least a whole sleep cycle, but avoided the possible bias of expecting an average hypnogram with e.g. wake epochs in the beginning and ending, and cyclic structure with more deep sleep in the first cycles and more REM sleep in the last cycles.

One probable reason for the low accuracy in distinguishing NREM stages is the underrepresentation of NREM 1 samples in the available data (only 3.8 % of training epochs). However, there are approximately as many samples for REM as for NREM 3 (approx. 14 % of training epochs), so with respect to distinguishing between L and D (resp. NREM 2 and NREM 3), obviously the used features themselves have limitations as seen in other models before [8,9,11]. A closer look into the literature

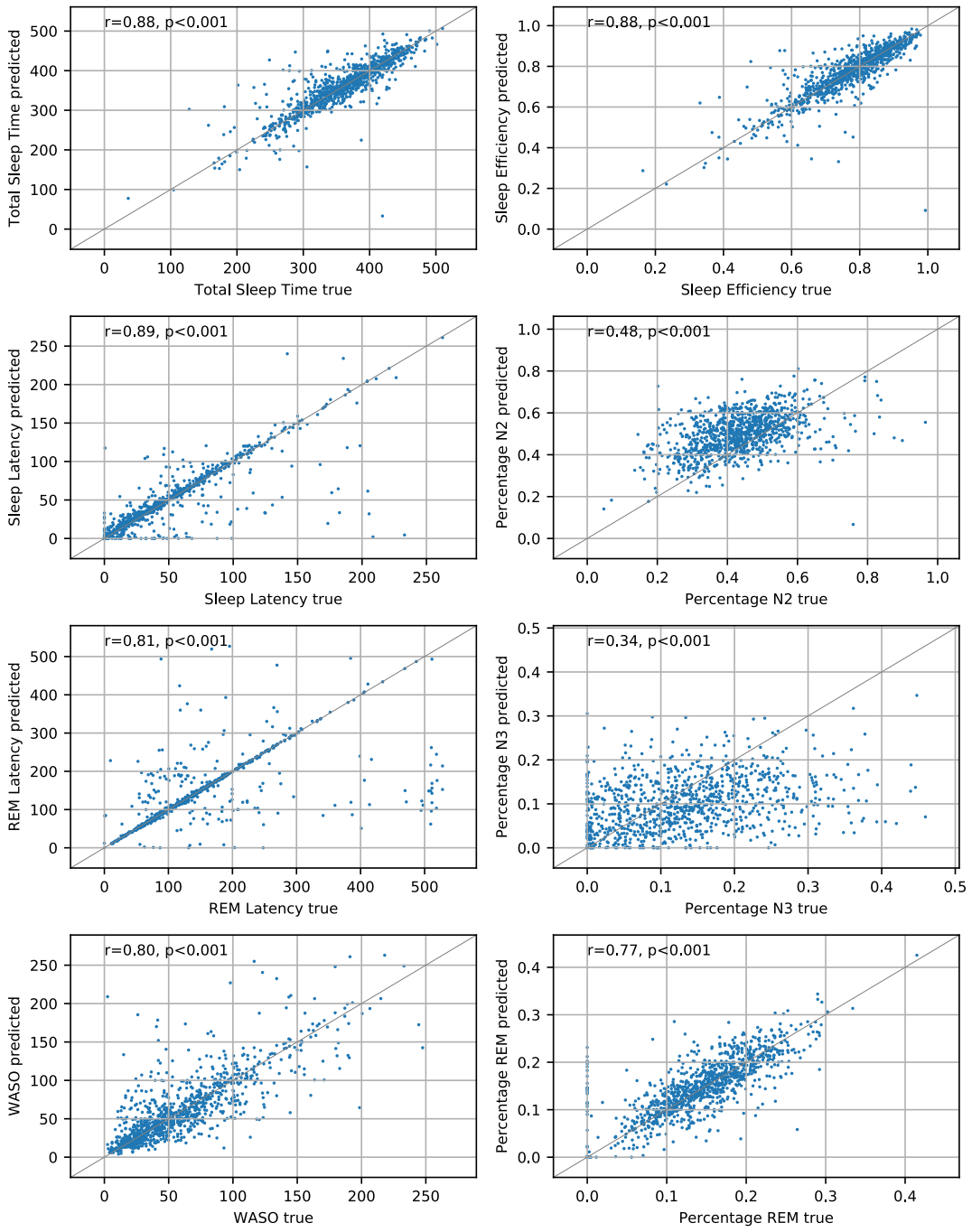

Fig. 4. Scatterplots for sleep metrics of hold-out test data calculated from true and predicted hypnograms. Spearman's rho correlation coefficient is printed into each graph as r. Model input: iRRI and iBBI, model output: AASM. WASO: wake after sleep onset, sleep stage groupings: W: Wakefulness, R: REM, L: Light Sleep (NREM 1 + NREM 2), D: Deep Sleep (NREM 3), iRRI: interpolated time series of intervals between subsequent R-peaks in the electrocardiogram, iBBI: interpolated time series of intervals between subsequent breathes in the respiratory effort.

reveals that models with better performance at detecting NREM 3 use downsampled raw signals as input: [8] combines a downsampled respiration signal with a QRS detection based time series yielding NREM 3 sensitivity 0.58 and precision 0.71, and [9] uses merely a downsampled PPG yielding NREM 3 sensitivity 0.54 and precision 0.75. In contrast, approaches that apply QRS detection and breath cycle detection first, yield e.g. NREM 3 sensitivity between 0.12 and 0.52 and precision between 0.58 and 0.68 [5,6,11,24], similar to our results of sensitivity 0.50 and precision 0.66. Therefore, there seems to be some general information loss concerning the difference between light sleep and deep sleep, when detecting QRS complexes