Development and Validation of Algorithms for Sleep Stage Classification and Sleep Apnea/Hypopnea Event Detection Using a Medical-Grade Wearable Physiological Monitoring System

Zhao Wang[1], Zhicheng Yang[2], Ke Lan[3], Peiyao Li[4], Yanli Hao[3], Ying Duan[5], Yingjia She[6], Yuzhu Li[6(✉)], and Zhengbo Zhang[7(✉)]

[1] Medical School of Chinese PLA, Beijing, China
[2] PAII Inc., Palo Alto, Santa Clara, CA, USA
[3] Beijing SensEcho Science and Technology Co., Ltd., Beijing, China
[4] Department of Computer Science and Technology, Tsinghua University, Beijing, China
[5] Sleep Medicine Division, Airforce Medical Center, Beijing, China
[6] Department of Respiratory Medicine, Chinese PLA General Hospital, Beijing, China
lyz301@163.com
[7] Center for Artificial Intelligence in Medicine, Chinese PLA General Hospital, Beijing, China
zhengbozhang@126.com

Abstract. Sleep is critical to the overall health of humans. Polysomnography (PSG) is the current gold standard for measuring sleep and diagnosing sleep-related breathing disorders. However, this method is labor-intensive, time-consuming, and confined to a sleep laboratory. In this paper, we leverage algorithms for sleep stage classification and sleep apnea/hypopnea event detection by using signals from single-lead electrocardiograph (ECG) and respiration. To validate the accuracy of the above two algorithms, two independent validation studies were conducted using a medical-grade wearable monitoring system to collect physiological data from patients in both clinical and home settings. In the validation study of sleep stage classification, the average accuracy of our four-class stage classification using the bi-directional long short-term memory (BLSTM) method is 77.83% on our in-house dataset of 30 enrolled patients. In the experiments of sleep apnea screening, the two-level apnea-hypopnea index (AHI) classification reports the overall accuracies of 96.67% and 91.43% in clinical and home environments,

Zhao Wang and Zhicheng Yang equally contributed to this work.
This work was done during Zhicheng Yang's internship at Beijing SensEcho Science & Technology Co., Ltd., Beijing, China, when he was a Ph.D. candidate at University of California, Davis, CA, USA.

X. Gao et al. (Eds.): MobiHealth 2021, LNICST 440, pp. 166–185, 2022.
https://doi.org/10.1007/978-3-031-06368-8_12

respectively. The results showed that the sleep analysis algorithms presented in this paper have good performance in both sleep stage classification and sleep event detection, either in clinical scenario and home settings, indicating that our device can be used along with the two algorithms for sleep analysis.

Keywords: Sleep stage classification · Sleep apnea/hypopnea event · Apnea-hypopnea index · Physiological monitoring · Wearable system · Polysomnography

1 Introduction

Sleep is critical to one's mood, cognition, physiological internal environmental balance and resilience [5,9]. To appropriately present one's sleep condition, sleep is commonly classified into multiple stages in which physiological signals have different patterns that indicate various physiological functions. According to the American Association of Sleep Medicine (AASM), sleep is divided into five stages: wake, rapid eye movement (REM) and three levels of non-rapid eye movement sleep (including N1, N2, N3) [1]. Among the numerous criteria for sleep stage classification, the four-class sleep stage criterion (W-N1/N2-N3-REM) is more commonly adopted for its adequacy in sleep architecture assessment than two-class (W-N1/N2/N3/REM) and three-class (W-N1/N2/N3-REM) classification [14,23,42]. In recent years, the prevalence of sleep disorders has been globally increasing and attracted more attention [18]. Compared with the other types of sleep disorders, sleep apnea/hypopnea is a potentially serious one that is likely to lead to sudden and severe medical conditions [4,39]. Accurate classification of sleep stages and detection of sleep apnea/hypopnea events are essential to the analysis of sleep architecture and identification of various sleep-related disorders.

For a typical sleep analysis, polysomnography (PSG) test is the gold standard, which involves the electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), electrocardiogram (ECG), respiratory effort signals, and other measurements. However, subjects have to wear manifold attachments which cause extra mental and physical burdens during the test, and this procedure is also time-consuming as well as labor-intensive for clinical specialists [27,29]. To overcome the above drawbacks, researchers are searching for methods to automatically analyze sleep based on cardiopulmonary physiological signals that can be relatively easily acquired by low-cost wearable devices. Studies have shown that ECG and respiratory signals can be used for sleep staging and apnea/hypopnea event detection. Pedro Fonseca et al. extracted 142 features from ECG and thoracic respiratory signal of 25 subjects, and applied a linear discriminant classifier with an accuracy of 69% for four-class sleep staging on a 30-second epoch basis [8]. Magnusdottir et al. used two algorithms (cardiopulmonary coupling and heart rate cyclic variation) to identify sleep apnea based on automated analysis of single-lead ECG data. The results showed that

the algorithms were as accurate as the automated scoring software for identifying patients with moderate to severe sleep apnea [15]. Recently, deep learning methods have become a center of attention in sleep analysis due to their advantages in handling time series over traditional machine learning methods. A long short-term memory (LSTM) network was applied to find out about sleep architecture by extracting 132 hand-engineering HRV features based on a comprehensive dataset (SIESTA [11]), and a five-fold cross-validation was performed on this dataset with satisfactory accuracy ($77.00 \pm 8.90\%$) [23]. Iwasaki A et al. adopted LSTM to distinguish patients with moderate-to-severe sleep apnea syndromes and from healthy subjects and achieved a sensitivity of 100% and specificity of 100% based on a threshold of apnea-hypopnea index (AHI) $\geq$ 15 events/h [10]. Despite the large number of sleep stage classification and event detection algorithms that have been developed, most of them are based on internal validation of the dataset itself, and the performance of these algorithms needs to be further validated with data collected from real clinical settings.

There have been some researches on validating commercial wearable devices in clinical scenarios. Pigeon et al. recruited 27 healthy adult subjects to validate their wrist-worn sleep monitoring actigraphy for sleep-wake scores with another commercially available actigraphy during a one-night PSG test [19]. GDL Pinheiro et al. proposed the validation of a wireless wearable oximeter on 58 patients for the diagnosis of obstructive sleep apnea (OSA) compared to the PSG test [20]. However, those wrist-based wearable devices are considered somewhat unreliable for clinical purposes. Xu et al. enrolled 80 subjects to validate the performance of a portable monitor (Nox-T3) to diagnose OSA in both laboratory and home settings [38]. Pion-Massicotte et al. validated the three-class (W-N1/N2/N3-REM) sleep stage classification of biometric shirts based on 2 nights of sleep lab recordings from 20 healthy adults, with a mean agreement rate of 77.4% when compared with PSG [21]. These validation studies are necessary if algorithms or systems are to achieve robust sleep monitoring for medical purposes.

