Noname manuscript No.
(will be inserted by the editor)

Enhanced Multi-Source Data Analysis for Personalized Sleep-Wake Pattern Recognition and Sleep Parameter Extraction

Sarah Fallmann · Liming Chen · Feng Chen

Received: date / Accepted: date

Abstract Sleep behavior is traditionally monitored with polysomnography, and sleep stage patterns are a key marker for sleep quality used to detect anomalies and diagnose diseases. With the growing demand for personalized healthcare and the prevalence of the Internet of Things, there is a trend to use everyday technologies for sleep behavior analysis at home, having the potential to eliminate expensive in-hospital monitoring. In this paper, we conceived a multi-source data mining approach to personalized sleep-wake pattern recognition which uses physiological data and personal information to facilitate fine-grained detection. Physiological data includes actigraphy and heart rate variability and personal data makes use of gender, health status and race information which are known influence factors. Moreover, we developed a personalized sleep parameter extraction technique fused with the sleep-wake approach, achieving personalized instead of static thresholds for decision-making. Results show that the proposed approach improves the accuracy of sleep and wake stage recognition, therefore, offers a new solution for personalized sleep-based health monitoring.

1 Introduction

Sleep stages are traditionally detected based on electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG) data. The process requires trained technicians to visually inspect data and score usually 30-second intervals based on guidelines, mainly the Rechtschaffen and Kales (R&K) [30]

Sarah Fallmann · Feng Chen
School of Computer Science and Informatics, De Montfort University, Leicester, U.K
E-mail: sarah.fallmann@gmail.com, fengchen@dmu.ac.uk

Liming Chen
School of Computing, Ulster University, Belfast, U.K
E-mail: l.chen@ulster.ac.uk

method or the American Academy of Sleep Medicine (AASM) [1]. The specific sleep stages can be divided into S1, S2, S3, S4, Rapid Eye Movement (REM), and wake [30] or N1, N2, N3, REM, and wake [1], where N3 is integrating S3 and S4 stages. Recently data analysis approaches to automatically detecting sleep stage are introduced to decrease the trained scoring technicians' workload, this has been proven helpful and is already adopted in certain settings. The approach to test against one technician is limited as technicians' scoring is subjective, i.e., have an individual component in the scoring process; therefore, the agreement is not always present [19]. This means that machine-learning approaches help to improve automation, but have also potentially learned one rater's style.

To collect sleep behavior data at home, and provide an unbiased data source for doctors, sensor technology is applied for detecting sleep stages [13, 9, 6]. Sleep parameters such as sleep onset latency (SOL) help to interpret the overall sleep quality; usually, extracted from sleep-wake behavior when monitored at home from acceleration data, but still lack in comparison to golden standards such as polysomnography (PSG). Home-monitoring sensors provide an unbiased data source, as data are collected in a natural environment and the number of body-attached sensors is minimal, therefore, less likely to influence sleep behavior compared to PSG [26]. In this context, a number of sensors such as actigraphy [17, 15, 27], photoplethysmography (PPG) [36], ballistocardiography [29], and non-contact microphones [7] have been applied to detect sleep-wake patterns and some progress has been made. The limitations of current methods are: (1) The use of an one-model-fits-all approach, trained and tested on (2) too small and (3) non-diverse datasets, (4) extracting too many features and (5) many factors influence sleep but have not been included in sleep stage detection. Non-diverse datasets contain, e.g., only healthy participants, but diversity is considered important during training [19] to provide a generalizable method. Especially important is to consider the performance of healthy and disease-affected subjects. These generalized methods follow an one-model-fits-all approach, even though, studies have shown sleep behavior differs for individuals based on factors such as biological factors, age, and lifestyle [32, 15]. Moreover, imbalanced data are usually applied in training, which can lead to restrictions for the model, being unable to reliably classify all stages as it is biased towards the majority class.

This work extends our previous work on sleep-wake behavior analysis [10]. To close the aforementioned gaps, firstly, we have developed an adaptive multi-source data learning approach for granular sleep-stage detection, which contributes to the existing sleep-wake classification by fusing and learning from multi-source data achieving novel features for sleep detection. Behavioral data are enhanced by establishing a flexible assessment based on different available data sources and by integrating personal information from medical history and genetic information. The approach is able to capture and characterize personal influence factors by the sleep-wake analysis at different levels of granularity. Specifically, this approach takes into consideration gender, race, and health status as influential factors. From the technological perspective, this study focuses on the use of HRV and actigraph data, as they are relatively easy to apply at home. In addition, we combine these two data sources to generate a rich feature set which has shown to improve the performance. Technically, we propose a multilayer perceptron (MLP), i.e., a feedforward artificial neural network. The approach has been evaluated on a large, diverse and balanced dataset, therefore, providing a representative investigation. This makes the model more robust and applicable for in-home usage. Moreover, we represent the impact of the individual features on the data model performance. The novelty lays in the capability of adapting to different users by incorporating multi-source data representing influence factors for sleep. Secondly, we contribute to knowledge with an adaptive sleep parameter extraction built upon the extracted sleep-wake stages, flexible towards individual needs. Personalized thresholds-based sleep parameter extraction improves reliability compared to current static threshold-based one-model fits all approaches.

The rest of the paper is organized as follows: Section 2 reviews related work, Section 3 describes the multi-source data learning approach for sleep stage recognition and parameter extraction and Section 4 outlines the experimental design. The results are presented in Section 5 and discussed in Section 6. Finally, overall conclusions and future work are given in Section 7.

2 Related Work

Sleep staging has been investigated in different complexity in home environments focusing on various problems: wake-sleep classification [17,21,36,37,27, 18], REM and NREM [31,5,7], 4 Stages, i.e., wake, REM, light and deep sleep [39] or 5 stages [16,38] [9]. Overall, methods usually deploy PSG as ground truth where a score is given ever 30-seconds. The basis for more complex sleep staging is a reliable sleep-wake assessment which is also relevant for sleep parameter extraction. Sleep parameter extraction, e.g., for sleep duration, means the extracted sleep-wake periods are analyzed in accordance to sleep episodes. In this work, we focus on sleep-wake staging and the extraction of sleep parameters thereof.

Many factors influence sleep and their impacts on sleep patterns can be observed. In the review of Johnson *et al.* [14], it was found that individual races show differences in sleep health and sleep disorders but research within-groups is still insufficient. Sleep disorders are defined over abnormal sleep behavior, therefore, influences sleep considerably [13]. Subjective measurement techniques found female gender to be a risk factor for poor sleep quality [33]; females have lower-quality sleep, higher SOL and wake up more often, whereas males often suffer from daytime sleepiness [24]. Influence factors are potential sources that can improve current sleep stage detection and are rarely investigated when sleep staging is performed. Current research on sleep-wake analysis mainly focuses on investigating one group of individuals, e.g. males only [27], or the combination of groups of people (healthy, elderly, and individuals with sleep restrictions) [15]. Tests on different groups of individuals are performed

[17] but the effects of influence factors on the sleep-wake analysis are not investigating and compared on the method level. Based on individual influence factors and health status changes, model parameters should be adjusted, and relevant measurements should be monitored to realize reliable sleep-wake analysis. The investigation of multi-source data would provide the possibility to incorporate influence factors and assess their influences on sleep stages in-depth, which we propose in this work.

Various sensors have been investigated for sleep staging. Relevant measurements which have proven promising in recent years for sleep-wake analysis are actigraphy [17,15,27], accelerometers body-worn and on the bed [37,21], PPG [36], ballistocardiography [29], and cameras [18]. Classifying sleep from wake using accelerometers is generally performing well, as movement is known as one of the main factors to distinguish them. Movement can also be monitored by cameras which has a more precise monitoring ability but comes with major privacy issues. HRV features can be extracted from electrocardiograms (ECG), PPG and ballistocardiography by analyzing heartbeats, which are changing in response to triggers such as rest and sleep. Changes in HRV with sleep quality have been used to diagnose sleep disorders and are also successful in sleep staging [36]. Research has investigated single data sources, but fusing multiple data sources that can easily be incorporated in one device has not been explored. This multi-source data can provide more in-depth information and the reason for specific data source performance can be discussed. In the presented work, we investigate the two most promising feature sources being HRV and actigraphy data.

In general, machine-learning approaches give accurate outcomes. In contrast, data-driven equations and thresholds can be considered to be alternatives [17]. A k-nearest neighbor (kNN) approach employed on PPG data reaches an accuracy of 77.35% for 10-fold cross-validation (CV). The investigation shows the best results by combining the extracted PPG and HRV features from 10 participants suffering from sleep apnea [36]. A convolutional neural network shows improvement compares to standard sleep-wake classification increasing specificity from 54% to 68%, but decreasing sensitivity from 82% to 80%. The study involves 22 elderly from which accelerometer-based night behavior is collected [37]. Instead of attaching wearable devices to the human body, they can be placed within the environment such as in [21], where five Shimmer sensors are positioned in the bed and validated towards a Philips Actiwatch. Random forests (RFs) are investigated on the down- and over-sampled data resulting in a sensitivity of 93% and specificity of 86%. Camera recordings are validated towards PSG and compared to actigraphy performance in [18]. To extract motion frame difference and motion history are extracted from 10 subjects. This video-based system reaches an accuracy of 92.13% [18] but comes with privacy concerns. Kuo *et al.* [17] propose four rules from density thresholds for raw one-axis accelerometer data to distinguish sleep from wake. The approach is tested on 81 subjects having poor and good sleep efficiency (SE) and results in an accuracy of 92.16%. Khademi *et al.* [15] show that personalized models trained per individual perform similarly compares to generalized models. The investigation is based on actigraph data collected from 54 subjects. These subjects include sleep-restricted individuals, acoustics disturbed, older adults on medication, and night workers. The overall model leads to an accuracy of 87% with extreme gradient boosting (XGB). But no significant difference in the performance could be found based on the individual characteristics investigated, i.e., gender, age, sleep disorder as well as time spent in bed. Actigraph data have also been investigated with a recurrence quantification analysis (RQA) including time aspects on 43 male participants reaching 85.3% accuracy [27].

