AUTOMATIC DETECTION OF OBSTRUCTIVE SLEEP APNEA BASED ON PHYSIOLOGICAL SIGNALS

A PREPRINT

Marino E. Gavidia[1,*], Arthur N. Montanari[1], and Jorge Goncalves[1,2,*]

[1]Luxembourg Center for Systems Biomedicine, University of Luxembourg, Belvaux L-4367, Luxembourg
[2]Department of Plant Sciences, Cambridge University, Cambridge CB2 3EA, United Kingdom
[*]Corresponding authors: marino.gavidia@uni.lu; jmg77@cam.ac.uk

September 28, 2022

ABSTRACT

Obstructive sleep apnea (OSA) is a common respiratory condition characterized by respiratory tract obstruction and breathing disorder. Early detection and treatment of OSA can significantly reduce morbidity and mortality. OSA is often diagnosed with overnight polysomnography (PSG) monitoring; however, continuous PSG monitoring is unfeasible as it is costly, time-consuming, and uncomfortable for patients. To circumvent these issues, we propose an automatic detection method of OSA events using only data available from easy-to-use wearables: electrocardiogram, respiratory, and oximetry data. We use data from three sleep studies from the National Sleep Research Resource (NSRR), the largest public repository, consisting of 10,878 recordings. The developed method is based on a combination of deep convolutional neural networks and a light-gradient-boos machine (LightGBM) for classification. On the test data, our model achieved the highest classification performance in the literature, with an accuracy and F1-score of 91%. Since the trained model is simple and computationally efficient, we expect that our method can be implemented for automatic detection of OSA in unsupervised home monitoring systems, reducing costs to healthcare systems, and improving patient care.

Keywords Obstructive sleep apnea · Artificial intelligence · Neural networks · Automatic detection

Obstructive sleep apnea (OSA) is a common sleep-related breathing disorder that affects 6% to 17% of adults, reaching up to 49% in older populations [1]. OSA is usually marked by pauses in breathing or shallow breathing. This happens when the soft tissue at the back of the throat collapses and closes at night [2]. Because blockage in the airways slows breathing, less oxygen is sent from the lungs to the heart and body. The CO_2 level in the blood then increases due to the impaired breathing, leading to episodes of sudden awakening or choking during sleep [3]. During OSA events, nasal airflow ceases for a short time, but the brain and body fight to keep breathing. OSA has been associated with an increased risk of heart disease, diabetes, chronic kidney disease, stroke, depression, and cognitive impairment [4]. As age and the likelihood of having sleep apnea are positively correlated [5], the growing number of patients with OSA is expected to increase pressure on healthcare systems.

Typically, the diagnosis and detection of OSA are based on polysomnography (PSG) tests conducted in a sleep facility [6]. PSG requires the overnight recording and monitoring of several physiological signals, including electroencephalogram (EEG), electrocardiogram (ECG), electrooculogram (EOG), chin muscle activity, leg movements, respiratory effort, nasal airflow, and oxygen saturation (SpO_2). Sleep specialists then examine these signals to provide a final diagnosis of OSA syndrome. PSG is time-consuming, expensive, and inconvenient. As a result, it is expected that more than 85% of people with OSA are not diagnosed [5]. PSG may also not be a suitable alternative to assess the severity of OSA because patients are tested for only one night in a strange and uncomfortable sleep laboratory [7]. Therefore, the development of portable, easy-to-use, reliable, and affordable OSA monitoring tools for home care applications is crucial to improve patient care.

Table 1: Data description.

Caracteristics	MESA	MROS	SHHS
Total PSG recordings	1516	2780	6584
Male	765	2780	3302
Female	751	0	3281
Age	69.6 (54–94)	78.3 (67–90)	74.4 (56–90)
BMI	28.7 (—)	27.1 (16.4–45.3)	27.7 (18.0–50.0)
Sleep time [min]	360.58 (181–599)	420 (184–739)	376.4 (242.0–473.5)
AHI [per hour]	24.3 (15.2–37.7)	24.0 (0.1–106.0)	22.0 (0.1–117.0)
Ethnicity			
White	35.1%	90.0%	85.7%
African	28.1%	3.4%	8.3%
Hispanic	23.7%	2.2%	—
Asian	13.1%	3.2%	—
Other	—	1.2%	6.0%

Data are reported in mean and range (in parenthesis).
Abbreviations: Apnea-Hypopnea Index (AHI), body mass index (BMI).

The automatic detection of OSA, considered the most severe and common type of sleep apnea [8], is a pressing problem in the literature [9–11]. Such methods can provide an efficient and accurate solution for the challenging, time-consuming diagnosis of diseases, relieving pressure on the healthcare systems. So far, detection tools for OSA have been based on supervised learning, which involves analyzing human-crafted features extracted from time and/or frequency domains of one or more physiological signals. These algorithms require expert knowledge for the design and selection of features to be extracted, which may still not include the most relevant features (information) for classification from data [12]. The procedure can be hard, time consuming, and often subject to bias or lack of generalizability [13, 14]. Deep learning, on the other hand, has shifted data modeling away from "expert-driven" feature engineering toward "data-driven" feature extraction [11, 15]. Widespread across many applications such as computer vision, natural language processing, and speech recognition [16, 17], deep learning methods have also been successfully adopted in biomedical applications, including for the detection of diseases such as atrial fibrillation [18], prediction of ventricular tachycardia [19], oxygen desaturation index estimation [20], and sleep stage classification [21].

In this paper, we propose a data-driven method for the automatic detection of OSA events from raw physiological data based on deep learning. Given that OSA affects millions of people worldwide, we expect small unbalanced datasets not to yield generalizable models. Unlike previous works [9–11], our study uses an extensive multicenter database containing 10,880 recordings of 8,444 patients belonging to multiple ethnicities and with an average age of 69.9 years, being the first data-driven method to use such an extensive database. The proposed model achieved the best classification performance available in the literature, even in the absence of certain physiological signals as input data. Cross-validation on out-of-sample data also shows that our model is highly generalizable, achieving high accuracy and balance on large external datasets not included during the training process. Given the high performance and small computational cost of our model, we expect that it could be implemented as part of a home OSA detection system.

Results

Data description. Three different datasets were considered in this work: the Multi-Ethnic Study of Atherosclerosis (MESA) [22], the Men Study of Osteoporotic Fracture (MrOS) [23], and the Sleep Heart Health Research (SHHS) [24]. These datasets are publicly available from the National Sleep Research Resource for Sleep-Related Studies (NSRR) [25]. In total, 14,370 PSG recordings were considered in our study. The PSG recordings were collected and annotated during typical PSG evaluations in healthcare facilities (Methods).