In this paper, we developed sleep analysis algorithms focusing on four-class sleep stage classification and apnea/hypopnea event detection. To validate the accuracy of the algorithms, two independent validation studies were conducted using a medical-grade wearable physiological monitoring system to collect physiological data from patients in both clinical and home settings. Our key contributions were summarized as follows:

1. For the sleep stage classification algorithm, 152 features were extracted from ECG and respiratory signals, in which three new features were proposed to effectively detect abrupt changes in the RR intervals. Then, the bi-directional long short-term memory (BLSTM) network was leveraged for four-class sleep stage classification, because the bi-directional architecture of BLSTM was able to learn from past and future information. This advantage was suitable for our proposed sleep stage classification task in which the valuable "context" of sleep stage can be leveraged by BLSTM well. As for the sleep event detection algorithm, a new method was established to automatically detect sleep apnea and hypopnea by thoracic and abdominal respiratory movements.

2. Unlike many previous studies, sleep analysis algorithms proposed in this study were first trained and validated on a large public dataset before external validation was performed with data collected both from clinical (in a sleep laboratory) and home settings.
3. Compared with existing literature based on a single type of sleep analysis algorithm only, our study included both types of sleep analytics, empowering us to achieve a holistic and robust analysis.
4. We adopted a medical-grade wearable physiological monitoring system *SensEcho* [13,35–37,40,41], to collect physiological data. Both the performance of the algorithm and the system was validated in real clinical and home settings. The preliminary study showed that the system equipped with the algorithm can be used for sleep measurement and analysis.

2 Material and Methods

2.1 Algorithm Design

Sleep Stage Classification Algorithm. We used a public sleep database, Sleep Heart Health Study (SHHS) for model training [22], which consisted of the PSG monitoring of 6,600 patients in the U.S., including the records of sleep stage classification (Wake, N1, N2, N3, N4 or REM) for each subject manually determined by clinical specialists using modified Rechtschaffen & Kales (R&K) criteria [6]. N3 and N4 were combined into a single N3 label to align with AASM. First-time admissions patients (SHHS1) were screened with high-quality ECG and respiratory signals. Finally, 4887 subjects were selected to construct the dataset from the SHHS1. To align our four-class sleep stage classification, we converted sleep stage data to four-class (Wake-Light sleep(N1/N2)-Deep sleep(N3/N4)-REM).

Our feature extraction was processed on either one 30-second epoch or a larger window consisting of several consecutive epochs. The moving step size was set at one epoch. When multiple consecutive epochs (a larger sliding window) were used, the number of epochs was odd so that the calculated features could be associated with the central epoch. A total of 152 features were extracted from RR intervals (including time-domain features (features commonly used for the HRV analysis, and conventional statistical features on RR intervals, such as the mean, and quantiles, we also extracted 6 non-linear features including sample entropy, zero-crossing analysis.) [8,24,34], frequency domain features(21 features of the frequency domain were extracted, such as the mean, spectrum power, entropy.) [34]), respiratory signals (We extracted 25 statistical features. For example, in the time domain, the mean and standard deviance of respiratory peak sequence, kurtosis and skewness were extracted; features in the frequency domain included the dominant peak and energy.) [8,24], and cardiopulmonary coupling (CPC) effect [33]. We also designed three novel features for RR intervals as follows:

$$f_1 = \overline{I_n^{\text{mid}}} - \overline{I_n} \tag{1}$$

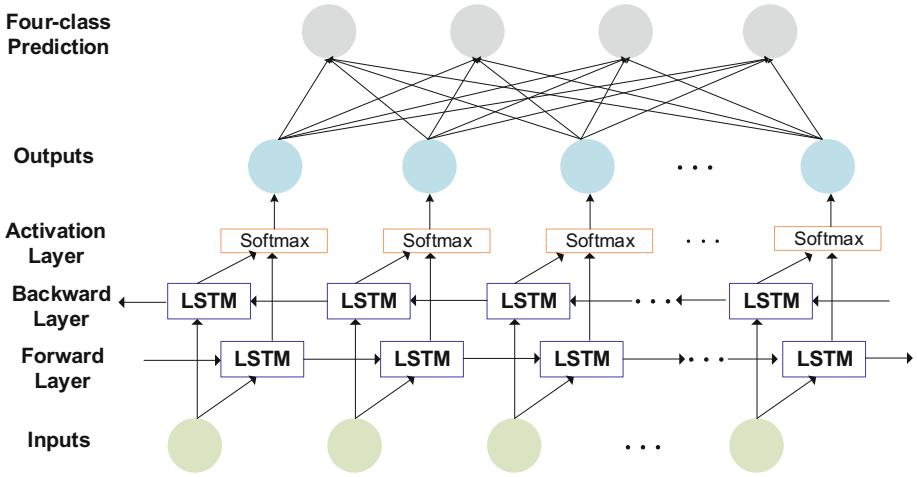


Fig. 1. Architecture of the BLSTM.

$$f_2 = \overline{I_n^{\text{mid}}} - \widetilde{I_n} \tag{2}$$

$$f_3 = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(I_n^k - \overline{I_n}\right)^2} \tag{3}$$

where I_n referred to raw RR intervals in the consecutive n epochs; I_n^k denoted the k-th epoch of I_n; I_n^{mid} was the middle epoch of I_n; $\overline{I_n}$ represented the average value of I_n; $\widetilde{I_n}$ denoted the median value of I_n. The three features investigated the impact of sudden variation of RR intervals in one epoch over the longtime series.

Unlike other conventional machine learning methods, the effectiveness of the LSTM approach has been proved in many studies for time series problems, because LSTM is able to effectively learn latent patterns from previous related information in a time series. Furthermore, the BLSTM is able to learn from past and future information. This advantage can facilitate our proposed sleep stage classification in which the valuable "context" of the sleep stage could be leveraged by BLSTM well. The structure of the BLSTM was illustrated in Fig. 1.

We randomly split the dataset into a training set and a validation set at a ratio of 4:1. No subject from the training set appeared in the validation set. All the feature vectors were fed into two 16-unit BLSTM layers and one fully connected layer with 4-unit outputs corresponding to the four sleep stage classes. The categorical cross-entropy loss was used as a loss function. In the process of training, the base learning rate was set at 1×10^{-3}, and the overall amount of training epochs was 5,000. The learning rate decay factor was 1×10^{-2}, and

the minimum learning rate was 1×10^{-5}. The Adam optimizer was used. The networks were trained with a batch size of 256, and the dropout was 0.2. Five-fold cross-validation was performed to test our model in the training phase with an average accuracy of 78.84% (±0.08) and a kappa coefficient of 0.67 (±0.13). The optimal model was selected to further validate the performance on our in-house dataset.