The used methods can be divided into deep learning algorithms (convolutional neural network), instance-based algorithms (kNN), ensemble algorithms (random forest, XGB), non-linear signal analysis methods (RQA) and thresholds-based analysis. RFs are a collection of multiple decision trees that can overcome the restriction of simple decision trees [4]. A large number of deep trees can have high computational costs and use a lot of memory, similar is true for all ensemble algorithms (e.g., XGB). KNN calculates the distance to every neighbor for each prediction step and performs slowly when many predictions must be made [2]. NNs are based on layers of artificial neurons. The training process is based on weighting, from which predictions can be made even for incomplete information. NNs need long training times influenced by the number of parameters used [12], but perform fast afterward. Multilayer perceptrons (MLPs) (Artificial NNs) have to our knowledge not been performed in sleep staging but performs well for similar time-series data in the medical domain [20]. NNs with more than three layers are considered deep learning methods used due to their reliable performance. However, it is not possible to see which features are important or how the outcome is produced; additionally, since a large quantity of training data is necessary, computational costs are quite high. Overall, the number of parameters that need to be assessed are related to the computational costs, therefore, more complex-structured models are not the best choice when similar performance can be achieved with simpler models.

Using sensors in a home environment is less intervening with the sleeping habit, but still needs investigations to provide a system being able to perform accurately for different groups of subjects. It is especially important to consider differences between monitoring healthy subjects and subjects with medical conditions [19]. Overall, research is based on one-model-fits-all approaches validated on small datasets that are not diverse, excluding the impact of potential personal influence factors on the performance and especially the models. Furthermore, imbalanced data are usually used during training, resulting in a biased model towards the majority class which can introduce disadvantages. We target on these issues by proposing a multi-source data learning approach for granular sleep-wake detection, investigating a balanced diverse dataset including health status, gender, and race. These factors are known influence factors for sleep which are relevant. Furthermore, we incorporate two data sources which are promising as such actigraphy and HRV features and combine them in our multi-source dataset. This provides the possibility to assess the data source and influence factors individual ability of classification, establish their performance differences and improvements. Furthermore, we propose a personalized sleep parameter extraction technique incorporating dynamic thresholds for sleep parameters using sleep-wake assessment.

3 Sleep-Wake Pattern Analysis

In this section, we describe the multi-source data learning, feature extraction, and preprocessing for sleep-stage recognition; and the personalized sleep parameter extraction technique fused with the sleep-wake recognition.

3.1 Multi-Source Data Learning for Sleep-Wake Recognition

We propose a multi-source data learning approach with a fine-grained structure to detect sleep and wake stages. Our aim is to move away from a one-model-fits-all approach towards a personalized approach. We consider a two-stage classification problem following the gold-standard of 30-second scoring intervals; these intervals are scored as sleep or wake stage.

A visual representation of our multi-source data learning approach is presented in Fig. 1. We incorporate two types of data: personal data, including gender, health status, and race; and physiological data, including actigraphy, HRV, or a combination of the two. The physiological information is fused with clinical history and genetic information (influence factors). Depending on the specifics of the personal data, a portion of this fused information is selected and a final dataset is assembled. The selected training datasets are pre-processed and features are extracted, which are explained in more detail in the following sections. The features are used to train MLP with one hidden layer that has the same size as the list of features. The structure is depicted in Fig. 2, which shows the number of input features $(F_1, F_2, ..., F_n)$ and nodes in the hidden layer $(N_1, N_2, ..., N_n)$. The overall model consists of three layers: an input layer, a hidden layer, and an output layer. The output layer describes the classification targets, which are wake (W) and sleep (S) stages. The number of nodes is adjusted to the accessible sensory data input features, either being 18 for HRV, 7 for actigraphy, or 25 for both (details in Section 3.3). We decided on the MLP empirically based on initial experiments comparing methods considered appropriate in the literature. These methods show similar outcomes with respect to accuracy; the best outcomes across several experiments are presented here: Generalized Matrix Learning Vector Quantization, 78.2%; kNN, 79.0%; MLP, 79.1%; ada-boosting, 79.2%; and RF, 80%. A model with a simple structure that is still able to learn from the data can counteract the possibility of overfitting. MLP contains three layers and still performs well; the tested ada-boosting and RF have a more complex structure—ada-boosting is tested on 100 estimators and RF is tested with 50 estimators and a depth of 30. The results from ada-boosting and MLP are not significantly different (t-test on two related samples). Furthermore, the standard deviation (SD)

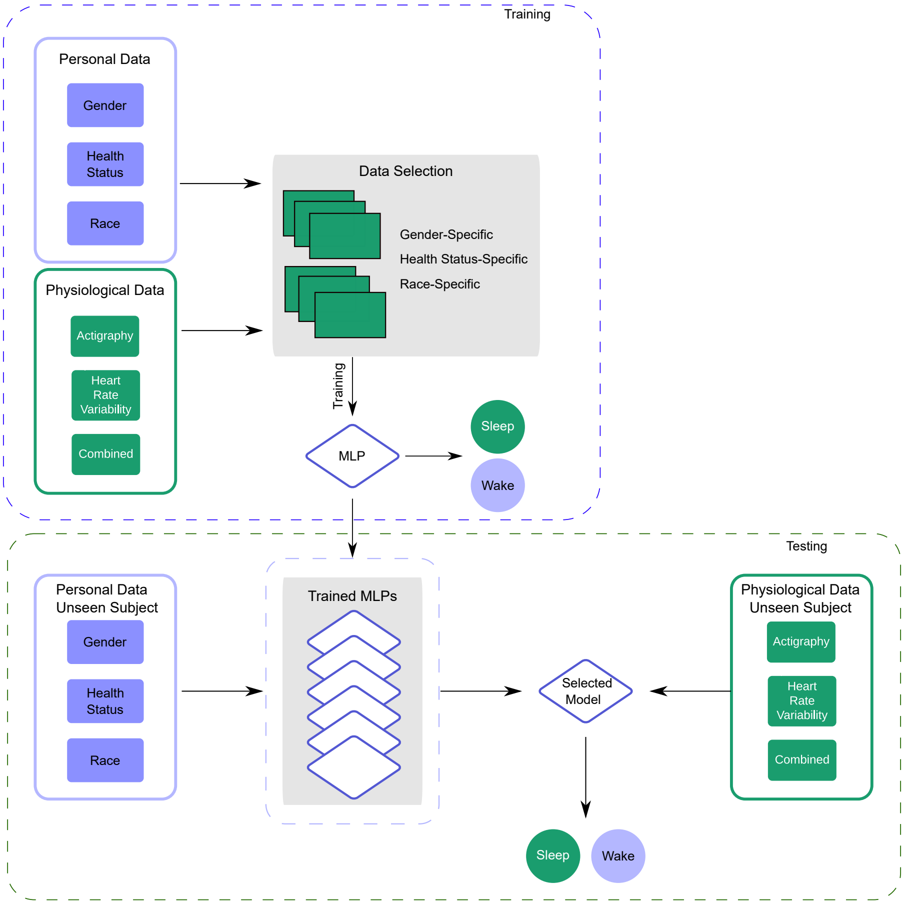

Fig. 1: The multi-source data learning approach for sleep-wake pattern detection

over the repeated experiments was lower for MLP compared to ada-boosting, therefore, was chosen, as it suggests a more stable performance. We investigated the mean squared error (MSE) for the training and test sets and found that RF was overfitting (MSE was high for testing and low for training). MLP did not show overfitting, as a consequence was chosen as being an appropriate model for further investigations.

For testing on unseen individuals, a model is selected based on the subject's personal information. Features from the physiological data are then calculated and given to the selected model to assess sleep and wake stages.

We illustrate the idea of a fine-grained adaptive structure against the traditional general approach in Fig. 3. The information used to reach a granular adaptive model details gender, race, and medical history data, i.e., health status. The health status reflects if an individual is affected by a sleep disor-

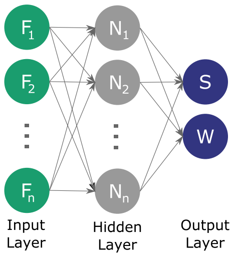

Fig. 2: Multilayer perceptron design.

der or healthy. In our tests, we include four races—White, Asian-American, African-American, and Hispanic—and three diseases—RLS, insomnia, and sleep apnoea. The standard data for general sleep-wake approaches are used to generate specific models. This data typically includes physiological data from a single data source (e.g., actigraphy or HRV). We expand upon this with multiple data sources by fusing actigraphy and HRV. Figure 3 shows the steps to achieve fine-grained models. First, gender information determines gender-specific models. Second, the health status of an individual—included if available—determines a health-status-specific classification. Third, race identification, if available, determines a race-specific model. This information can also be combined to select a model with two or all three influence factors. Each of these steps is flexible; if at least one is included, a fine-grained sleep-wake recognition approach is reached. The advantage of a fine-grained approach is that influence factors can be incorporated into individual models.

The proposed approach is designed to be adaptable to (1) available personal information and (2) available sensory data. First, the availability of gender, health status, and race is not guaranteed. Therefore, only one is necessary to reach a somewhat personalized approach. The different integrated medical history and genetic aspects can deal with different groups of individuals, e.g., female healthy Asians. Second, the approach can incorporate multiple data sources by combining calculated features through a 30-second raw data structure. This is feasible by adapting the size and overlap of the windows when calculating individual features. Generally, this approach can be applied to any data source that provides information on sleep in a similar structure. The design of our approach allows for varying levels of granularity, ranging from gender-specific to race-health-gender-specific. The system can work with the granularity of genders, health status states, and races, concrete validation details are given in Section 4.1 and presented in the 'tested on' parts.