We consider only measurement channels that can be easily implemented in a home environment and are accessible in the three datasets: pulse oximetry (SpO_2), electrocardiogram (ECG), thoracic movement (ThorRes), abdominal movement (AbdoRes), and nasal airflow. Note that, instead of using ECG data as input to our model, we use R-to-R interval (RRI) data, which is widely available in wearable devices such as smartwatches. The RRI data is inferred from ECG data by measuring the time difference between heartbeats (from one R peak to the next R peak) using the Pan-Tompkins algorithm [26, 27]. Moreover, since data were collected from various healthcare facilities using different sampling frequencies, all channels were resampled to ensure that they were consistent with the maximum sampling

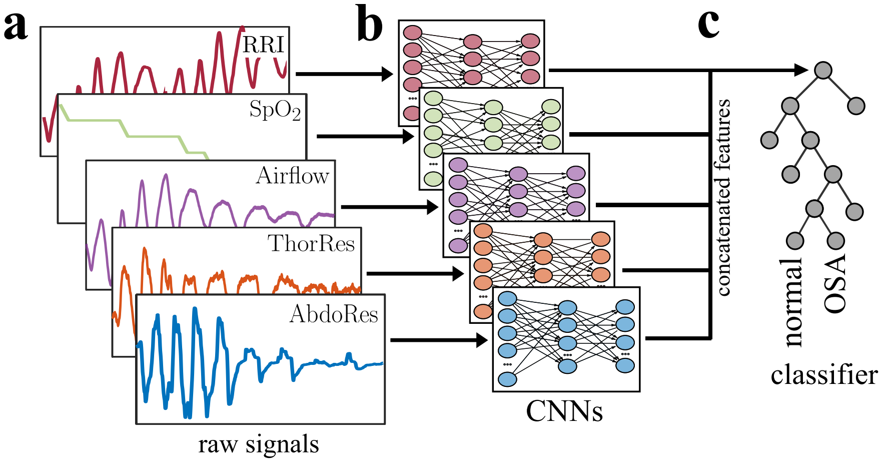

Figure 1: Pipeline of the proposed method. (a) Data are separated by channels (RRI, SpO_2, among others) and segmented into 60s windows. (b) For each channel, a distinct trained deep CNN extracts features (outputs) from the raw signal (input). (c) The extracted features are concatenated and fed to a LightGBM that classifies the input data between normal and OSA events.

frequency. The cubic spline was used as the interpolation method [28]; it is simple to implement and, at the same time, does not attenuate the higher frequency components of the signal. This procedure allows us to standardize the data to the same length, which is required to train the neural network.

PSG recordings were included in our study if none of the following exclusion criteria were met: (1) total sleep time for PSG less than 1.5 hours; (2) poor-quality PSG recordings given by the presence of “unsure” or “noise” labels in more than a third of the total sleep time; and (3) the absence of one or more channels from the five channels selected for the study. After applying the exclusion criteria, a total of 10,880 sleep recordings were obtained to develop and evaluate the proposed method. Table 1 summarizes the databases.

Deep learning-model for the detection of OSA events. We developed a hybrid deep-learning model for the automatic classification of normal and OSA events from data. The inputs of the model are short segments of time-series data (e.g., 60s) recorded during the patients’ sleep, which includes the physiological signals available in the considered datasets: RRI, SpO_2, ThorRes, AbdoRes, and/or nasal airflow. The model then classifies whether a given input corresponds to a normal or an OSA event. Fig. 1 illustrates the method’s pipeline. For each available physiological signal, a distinct deep convolutional neural network (CNN) is trained to automatically extract global features from the raw 1-dimensional signals. The extracted features by all CNNs are then combined on a light gradient-boos machine (LightGBM) [29] to perform the classification between normal and OSA events. The LightGBM yields the probability of a given sample belonging to the OSA class. The default classification threshold of 0.5 is used to perform the binary classification betwen OSA and normal samples. See Methods for details on data pre-processing and model training.

We train and assess the proposed method on raw physiological signals to reduce complexity in the design and implementation of our model and to eliminate the need for human-created features. The performance is assessed via a 10-fold cross-validation strategy, separating data by patient record to avoid data leakage from samples taken from the same patient. To overcome the bias and inaccuracy associated with classification using imbalanced data between OSA and normal events, the undersampling approach was used to balance the number of samples from each individual record. OSA labeled events were sampled every 60s with a 10s overlap. An equal number of normal events are randomly sampled from the same PSG record, yielding a total of 949,428 balanced samples.

Performance of the hybrid model. The performance of the proposed method in the classification task between normal and OSA events is evaluated in this section. Results are shown in Fig. 2 assuming that all five physiological signals are available for training and testing. The area under the receiver operating characteristic curve (AUROC) and the precision-recall curve (AUPRC) of 0.96 and 0.97, respectively (Fig. 2a,b). This demonstrates that the separation between OSA and normal events is highly accurate and has robust discriminative power concerning the positive class (OSA) and the negative class (normal). Fig. 2c shows the confusion matrix of the 10-folds on 949,428 total samples, containing the number of correct and incorrect predictions made by the model for positive (OSA) and negative (normal) events. Across all samples, the percentage of false negatives and false positives is under 10%. This indicates that the method is not perfect, but attains high performance in the classification of both normal and OSA events.

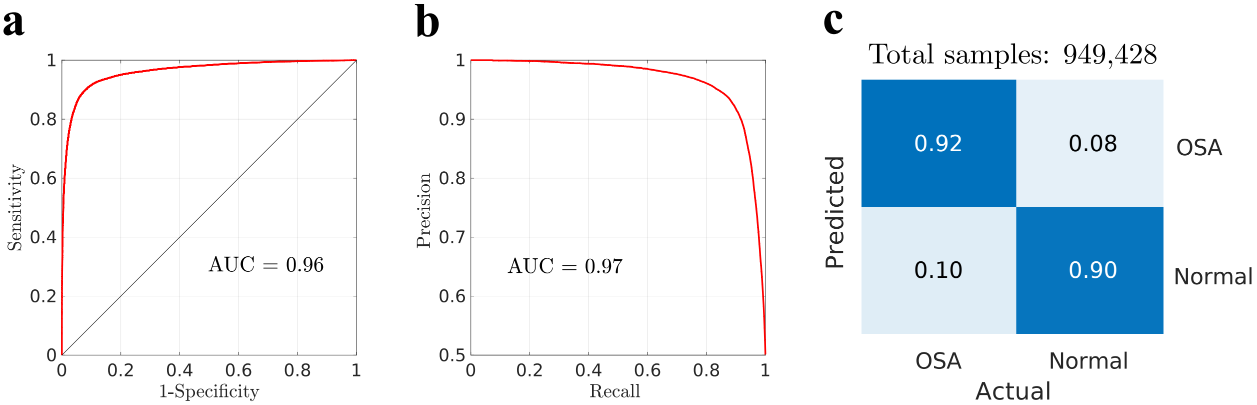

Figure 2: Performance of the proposed method. (a) Receiver-operator characteristic curve. (b) Precision-recall curve. (c) Confusion matrix.