Sleep Apnea/Hypopnea Event Detection Algorithm. A sleep event detection algorithm was used to determine sleep apnea/hypopnea through thoracic and abdominal respiratory movements. In the signal pre-processing step, we apply a median filter to the collected respiratory signals to remove outliers, remove the wavelet variation from the baseline of the respiratory signals, and adopt a band-pass filter with the frequency of 0.1∼0.5 Hz to remove noise. The denoised thoracic and abdominal respiratory signals are used to calculate relative tidal volume. Based on the detected peaks and troughs of relative tidal volume, we calculate the amplitude of respiration series and eliminate pseudo-peaks (the amplitude of respiration ≤ 0.15). The processed amplitude of respiration series is used to generate our key value *baseline respiratory amplitude*, which was defined as the median of the second to fourth highest respiratory amplitude within the first two minutes. The respective thresholds for apnea θ_a and hypopnea θ_h were then defined below:

$$\theta_a = A_{\text{base}} \times \alpha_a \tag{4}$$

$$\theta_h = A_{\text{base}} \times \alpha_h \tag{5}$$

where A_{base} denotes the baseline respiratory amplitude, α_a and α_h represent the scale factors for apnea and hypopnea thresholds, respectively. In our cases, we set α_a as 0.35 and α_h as 0.70. The following criteria were used to determine sleep apnea/hypopnea events:

- Central sleep apnea: (i) The interval between the two peaks exceeds 10 s and there is no respiratory movement in either thoracic or abdominal; (ii) The amplitude of respiration less than θ_a for more than 10 s with no peaks detected, or the interval between the two peaks is more than 10 s.
- Obstructive sleep apnea: (i) The interval between the two peaks exceeds 10 s and at least one of the thoracic and abdominal respiratory movements shows respiratory effort; (ii) The amplitude of respiration less than θ_a for more than 10 s and more than 6 peaks are detected, or no more than 6 peaks are detected, and the interval between peaks is not more than 10 s.
- Mixed sleep apnea: If the above two criteria are satisfied at the same time, and the central sleep apnea appears before the obstructive sleep apnea.
- Hypopnea: The amplitude of respiration less than θ_h but greater than θ_a signal for more than 10 s, accompanied by a decrease in oxygen saturation of more than 4%.

Fig. 2. Hardware and signal acquisition of SensEcho.

To evaluate the algorithm performance, 600 subjects were randomly selected from the SHHS database. There were 300 cases each with AHI < 5 (normal) and AHI ≥ 5. The accuracy for the two-class AHI classification was 92.18%.

2.2 Wearable Device Used in Validation Studies

Our medical-grade wearable device *SensEcho* is a wearable vest embedded with multiple biosensors to monitor various vital signs [13,35–37,40,41]. The system consists of three parts, including the monitoring terminal of accompanying physiological parameters, wireless networking and data transmission, and the central monitoring system. It has three comfortable electrode patches to capture the single-lead ECG signals at a sampling rate 200 Hz. Two sensing bands are embedded at the locations of the thorax and abdomen for monitoring the two types of breathing behaviors (thoracic and abdominal respiratory movements) at 25 Hz sampling rate. The error of heart rate and respiratory rate measurement is ± 2 and ± 0.15 beats per minute, respectively. SensEcho also has an ultra-low-power, 3-axis accelerometer MMA7260 (Freescale Inc., TX, USA) that is integrated into a data acquisition unit to capture posture and motility information. The accuracy of the 3-axis accelerometer measurement is 8 mg/LSB (Least Significant Bit). The main control chip of the system is ultra-low power ARM cortex-m3 MCU (EFM32GG330, Silicon Labs, USA). The power consumption of SensEcho is 100 mW. The device and its mainboard are shown in Fig. 2. Additionally, a wrist oximeter communicates with SensEcho via Bluetooth, whose sampling rate 1 Hz. SensEcho also provides local and cloud data storage options. When the cloud storage is neither stable nor available, the local storage can be activated to save the raw data in a 2-GB integrated flash drive.

The quality of acquired signals in wearable devices is very important for further data analyses and applications. SensEcho's signal quality can be mainly

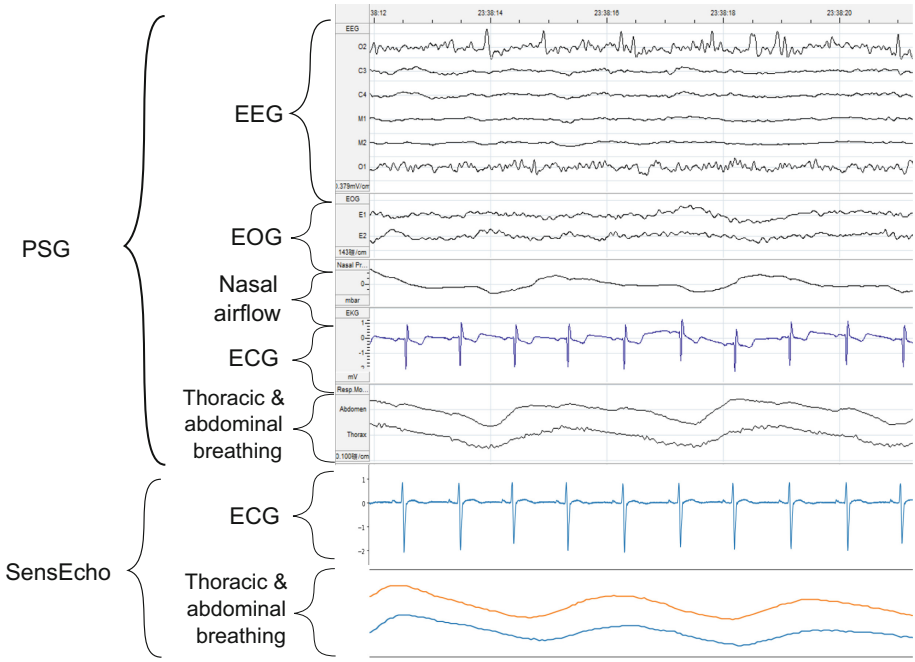


Fig. 3. A comparison of acquired signals between PSG and SensEcho.

interpreted in terms of device signal acquisition and wireless signal communication.

- Device signal acquisition: A comparison of signals of SensEcho and PSG records during an example period is depicted in Fig. 3. We can see that the consistency of the ECG signal and amplitude is higher between SensEcho and PSG, and the respiratory signal is also highly consistent with the PSG device. Compared with the gold standard, SensEcho can acquire signals with few errors.
- Wireless signal communication: The wireless physiological signal transmission unit is a network system based on Wi-Fi technology, including an ultra-low-power Wi-Fi module and WLAN system, which is capable of mobile monitoring, wireless network and roaming data transmission of multiple patients in the ward. The average packet loss rate between SensEcho and access points between 17 wards in our hospital was less than 0.1% for 16 months. The successful re-transmission rate was 100%. The data can be also successfully re-transmitted after the device is powered on next time, thus ensuring data integrity.