and breath cycles first. But as these approaches, that apply QRS detection as our model, show a very good overall performance for sleep staging, there seems to be some gain of information, too. Concluding, a combination of downsampled raw signals and feature detection based time series might fuse the advantages of both approaches and should be investigated further. Still, assuming the transition from light to deep sleep is a more continuous shift of signal characteristics rather than a series of abrupt changes, then these changes in the labels are partly due to thresholds in the scoring rules (e.g. amount of delta activity). Therefore, if intermediate epochs between light and deep sleep are predicted as stable phases of either light or deep sleep, this might reflect the cardiorespiratory state of the

Table 5
Mean Cohen's kappa results on hold-out test data for our model trained on modifications of the data (lines) and tested on those modifications (columns). Model input: iRRI (with diverse modifications), model output: W/L/D/R. RRI: interval between two subsequent R-peaks in the electrocardiogram, iRRI: RRI interpolated at 4 Hz.

		Test on									
	Modification	None	no filter	90% RRIs	80% RRIs	70% RRIs	5% Noise	1 s mean	5 s mean	10 s mean	30% gaps
Train on	None	0.64	0.48	0.63	0.63	0.62	0.57	0.64	0.54	0.40	0.52
	no filter	0.61	0.63								
	90% RRIs	0.63		0.62							
	80% RRIs	0.62			0.62						
	70% RRIs	0.62				0.62					
	5% Noise	0.55					0.60				
	1 s mean	0.63						0.63			
	5 s mean	0.53							0.62		
	10 s mean	0.50								0.60	
	30% gaps	0.61									0.59

Table 6
Mean Cohen's kappa results on hold-out test data for classifying W/N/R from different inputs. Results are listed by number of patients used for training in each fold. Results are from 3fcv (*5fcv, **10fcv). Model input: see column headings, model output: W/N/R. iRRI: interval between two subsequent R-peaks in the electrocardiogram interpolated to 4 Hz, iBBI: interval between two subsequent breaths in the respiratory effort signal interpolated to 4 Hz.