Table 2: Comparison of the proposed method with state-of-the-art OSA detection methods using the same databases.

Year	Reference	Database	Records	Channels used	Classifiers	Acc.	Sens.	Spec.	F1-Score
2017	[30]	MESA	100	Airflow	CNN1D	0.74	0.74	0.74	0.74
2017	[31]	SHHS	2,100	AbdoRes+ThorRes	LSTM	0.71	0.62	0.80	0.48
2018	[32]	MESA	1,507	Airflow	CNN2D	0.79	0.79	0.79	0.79
2020	[33]	MROS	545	ECG	CNN1D-LSTM	0.79	0.77	0.85	0.79
2022	[34]	SHHS	8,444	ECG+AbdoRes+ThorRes+Airflow+SpO_2	Wavelet decomposition	0.84	0.84	0.85	0.85
2022	Proposed Method	All	10,880	RRI+AbdoRes+ThorRes+Airflow+SpO_2	Hybrid CNN classifier	0.91	0.90	0.92	0.91

Table 2 compares the performance of our method with previously published methods in OSA detection in the literature using the same databases and at least 100 records. Our results achieve the highest score in terms of accuracy (91%), sensitivity (90%), specificity (92%), and F1-score (91%), with a good generalization of the data across the ten-fold cross-validation (as represented by a small standard deviation of 0.0053, see Table S7). This illustrates that the designed hybrid deep CNN classifier as well as the extensive dataset (consisting of three large databases) lead to a high improvement in the performance of our algorithm while keeping its computational complexity and data pre-processing step reasonably simple.

So far, the performance of the method was evaluated assuming that all measurement channels were available as inputs. However, we expect that some of the physiological signals may contain features that have a higher contribution in the classification task than other signals. Moreover, some physiological signals may not be available in a homecare monitoring application of sleep apnea. Therefore, it is interesting to evaluate the performance of our method for different combinations of input types (physiological signals). For each combination of input channels, the hybrid model is trained and tested as previously described.

Fig. 3a shows the performance results for different combinations of input channels. As expected, the best performance was obtained when all five channels were used in conjunction, and the performance decreases in general as the number of input channels reduces. Overall, models trained and tested with AbdoRes and/or ThorRes as inputs achieve higher performance. On the other hand, performance seems to be less dependent on RRI data; the model trained on RRI data alone was also the one with the smallest performance. The results show that, independently of the number and type of input channels, our results are balanced. For models trained on two or more input channels, as well as the model trained on AbdoRes data, performance was consistently higher than in previous publications.

Out-of-distribution performance. To evaluate how generalizable the method's performance is to out-of-sample data (that is, data not collected in the same study that was used to train the model), we employ a "leave-one-database-out" validation. In this validation, all but one database reported in Table 1 are used as training data, while the remaining one is used to evaluate the quality of the predictions on populations; hence the database used for testing is not used for training, and vice-versa.

Fig. 3b reports the out-of-distribution performance tested on each one of the available databases. The best performance was obtained on the MESA and MROS databases when all five channels were used as inputs, achieving an accuracy (F1-score) of 87.98% (88.62%) and 90.23% (90.84%) for the MESA and MROS database, respectively. Interestingly, in these cases, the out-of-distribution performance of the trained models achieved better results than previous methods

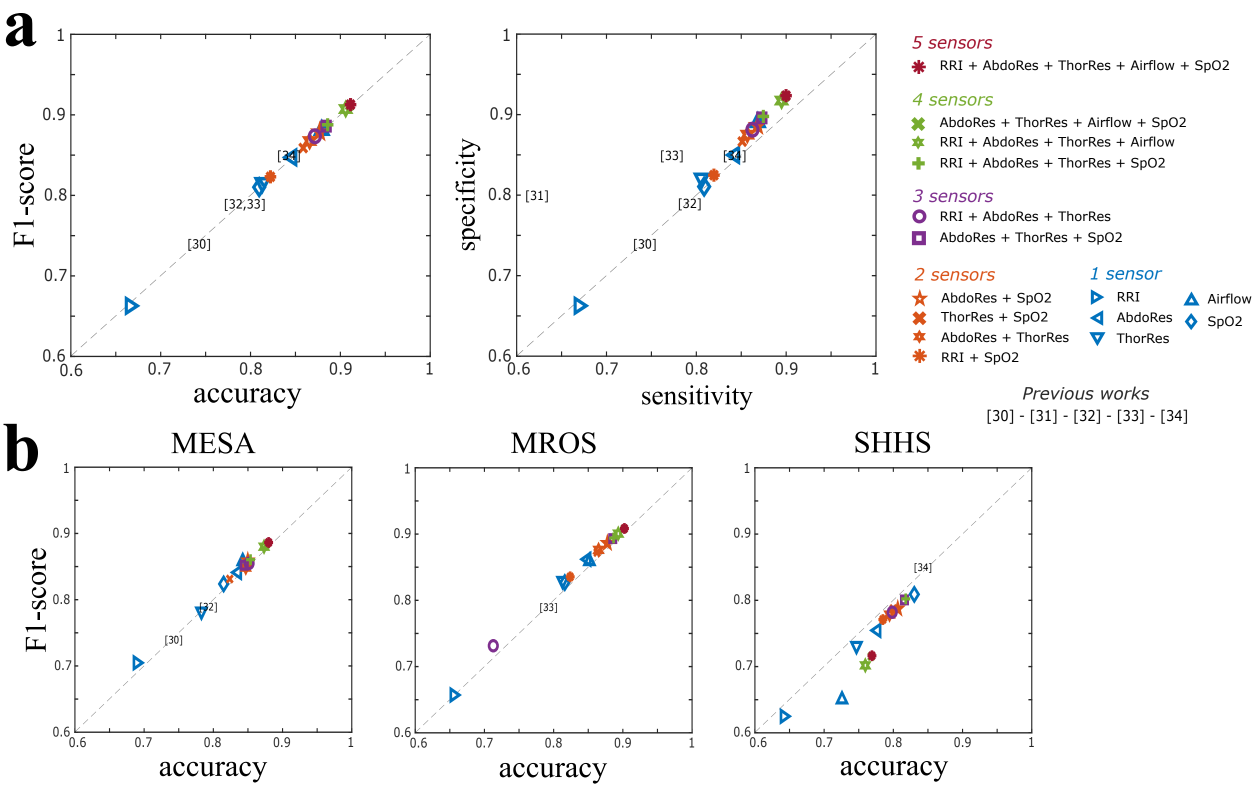

Figure 3: (a) Performance metrics of the proposed method for different combinations of input channels (physiological signals). Performance metrics of previous works in the literature using one of the databases studied in this work are included for comparison purposes. (b) Performance of the proposed method on out-of-distribution data using the "leave-one-database-out" methodology. Each plot shows the performance metrics of a model on a given database in which this database was not included in the training stage (e.g., performance metrics on MESA database are shown for models trained on MROS and SHHS data only). Performance metrics of previous works *trained and tested* on the same database were included for comparison purposes. Tables S1– S4 in Supplementary Material contain the specific values reported in these plots.