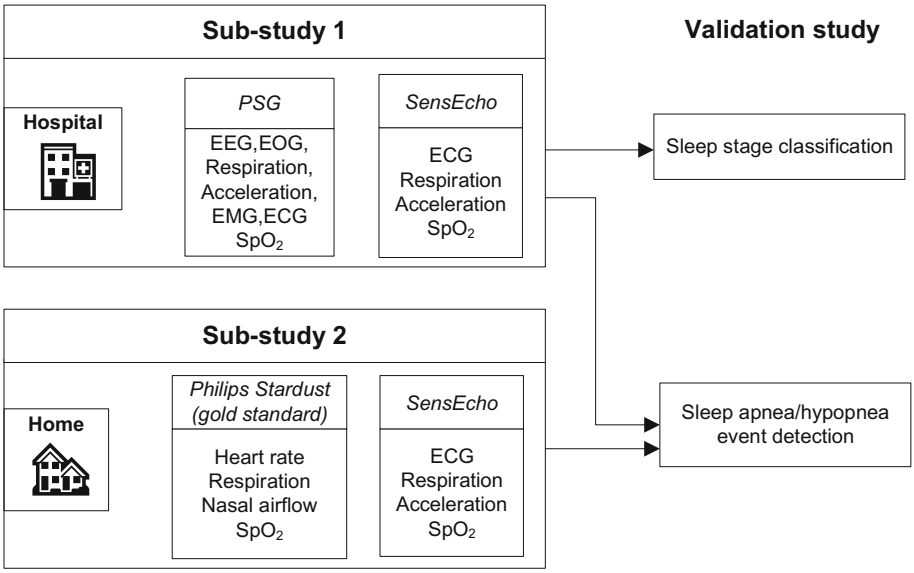


Fig. 4. Overview of validation studies.

2.3 Validation Study Design

Our study consisted of two independent sub-studies to validate sleep stage classification and sleep apnea/hypopnea event detection in clinical and home settings, respectively. Sub-study 1 was conducted in the sleep lab of Chinese PLA General Hospital, Beijing, China. All the research participants enrolled in this study wore both SensEcho and PSG. Aside from a well-controlled clinical laboratory environment, we also conducted another sub-study 2 in the home environment. All enrolled participants were asked to wear SensEcho and a CFDA-approved device (as the gold standard) for sleep apnea/hypopnea event detection due to the unavailability of PSG outside the hospital. An overview of two sub-studies was illustrated in Fig. 4. Every participant in both sub-studies complied with the protocol approved by the IRB review board (IRB number: S2018-095-01) and signed the written informed consent, and this study was conducted in accordance with the Declaration of Helsinki. Demographic information was collected by a questionnaire survey, including age, gender, weight and height.

Sub-study 1: SensEcho Compared with PSG for Sleep Stage Classification and Sleep Apnea/Hypopnea Event Detection. A total of 30 participants, who were suspected of sleep apnea in a respiratory clinic and met

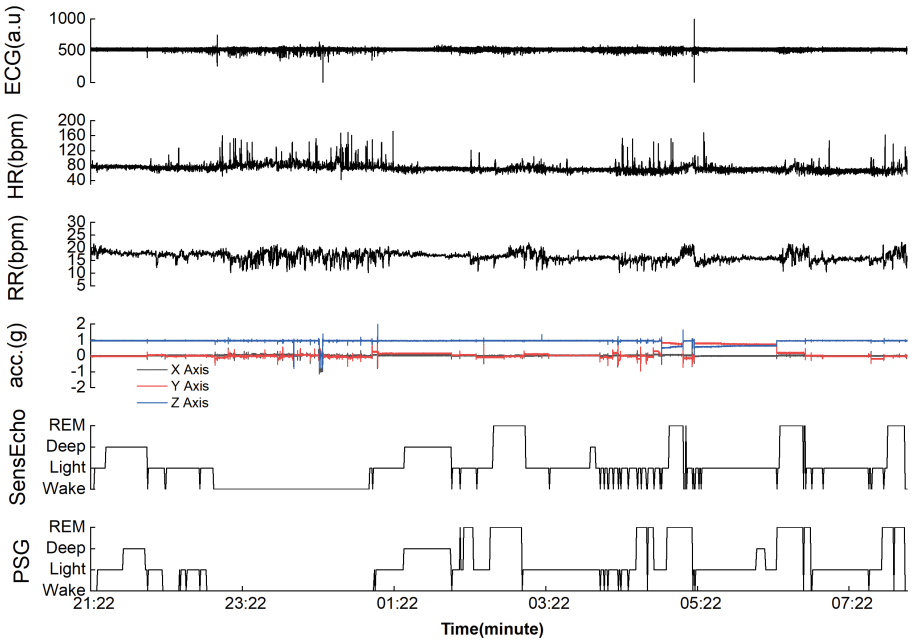


Fig. 5. Example raw ECG, heart rate, respiratory rate, acceleration of SensEcho, sleep stage classification of SensEcho and PSG for one night. "acc" stands for acceleration.

the inclusion and exclusion criteria, were enrolled in sub-study 1. The inclusion criteria included: (1) 18 years of age and over [17,19]; (2) willingness to cooperate with clinicians and to provide informed consent as determined. The exclusion criteria included: (1) current pregnancy; (2) recent healthy history of major psychiatric disorders or drug dependency or history of schizophrenia. Such criteria were endorsed by clinicians according to the recommendation of previous research [19]. PSG recordings of each subject were collected by Embla N7000 [16] The device settings and the electrode placement followed the regulation of the AASM [2]. The time-series data in a PSG test was segmented into 30-second epochs, and each epoch was scored by several certified sleep specialists' consent following the commonly adopted guideline [2]. During the PSG test, each participant wore SensEcho under the guidance of doctors. SensEcho synchronously collected his/her physiological data of ECG, respiration, posture/activities, and SpO_2 for one night depending on the actual sleep time. The configuration of SensEcho strictly followed the product's manual. Figure 8a showed the actual set-up of a participant who had signed written informed consent for use of his image in sub-study 1.

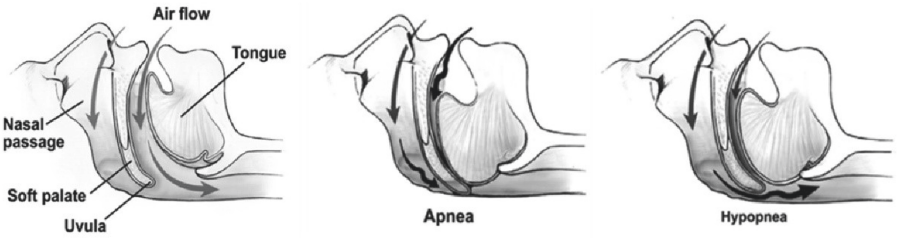


Fig. 6. The airways of normal breathing, apnea, and hypopnea.