3.2 AI-enabled Sleep Parameter Extraction Technique

We developed a personalized AI-enabled sleep parameter extraction technique which is fused with the sleep-wake approach described in the previous section. The steps of the technique are summarized in Fig. 4. Firstly, data coming from

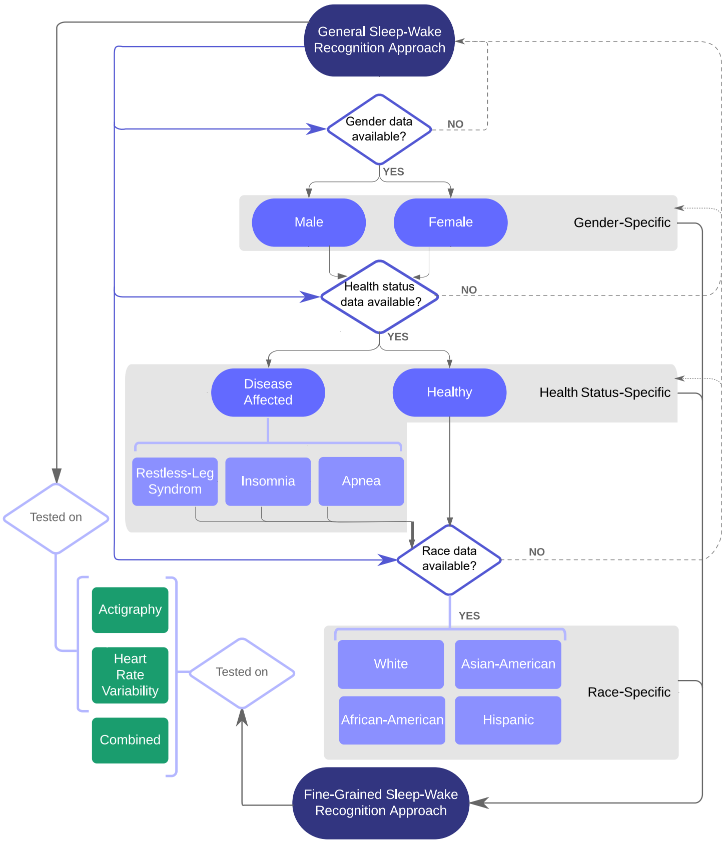

Fig. 3: Illustration of the steps from a general to a fine-grained approach.

behavioral and personal information are prepared for the sleep-wake pattern extraction. This means individuals are categorized by personal information and features are extracted, details are given in the next section. Secondly, the extracted wake-sleep patterns are standardized towards 30-second epochs (the standard interval for medical assessment). The wake-sleep episodes are the basis for sleep parameter extraction. This means, e.g., for sleep duration, the extracted sleep-wake periods are analyzed in accordance to sleep episodes. Generally, static thresholds are used to define sleep parameters not personalized to groups of individuals; e.g., the 15 minutes rule is applied [23] for SOL investigation, but research has shown that age-based variable thresholds improve the correspondence to PSG [22]. Therefore, we incorporate dynamic thresholds that change depending on personal factors. We propose to personalize the thresholds for sleep parameters based on subcategories, i.e., health status, race, and gender. Therefore, thresholds need to be adapted according to the personal information available. In this investigation, we follow an empirical threshold determination applying the previously described sleep-wake pattern recognition.

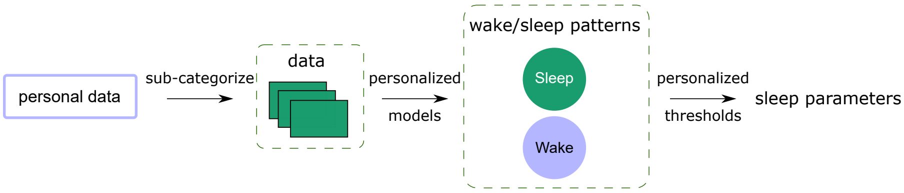

Fig. 4: Personalized sleep parameter extraction.

3.3 Feature Extraction and Preprocessing

Feature extraction and preprocessing are essential for our approach especially for being adaptable to accessible data source information. Therefore, different sampling rates need to be adapted to a 30-second interval structure. Furthermore, we consider a balanced training and test set for sleep stages and gender.

For actigraph data, the provided data include activity counts and light data in 30-second intervals. Furthermore, we calculate the mean and SD features over 5-minute overlapping windows. Activity counts summarise the intensity of activeness at a time [8]. While light gives information about the ambient light divided into white-, red-, green- and blue-light. These light features can help to judge the day-night-rhythm [8]. Light exposure at night has shown effects on sleep, blue light increased alertness and red light increased cortisol levels [11]. Furthermore, blue light is emitted by smartphones and can help to judge a by movement unrelated wakefulness with the blue light level. Both can cause awakenings or difficulties falling asleep, therefore, they are important aspects to assess sleep and wake stages. When using light as features patterns can be learned by the MLP which reflect the influences and help better judge sleep and wake stages. Overall seven features are extracted: activity counts, mean, SD and four light features.

To extract the HRV features, an analysis based on ECG is performed by the NSRR [25] to extract QRS complexes and, therefore, R-points. The QRS complex describes three of the graphic deflections in a typical ECG. From these R-points, normal sinus beat intervals (NNs) and cardiac inter-beat intervals (RRs)—the intervals between adjacent QRS complexes—are determined (for details compare [35]). Certain intervals are excluded based on the exclusion criteria presented in [25] to remove artifacts: Excluding individuals (1) with sleep stages of <2 hours, (2) with normal-to-normal-intervals (NN) <1.000, (3) NN $<$ 0.35 sec or $>$ 2.5 sec, (4) <180 NN within 5-minute windows. To receive the HRV features the information is fused into features getting values for 30-seconds each coming from 256Hz sampling rate. The idea is based on

Table 1: HRV features which are calculated and excluded (x) based on their pre-assessed pairwise correlation.

Feature	Description	x	Feature	x
	Time-Domain			
	30-sec Intervals		Whole Sleep Period	
NN_RR	ratio of consecutive NNs* over all RRs*	x	Tot_NN_RR	x
AVNN	average of all NNs*		Tot_AVNN	
IHR	average instantaneous heart rate		Tot_IHR	
SDNN	standard deviation of all NNs*	x	Tot_SDNN	
rMSSD	square root of the mean of the squares of difference between adjacent NNs*		Tot_rMSSD	
	percentage of differences between adjacent NNs* that are			
pNN10	>10 ms		Tot_pNN10	x
pNN20	>20 ms		Tot_pNN20	x
pNN30	>30 ms	x	Tot_pNN30	x
pNN40	>40 ms	x	Tot_pNN40	
pNN50	>50 ms	x	Tot_pNN50	
	5-min Intervals			
SDANN	standard deviation of the averages of NNs*			
SDANNINDX	mean of the standard deviations of NNs*			
	Frequency-Domain			
	5-min Intervals		Whole Sleep Period	
TOTPWR	total NNs* spectral power up to 0.4 Hz	x	Tot_TOTPWR	x
ULF	ultra-low FP*: [0, 0.003 Hz]		Tot_ULF	
VLF	very low FP*: [0.003, 0.04 Hz]		Tot_VLF	x
LF	low FP*: [0.04, 0.15 Hz]	x	Tot_LF	
HF	high FP*: [0.15, 0.4 Hz]		Tot_HF	x
LFn	normalized LF	x	Tot_LFn	x
HFn	normalized HF	x	Tot_HFn	x
LF-HF-ratio	the ratio of low to high FP*	x	Tot_LFHF	

*NNs-normal sinus beat intervals; RRs-cardiac inter-beat intervals; FP-frequency power

the amount of extracted beats within epochs of 30 seconds. In Table 1, the extracted features are provided, divided into time- and frequency-domain features. The HRV features are calculated over 30 seconds, the whole recorded sleep period, and over 5-minute intervals according to the NSRR formulation [25,34]. Whole night values are treated as a constant feature for data from a specific subject. Traditional resampling can cause a reduction in high-frequency components and Fast Fourier transformation needs balanced distributed samples; Lomb periodograms are used for frequency-domain spectra calculation of irregular data samples [34,25]. Furthermore, pairwise Pearson correlated features are excluded reducing the number of features from 38 to 18 (excluded features are marked as x in Table 1).

Moreover, we apply down-sampling to reach a balanced dataset based on gender and sleep stages. As the used MLP does not look at a time series, we delete randomly different values from the majority class per repeat of CV. Moreover, the data are normalized per feature. For the case of non-gender specific training, we keep a ratio of 1/1 for male/female and sleep/wake stages, creating a gender and sleep stage balanced dataset and reject those subjects which less than 400 examples of sleep-wake stages, details in Section 4.2.

4 Experimental Design

To validate our approach and test the necessity of personalized models tangled with granular adaptive data analysis, certain experiments are performed. We are especially interested in sleep stage recognition with actigraphy and HRV data and sleep parameter extraction. Therefore, data are used which provides actigraphy and HRV features. In this section, we present the considered experimental data and the experimental settings to investigate the performance and influential aspects on sleep-wake stage detection.

4.1 Experimental Data

To explore our approach, we use an existing dataset from the National Sleep Research Resource (NSRR) [25]. More concrete data from the Multi-Ethnic Study of Atherosclerosis (MESA). The dataset includes PSG, actigraphy and sleep questionnaires data measured at home in an unsupervised setting [3] and was made available by the [25] [34]. The advantage of the dataset is the diversity which comes from the inclusion of participants (1) from different ethnic backgrounds including African-American (B), White-American (W), Hispanic (H), and Asian-American (A), (2) with both genders, (3) aged between 45 and 84 and (4) with different health status (HS) either healthy (He) or diagnosed with sleep disorders, in concrete, sleep apnea (Apn), insomnia (Ins), and restless leg syndrome (RLS) [3]. The segmentation of the dataset which is used during the investigation for training and testing is presented in Table 2 shows the number of people and average age with SD per data source and diversity factors. An overview of the data segmentation on gender and ethnicity can be seen in Fig. 5. We can see the unequal distribution which needs to be taken into account. For the case of females and males together, we provide the number of available data used in the two different experimental settings, i.e., gender-balanced and gender-imbalanced case, details are given in Section 4.2. The data collection was performed in the United States.

The initial objective of the study was to examine disparity of sleep disorders and behavior across ethnic groups and gender as well as relate sub-clinical atherosclerosis [3,25]. In this investigation, we use the first night of actigraphy which was collected simultaneously with PSG [3,34] to validate our proposed personalized multi-source data learning approach for diverse subjects and influence factors. Recordings from PSG and actigraphy were scored by the help of trained technicians, but have only a 76.71% match with the PSG scores, while down-sampling and averaging shows only 73.45% coherence, suggesting a majority class bias, compare Section 5. This shows the necessity of improving current recognition rates and investigating a granular adaptive approach. The

Table 2: Diversity-factor-specific data segmentation incorporating gender, race, and health status.