published in the literature that were *trained and tested* on the same database. For example, the performance of our model trained using the MROS and SHHS databases achieved better performance on the MESA database than Refs. [30, 32] that were trained on MESA. Such results demonstrate the high generalizability of our model to other types of data and different populations, especially when the SHHS is incorporated in the training phase (which comprises 60% of the total number of records across all databases). Results tested on SHHS data without including it in the training phase achieved relatively lower accuracy and F1-score (ranging between 60% and 85%), whereas testing on MESA and MROS databases had a higher performance, with an accuracy and F1-score ranging between 80% and 90% for most combinations of channel inputs.

In general, the more channels available as inputs to the model, the better the (out-of-distribution) performance of our method (Fig 3). As an exception, however, the best out-of-distribution performance on the SHHS database was obtained when using a combination of the AbdoRes, ThorRes and SpO_2 channels, achieving an accuracy of 81.60% and F1-Score of 81.60%. In this case, results employing RRI and airflow data as input channels led to poor performance.

Continuous monitoring of sleep apnea. We expect that the developed algorithm can also be implemented for continuous (real-time) monitoring of sleep apnea events during PSG examinations, or as part of a home OSA detection system. To simulate a continuous monitoring scenario and investigate the number of true positive and false positive OSA events given by our detection algorithm, a retrospective analysis is conducted with data previously collected in the MESA, MROS, and SHHS studies. We evaluate the performance of our OSA detection algorithm when applied to the full record of a patient sleep (comprising of time-series data of 8.78h, on average). Fig. 4 shows the overall performance across all patients. As in a real-time monitoring scenario, a moving window is implemented to consecutively sample short segments of data from the measured physiological signals (Fig. 4a, top), which are then fed to the hybrid model to perform the classification task of OSA and normal events. The hybrid model outputs the probability of a sampled data

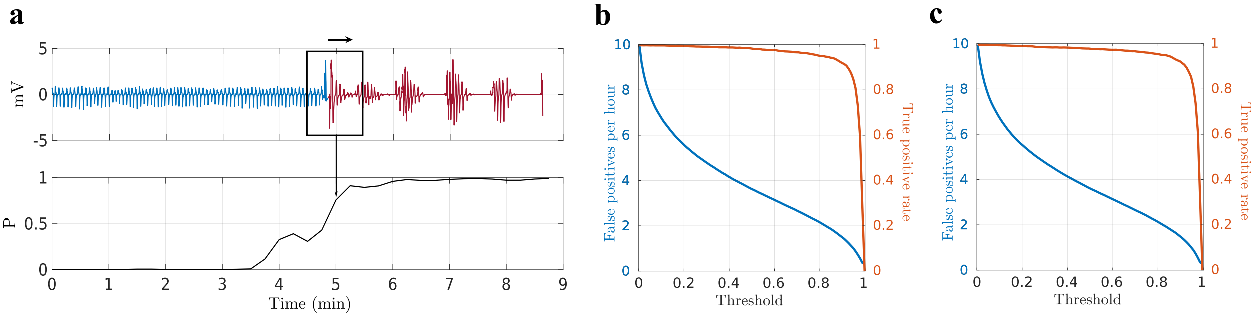

Figure 4: Continuous monitoring of sleep apnea events. (a) A moving window continuously samples short segments of time-series data (from all sensor channels) during a sleep study (top) and feeds them to the hybrid model, which outputs the probability of an OSA event (bottom). (b,c) True-positive-rate and number of false positives per hour as a function of the classification threshold on the (b) validation set and (c) test set.

belonging to the OSA class, which changes over time depending how far or close a patient is to an OSA event (Fig. 4a, bottom).

For the performance evaluation, we choose one of the trained models in the ten-fold cross-validation (Table S7), considering that all input channels are available. To tune the model for this task, we strive to maximize the average *ratio* between the true-positive rate and the number of false positives per hour across all records in the validation set. This procedure is designed as follows. First, for each patient's record, sequential segments of 60s are sampled consecutively with time steps of 15s and fed to the hybrid model, which computes the probability of each sample belonging to the OSA or normal class (Fig. 4a). Second, a parameter search is conducted to find the optimal classification threshold that maximizes the ratio in the validation set. Fig. 4b shows the true-positive rate and the number of false positives per hour as a function of the classification threshold. The threshold is set as 0.98, which maximizes the ratio to 0.94.

After selecting the optimal threshold on the validation set, we evaluate the performance of the model on the testing set for continuous monitoring (Fig. 4c). We obtained an average number of false positives (per hour) of 0.34 and a true-positive rate of 0.51 across all records. The results are meaningful, considering that the average Apnea-Hypopnea index (AHI) of the patients is 23.2 events per hour. Fig. 4c shows that there is an inverse trade-off between the true-positive rate and the number of false positives per hour. This opens up the possibility of tuning the model in a patient-specific manner, personalizing the choice of threshold according to the severity of AHI for each given patient.

Discussion

In this paper, we introduced a method for detecting OSA events using RRI, respiratory signals, and SpO_2 data. We showed that a deep neural network could be effectively trained using an extensive database to extract relevant features from raw physiological data. The detection was enhanced by combining it with a LightGBM classifier to obtain state-of-the-art performance.

We investigated the idea of combining respiratory, RRI, and SpO_2 signals rather than using them independently. Our results suggest that the combination of multiple input sources produced better results than the use of individual sources. We examine individual performance and ten combinations of these signals, as shown in Table S1. The individual airflow signal can perform better than the individual RRI and SpO_2 signals. However, combining these signals significantly improved performance.

The classification performance of the model obtained from the four datasets is consistent, as evidenced by a high area under the receiver operator characteristic curve, as well as the precision and recall curve, with values of 0.96 and 0.97, respectively, indicating that our model is not biased towards any class. To generate a robust detection method and prevent overfitting of the model, we employ a 10-fold cross-validation strategy.