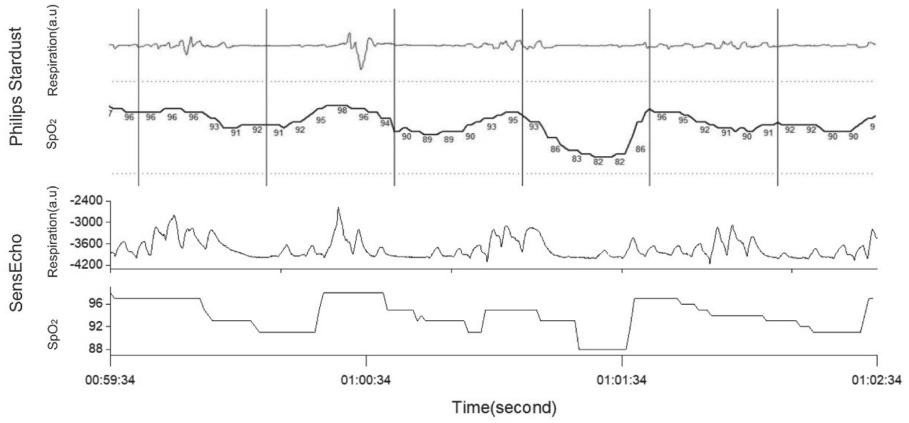


Fig. 7. An episode of sleep events lasting 3 min extracted from SensEcho recordings from one participant.

Before further analysis, clinicians first determined the sleep and awake points from the PSG data to extract the valid sleep duration. Every 30-s epoch of the sleep duration was then manually classified by clinicians into four categories (Wake-Light sleep(N1/N2)-Deep sleep(N3)-REM). The acquired physiological signals and the calculated vital signs by SensEcho of a subject and his four-class sleep stage classification are presented in Fig. 5.

Apart from sleep stage classification, we also compared sleep apnea/hypopnea event detection by SensEcho with PSG (shown in Fig. 3). Among various sleep disorder symptoms, we focused on apnea and hypopnea, which were the two most common ones [25,30]. Apneas were defined as more than 90% reduction in airflow from baseline for at least 10 s. Hypopneas referred to a decrease in airflow greater than 30% of baseline for at least 10 s duration accompanied by a decrease in blood oxygen saturation more than 3% and/or arousal, or a decrease in airflow greater than 30% of baseline for at least 10 s duration associated with more than 4% oxygen desaturation. Figure 6 shows the airways of normal breathing, apnea, and hypopnea(re-organized from Fig. 1 in the article [28]). We here leveraged the metric of AHI (the average number of apnea and hypopnea

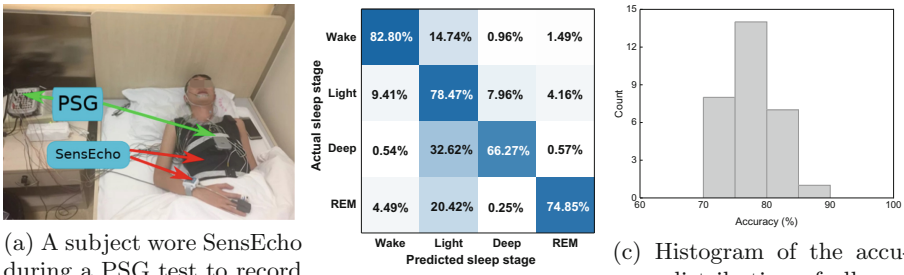
Table 1. Participant characteristics.

Participants	In-house dataset in Sub-study 1 (sleep stage and apnea/hypopnea event detection)	In-house dataset in Sub-study 2 (apnea/ hypopnea event detection)
Number	30	35
Sex (F/M)	8/22	6/29
Night (number)	30	35
Age (year)		
Mean (std)	66.10 ± 12.85	43.74 ± 10.03
Maximum	90	64
Minimum	42	20
BMI (kg/m^2)		
Mean (std)	29.64 ± 4.74	26.89 ± 3.67
Maximum	41.15	34.40
Minimum	23.29	18.73
AHI		
< 5	5	6
≥ 5	25	29

events per hour) to detect apnea/hypopnea events. A cutoff of AHI = 5 events/h was used as a threshold in clinical practice for determining whether a subject had sleep apnea/hypopnea (AHI $\geq$ 5) or not (AHI $<$ 5) [25].

Sub-study 2: SensEcho Compared with a CFDA-approved Device for Sleep Apnea/Hypopnea Event Detection. A total of 35 patients who met the eligibility criteria of our hospital's respiratory department were enrolled in the study. The inclusion criteria included: (1) 18 years old and over; (2) patients with suspected symptoms of sleep apnea. The exclusion criteria were the same as in sub-study 1.

All participants were required to wear SensEcho and a CFDA-approved device, Philips Stardust (PS), simultaneously for one-night sleep apnea/hypopnea event detection at home. The heart rate, respiratory effort, nasal airflow, and SpO_2 readings were recorded in PS, while the ECG signals, respiratory effort, body movements, SpO_2 data of SensEcho were collected. Figure 7 depicts the example signals of PS and SensEcho of a subject suspected of OSA. The process of wearing SensEcho was identical to that in sub-study 1. The AHI values of SensEcho and PS were compared to validate SensEcho's sleep apnea/hypopnea event detection performance, which was presented in the result.



(a) A subject wore SensEcho during a PSG test to record his sleep behavior

(b) Confusion matrix

(c) Histogram of the accuracy distribution of all participants

Fig. 8. 4-class sleep stage classification on our in-house dataset in sub-study 1.

3 Experiment Results

Throughout our study, no unexpected incidents or study withdrawals were reported. All collected data were qualified for analysis: a total of 30 participants with 27,669 scored epochs in sub-study 1, and 35 participants with complete night data of SensEcho and PS in sub-study 2. Participant characteristics of the entire samples in both sub-studies are provided in Table 1.

3.1 Sub-study 1: SensEcho Compared with PSG for Sleep Stage Classification and Sleep Apnea/Hypopnea Event Detection

Results of Sleep Stage Classification. The average accuracy of BLSTM was 77.83% (±0.04%), and the kappa coefficient was 0.63 (±0.19). Figure 8b presents the confusion matrix of four-class sleep stage classification in sub-study 1. The accuracy of BLSTM for four-class sleep stage classification was 82.80%, 78.47%, 66.27%, 74.85%, respectively. The classification of wake had the best performance (82.80%) due to its distinct patterns against other sleep stages. As shown in Fig. 8c, 8 out of 30 had an accuracy of more than 80%, and there were no cases with an accuracy below 70%. The best case by SensEcho was 87.73% and the worst case was 71.47%. Figure 9 presents the examples of the best case and the worst case, respectively. Determining sleep stages based on an EEG of patients with sleep apnea is believed to be challenging for specialists while following normal staging rules [3]. Our results were relatively satisfactory.