Number of participants in each training fold	iRRI	iBBI	iRRI & iBBI
234	0.62	0.57	0.69
476	0.66	0.62	0.72
915	0.70	0.63	0.76
1346	0.72	0.65	0.77
1802	0.73	0.66	0.78
2297	0.74	0.67	0.78
2692	0.74	0.67	0.79
3230	*0.75	*0.68	*0.79
3634	**0.76	**0.68	**0.80

subject. Then again, while we can see from our results that we can identify changes in cardiorespiratory activity between W, NREM and REM, the physiological differences between light and deep sleep might be less distinct.

Concerning differences in classification quality among participant subgroups, our model generally shows the same slight influences of sex, age, BMI and AHI as stated in the literature. In contrast to the work of Sun et al. [8], that shows a constant decrease in performance by increasing AHI, our model shows best performance for AHI 5–15 (see Table 4 and Fig. 3). However, this is most likely due to imbalanced data used for training, as in our dataset participants with AHI 5–15 represent 43% of the data while in Sun et al.'s data, 39% of participants had AHI 0–5 [8]. Therefore, the models each performed best on data of participants that were prevalent in the respective training data.

As confirmed by ρ_s in Fig. 4, our model predicts many sleep metrics with very strong confidence. However, it is noticeable that WASO and the percentage of REM during the night show more general deviation from the true values and only moderate correlation, even though κ and the confusion matrix suggest that they are distinguished very well. Furthermore, confusion matrices that show the misclassification between L and D as a sum of all patients' epochs (Table 3) or overall statistics might be positively interpreted as just over- and underestimating L and D equally for each patient. An effect that might compensate in the end, or an overall bias that can be taken into account. But when calculating sleep metrics from the predictions, the scatter plots (Fig. 4) show clearly that our model predicts percentages of L and D very differently for each patient. There is no general bias of overestimating L and underestimating D. All this underlines the need to further evaluate classifiers by their application (here, i.e. clinical sleep scoring data) and not just generic metrics.

Investigating sleep metrics and participants' characteristics, we noticed that both, model performance and age, are correlated with time in REM. Our model's performance decreases with decreasing time in REM and specifically shorter durations of REM episodes. In our data, time in REM and duration of REM episodes also decrease with age. Moreover, this should be taken into account when analysing pathological PSG data, e.g. sleep related breathing disorders which are typically accompanied by a selective loss of REM-sleep. Therefore, the hypnograms of older participants and patients with severe sleep disorders strike a weakness of our model, which needs to be addressed in further optimizations steps.

From Table 5 we conclude that our architecture is applicable to a variety of signals, devices and preprocessing to classify sleep stages from heart rate, if there is enough training data available. Nevertheless, a model trained and tested on the complete filtered iRRI performs best. But even though our variations were inspired by real world scenarios, like using a different device for signal acquisition or using previously preprocessed data, they are still just basic simulations of these scenarios. Further experiments on real data will show whether our assumption regarding robustness and transferability is correct. Nevertheless, as SHHS1 was specifically designed to research risk factors for cardiovascular disease [13], there is a low prevalence for healthy heart patients and we infer general robustness towards most common cardiovascular diseases. This leaves the prospect that there should be a wide range of future applications for this kind of model, and that further research in this direction is necessary.

As we see some saturation in the performance with 2500 patients and more (Table 6), this is our suggestion for the minimally necessary amount of training data (validation and hold-out test data not included). This number gives an orientation regarding the necessary data for actually transferring the architecture. Nevertheless, further exploration into transfer learning from pretrained models might show that even much smaller numbers will suffice, as the general patterns learned from the data should be similar.

Concerning imminent clinical application, we see the main use cases of our model in pre-screening and home monitoring. Using established clinical tools like polygraphy or long-term ECG, our model could enhance the standard clinical evaluation by adding hypnogram and sleep metrics.

5. Conclusion

Main limitations of our work are, firstly, that the model was not tested on external test data (i.e. on a different dataset than SHHS1). This will be one next step in our research. And secondly, we cannot generally assume that other models show the same confidence for calculating sleep metrics, as these metrics are rarely reported for models. Therefore, the benefits and limitations we found are only first observations and need to be considered, assessed and compared by other researchers.