The proposed method demonstrated robust generalization as measured by the out of distributions performance. The best results were obtained when testing on MESA and MROS, with an accuracy of 87.98% and 90.23%, respectively. The proposed method outperformed previous studies that trained and tested performance in the same databases (Table 2). When testing on the SHHS, the performance dropped to 81.60%, which can be attributed to the fact that 60% of the total available records come from this database and were excluded. Therefore, the number of samples for training is significantly reduced. Despite this, the best results were obtained when only using AbdoRes, ThorRes, and SpO_2 were used. This performance is still comparable to those obtained in the literature (see Table 2), although the databases were not used for training.

Table 2 compares the proposed method with other studies conducted using similar databases and more than 100 PSG recordings. As shown in the table, our proposed method outperformed previous studies. Haidar et al. [30] used a 1D CNN to classify apnea-hypopnea events from the MESA database; using raw airflow data, they achieved a 74% accuracy al. [31] used the SHHS data set for the detection of sleep apnea, using LSTM to automatically learn and extract relevant features, as well as detect possible sleep apnea events from raw physiological respiratory signals, with a 70% accuracy. McCloskey et al. [32] employed only MESA to perform a multiclass classification (normal, apnea, and hypopnea) and used a 2D CNN to achieve an accuracy of 77% by examining nasal airflow spectrograms. Banluesombatkul et al. [33] combined CNN1D and LSTM to detect extremely severe OSA patients in normal subjects using ECGs from the MrOS dataset. They achieved a 79% accuracy rate. Sharma et al. [34] used the SHHS dataset to detect OSA events. They trained a GentleBoost using features of the frequency domain of pulse oximetry (SpO_2) and respiratory data. Obtaining a 70% accuracy.

The proposed method outperformed the previous state-of-the-art method that used the same physiological signals by 8.3% and is the first of its type to consider a large multicenter cohort. Furthermore, our method is simpler than others because we only used raw physiological data to detect OSA events, which requires minimal complexity. As a result, it could be used as part of a home sleep monitoring system and perform continous monitoring in real-time.

The fact that the data used in this study were collected from controlled sleep laboratories is one of the study's weaknesses; in the future, similar research should be performed from the data collected in home settings.

Methods

Data specification. The databases are available in European Data Format (EDF) [35] and XML files that have been annotated for each sleep stage using the Rechtschaffen & Kales (R&K) criteria [36]. The XML file includes annotations for every 30-second sample of sleep stages and instances of sleep problems. Databases are summarized as follows.

1. The MrOS Sleep Study: MrOS is a substudy of the Men Study of Osteoporotic Fractures. Between 2000 and 2002, a baseline evaluation was performed on 594 elderly men 65 years of age or older taken from six clinical institutions. Between December 2003 and March 2005, a total of 3,135 of these participants were subjected to complete unattended polysomnography and 3 to 5-day actigraphy tests as part of the sleep study. The purpose of the Sleep Study was to determine the extent to which sleep disorders are linked to adverse health outcomes, such as the increased risk of death, fractures, falls, and cardiovascular disease.
2. The MESA Sleep Study: The Multi-Ethnic Study of Atherosclerosis (MESA) is a collaborative 6-center longitudinal investigation of factors associated with the development and progression of subclinical to clinical cardiovascular disease in 6,814 black, white, Hispanic, and Chinese American men and women aged 45 to 84 years in 2000-2002. The participants were all between the ages of 45 and 84 at the time of the study. Four follow-up examinations have been performed, one each in the years 2003–2004, 2004–2005, 2005–2007, and 2010–2011. Furthermore, 227 people participated in a Sleep Exam conducted by MESA Sleep between 2010 and 2012. This exam included an unattended overnight polysomnogram, wrist-worn actigraphy for seven days, and a sleep questionnaire. The sleep study aims to determine whether there is a correlation between subclinical atherosclerosis and sex, ethnicity, or other demographic differences in sleep and sleep disorders.
3. The SHHS Sleep Study: The Sleep Heart Health Study is a multicenter cohort study conducted by the National Institute of Heart, Lung, and Blood. The purpose of the study was to investigate the effects of sleep-disordered breathing on the cardiovascular system and also on other aspects of a person's health and to determine whether breathing problems that occur during sleep are related to an increased risk of coronary heart disease, stroke, death from all causes, and hypertension. Between November 1995, and January 1998, SHHS Visit 1 research focused on participants who were at least 40 years old and included 6,441 men and women. A second polysomnogram, known as SHHS Visit 2, was performed on 3,295 participants during the third exam cycle (January 2001 to June 2003).

Data pre-processing and model training. First, data is split by measurement channel (RRI, SpO_2, etc) and segmented into 60s windows of "normal" and "OSA" events. Data windows correspond to the same time instance across all channels (Fig. 1a). Second, all signals are standardized and normalized by calculating their z-scores and applying min-max normalization to eliminate the bias of mean and variance of the raw one-dimensional signal and speed up training [37]. Normalized samples of 60s from each channel are then fed in parallel into distinct deep CNNs. Each DNN is independently trained on a specific channel and then used for automatic feature extraction (Fig. 1b). Once the networks have been trained, features are extracted and concatenated to integrate information from different physiological biomarkers. This is achieved by using the concatenated features to train a LightGBM classifier for binary classification between normal and OSA events (Fig. 1c).

Neural network architecture and training. The CNNs were trained and cross-validated in 949,428 samples from 10,880 PSG recordings. We use the EfficientNetV2 architecture, a deep CNN with 479 layers developed by Google in 2021 [38]. It is a modified and optimized version of EfficientNet [39], a popular image classification algorithm that won the ImageNet 2019 competition [40]. The architecture used in this paper has been modified to handle unidimensional data and perform binary classification. Each CNN was independently trained using raw unidimensional physiological data from pulse oximetry (SpO2), electrocardiogram (ECG), thoracic movement (ThorRes), abdominal movement (AbdoRes), or airflow. Categorical cross-entropy was used as the loss function, ADAM as the optimizer [41], and stochastic gradient descent as the objective function optimizer [42]. If the validation loss did not decrease after eight consecutive epochs, the training was terminated. Once the networks were trained, the final layer was removed, and the last global average pooling layer is used to yield 1,280 features from each data channel.

Feature classifier. After the neural networks are trained, and features are extracted, the next step is to concatenate the features extracted by all CNNs and use them for training a LightGBM classifier. In contrast to a large number of other well-known algorithms, such as XGBoost [43] and GBDT [44], LightGBM employs the classification algorithm for growing trees in a leaf-wise manner rather than in a depth-wise manner. The leaf-wise algorithm can converge significantly more quickly than the depth-wise growth method, although its growth can be subject to overfitting if the appropriate hyperparameters are not used [29]. We use a random search method within a specified set of parameters to optimize the training and performance of the LightGBM. This allows a fixed number of parameters from a particular distribution to be sampled instead of testing all the values of potential parameters [45].