		ACTIGRAPHY						HEART RATE VARIABILITY					
		Together		Male		Female		Together		Male		Female	
RACE	HS	No.	No.G.*	No.	Age	No.	Age	No.	No.G.*	No.	Age	No.	Age
All	All - HS	1641	1458	896	69.1±9.0	745	69.1±9.0	1576	1236	862	68.9±8.9	714	69±8.9
All	He	1367	1226	740	69.3±9.1	627	69.2±9.0	1313	1030	712	69.2±9.0	601	69.1±8.9
All	Apn	120	96	50	67.8±7.7	70	66.6±7.3	116	82	50	67.8±7.7	66	66.5±7.3
All	Ins	99	64	66	66.6±8.7	33	67.6±9.5	94	56	63	66.7±8.8	31	67.7±9.6
All	RLS	73	54	46	67.9±9.2	27	68.1±9.2	69	52	43	67.1±8.7	26	68.5±9.2
W	All - HS	612	566	325	69.2±9.1	287	69.3±9.0	591	492	315	68.9±9.0	276	69.3±9.0
W	He	518	474	277	69.6±9.3	241	69.6±9.1	500	410	268	69.3±9.2	232	69.5±9.1
W	Apn	46	32	17	66.5±9.0	29	65.9±6.9	44	24	17	66.5±9.0	27	66.1±7.0
W	Ins	24	16	16	64.5±5.9	8	64.9±9.4	23	16	15	63.8±5.4	8	64.9±9.4
W	RLS	26	24	14	67.8±7.9	12	69.7±10.2	26	24	14	67.8±7.9	12	69.8±10.2
A	All - HS	176	156	97	69.2±8.4	79	68.0±9.2	166	132	91	69.7±8.5	75	67.6±8.8
A	He	129	144	71	69.5±8.1	58	67.9±9.0	122	92	67	69.9±8.2	55	67.4±8.2
A	Apn	8	2	1	77.0±0.0	7	64.1±10.0	7	2	1	77.0±0.0	6	62.7±10.2
A	Ins	28	22	17	66.8±7.9	11	69.4±9.5	25	20	15	67.5±8.1	10	69.1±9.9
A	RLS	8	8	4	65.8±12.0	4	62.8±8.02	8	8	4	65.8±12.1	4	62.8±8.0
B	All - HS	453	382	256	69.0±8.8	197	69.5±8.7	435	312	249	69.1±8.7	186	69.3±8.5
B	He	382	332	211	69.2±8.8	171	69.6±8.7	367	272	205	69.3±8.8	162	69.4±8.6
B	Apn	40	38	20	69.6±7.0	20	68.2±7.4	39	32	20	69.5±7.0	19	67.8±7.4
B	Ins	20	6	17	66.8±8.9	3	66.7±14.2	19	2	17	66.8±8.9	2	71.0±17.0
B	RLS	16	10	11	68.9±12.2	5	69.0±11.3	14	8	10	67.1±11.2	4	71.8±10.9
H	All - HS	400	354	218	68.9±9.5	182	68.7±9.2	384	300	207	68.5±9.4	177	68.7±9.1
H	He	338	306	181	68.9±9.5	157	68.7±9.4	324	256	172	68.4±9.3	152	68.6±9.2
H	Apn	26	22	12	65.8±6.3	14	67.0±6.8	26	22	12	65.8±6.3	14	67.0±6.8
H	Ins	27	20	16	68.4±11.7	11	67.9±9.4	27	18	16	68.4±11.7	11	67.9±9.4
H	RLS	23	12	17	67.8±8.1	6	67.8±6.5	21	12	15	66.7±7.6	6	67.8±6.5

* No.G. - Number of subjects in gender balanced case

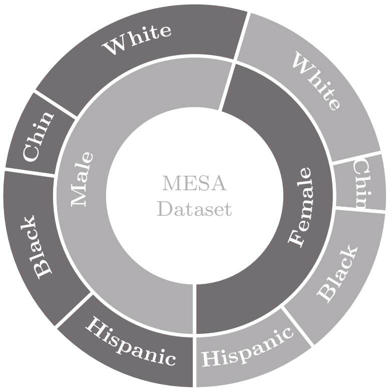

Fig. 5: Data segmentation overview on Gender and Ethnicity.

ground truth (PSG) of sleep stages are following the R&K method, which we translated to wake and sleep stages. In the investigation, sensors are used that can easily be applied at home, as such actigraphy and HRV features from ECG measurements. To use the dataset, we need to extract features and pre-process in line with the process described in Section 3.3.

4.2 Experiment Settings

The validation of our proposed personalized sleep-wake recognition approach is performed in various scenarios that are described in detail in this section. As already stated, we conducted health-status-specific and race-specific experiments. Health status categories include: all health statuses (shown as All-HS); healthy (He); diagnosed with apnea (Apn); diagnosed with insomnia (Ins); and diagnosed with the restless-leg syndrome (RLS). Race categories include: all races (All-Races); African-American (B); White (W); Hispanic (H); and Asian-American (A).

The training and test sets are the individual segmentation per experiment shown in Fig. 5. All investigations are validated with six times 10-fold CV to ensure representative outcomes and we report averaged accuracies with SD in Section 5. This means that each data segmentation in Fig. 5 is separated into 10 sub-sets were each segment is once used for testing, while the other 9 segments are used for training, which is repeated six times. Accuracy is considered sufficient, as data are always balanced for sleep stages (this implies recall is the same as the accuracy). Based on changes of SD across different repeats of CV, we find the influences that the used training data has; if the SD is higher, the approach works better for certain training-/testing-splits than for others. The experiments are investigated for different accessible data sources: actigraphy, HRV, and the two fused together (fusion is done on the feature level).

The structures of the sleep-wake analysis scenarios include the following steps:

First, we compare our fine-grained approach on actigraphy data with the original sleep-wake classification, which is part of the used dataset [25]. The original sleep-wake classification is denoted as 'Original Sample' and is based on the imbalanced dataset with respect to sleep-wake stages. Additionally, we explore the performance in the down-sampled case, referred to as the 'Undersampled'. This scenario is further referred to as scenario (1).

Second, we analyze data for females (shown as Female in the result section) and males (Male); all data from the specific gender groups are used in the model supporting a gender-specific model. The average of these outcomes is also given (Avg. Gender) in order to be comparable to the overall performance when personalization is not integrated. Furthermore, the race-specific and health status-specific analyzes are performed, which are using the available race and health status information to develop a fine-grained adaptive model. We will refer to 'Avg. Races' as the average of all individual race-specific trained models and 'Avg. Health Status' as the averaged performance of all health status-specific models. Combinations of, e.g., gender and race information in a model are referred to as 'Avg. Gender and Race'. These scenarios are further referred to as scenarios (2).

Third, we compare the outcomes with the case in which the data from both genders are taken together to train a more general model. This general approach is explored with a 1/1 ratio for female/male and sleep-wake stages per individual (All; (3a)) or imbalanced towards gender using all data from all individuals (All-biased; (3b)). These scenarios are further referred to as scenarios (3a) and (3b).

Fourth, we consider gender as an extra feature in the data, either balanced sleep stages and gender (Incl. Gender; (4a)) or imbalanced towards gender (Incl. Gender-Biased; (4b)). These scenarios are further referred to as scenarios (4a) and (4b). Experiments (3a) and (4a) are trained on six times 10-fold CV based on subjects, making sure the sleep-wake ratio is 1/1 for each subject set.

Fifth, we determine the best approach in terms of training data. Models are trained on individuals with a medical condition and tested on both healthy individuals and individuals with a medical condition. Additionally, models are trained on healthy individuals and tested on both healthy individuals and individuals with a medical condition. In the results section, we refer to this analysis as 'Train Disease - Test Healthy' (i.e., train model on diseased individuals and test on healthy individuals).

Lastly, we investigate the performance of the variable-threshold-based parameter extraction process built upon our wake-sleep assessment approach. We use the lights-off time from PSG provided in the dataset to make a direct comparison with the actigraphy. This means that both data sources are used from the lights-off time onwards. The scope of the functionality investigation lies in SOL, because SOL is important to show abnormalities of sleep behavior and to indicate insomnia.

5 Results

In this section, the results of the experiments are presented investigating actigraphy and HRV features.

5.1 Comparison of Different Approaches

In Fig. 6, the outcomes of the original sleep-wake classification is compared with the outcomes of our fine-grained approach on actigraph data. For actigraph data, the original classification for sleep-wake stages from [25] can be seen as 'Original Sample' which is around 76%. In the downsampled case, shown as 'Under-sampled' the repeated-averaged accuracy decreases by around 3%. For example, for 'All-HS' a decrease can be observed from 76.71% to 73.45%.

Furthermore, the results are given for: 'All' (compare Section 4.2 3a), ' Gender' (4a), 'Avg. Gender', 'Avg. Races', 'Avg. Gender and Race', and 'Avg. Health Status'.

In Fig. 6, actigraphy based recognition rate can be improved for around 5% when using MLP, compared to the down-sampled currently used recognition methodology. For example, 'All-HS' reached 78.98% for 'Avg. Gender' compared to the downsampled 73.45%.

The 'Avg. Health Status' models reach an accuracy of 77.88%, while gender-specific personalized models on top of the health-specific classification can achieve an averaged accuracy of 78.47% [79.16, 79.38, 78.06, 77.29] and even 78.75% [78.87, 78.93, 79.05, 78.15] if also the race is specified. Overall, the gender-specific approach reaches the best outcome with 78.98% in the 'All-HS' case.

We can especially see the effect of inclusion of race- and gender-specifics in the disease-affected cases. For example, 79.05% accuracy for insomnia 'Avg. Gender and Race' compares to 78.06% for 'Avg. Gender'. The results show that a fine-grained model improves current one-model-fits-all approaches, especially for subjects affected by a disease compare 'Avg. Gender and Race' (79.05%) and 'All' (77.01%).

Furthermore, it can be observed that including gender ('Incl. Gender') as a feature during training does not change the recognition rate towards 'All' mentionable. This can be seen, e.g., in 'Ins' which reaches an accuracy of 77.01% for 'All' which is basically the same as 77.05% for 'Incl. Gender'.

5.2 Investigating Different Data Sources

In Fig. 8, 9 and 10, the results from experiments (2)-(4) are presented for 'Male', 'Female', 'Avg. Gender' (compare Section 4.2 2), 'All' (3a) and 'Incl. Gender' (4a).

The SD is not given in 'All' and 'Incl. Gender' for visual reasons, overall the SD for 'All' and 'Incl. Gender' are similar, and 1-5% higher than in the

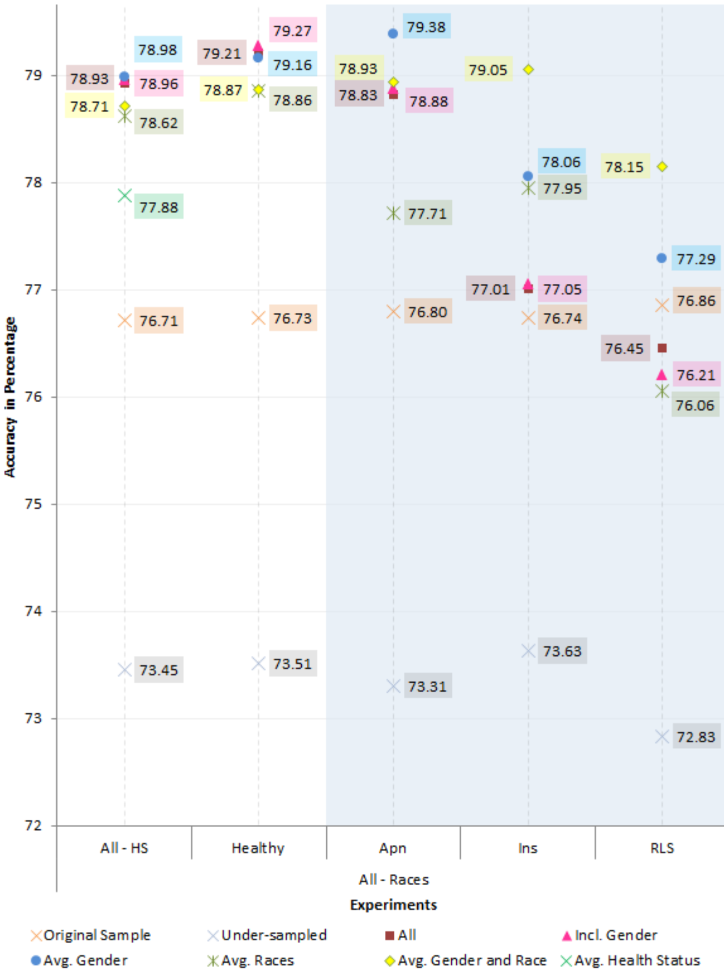

Fig. 6: Performance overview of original sleep-wake classification and the proposed approach.

individual trained models. For example, 'White, RLS, All' has a SD from 7.8% and 'White, RLS, Incl. Gender' 7.6% while the gender-specific 'Avg. Gender' has only 0.95%.