Concluding, we present a model that was thoroughly tested on PSGs from 998 hold-out test patients. It shows high classification quality for differentiating W/N/R and only slight common biases by sex, age, AHI

or BMI. The sleep stages predicted by this model can be reliably summarized in many sleep metrics like total sleep time, sleep efficiency and sleep latency but not as reliably in some others like percentage of light sleep and deep sleep. We showed that the architecture is robust to different scenarios that include variations and errors in the input data. To round off, we explored the dependence on the amount of data that is necessary to successfully transfer our architecture and train the model from scratch. Prospectively, our model allows for confident classification of sleep macrostructure based on ECG signal and respiratory effort data that can be collected in a much broader clinical context than PSG-data and may significantly improve medical care of patients with sleep disorders.

CRediT authorship contribution statement

Miriam Goldammer: Methodology, Software, Validation, Formal analysis, Writing – original draft, Visualization. **Sebastian Zaunseder:** Conceptualization, Software, Writing - review & editing. **Moritz D. Brandt:** Formal analysis, Writing - review & editing. **Hagen Malberg:** Conceptualization, Resources, Supervision, Project administration, Funding acquisition. **Felix Gräßer:** Conceptualization, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the Centre for Information Services and High Performance Computing (ZIH HPC) at TU Dresden for generous allocations of computer time.

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute.

Funding

European Union

Europe funds Saxony.
EFRE
European Regional Development Fund

This research was partly funded by the European Regional Development Fund with the project 100346021 “Tele-Schlaf-Medizin”.

Disclosure statement

This research was partly funded by the European Regional Development Fund (EFRE). EFRE was not involved in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

References

[1] J. Malik, Y.-L. Lo, H.-T. Wu, Sleep-wake classification via quantifying heart rate variability by convolutional neural network, Physiol. Meas. 39 (8) (2018) 085004, https://doi.org/10.1088/1361-6579/aad5a9.

[2] S. Zaunseder, A. Henning, D. Wedekind, A. Trumpp, H. Malberg, Unobtrusive acquisition of cardiorespiratory signals: Available techniques and perspectives for sleep medicine Kontaktlose Erfassung kardiorespiratorischer Signale: Verfügbare Verfahren und Perspektiven für die Schlafmedizin, Somnologie 21 (2) (2017) 93–100, https://doi.org/10.1007/s11818-017-0112-x.

[3] M.M. Schade, C.E. Bauer, B.R. Murray, L. Gahan, E.P. Doheny, H. Kilroy, A. Zaffaroni, H.E. Montgomery-Downs, Sleep validity of a non-contact bedside movement and respiration-sensing device, J. Clin. Sleep Med. JCSM Off. Publ. Am. Acad. Sleep Med. 15 (07) (2019) 1051–1061, https://doi.org/10.5664/jcsm.7892.

[4] C. Veauthier, J. Ryczewski, S. Mansow-Model, K. Otte, B. Kayser, M. Glos, C. Schöbel, F. Paul, A.U. Brandt, T. Penzel, Contactless recording of sleep apnea and periodic leg movements by nocturnal 3-D-video and subsequent visual perceptive computing, Sci. Rep. 9 (2019) 16812, https://doi.org/10.1038/s41598-019-53050-3.

[5] P. Fonseca, M.M. Van Gilst, M. Radha, M. Ross, A. Moreau, A. Cerny, P. Anderer, X. Long, J.P. Van Dijk, S. Overeem, Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population, Sleep 43 (2020), https://doi.org/10.1093/sleep/zsaa048.

[6] Q. Li, Q. Li, C. Liu, S.P. Shashikumar, S. Nemati, G.D. Clifford, Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram, Physiol. Meas. 39 (12) (2018) 124005, https://doi.org/10.1088/1361-6579/aaf339.

[7] D.-Y. Geng, J. Zhao, C.-X. Wang, Q. Ning, A decision support system for automatic sleep staging from HRV using wavelet packet decomposition and energy features, Biomed. Signal Process. Control 56 (2020), 101722, https://doi.org/10.1016/j.bspc.2019.101722.