Models comparison. We compared the performance of the proposed method with two other CNN architectures that serve as a benchmark, the LSTM and Resnet [46], which are commonly used for classification tasks related to the detection of apnea and hypopnea events [11]. The EfficientNet architecture outperformed Resnet and LSTM by 10% and 28%, respectively (Table S8). We also compared the performance of various classifiers on the automatically extracted features by EfficientNet. The LightGBM classifier achieved the best performance, with a 15% improvement compared to a logistic regression (Table S9). Furthermore, we compared the influence on the model performance for time-series windows with different lengths (ranging from 10s to 120s, see Table S6). A small window, such as 10 seconds, resulted in poor performance and could lead to information loss [47]. The performance did not improve when the window length increased from 60 to 120 seconds. Therefore, we fixed the window length to 60 seconds, a value commonly used in the literature [9–11]. Finally, we also evaluated the performance of a single neural network model with multiple sensor inputs, instead of multiple neural networks with a single sensor input each as in the proposed pipeline. Table S5 shows the performance metrics of this model for different combinations of sensor inputs. Performance decreased by 5.5% when using the five channels and 4.3% for the four channels.

Performance metrics. The performance of the models was evaluated based on the following metrics. Let true positive (TP) represent the number of OSA events that are accurately predicted. True negative (TN) represents the number of normal events that are accurately predicted. False negative (FN) represents the number of OSA events incorrectly predicted as normal events, and false positive (FP) represents the number of normal events incorrectly predicted as OSA events. The accuracy of the model is given by $(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})$, indicating the probability of correctly identifying OSA and normal events; the sensitivity is given by $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, indicating the probability of identifying OSA events; the specificity is given by $\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$, indicating the probability of detecting OSA events, the precision is given by $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, indicating the ratio of patients undergoing OSA among all OSA cases; and the F1-score $2(\mathrm{precision}\times\mathrm{sensitivity})/(\mathrm{precision}+\mathrm{sensitivity})$, indicating the harmonic mean of precision and sensitivity.

Acknowledgements The authors want to thank the support from the Luxembourg National Research Fund (FNR) through PRIDE15/10907093/CriTiCS.

Contributors. J.G. conceptualized the research; M.E.G., J.G., and A.N.M. developed the AI model; M.E.G. implemented the codes and validated the results; M.E.G., J.G., and A.N.M. analyzed the results; J.G. supervised the work; M.E.G., A.N.M., and J.G. wrote and revised the manuscript.

Data availability. The data was provided by the National Sleep Research Resource and are publicly available on request at `https://sleepdata.org/`.

Code availability. Data pre-processing and segmentation were implemented using MATLAB software. The neural network was implemented on the Keras framework with Tensorflow backend on Python 3.7. The codes have been deposited on GitHub at .

References

[1] Chamara V Senaratna, Jennifer L Perret, Caroline J Lodge, Adrian J Lowe, Brittany E Campbell, Melanie C Matheson, Garun S Hamilton, and Shyamali C Dharmage. "Prevalence of obstructive sleep apnea in the general population: a systematic review". In: *Sleep medicine reviews* 34 (2017), pp. 70–81.

[2] Patrick Lévy, Malcolm Kohler, Walter T McNicholas, Ferran Barbé, R Doug McEvoy, Virend K Somers, Lena Lavie, and Jean-Louis Pépin. "Obstructive sleep apnoea syndrome". In: *Nature reviews Disease primers* 1.1 (2015), pp. 1–21.

[3] Massimo R Mannarino, Francesco Di Filippo, and Matteo Pirro. "Obstructive sleep apnea syndrome". In: *European journal of internal medicine* 23.7 (2012), pp. 586–593.

[4] Virend K Somers, David P White, Raouf Amin, William T Abraham, Fernando Costa, Antonio Culebras, Stephen Daniels, John S Floras, Carl E Hunt, Lyle J Olson, 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)". In: *Journal of the American College of Cardiology* 52.8 (2008), pp. 686–717.

[5] Naresh M Punjabi. "The epidemiology of adult obstructive sleep apnea". In: *Proceedings of the American Thoracic Society* 5.2 (2008), pp. 136–143.

[6] Walter T McNicholas. "Diagnosis of obstructive sleep apnea in adults". In: *Proceedings of the American thoracic society* 5.2 (2008), pp. 154–160.

[7] Kimberly N Hutchison, Yanna Song, Lily Wang, and Beth A Malow. "Analysis of sleep parameters in patients with obstructive sleep apnea studied in a hospital vs. a hotel-based sleep center". In: *Journal of clinical sleep medicine* 4.2 (2008), pp. 119–122.

[8] Dai Yumino, Takatoshi Kasai, Derek Kimmerly, Vinoban Amirthalingam, John S Floras, and T Douglas Bradley. "Differing effects of obstructive and central sleep apneas on stroke volume in patients with heart failure". In: *American journal of respiratory and critical care medicine* 187.4 (2013), pp. 433–438.

[9] Anita Ramachandran and Anupama Karuppiah. "A survey on recent advances in machine learning based sleep apnea detection systems". In: *Healthcare*. Vol. 9. 7. MDPI. 2021, p. 914.

[10] Fabio Mendonca, Sheikh Shanawaz Mostafa, Antonio G Ravelo-Garcia, Fernando Morgado-Dias, and Thomas Penzel. "A review of obstructive sleep apnea detection approaches". In: *IEEE journal of biomedical and health informatics* 23.2 (2018), pp. 825–837.

[11] Sheikh Shanawaz Mostafa, Fábio Mendonça, Antonio G. Ravelo-Garcıa, and Fernando Morgado-Dias. "A systematic review of detecting sleep apnea using deep learning". In: *Sensors* 19.22 (2019), p. 4934.

[12] Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. "Deep learning for healthcare applications based on physiological signals: A review". In: *Computer methods and programs in biomedicine* 161 (2018), pp. 1–13.

[13] Jiliang Tang, Salem Alelyani, and Huan Liu. "Feature selection for classification: A review". In: *Data classification: Algorithms and applications* (2014), p. 37.

[14] Nagarajan Ganapathy, Ramakrishnan Swaminathan, and Thomas M Deserno. "Deep learning on 1-D biosignals: a taxonomy-based survey". In: *Yearbook of medical informatics* 27.01 (2018), pp. 098–109.

[15] Thomas M Bury, RI Sujith, Induja Pavithran, Marten Scheffer, Timothy M Lenton, Madhur Anand, and Chris T Bauch. "Deep learning for early warning signals of tipping points". In: *Proceedings of the National Academy of Sciences* 118.39 (2021), e2106140118.

[16] Sheena Angra and Sachin Ahuja. "Machine learning and its applications: A review". In: *2017 international conference on big data analytics and computational intelligence (ICBDAC)*. IEEE. 2017, pp. 57–60.