Furthermore, it should be noted that the SD is lower in the (3b), (4b) gender-imbalanced experiments compared to the (3a), (4a) balanced case, refer to Fig. 7. More concrete, for (3b): 'All-Races, Apn, All' the SD is 0.5% and for (4b): 'All-Races, Apn, Incl. Gender', 0.4% compared to (3a): 3.4% and (4a): 3.7%. There is an accuracy difference between (3a) and (3b) as well as (4a) and (4b) for HRV, which is not as obvious present in actigraphy. For example, (3a) 'A, He, All' reaches an accuracy of 79.1% while (3b) reaches an accuracy of 71.6% in the HRV case.

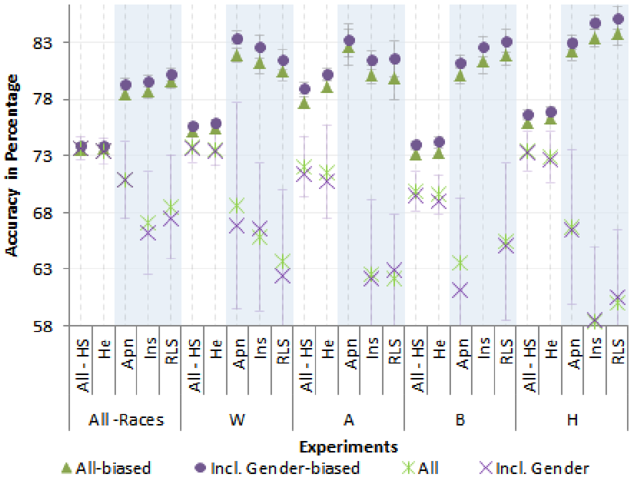

Fig. 7: Gender-balanced and gender-imbalanced HRV training results.

In Fig. 8, accuracy is usually higher for male participants compared to females. For example, 'All-HS' in the 'All-Races' case can reach 80.04% accuracy for males and 77.92% for females. In contrast, for HRV in Fig. 9 the outcomes are more similar, with the individual female models performing better overall with 80.60% averaged accuracy compared to 79.64% for males. In the combined case in Fig. 10, the males (85.77%) outperform the female models (84.65%) again.

The effect of fewer subjects is represented in higher SD, compare Table 2. For example, 'HRV, Female, A, Apn' with six subjects has a SD of 1.6% and 'HRV, Female, W, He' with 601 subjects has a SD of 0.2%. Be aware that the 'Male, Asian, Apnea' case simply contains one participant, which makes it not representative to make conclusions from this setting, therefore, it is excluded in the presented results.

For 'He' individuals (e.g in Fig. 8 'White, He, Male' with 0.2%) and 'All-HS' data (0.2%), the SD is normally lower than for the disease models ('Apn': 1.1%, 'Ins': 0.8% and 'RLS': 0.8%).

The average SD for HRV is in general higher with 1.08% for 'Avg. Gender', 4.27% for 'All' and 0.68% for 'All-biased', while it is 0.79% for 'Avg. Gender', 3.42% for 'All' and 0.59% for 'All-biased' for actigraphy, and 0.86% for 'Avg. Gender', 4.09% for 'All' and 0.6% for 'All-biased' in the combined case.

Overall, the accuracy for actigraph data is less diverted over races and health statuses (ranging from 76.3% to 82.2% for 'Avg. Gender') compared to the HRV case (ranging from 74.2% to 86.2%), as can be seen in Fig. 8 and 9. For the actigraph case, it is more difficult to distinguish wake and sleep stages in the Hispanic group (averaged 'Avg. Gender' over Hispanic individuals 76.65%), while for HRV it is very accurate (81.22%).

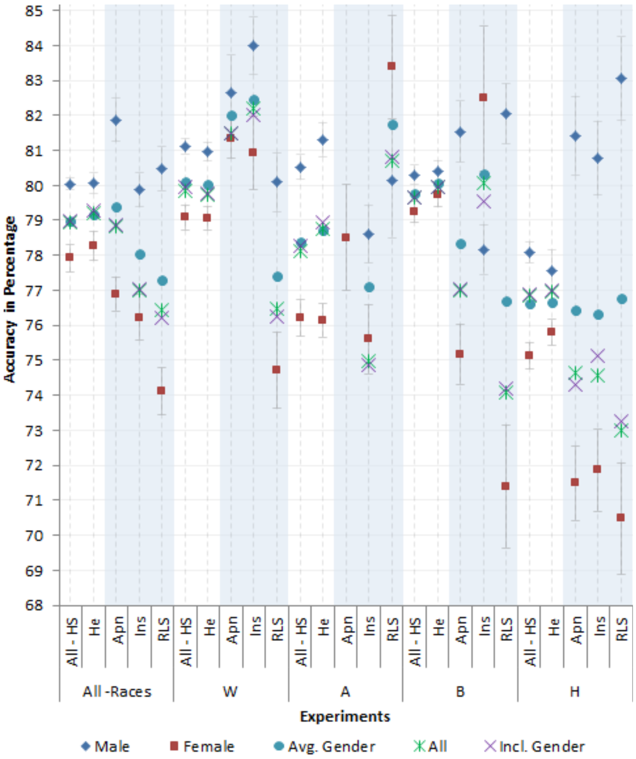

Fig. 8: Outcomes for actigraphy data: Accuracy with SD for female (Female), male (Male), and both together (All); averaged female and male (Avg. Gender); and both together including gender as a feature (Incl. Gender).

If we look at disease-affected and healthy subjects, HRV in Fig. 9 shows a trend, performing better for disease participants compared to healthy. For example, 'White' averaged disease-affected results reach 82.05% accuracy compared to healthy with 76.64% accuracy. This trend is also visible when combining the data sources in Fig. 10. Actigraphy shows closer accuracy rates while male and female models seem more diverted. The disease-affected cases perform slightly worse than the healthy cases. For example, 'B' averaged disease-affected outcomes reach 78.46% while for healthy cases 80.08% accuracy can be reached.

Based on these trends, we investigated the combined cases (compare Fig. 10), which improved the accuracy around 5% compared to HRV (e.g. 74.5% for 'All-Races, He' using HRV to 80.6% in the combined case) keeping the general trend of the HRV models. For actigraphy, only 1% increase in the healthy subjects (from 79.21% to 80.87% for 'All-Races, He') but also 5-8% in the disease-affected subjects can be recognized (e.g. 'All-Races, RLS' from 77.29% to 85.47%).

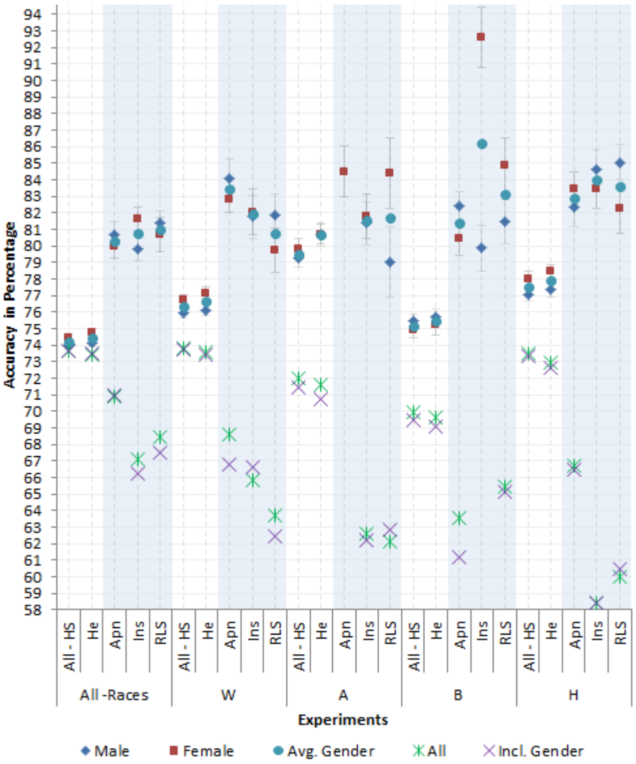

Fig. 9: Outcomes for HRV data: Accuracy with SD for female (Female), male (Male), and both together (All); averaged female and male (Avg. Gender); and both together including gender as a feature (Incl. Gender).

5.3 Exploring Healthy and Disease-Affected Performance

In Fig. 11, 12, and 13, the experiments from setting (5) are given for actigraph data, HRV data and the combination of HRV data and actigraphy measurements. The results are given for the trained on individuals with a medical condition and tested on healthy subjects ('Train Disease - Test Healthy') and tested on subjects with a medical condition ('Trained Disease - Test Healthy'). Furthermore, the average of the outcomes ('Avg. Disease Trained') is given. The fields are colored in shades of red for the model trained on subjects affected by a medical condition and colored in shades of blue for on healthy subjects trained models. Furthermore, the average of the individual models trained on specific diseases is given ('Avg. Ind. Disease Models') and the average of all individual health statuses models ('Avg. Ind. Models'), i.e., healthy and diseased-affected.