Results of Sleep Apnea/Hypopnea Event Detection. In the left figure of Fig. 10a, the middle horizontal dashed line indicated the average bias (mean of AHI = –2.94 events/h), the top and bottom dashed lines were the 1.96 standard deviation limits. The right figure of Fig. 10a depicts the reference AHI of PSG (x-axis) versus SensEcho AHI (y-axis). Two black dashed lines indicated the reference threshold (AHI = 5 events/h) and SensEcho's threshold (AHI = 5 events/h). A confusion matrix for the classification of all participants into two

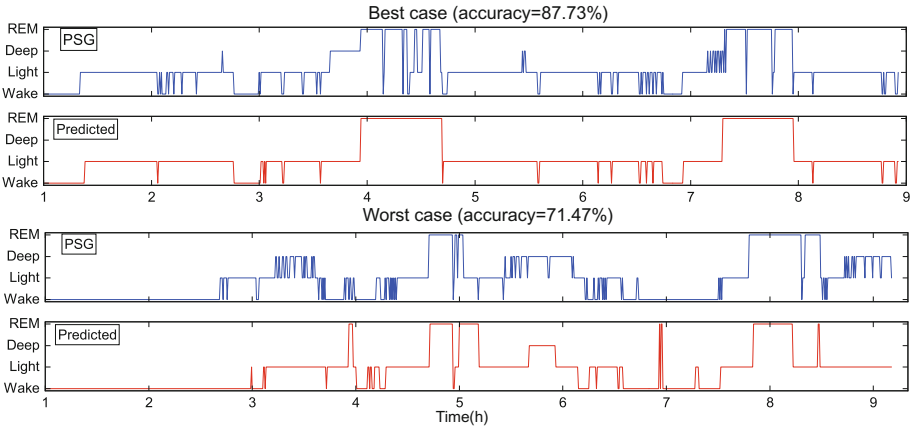


Fig. 9. Examples of the best and the worst cases of sleep stage classification in sub-study 1.

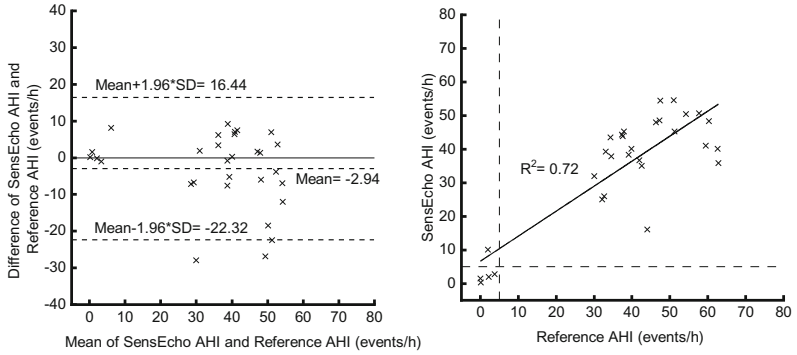
Table 2. Confusion matrix of sleep apnea detection using SensEcho in sub-study 1.

		PSG		
		AHI ≥ 5	AHI < 5	
SensEcho	AHI ≥ 5	25	1	PPV 96.15%
	AHI < 5	0	4	NPV 100%
		Sensitivity 100%	Specificity 80%	Accuracy 96.67%

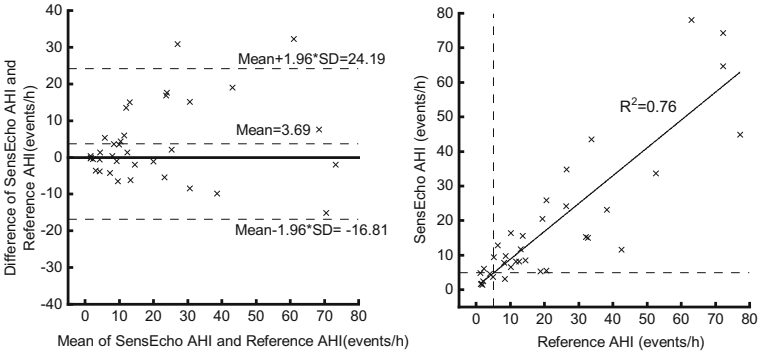
categories is presented in Table 2. SensEcho had a sensitivity of 100% in detecting AHI of 5 and above. The specificity, positive predictive value (PPV), negative predictive value (NPV), and total accuracies were 80%, 96.15%, 100%, 96.67%, respectively.

3.2 Sub-study 2: SensEcho Compared with a CFDA-approved Device for Sleep Apnea/Hypopnea Event Detection

The dashed lines in the left figure of Fig. 10b referred to the average bias and the 1.96 standard deviation limits, and the dashed lines in the right figure of Fig. 10b denoted the reference threshold (AHI = 5 events/h) and SensEcho's threshold (AHI = 5 events/h). A confusion matrix for the classification of all participants into two categories is illustrated in Table 3. SensEcho had a sensitivity of 93.10% in detecting AHI of 5 and above. In addition, specificity, PPV, NPV, and the total accuracy was 83.33%, 96.43%, 71.43%, 91.43%, respectively.



(a) Bland-Altman (left) and correlation (right) plots of SensEcho AHI and the reference AHI in sub-study 1



(b) Bland-Altman (left) and correlation (right) plots of SensEcho AHI and the reference AHI in sub-study 2

Fig. 10. Sleep apnea/hypopnea event detection in sub-study 1 and 2.

4 Discussion

In this study, we proposed two algorithms for sleep analysis with sleep stage classification and apnea/hypopnea event detection. We also validated our algorithms in different environmental settings via two independent sub-studies, enabling us to provide a comprehensive validation study on the two algorithms in both hospital and home scenarios. Compared with the gold standard, the results in both sub-studies indicate that the algorithms along with the wearable physiological monitoring system provide a reliable approach to sleep analysis. The satisfactory overall accuracy is attributed to the stability and robustness of the wearable physiological detection system. Based on a user survey of 30 subjects in sub-study 1, 85% of the subjects responded that they felt comfortable with SensEcho and

Table 3. Confusion matrix of sleep apnea detection using SensEcho in sub-study 2.

		PS		
		AHI ≥ 5	AHI < 5	
SensEcho	AHI ≥ 5	27	1	PPV 96.43%
	AHI < 5	2	5	NPV 71.43%
		Sensitivity 93.10%	Specificity 83.33%	Accuracy 91.43%

did not feel tight in their thoraxes. SensEcho is regarded as patient-friendly for long-term sleep monitoring.