[17] Raghavendra Chalapathy and Sanjay Chawla. "Deep learning for anomaly detection: A survey". In: *arXiv preprint arXiv:1901.03407* (2019).

[18] Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M.M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P.S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Meira Wagner, Thomas B. Schön, and Antonio Luiz P. Ribeiro. "Automatic diagnosis of the 12-lead ECG using a deep neural network". In: *Nature Communications* 11 (2020), p. 1760.

[19] Hyojeong Lee, Soo-Yong Shin, Myeongsook Seo, Gi-Byoung Nam, and Segyeong Joo. "Prediction of ventricular tachycardia one hour before occurrence using artificial neural networks". In: *Scientific reports* 6.1 (2016), pp. 1–7.

[20] Sami Nikkonen, Isaac O Afara, Timo Leppänen, and Juha Töyräs. "Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea". In: *Scientific reports* 9.1 (2019), pp. 1–9.

[21] Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M Aarts. "Sleep stage classification from heart-rate variability using long short-term memory neural networks". In: *Scientific reports* 9.1 (2019), pp. 1–11.

[22] Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L Jackson, Michelle A Williams, and Susan Redline. "Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA)". In: *Sleep* 38.6 (2015), pp. 877–888.

[23] Terri Blackwell, Kristine Yaffe, Sonia Ancoli-Israel, Susan Redline, Kristine E Ensrud, Marcia L Stefanick, Alison Laffan, Katie L Stone, and Osteoporotic Fractures in Men Study Group. "Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study". In: *Journal of the American Geriatrics Society* 59.12 (2011), pp. 2217–2225.

[24] Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. "The sleep heart health study: design, rationale, and methods". In: *Sleep* 20.12 (1997), pp. 1077–1085.

[25] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. "The National Sleep Research Resource: towards a sleep data commons". In: *Journal of the American Medical Informatics Association* 25.10 (2018), pp. 1351–1358.

[26] Jiapu Pan and Willis J Tompkins. "A real-time QRS detection algorithm". In: *IEEE Transactions on Biomedical Engineering* 3 (1985), pp. 230–236.

[27] Hooman Sedghamiz. "Matlab implementation of Pan Tompkins ECG QRS detector". In: *Code Available at the File Exchange Site of MathWorks* (2014).

[28] Robert Keys. "Cubic convolution interpolation for digital image processing". In: *IEEE transactions on acoustics, speech, and signal processing* 29.6 (1981), pp. 1153–1160.

[29] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017).

[30] Rim Haidar, Irena Koprinska, and Bryn Jeffries. "Sleep apnea event detection from nasal airflow using convolutional neural networks". In: *International Conference on Neural Information Processing*. Springer. 2017, pp. 819–827.

[31] Tom Van Steenkiste, Willemijn Groenendaal, Dirk Deschrijver, and Tom Dhaene. "Automated sleep apnea detection in raw respiratory signals using long short-term memory neural networks". In: *IEEE journal of biomedical and health informatics* 23.6 (2018), pp. 2354–2364.

[32] Stephen McCloskey, Rim Haidar, Irena Koprinska, and Bryn Jeffries. "Detecting hypopnea and obstructive apnea events using convolutional neural networks on wavelet spectrograms of nasal airflow". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2018, pp. 361–372.

[33] Nannapas Banluesombatkul, Thanawin Rakthanmanon, and Theerawit Wilaiprasitporn. "Single channel ECG for obstructive sleep apnea severity detection using a deep learning approach". In: *TENCON 2018-2018 IEEE Region 10 Conference*. IEEE. 2018, pp. 2011–2016.

[34] Manish Sharma, Divyash Kumbhani, Jainendra Tiwari, T Sudheer Kumar, and U Rajendra Acharya. "Automated detection of obstructive sleep apnea in more than 8000 subjects using frequency optimized orthogonal wavelet filter bank with respiratory and oximetry signals". In: *Computers in Biology and Medicine* 144 (2022), p. 105364.

[35] Bob Kemp and Jesus Olivan. "European data format 'plus'(EDF+), an EDF alike standard format for the exchange of physiological data". In: *Clinical neurophysiology* 114.9 (2003), pp. 1755–1761.

[36] Allan Rechtschaffen. "A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects". In: *Brain information service* (1968).

[37] Lei Huang, Jie Qin, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. "Normalization techniques in training dnns: Methodology, analysis and application". In: *arXiv preprint arXiv:2009.12836* (2020).

[38] Mingxing Tan and Quoc Le. "Efficientnetv2: Smaller models and faster training". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10096–10106.

[39] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[40] Kaiming He, Ross Girshick, and Piotr Dollár. "Rethinking imagenet pre-training". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4918–4927.

[41] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[42] Grégoire Montavon, Geneviève Orr, and Klaus-Robert Müller. *Neural networks: tricks of the trade*. Vol. 7700. springer, 2012.

[43] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* 1.4 (2015), pp. 1–4.

[44] Jerry Ye, Jyh-Herng Chow, Jiang Chen, and Zhaohui Zheng. "Stochastic gradient boosted distributed decision trees". In: *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009, pp. 2061–2064.

[45] James Bergstra and Yoshua Bengio. "Random search for hyper-parameter optimization." In: *Journal of machine learning research* 13.2 (2012).

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[47] K Hauke Kraemer, Reik V Donner, Jobst Heitzig, and Norbert Marwan. "Recurrence threshold selection for obtaining robust recurrence characteristics in different embedding dimensions". In: *Chaos* 28.8 (2018), p. 085720.

SUPPLEMENTARY MATERIAL

A PREPRINT

Marino E. Gavidia[1,*], Arthur N. Montanari[1], and Jorge Goncalves[1,2,*]

[1]Luxembourg Center for Systems Biomedicine, University of Luxembourg, Belvaux L-4367, Luxembourg
[2]Department of Plant Sciences, Cambridge University, Cambridge CB2 3EA, United Kingdom
[*]Corresponding authors: marino.gavidia@uni.lu; jmg77@cam.ac.uk

September 28, 2022

Table S1: Performance of the proposed method on all databases.