For Fig. 11 a very clear performance difference between male and female groups is visible over all experiments, e.g., 'All-Races, Males' reach 80.4% while

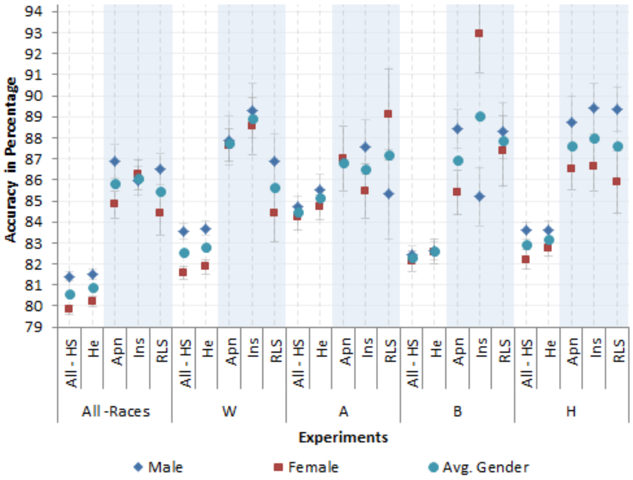

Fig. 10: Outcomes for the combined data of actigraphy and HRV: Accuracy with SD for female (Female), male (Male), and both together (All); averaged female and male (Avg. Gender); and both together including gender as a feature (Incl. Gender).

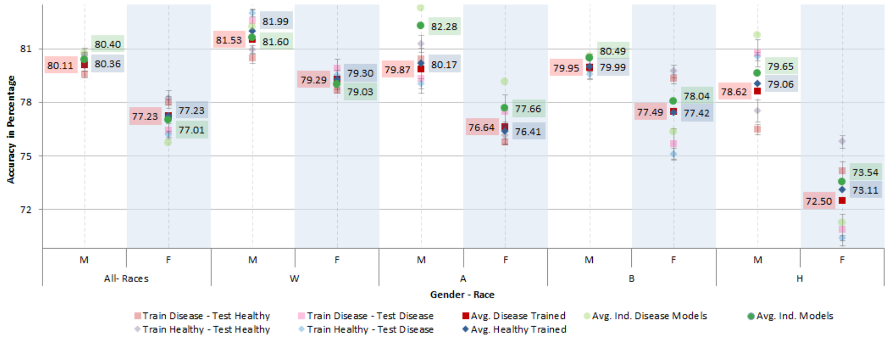

Fig. 11: Diseased-individual- and healthy-individual-trained models tested on diseased and healthy individuals for actigraphy data.

'Females' reach 77.01%. In the Hispanic group, it is difficult to distinguish the sleep-wake stages for females.

When comparing Fig. 12 and 13, a similar trend can be recognized from HRV and the combined case. The trained models are most distinctive in the HRV and combined case Fig. 13, meaning individual trained models fine-grained on disease and health status are preferable. For example, for white males '85.85%' are reached with the fine-grained model compared to 82.73% and 81.43% for 'Avg. Disease Trained' and 'Avg. Healthy Trained'. This trend is also present in Fig. 11 when races are considered, such as for Asians, males

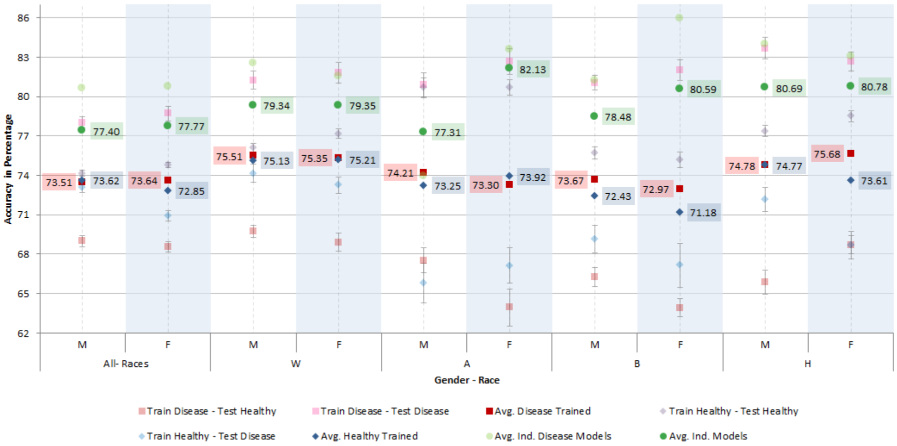

Fig. 12: Diseased-individual- and healthy-individual-trained models tested on diseased and healthy individuals for HRV data.

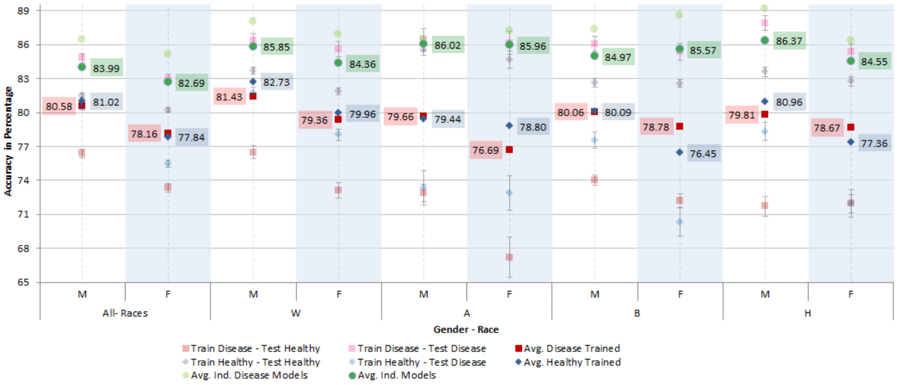

Fig. 13: Diseased-individual- and healthy-individual-trained models tested on diseased and healthy individuals for combined actigraphy and HRV data.

72.28% in 'Avg. Ind. Models' compared to 80.17% for 'Avg. Healthy Trained' and 79.87% for 'Avg. Disease Trained'.

5.4 Comparison to Existing Methods

In Table 3, we compare our proposed model to standard sleep-wake recognition (presented in Section 2). Note that we calculated the macro-average for recall on a balanced dataset, meaning that recall and accuracy are the same. It is not possible to accurately compare our outcomes with Wolz *et al.* [37] and McDowell *et al.* [21], as their ground truth is provided by an actigraph unit while ours is based on PSG. Only Uçar *et al.* [36] provide a 10-fold CV.

However, we also repeat this analysis six times and on a bigger dataset to provide results that more accurately represent the general population. Our model is trained on the highest number of individuals and with the highest level of diversity of any study done before on this topic. Overall, we can reach the highest Cohen's Kappa (κ) of 0.69 using our fine-grained approach on gender and health statuses by combining two data sources.

Evidence shows that camera recordings lead to the highest levels of accuracy [18] which comes with certain privacy-preservation issues. The studies, conducted by Kuo *et al.* [17] and Khademi *et al.* [15], are very similar in accuracy, though Khademi *et al.* [15] reach a recall of 38%, which is considered low. Our accuracy is lower than that of Kuo *et al.* [17] but we use a balanced dataset in terms of gender and sleep-wake stages with more participants. We also use processed activity counts and light levels instead of raw accelerometer data. Additionally, our approach reaches higher κ values than Parro and Valdo [27] despite ours including both males and females while they only included males.

Table 3: Comparison with other methods

	No	Health Status	Sensor	F*	Acc*	Rec*	κ	M*
[36]	10	Apnea	PPG	28	77.4	79.0	0.59	kNN
[18]	10	-	Cam.	-	92.1	-	-	-
[17]	81	poor/good SE	Acti.	2	89.7	92.9	0.62	-
[15]	54	diverse	Acti.	39	87.0	38.0	-	XGB
[27]	43	male	Acti.	-	85.3	65.4	0.53	RQA
Ours	1641	diverse	ECG	18	79.1	79.1	0.58	MLP
	1576	diverse	Acti.	7	79.0	79.0	0.58	MLP
	1576	diverse	Com.	25	84.6	84.6	0.69	MLP

*F-Features; Acc-Accuracy; Rec-Recall; M-Method

5.5 Personalized Sleep Parameter Extraction

In Table 4, the original SOL calculation using the 15 minutes rule on the original sleep-wake stages is compared to our proposed personalized variable-threshold technique fused with the proposed sleep-wake approach. It can be seen that the mean difference is lower for our sleep-wake analysis approach with 20.16 compared to 20.74 for the original sleep-wake stages. The personalization of the thresholds is relevant, we explored various threshold values between 2 and 50. The personalization of gender, health status, and race produces a better correlation with PSG with 17.78 mean difference compared to 20.74 in the original. Furthermore, we investigate the Pearson ad Spearman correlation with PSG and reach a lower correlation for the Spearman correlation, but overall better mean differences, which we use as the main indicator. For different personal information, thresholds need to be differently

Table 4: Variable threshold performance compared to original 15-minute rule for SOL

Approach	Threshold	Correlation		MD
		Pearson	Spearman	
Original	15	0.68	0.73	20.74
Our	25	0.63	0.57	20.16
Original M*	15	0.63	0.73	26.50
Our M*	15	0.63	0.62	24.08
Personalized Thresholds				
Gender-Specific				
Female	25	0.60	0.50	20.29
Male	25	0.72	0.69	19.94
			Mean	20.16
Health Status-Specific				
Healthy	25	0.67	0.60	19.98
Apnea	20	0.61	0.56	14.20
Insomnia	20	0.62	0.50	27.43
RLS	25	0.65	0.64	18.09
			Mean	19.78
	HS	F/M	Mean	19.67
Race-Specific				
White	25	0.60	0.54	17.84
Asian	20	0.77	0.62	17.28
African	20	0.66	0.58	22.87
Hispanic	25	0.58	0.60	23.56
			Mean	20.05
	Race	F*/M*	Mean	19.69
HR	Race	F*/M*	Mean	17.78

*MD-mean difference; F-Female; M-Male

set to reach the best match with PSG data. A comparison between the original 15-minute rule and our variable personalised threshold (variable-minute approach) is depicted in Fig. 14. For male participants (shown in green), improvement is visible. Overall, our approach tends to underestimate SOL while the original tends to overestimate it.