In terms of sleep stage classification, numerous researches relied on PSG tests where EEG, EOG, and EMG were recorded for an accurate sleep stage classification [12,26]. However, those signal sources (EEG, EOG, EMG) were not available for typical wearable devices, making the sleep stage classification task more challenging. Some researchers utilized the public dataset (SIESTA, SHHS) and laboratory data to classify four sleep stages and reported an satisfactory accuracy [8,32]. However, most of them did not conduct further clinical validation studies. Asher Tal et al. validated a contact-free sleep monitoring device (EarlySense) in 2017 [31]. EarlySense achieved relatively high accuracy in two-class sleep stage classification (sleep and wake). However, the overall accuracy of four-class sleep stage classification was moderate (about 63.5%, calculated from the confusion matrix provided by the authors). Unlike the researches mentioned above, we performed validation studies in clinical settings. The average accuracy of our four-class sleep stage classification was satisfactory on the dataset collected from the enrolled patients.

Considering the performance of sleep apnea detection using wearable systems in existing studies, based on a threshold of AHI ≥ 5 events/h, Nox-T3 portable monitor had a sensitivity of 95%, a specificity of 69%, PPV of 94% and NPV of 75% when used to detect AHI compared to PSG [38], a photoplethysmography (PPG)-based device detected OSA with a sensitivity of 100%, specificity of 44%, PPV of 62%, and NPV of 100% compared to PSG on 48 patients [7]. Compared to those results, SensEcho had a satisfactory sensitivity, specificity, and accuracy in terms of sleep apnea/hypopnea event detection. The confident results could be attributed to the robust sleep apnea/hypopnea event detection algorithm presented in this study, as well as accurate measurement of respiratory signals from SensEcho. The sensitivity of the AHI detection in sub-study 2 was slightly lower than that in sub-study 1 possibly for the following reasons. Firstly, patients who wore PSG usually suffered from worse sleep problems than those who wore PS (shown in Fig. 10a and Fig. 10b). Secondly, PS was specifically designed for portable sleep apnea/hypopnea event detection. Its results might be influenced by factors such as wearing conditions of the patients and device performance so that the results might be not so reliable as those of PSG.

Admittedly, there are some limitations to our study. First, the participants enrolled in the two sub-studies were not identical. They were allowed to choose

to sleep in clinical or home settings. However, since each sub-study was independent, there was no substantial impact on our validation results. Second, the number of patients and healthy individuals in sub-study 1 was not matched, which was why the sleep apnea/hypopnea event detection algorithm showed higher accuracy on the external test dataset than on the validation dataset. Third, since the experiments in sub-study 2 were performed in a home setting, there was some uncertainty about the device. For instance, some patches were displaced or not tightly connected by accident while the participants were not realizing such events, even though every participant was given instructions as to how to use the device by specialists before he/she conducted the experiment at home. Fourth, while our study significantly improved the accuracy of sleep stage classification, the accurate classification between N3 and N1/N2 remained challenging.

In our ongoing and future research, we will keep expanding the clinical sample size, recruiting patients who have sleep-related symptoms or diseases to continuously constitute our comprehensive sleep database. A large in-house database will also mitigate the possible bias introduced by ethnicity, age, and disease, etc. when a public database is used for model training. Second, we will optimize and fine-tune the sleep stage classification and sleep apnea/hypopnea event detection models based on large sample datasets to improve the accuracy and robustness. The current sleep event detection algorithm can effectively identify whether a patient has sleep apnea/hypopnea. In future studies, we will continue to explore sleep event detection algorithms based on the different severity levels of AHI to provide patients with a severity degree screening for sleep apnea/hypopnea events. Finally, despite the inability of SensEcho to fully achieve the accuracy of PSG's sleep analysis, it still provides sleep architecture and apnea/hypopnea event detection. We will apply the wearable system with algorithms to monitor patients' sleep conditions in both clinical and home scenarios to further validate the feasibility of the system and algorithms.

5 Conclusions

In this study, we developed algorithms for sleep stage classification and sleep event detection using cardiopulmonary signals based on a large sample dataset, and adopted a medical-grade wearable system to collect physiological data. To further validate the performance of the algorithms, we conducted validation studies in clinical and home settings, and the results showed high agreement with PSG for four-class sleep stage classification and sleep apnea/hypopnea detection in clinical settings, and good performance for sleep apnea/hypopnea screening in home settings. The results demonstrate that the sleep analysis algorithms proposed in this paper perform well in both sleep stage classification and sleep event detection, either in clinical scenarios and home settings, indicating that the wearable system of SensEcho can be used along with the two algorithms for sleep measurement and analysis.

Acknowledgment. This work is supported by The National Natural Science Foundation of China (62171471); Beijing Municipal Science and Technology (Z181100001918023); Big Data Research & Development Project of Chinese PLA General Hospital (2018MBD-09).

References

1. Berry, R.B., et al.: Aasm scoring manual updates for 2017 (version 2.4). J. Clin. Sleep Med. **13**(5), 665–666 (2017)
2. Berry, R.B., et al.: Aasm scoring manual version 2.2 updates: new chapters for scoring infant sleep staging and home sleep apnea testing. J. Clin. Sleep Med. **11**(11), 1253–1254 (2015)
3. Carskadon, M.A., Rechtschaffen, A.: Monitoring and staging human sleep. Principles Pract. Sleep Med. **5**, 16–26 (2011)
4. Chaiard, J., Weaver, T.E.: Update on research and practices in major sleep disorders: part ii-insomnia, willis-ekbom disease (restless leg syndrome), and narcolepsy. J. Nurs. Sch. **51**(6), 624–633 (2019)
5. Cheng, W., Rolls, E.T., Ruan, H., Feng, J.: Functional connectivities in the brain that mediate the association between depressive problems and sleep quality. JAMA Psychiatry **75**(10), 1052–1061 (2018)
6. Dean, D.A., et al.: Scaling up scientific discovery in sleep medicine: the national sleep research resource. Sleep **39**(5), 1151–1164 (2016)
7. Faßbender, P., Haddad, A., Bürgener, S., Peters, J.: Validation of a photoplethysmography device for detection of obstructive sleep apnea in the perioperative setting. J. Clin. Monit. Comput. **33**(2), 341–345 (2018). https://doi.org/10.1007/s10877-018-0151-2
8. Fonseca, P., Long, X., Radha, M., Haakma, R., Aarts, R.M., Rolink, J.: Sleep stage classification with ecg and respiratory effort. Physiol. Meas. **36**(10), 2027 (2015)
9. Irwin, M.R.: Sleep and inflammation: partners in sickness and in health. Nat. Rev. Immunol. **19**(11), 702–715 (2019)
10. Iwasaki, A., et al.: Screening of sleep apnea based on heart rate variability and long short-term memory. Sleep Breathing **25**(4), 1821–1829 (2021). https://doi.org/10.1007/s11325-020-02249-0
11. Klosh, G., et al.: The siesta project polygraphic and clinical database. IEEE Eng. Med. Biol. Mag. **20**(3), 51–57 (2001)
12. Lajnef, T., et al.: Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines. J. Neurosci. Meth. **250**, 94–105 (2015)
13. Li, P., et al.: Mobicardio: a clinical-grade mobile health system for cardiovascular disease management. In: 2019 IEEE International Conference on Healthcare Informatics (ICHI), pp. 1–6. IEEE (2019)
14. Long, X., et al.: Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging. Physiol. Meas. **35**(12), 2529 (2014)
15. Magnusdottir, S., Hilmisson, H.: Ambulatory screening tool for sleep apnea: analyzing a single-lead electrocardiogram signal (ecg). Sleep Breathing **22**(2), 421–429 (2018). https://doi.org/10.1007/s11325-017-1566-6
16. Myllymaa, S., et al.: Assessment of the suitability of using a forehead eeg electrode set and chin emg electrodes for sleep staging in polysomnography. J. Sleep Res. **25**(6), 636–645 (2016)