[8] H. Sun, W. Ganglberger, E. Panneerselvam, M.J. Leone, S.A. Quadri, B. Goparaju, R.A. Tesh, O. Akeju, R.J. Thomas, M.B. Westover, Sleep staging from electrocardiography and respiration with deep learning, Sleep 43 (2020) 2341–2386, https://doi.org/10.1093/sleep/zsz306.

[9] H. Korkalainen, J. Aakko, B. Duce, S. Kainulainen, A. Leino, S. Nikkonen, I. O. Afara, S. Myllymaa, J. Töyräs, T. Leppänen, Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea, Sleep 43 (2020), https://doi.org/10.1093/sleep/zsaa098.

[10] R. Casal, L.E. Di Persia, G. Schlotthauer, Classifying sleep–wake stages through recurrent neural networks using pulse oximetry signals, Biomed. Signal Process. Control. 63 (2021), 102195, https://doi.org/10.1016/j.bspc.2020.102195.

[11] N. Sridhar, A. Shoeb, P. Stephens, A. Kharbouch, D. Ben Shimol, J. Burkart, A. Ghoreyshi, L. Myers, Deep learning for automated sleep staging using instantaneous heart rate, Npj Digit Med. 3 (2020) 106, https://doi.org/10.1038/s41746-020-0291-x.

[12] M. Goldammer, S. Zaunseder, H. Malberg, F. Gräßer, Specializing CNN Models for Sleep Staging based on Heart Rate, Comput. Cardiol. Conf. (CinC) 47 (2020), https://doi.org/10.22489/CinC.2020.105.

[13] S.F. Quan, B.V. Howard, C. Iber, J.P. Kiley, F.J. Nieto, G.T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J.M. Samet, P.W. Wahl, The sleep heart health study: design rationale, and methods, Sleep 20 (1997) 1077–1085, https://doi.org/10.1093/sleep/20.12.1077.

[14] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, S. Redline, The national sleep research resource: towards a sleep data commons, J. Am. Med. Informatics Assoc. 25 (2018) 1351–1358, https://doi.org/10.1093/jamia/ocy064.

[15] V.X. Afonso, W.J. Tompkins, T.Q. Nguyen, Shen Luo, ECG beat detection using filter banks, IEEE Trans. Biomed. Eng. 46 (2) (1999) 192–202, https://doi.org/10.1109/10.740882.

[16] C. Vidaurre, T.H. Sander, A. Schlögl, BioSig: the free and open source software library for biomedical signal processing, Comput. Intell. Neurosci. 2011 (2011) 1–12, https://doi.org/10.1155/2011/935364.

[17] D. Wichterle, J. Simek, M.T. La Rovere, P.J. Schwartz, A.J. Camm, M. Malik, Prevalent low-frequency oscillation of heart rate: novel predictor of mortality after myocardial infarction, Circulation 110 (10) (2004) 1183–1190, https://doi.org/10.1161/01.CIR.0000140765.71014.1C.

[18] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, (2015). http://download.tensorflow.org/paper/whitepaper2015.pdf.

[19] F. Chollet and others, Keras, (2015). https://github.com/fchollet/keras.

[20] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: ICLR'15, San Diego, 2015. arXiv:1412.6980.

[21] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37–46, https://doi.org/10.1177/001316446002000104.

[22] J.R. Landis., G.G. Koch, The Measurement of Observer Agreement for Categorical Data, Biometrics 33 (1) (1977) 159–174, https://doi.org/10.2307/2529310.

[23] R.B. Berry, S.F. Quan, A.R. Abreu, et al.; for the American Academy of Sleep Medicine. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.6, American Academy of Sleep Medicine, Darien IL, 2020.

[24] P. Fonseca, X. Long, M. Radha, R. Haakma, R.M. Aarts, J. Rolink, Sleep stage classification with ECG and respiratory effort, Physiol. Meas. 36 (10) (2015) 2027–2040, https://doi.org/10.1088/0967-3334/36/10/2027.