6 Discussion

6.1 Comparison of Different Approaches

The difference between the original and down-sampled case shows that sleep, the majority class which is down-sampled, is more often matching with the

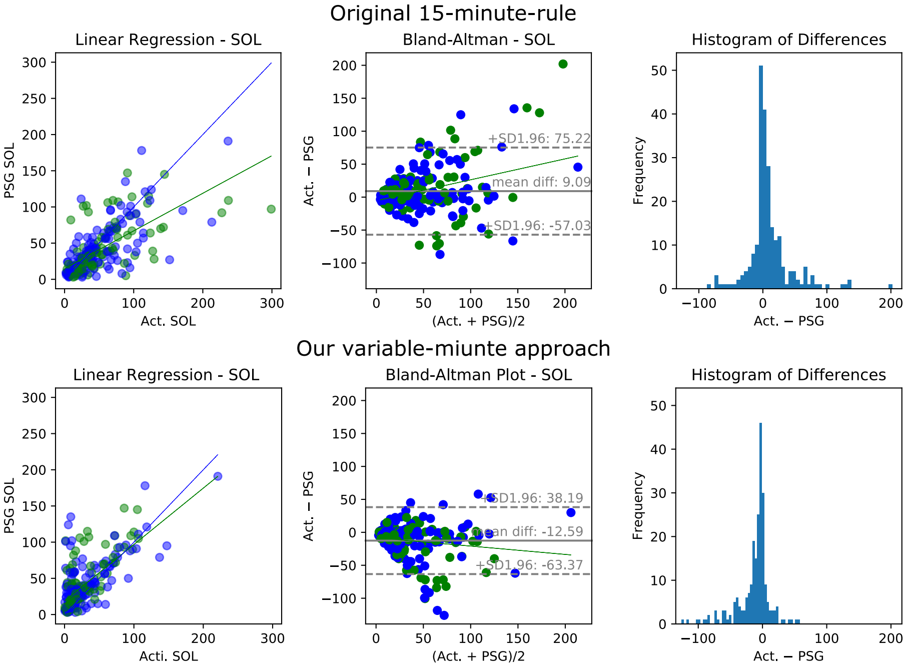

Fig. 14: Comparing SOL from original sleep-wake algorithm with 15-minute-rule and our fine-grained model with variable-minute approach with PSG. Linear regression, Bland-Altman Plot and Histogram of Differences is presented. Green indicates male participants, while white indicates females.

sleep stages in the ground truth (PSG) than the wake stages. This shows the better fit between the majority class and the ground truth from PSG.

The lower outcome for 'Avg. Health Status' shows the influence of the training data on the recognition rate and the importance of fine-grained approaches, as we can increase the performance by personalizing the model in terms of gender and race.

The findings show that fine-grained models have an advantage over current one-model-fits-all approaches. Especially, subjects with a medical condition show an improvement in performance, e.g., 'Avg. Gender and Race' (79.05%) and 'All' (77.01%). This is a positive result as disease-affected individuals are usually the target group for sleep behavior analysis. This effect comes from the fact that gender, race, and health status are influence factors for sleep.

Another finding is the importance of how to include the information during the training process. As we tested the inclusion of gender as a feature 'Inc. Gender' and in the personalized approach in the sub-categorization. The model can train the differences better if it is individually trained compared to the knowledge inclusion as a feature.

6.2 Investigating Different Data Sources

The differences in SD, between the general approach and the personalized gender-specific case, shows the stability and reliability of a granular perspective in sleep-wake recognition, as the SD is lower for repeated experiments.

The findings that SD for the HRV data, i.e., imbalanced is lower than in the balanced case, suggest the effect that the biased case learns the majority class better, but also that patterns between individuals and time in sleep are present. This can be explained by HRV changes between different sleep stages; therefore, it is also used to investigate higher granular stages; while actigraphy is based on movement being more consistent over time; therefore, this effect is not as obviously present.

The finding that male participants perform usually better than females suggests that male participants and their movement are more consistent over the population, while females have more individual components, suggesting individual trained models have an advantage.

Subject size affects the performance and needs to be taken into account when setting the experimental outcomes into context.

The result that the SD is lower for healthy individuals and all health statuses compared to disease-specific models, can either be caused by the influence of the subject numbers or the classification task at hand.

The average SD for HRV is in general higher than for actigraphy, and the combined case. These findings also show that actigraphy has a positive influence on the SD in the combined case.

There is a difference in which data source is used and which is ideal for which case. For actigraph data, outcomes are less diverted over all fine-grained models, while HRV data usage shows a higher range from 74.2% to 86.2%. There is also a difference between certain subject groups which shows difficulties for actigraph while it shows good performance with HRV, e.g., Hispanic group. The outcomes show that it is relevant which information, i.e., data source, is used to distinguish sleep and wake stages.

We found that HRV performs better for disease participants compared to healthy. This suggests that the HRV differences between sleep and wake stages are more significant in disease-affected cases. This trend is consistent when the data sources are combined. On the other side, actigraphy shows more differences between male and female models, and the disease-specific models perform slightly worse than the healthy cases.

Overall, the combination of both data sources performs best. Presumably, actigraphy pushes the recognition rate in the healthy case while HRV pushes it in the disease-affected cases. Most likely the actigraphy and HRV data are independent of each other being able to capture different behavior, therefore a combination of the data sources features is a promising approach. This is an important finding which has to our knowledge not been investigated before in this context.

6.3 Exploring Healthy and Disease-Affected Performance

In the explored experiments male and female groups are showing differences in performance for actigraphy data, which relates to the argument of consistent body movement over the male population mentioned before. For some disease groups sleep and wake episodes are more difficult to separate for females, e.g., the Hispanic group. This can be based on the smaller subject set (157) than in the male case (181), compare Table 2 or more likely based on existing underlying differences in sleep-wake patterns based on gender. This again makes the personalized fine-grained models an important solution for more reliable sleep-wake stage recognition.

Our investigation showed that individual trained models fine-grained on disease and health status are preferable. The best choice is to follow the proposed fine-grained sleep-wake behavior analysis approach, apart from this a disease model is preferred in HRV, and use case dependent models for actigraphy and the combined case.

6.4 Comparison to Existing Methods

Our investigation is the only one provided in Table 3 which considers a balanced dataset for training (see figures 6 and 7); therefore, there is no majority class and subsequently, no bias is introduced. We also investigate, in a study on this subject, the combination of multiple data sources and the relevance of clinical history and personal information. As we saw in our investigations, multi-data sources are highly relevant, as influence factors affect sleep-stage pattern detection and different data sources have individual advantages which can be fused. We incorporate the highest number of individuals and the highest level of diversity while using repeated CV, which provides an accurate representation of the investigated population. We use sensors that are easy to apply at home and do not interfere much with the privacy of individuals; video recordings, used in studies like Liao and Yang [18], continue to raise privacy concerns. Characteristics of our validation data—fewer features and higher accuracy—make our HRV model a better choice than other methods, such as in [36]. Overall, we reach the highest Cohen's kappa value, 0.69, by using our fine-grained approach on gender and health statuses and combining two data sources. Many approaches test against just one technician, making that technician's style too influential [19]. We avoid the testing against just one technician's scoring by using a dataset scored by multiple technicians. The problem with one technician is, that they have individual styles and machine-learning algorithms learn them. This issue can be reduced by including scoring by multiple technicians and, therefore, potentially reduce variation in the trained models. A direct method would be to train one model per technician and combine the answers of the different models. However, the dataset provides an averaged scoring and, therefore, indirectly deals with the issue. In addition, we investigate the combination of actigraphy and HRV as a means to enhance model accuracy, which was achieved.

Compared to other approaches, we simply use five features, one from actigraphy and four from light levels, and reach the highest κ result for a balanced dataset, but we do not reach the same accuracy rates. Usually, the higher accuracy rates were related to lower recall values, which is a drawback which we can overcome by more stable results. This gives us the impression that the larger and more diverse dataset used in our study influences the performance and, in turn, better represents the general population. Overall, a balanced dataset represents a realistic recognition rate but leads to a decrease in performance and higher variability in SD, such as in the original sleep-wake recognition model (see figures 6 and 7). However, our multi-source data learning approach for granular sleep-wake behavior detection adapts to these effects and provides comparable accuracy, improving when multiple data sources are available.

6.5 Personalized Sleep Parameter Extraction

The personalized sleep parameter extraction shows a better mean difference to the golden standard compared with the state-of-the-art method with static thresholds. It indicates that ideal thresholds need to be adapted for different personal factors. For example, sleep apnea and healthy individuals require different thresholds to represent the ground truth. Furthermore, the results show that our multi-source data learning approach for sleep-wake stages applied in a real-world application brings improvement in sleep parameter extraction. The variable threshold technique shows advantages in the Bland-Altman plot with a better fit but tends to underestimate the SOL. Especially, SOL from male participants can be improved by our method which is likely caused by the more consistent behavior of the sleep-wake patterns for the males, as also indicated in the sleep-wake recognition investigation. Overall, we can see that a personalized approach is necessary but still needs further analysis to fully investigate the best thresholds levels.

6.6 Limitations

Limitations in our experiments are (1) the need of available medical history and genetic data, such as gender, (2) downsampling, (3) not considering comorbidity and (4) number of subjects vary and (5) HRV is calculated from ECG. Firstly, information from medical history and gender needs to be available for the basis of the granular adaptive approach. Secondly, when downsampling is performed, likely not all the data of the majority class are used even if repeated. To reach a balance in sleep and wake stages the majority class needs to be reduced. We did choose them randomly within ever repetition, and did not cover for already-used or not-used periods. Third, comorbidity is not considered in our approach; it may have an influence on the recognition rate that we are unaware of. Fourthly, as mentioned throughout the paper the number of subjects vary within the groups and differences in subject count can influence the performance. Fifthly, HRV features are extracted from ECG which has more contact points with the human body and therefore is considered less comfortable to monitor than with PPG. Monitoring PPG with pulse rate variability is a promising alternative to ECG for HRV feature extraction with good coherence [28].

6.7 Implications of Findings for Future Studies

We provided evidence that personalized models are an important direction for more reliable sleep-wake stage recognition and sleep parameter extraction.

These findings can influence future research on higher granular sleep stage investigation, as the incorporation of this knowledge can lead to new insights. This means the personalized approach has shown successful for sleep-wake stages and has the potential to show the same trends when sleep is divided by the guideline specific stages, e.g., N1, N2, N3, and REM. A potential research area is to investigate multiple two-class classification problems that would ultimately be fused together into an overall model. One would distinguish between REM from NREM sleep, one would distinguish between light (N1, N2) and deep sleep (N3); and one would distinguish between N1 and N2. This method includes multiple personalized models for pairs of sleep stages, each optimised for an individual problem.

Furthermore, the incorporation of multiple data sources suggests a promising direction for boosting recognition rates. Future research could look at new combinations of already-investigated data and potentially open up new avenues of investigation for sleep-wake classification.

The variable-threshold technique provides the possibility for personalized sleep health monitoring in a home environment which is reliable with less variation to the ground truth.

7 Conclusion

This paper introduces a multi-source data learning approach for granular adaptive sleep-wake pattern analysis adaptable to gender, race, and health-status information and accessible data sources. It describes in detail an AI-enabled sleep parameter extraction technique and the learning algorithms incorporating multi-source data features into the MLP model. The developed approach and methods are validated using HRV and actigraph data. The approach improves recognition rates for in-home usage, based on specification-dependent recognition. Furthermore, the combination of HRV and actigraph features can improve the recognition rate considerably, to match better with the gold standard of PSG. Using computational methods to assess and interpret available sleep behavior data is key for the development of predictive and personalized healthcare methods. As more individuals than ever are affected by sleep problems, demand is rapidly increasing for home-based sleep assessment that delivers adequate, reliable, and personalized sleep evaluation.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676157.