17. Cho, J.H., Kim, H.J.: Validation of apnealink ox^{TM} plus for the diagnosis of sleep apnea. Sleep Breathing **21**(3), 799–807 (2017). https://doi.org/10.1007/s11325-017-1532-3
18. Peppard, P.E., Young, T., Barnet, J.H., Palta, M., Hagen, E.W., Hla, K.M.: Increased prevalence of sleep-disordered breathing in adults. Am. J. Epidemiol. **177**(9), 1006–1014 (2013)
19. Pigeon, W.R., Taylor, M., Bui, A., Oleynk, C., Walsh, P., Bishop, T.M.: Validation of the sleep-wake scoring of a new wrist-worn sleep monitoring device. J. Clin. Sleep Med. **14**(6), 1057–1062 (2018)
20. Pinheiro, G., Cruz, A., Genta, P., Lorenzi-Filho, G.: Validation of a wireless wearable oximeter using mobile technology and cloud computing for the diagnosis of obstructive sleep apnea. In: B68. Diagnosis and Treatment of Sleep Disordered Breathing, pp. A3976–A3976. American Thoracic Society (2018)
21. Pion-Massicotte, J., Godbout, R., Savard, P., Roy, J.F.: Development and validation of an algorithm for the study of sleep using a biometric shirt in young healthy adults. J. Sleep Res. **28**(2), e12667 (2019)
22. Iber, C., et al.: The sleep heart health study: design, rationale, and methods. Sleep **20**(12), 1077–1085 (1997)
23. Radha, M., et al.: Sleep stage classification from heart-rate variability using long short-term memory neural networks. Sci. Rep. **9**(1), 1–11 (2019)
24. Redmond, S.J., de Chazal, P., O'Brien, C., Ryan, S., McNicholas, W.T., Heneghan, C.: Sleep staging using cardiorespiratory signals. Somnologie-Schlafforschung und Schlafmedizin **11**(4), 245–256 (2007). https://doi.org/10.1007/s11818-007-0314-8
25. Ruehland, W.R., Rochford, P.D., O'Donoghue, F.J., Pierce, R.J., Singh, P., Thornton, A.T.: The new aasm criteria for scoring hypopneas: impact on the apnea hypopnea index. Sleep **32**(2), 150–157 (2009)
26. Shi, P., Zheng, X., Du, P., Yuan, F.: Automatic sleep stage classification based on LSTM. In: Sun, Y., Lu, T., Xie, X., Gao, L., Fan, H. (eds.) ChineseCSCW 2018. CCIS, vol. 917, pp. 478–486. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-3044-5_35
27. Silber, M.H., et al.: The visual scoring of sleep in adults. J. Clin. Sleep Med. **3**(02), 121–131 (2007)
28. Somers, V.K., et al.: Sleep apnea and cardiovascular disease: an American heart association/American college of cardiology foundation scientific statement from the american heart association council for high blood pressure research professional education committee, council on clinical cardiology, stroke council, and council on cardiovascular nursing in collaboration with the national heart, lung, and blood institute national center on sleep disorders research (national institutes of health). J. Am. Coll. Cardiol. **52**(8), 686–717 (2008)
29. Stephansen, J.B., et al.: Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nat. Commun. **9**(1), 1–15 (2018)
30. Strollo, P.J., Jr., Rogers, R.M.: Obstructive sleep apnea. N. Engl. J. Med. **334**(2), 99–104 (1996)
31. Tal, A., Shinar, Z., Shaki, D., Codish, S., Goldbart, A.: Validation of contact-free sleep monitoring device with comparison to polysomnography. J. Clini. Sleep Med. **13**(3), 517–522 (2017)
32. Tataraidze, A., Korostovtseva, L., Anishchenko, L., Bochkarev, M., Sviryaev, Y.: Sleep architecture measurement based on cardiorespiratory parameters. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3478–3481. IEEE (2016)

33. Thomas, R.J., Mietus, J.E., Peng, C.K., Goldberger, A.L.: An electrocardiogram-based technique to assess cardiopulmonary coupling during sleep. Sleep **28**(9), 1151–1161 (2005)
34. Xiao, M., Yan, H., Song, J., Yang, Y., Yang, X.: Sleep stages classification based on heart rate variability and random forest. Biomed. Sign. Proces. Control **8**(6), 624–633 (2013)
35. Xu, H., et al.: Study on the accuracy of cardiopulmonary physiological measurements by a wearable physiological monitoring system under different activity conditions. Sheng wu yi xue gong cheng xue za zhi=J. Biomed. Eng.=Shengwu yixue gongchengxue zazhi **37**(1), 119–128 (2020)
36. Xu, H., et al.: Construction and application of a medical-grade wireless monitoring system for physiological signals at general wards. J. Med. Syst. **44**(10), 1–15 (2020). https://doi.org/10.1007/s10916-020-01653-z
37. Xu, H., et al.: Assessing electrocardiogram and respiratory signal quality of a wearable device (sensecho): semisupervised machine learning-based validation study. JMIR mHealth uHealth **9**(8), e25415 (2021)
38. Xu, L., et al.: Validation of the nox-t3 portable monitor for diagnosis of obstructive sleep apnea in chinese adults. J. Clin. Sleep Med. **13**(5), 675–683 (2017)
39. Yang, Z., Pathak, P.H., Zeng, Y., Liran, X., Mohapatra, P.: Vital sign and sleep monitoring using millimeter wave. ACM Trans. Sens. Netw. (TOSN) **13**(2), 1–32 (2017)
40. Zhang, Y., et al.: Breathing disorder detection using wearable electrocardiogram and oxygen saturation. In: Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems, pp. 313–314 (2018)
41. Zhang, Y., et al.: Automated sleep period estimation in wearable multi-sensor systems. In: Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems, pp. 305–306 (2018)
42. Zhao, M., Yue, S., Katabi, D., Jaakkola, T.S., Bianchi, M.T.: Learning sleep stages from radio signals: a conditional adversarial architecture. In: International Conference on Machine Learning, pp. 4100–4109. PMLR (2017)