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO$_2$	0.9113	0.8999	0.9235	0.9257	0.9126
AbdoRes+ThorRes+Airflow+SpO$_2$	0.9104	0.8990	0.9225	0.9247	0.9117
RRI+AbdoRes+ThorRes+Airflow	0.9059	0.8949	0.9174	0.9197	0.9071
RRI+AbdoRes+ThorRes+SpO$_2$	0.8858	0.8746	0.8977	0.9008	0.8875
RRI+AbdoRes+ThorRes	0.8716	0.8625	0.8812	0.8842	0.8732
AbdoRes+ThorRes+SpO$_2$	0.8841	0.8729	0.8959	0.8990	0.8858
AbdoRes+SpO$_2$	0.8769	0.8681	0.8862	0.8889	0.8784
ThorRes+SpO$_2$	0.8586	0.8511	0.8664	0.8692	0.8586
AbdoRes+ThorRes	0.8657	0.8568	0.8750	0.8781	0.8673
RRI +SpO$_2$	0.8221	0.8196	0.8247	0.8260	0.8228
SpO$_2$	0.8096	0.8087	0.8105	0.8110	0.8099
Airflow	0.8792	0.8685	0.8904	0.8936	0.8809
ThorRes	0.8131	0.8060	0.8206	0.8248	0.8153
AbdoRes	0.8463	0.8430	0.8497	0.8511	0.8470
RRI	0.6657	0.6692	0.6624	0.6554	0.6623

Table S2: Performance of the proposed method on the MESA database.

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO_2	0.8798	0.8414	0.9280	0.9361	0.8862
AbdoRes+ThorRes+Airflow+SpO_2	0.8766	0.8405	0.9214	0.9297	0.8828
RRI+AbdoRes+ThorRes+Airflow	0.8732	0.8363	0.9191	0.9280	0.8798
RRI+AbdoRes+ThorRes+SpO_2	0.8537	0.8236	0.8900	0.9003	0.8602
RRI+AbdoRes+ThorRes	0.8516	0.8375	0.8670	0.8725	0.8546
AbdoRes+ThorRes+SpO_2	0.8446	0.8141	0.8817	0.8932	0.8518
AbdoRes+SpO_2	0.8497	0.8101	0.9009	0.9135	0.8587
ThorRes+SpO_2	0.8236	0.7976	0.8547	0.8674	0.8310
AbdoRes+ThorRes	0.8468	0.8408	0.8531	0.8557	0.8482
RRI +SpO_2	0.8516	0.8375	0.8670	0.8725	0.8546
SpO_2	0.8148	0.7865	0.8493	0.8641	0.8235
Airflow	0.8424	0.7815	0.9371	0.9507	0.8578
ThorRes	0.7827	0.7850	0.7806	0.7788	0.7819
AbdoRes	0.8355	0.8139	0.8602	0.8698	0.8409
RRI	0.6892	0.6717	0.7107	0.7403	0.7043

Table S3: Performance of the proposed method on the MROS database.

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO_2	0.9023	0.8552	0.9638	0.9686	0.9084
AbdoRes+ThorRes+Airflow+SpO_2	0.9018	0.8546	0.9637	0.9685	0.9080
RRI+AbdoRes+ThorRes+Airflow	0.8934	0.8433	0.9605	0.9662	0.9006
RRI+AbdoRes+ThorRes+SpO_2	0.8866	0.8393	0.9493	0.9564	0.8940
RRI+AbdoRes+ThorRes	0.7128	0.8149	0.6187	0.6633	0.7313
AbdoRes+ThorRes+SpO_2	0.8854	0.8365	0.9508	0.9579	0.8931
AbdoRes+SpO_2	0.8776	0.8284	0.9442	0.9526	0.8862
ThorRes+SpO_2	0.8628	0.8169	0.9242	0.9352	0.8721
AbdoRes+ThorRes	0.8649	0.8090	0.9454	0.9553	0.8761
RRI +SpO_2	0.8239	0.7844	0.8762	0.8934	0.8354
SpO_2	0.8161	0.7815	0.8603	0.8775	0.8267
Airflow	0.8517	0.8189	0.8921	0.9032	0.8590
ThorRes	0.8129	0.7623	0.8876	0.9092	0.8293
AbdoRes	0.8487	0.7925	0.9316	0.9447	0.8619
RRI	0.6555	0.6542	0.6568	0.6596	0.6569

Table S4: Performance of the proposed method on the SHHS database.

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO$_2$	0.7690	0.9301	0.6957	0.5817	0.7158
AbdoRes+ThorRes+Airflow+SpO$_2$	0.7678	0.9266	0.6952	0.5817	0.7147
RRI+AbdoRes+ThorRes+Airflow	0.7594	0.9248	0.6867	0.5647	0.7012
RRI+AbdoRes+ThorRes+SpO$_2$	0.8178	0.8790	0.7736	0.7371	0.8018
RRI+AbdoRes+ThorRes	0.7983	0.8521	0.7587	0.7218	0.7816
AbdoRes+ThorRes+SpO$_2$	0.8160	0.8754	0.7728	0.7369	0.8002
AbdoRes+SpO$_2$	0.8063	0.8749	0.7589	0.7148	0.7868
ThorRes+SpO$_2$	0.7997	0.8402	0.7679	0.7402	0.7871
AbdoRes+ThorRes	0.7944	0.8453	0.7565	0.7206	0.7780
RRI +SpO$_2$	0.7848	0.8252	0.7533	0.7226	0.7705
SpO$_2$	0.8302	0.9262	0.7694	0.7175	0.8086
Airflow	0.7257	0.8990	0.6574	0.5086	0.6496
ThorRes	0.7466	0.7794	0.7206	0.6878	0.7307
AbdoRes	0.7767	0.8391	0.7337	0.6847	0.7541
RRI	0.6416	0.6561	0.6296	0.5952	0.6242

Table S5: Performance of a single neural network model with multiple input channels.

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO$_2$	0.8607	0.8493	0.8728	0.8770	0.8629
RRI+AbdoRes+ThorRes+Airflow	0.8715	0.8603	0.8834	0.8870	0.8735
AbdoRes+ThorRes+SpO$_2$	0.8392	0.8342	0.8444	0.8467	0.8404
RRI+AbdoRes+ThorRes	0.8428	0.8380	0.8478	0.8500	0.8428
AbdoRes+ThorRes	0.8513	0.8413	0.8620	0.8661	0.8535
RRI +SpO$_2$	0.6974	0.8071	0.6455	0.5189	0.6317

Table S6: Performance for different lengths of the sampling window.

Length (seconds)	Accuracy
10	0.82
20	0.89
30	0.90
60	0.91
120	0.91

Table S7: 10-fold cross-validation accuracy on the testing sets after training the EfficientNetV2 and the Light Gradient Boosting Machine using all channels.

Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Average
0.91	0.91	0.91	0.90	0.91	0.90	0.90	0.91	0.90	0.90	0.91

Table S8: Performance for different network architectures.

Model	Accuracy
LSTM	0.71
1D Resnet	0.82
1D EfficientNetV2	0.91

Table S9: Performance for different classifiers.

Model	Accuracy
Logistic regression	0.79
Support vector machine	0.80
XGBoost	0.81
Random forest	0.90
Light Gradient Boosting Machine	0.91