References

1. American Academy of Sleep Medicine: The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. American Academy of Sleep Med. (2007)
2. Bhatia, N., Vandana: Survey of nearest neighbor techniques. ArXiv (2010). URL `https://arxiv.org/abs/1007.0085`
3. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., JacobsJr., D.R., Kronmal, R., Liu, K., Nelson, J.C., O'Leary, D., Saad, M.F., Shea, S., Szklo, M., Tracy, R.P.: Multi-Ethnic Study of Atherosclerosis: Objectives and Design. Am. J. Epidemiol. **156**(9), 871–881 (2002). DOI https://doi.org/10.1093/aje/kwf113
4. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001). DOI https://doi.org/10.1023/a:1010933404324
5. Chung, K.Y., Song, K., Shin, K., Sohn, J., Cho, S.H., Chang, J.H.: Noncontact sleep study by multi-modal sensor fusion. Sensors **17**(7), 1–17 (2017). DOI https://doi.org/10.3390/s17071685
6. Crivello, A., Barsocchi, P., Girolami, M., Palumbo, F.: The meaning of sleep quality: A survey of available technologies. IEEE Access **7**, 167374–167390 (2019). DOI https://doi.org/10.1109/ACCESS.2019.2953835
7. Dafna, E., Tarasiuk, A., Zigel, Y.: Sleep staging using nocturnal sound analysis. Sci. Rep. **8** (2018). DOI https://doi.org/10.1038/s41598-018-31748-0
8. Fallmann, S., Chen, L.: Detecting chronic diseases from sleep-wake behaviour and clinical features. In: Proc. IEEE 5th Int. Conf. Syst. and Inform., pp. 1076–1084 (2018). DOI https://doi.org/10.1109/ICSAI.2018.8599388
9. Fallmann, S., Chen, L.: Computational sleep behavior analysis: A survey. IEEE Access **7**, 142421–142440 (2019). DOI https://doi.org/10.1109/ACCESS.2019.2944801
10. Fallmann, S., Chen, L., Chen, F.: Fine-grained sleep-wake behaviour analysis. In: Proc. IEEE SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI, pp. 667–674 (2019). DOI https://doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00150
11. Figueiro, M., Sahin, L., Roohan, C., Kalsher, M., Plitnick, B., Rea, M.: Effects of red light on sleep inertia. Nature and Science of Sleep **Volume 11**, 45–57 (2019). DOI https://doi.org/10.2147/NSS.S195563
12. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco, CA (2011)
13. Ibàñez, V., Silva, J., Cauli, O.: A survey on sleep questionnaires and diaries. Sleep Medicine **42**, 90–96 (2018). DOI https://doi.org/10.1016/j.sleep.2017.08.026
14. Johnson, D., Jackson, C., Williams, N., Alcántara, C.: Are sleep patterns influenced by race/ethnicity – a marker of relative advantage or disadvantage? evidence to date. Nature and Science of Sleep **Volume 11**, 79–95 (2019). DOI https://doi.org/10.2147/NSS.S169312
15. Khademi, A., El-Manzalawy, Y., Buxton, O.M., Honavar, V.: Toward personalized sleep-wake prediction from actigraphy. In: Proc. IEEE EMBS Int. Conf. Biomed. Health Inform., pp. 414–417 (2018). DOI https://doi.org/10.1109/BHI.2018.8333456

16. Koushik, A., Amores, J., Maes, P.: Real-time sleep staging using deep learning on a smartphone for a wearable EEG. NIPS ML4H (2018). URL `http://arxiv.org/abs/1811.10111`
17. Kuo, C.E., Liu, Y.C., Chang, D.W., Young, C.P., Shaw, F.Z., Liang, S.F.: Development and Evaluation of a Wearable Device for Sleep Quality Assessment. IEEE Trans. Biomed. Eng. **64**(7), 1547–1557 (2017). DOI https://doi.org/10.1109/TBME.2016.2612938
18. Liao, W.H., Yang, C.M.: Video-based activity and movement pattern analysis in overnight sleep studies. In: Proc. 19th Int. Conf. Pattern Recognition, pp. 1–4 (2008). DOI https://doi.org/10.1109/ICPR.2008.4761635
19. Malafeev, A., Laptev, D., Bauer, S., Omlin, X., Wierzbicka, A., Wichniak, A., Jernajczyk, W., Riener, R., Buhmann, J., Achermann, P.: Automatic human sleep stage scoring using deep neural networks. Frontiers in Neuroscience **12**(November), 1–15 (2018). DOI https://doi.org/10.3389/fnins.2018.00781
20. Marcos, J.V., Hornero, R., Álvarez, D., del Campo, F., Zamarrón, C., López, M.: Utility of multilayer perceptron neural network classifiers in the diagnosis of the obstructive sleep apnoea syndrome from nocturnal oximetry. Computer Methods and Programs in Biomedicine **92**(1), 79 – 89 (2008). DOI https://doi.org/10.1016/j.cmpb.2008.05.006
21. McDowell, A., Donnelly, M.P., Nugent, C.D., Galway, L., McGrath, M.J.: Addressing the challenges of sleep/wake class imbalance in bed based non-contact actigraphic recordings of sleep. In: Conf. Proc. IEEE Eng. Med. Biol. Soc., pp. 4654–4657 (2013). DOI https://doi.org/10.1109/EMBC.2013.6610585
22. Meltzer, L.J., Walsh, C.M., Peightal, A.A.: Comparison of actigraphy immobility rules with polysomnographic sleep onset latency in children and adolescents. Sleep and Breathing **19**, 1415–1423 (2015). DOI https://doi.org/10.1007/s11325-015-1138-6
23. Meltzer, L.J., Westin, A.M.: A comparison of actigraphy scoring rules used in pediatric research. Sleep Medicine **12**(8), 793 – 796 (2011). DOI https://doi.org/10.1016/j.sleep.2011.03.011
24. Middelkoop, H., A Smilde-van den Doel, D., Neven, A., A. C. Kamphuisen, H., P Springer, C.: Subjective sleep characteristics of 1,485 males and females aged 50-93: Effects of sex and age, and factors related to self-evaluated quality of sleep. The Journals of Gerontology. Series A, Biological sciences and medical sciences **51**, 108–15 (1996). DOI https://doi.org/10.1093/gerona/51a.3.m108
25. National Sleep Research Resource: Multi-Ethnic Study of Atherosclerosis, HRV Analysis Overview. `https://sleepdata.org/datasets/mesa/pages/hrv-analysis.md`. Accessed: 2019-03-11
26. Newell, J., Mairesse, O., Verbanck, P., Neu, D.: Is a one-night stay in the lab really enough to conclude? First-night effect and night-to-night variability in polysomnographic recordings among different clinical population samples. Psychiatry Res. **200**(2), 795–801 (2012). DOI https://doi.org/10.1016/j.psychres.2012.07.045
27. Parro, V., Valdo, L.: Sleep-wake detection using recurrence quantification analysis. Chaos **28**(8), 085706 (2018). DOI https://doi.org/10.1063/1.5024692
28. Pinheiro, N., Couceiro, R., Henriques, J., Muehlsteff, J., Quintal, I., Gonçalves, L., Carvalho, P.: Can PPG be used for HRV analysis? In: Proc. 38th Ann. Int. Conf. IEEE Eng. Med. Biol. Soc., pp. 2945–2949 (2016). DOI https://doi.org/10.1109/EMBC.2016.7591347
29. Rao, S., Ali, A.E., Cesar, P.: Deepsleep: A ballistocardiographic deep learning approach for classifying sleep stages. In: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, UbiComp/ISWC '19 Adjunct, p. 187–190. Association for Computing Machinery, New York, NY, USA (2019). DOI https://doi.org/10.1145/3341162.3343758
30. Rechtschaffen, A., Kales, A.: A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. Nat. Inst. of Health Publication **204**, 976–977 (1968)
31. Renevey, P., Delgado-Gonzalo, R., Lemkaddem, A., Proença, M., Lemay, M., Solà, J., Tarniceriu, A., Bertschi, M.: Optical wrist-worn device for sleep monitoring. In: Proc. EMBEC & NBC, pp. 615–618 (2017). DOI https://doi.org/10.1007/978-981-10-5122-7_154

32. Rusterholz, T., Tarokh, L., Van Dongen, H.P.A., Achermann, P.: Interindividual differences in the dynamics of the homeostatic process are trait-like and distinct for sleep versus wakefulness. J. Sleep Res. **26**(2), 171–178 (2017). DOI https://doi.org/10.1111/jsr.12483
33. Shim, J., Kang, S.W.: Behavioral factors related to sleep quality and duration in adults. Journal of lifestyle medicine **7**(1), 18–26 (2017). DOI https://doi.org/10.15280/jlm.2017.7.1.18
34. Tao, S., Cui, L., Zhang, G.Q., Mobley, D., Kim, M., Rueschman, M., Mueller, R., Mariani, S., Redline, S.: The National Sleep Research Resource: towards a sleep data commons. J. Am. Med. Inform. Assoc. **25**(10), 1351–1358 (2018). DOI https://doi.org/10.1093/jamia/ocy064
35. Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology : Heart rate variability. Circulation **93**(5), 1043–1065 (1996). DOI https://doi.org/10.1161/01.CIR.93.5.1043
36. Uçar, M.K., Bozkurt, M.R., Bilgin, C., Polat, K.: Automatic sleep staging in obstructive sleep apnea patients using photoplethysmography, heart rate variability signal and machine learning techniques. Neural Comput. Appl. **29**(8), 1–16 (2018). DOI https://doi.org/10.1007/s00521-016-2365-x
37. Wolz, R., Munro, J., Guerrero, R., Hill, D.L., Dauvilliers, Y.: Predicting Sleep/Wake Patterns From 3-Axis Accelerometry Using Deep Learning. Alzheimers Dement. **13**(7), 1012 (2017). DOI https://doi.org/10.1016/j.jalz.2017.06.1412
38. Zhang, X., Kou, W., Chang, E.I.C., Gao, H., Fan, Y., Xu, Y.: Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device. Comput. Biol. Med. **103**, 71–81 (2018). DOI https://doi.org/10.1016/j.compbiomed.2018.10.010
39. Zhao, M., Yue, S., Katabi, D., Jaakkola, T.S., Bianchi, M.T.: Learning sleep stages from radio signals: A conditional adversarial architecture. In: Proc. 34th Int. Conf. Mach. Learning, vol. 70, pp. 4100–4109 (2017)