DETECTION OF OBSTRUCTIVE SLEEP APNEA FROM WEARABLE PHYSIOLOGICAL DEVICES

A PREPRINT

Marino E. Gavidia[1,*], Arthur N. Montanari[1,2], and Jorge Goncalves[1,3,*]

[1]Luxembourg Center for Systems Biomedicine, University of Luxembourg, Belvaux L-4367, Luxembourg
[2]Department of Physics and Astronomy, Northwestern University, Evanston, IL 60208, USA
[3]Department of Plant Sciences, Cambridge University, Cambridge CB2 3EA, United Kingdom
*Corresponding authors: marino.gavidia@uni.lu; jmg77@cam.ac.uk

February 15, 2023

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 and uncomfortable for patients. To circumvent these issues, we propose a detection method of OSA events, named DRIVEN, using only two signals that can be easily measured at home: abdominal movement and pulse oximetry. On test data, DRIVEN achieves an accuracy and F1-score of 88%, a reasonable trade-off between the model's performance and patient's comfort. We use data from three sleep studies from the National Sleep Research Resource (NSRR), the largest public repository, consisting of 10,878 recordings. DRIVEN is based on a combination of deep convolutional neural networks and a light-gradient-boost machine for classification. Since DRIVEN is simple and computationally efficient, we expect that it can be implemented for automatic detection of OSA in unsupervised long-term 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 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 [2]. As age and the likelihood of having sleep apnea are positively correlated [4], 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 [5]. 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 [4]. PSG may also not be suitable to assess the severity of OSA because patients are monitored for only one night in a strange and uncomfortable sleep laboratory [6]. Hence, the development of portable, easy-to-use, reliable, and affordable OSA monitoring tools for home care applications is crucial to improve patient care.

The automatic detection of OSA, considered the most severe and common type of sleep apnea [7], is a pressing problem in the literature [8–10]. 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 [11]. The procedure can be hard, time consuming, and often subject to bias or lack of generalizability [12, 13]. Deep learning, on the other hand, has shifted data modeling away from "expert-driven" feature engineering toward "data-driven" feature extraction [10, 14]. Widespread across many applications such as computer vision, natural language processing, and speech recognition [15, 16], deep learning methods have also been successfully adopted in biomedical applications, including for the detection of diseases such as atrial fibrillation [17], prediction of ventricular tachycardia [18], oxygen desaturation index estimation [19], and sleep stage classification [20].

This paper proposes a data-driven method, named DRIVEN for Detection of obstRuctIVe slEep apNea, based on deep learning to detect OSA events from raw physiological data. We focus on devices that can be implemented in a home care environment and were available in the considered datasets, including abdominal and thoracic pressure, nasal airflow, oxygen saturation, and R-to-R interval data. Our study employs 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. In contrast, most previous works [8–10] used small datasets with 100 or fewer recordings, which can lead to poorly generalizable models. DRIVEN achieved the best classification performance available in the literature using data from all five sensors. Even in the absence of certain physiological signals as input data, DRIVEN's performance remained fairly high, achieving a good trade-off between the model's performance and the patient's comfort (related to the number and type of wearable devices used during sleep). Cross-validation on out-of-sample data shows that DRIVEN 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 DRIVEN, 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) [21], the Men Study of Osteoporotic Fracture (MrOS) [22], and the Sleep Heart Health Research (SHHS) [23]. These datasets are publicly available from the National Sleep Research Resource for Sleep-Related Studies (NSRR) [24]. In total, 14,370 PSG recordings were considered in our study. Each PSG recording consists of data collected from several physiological sensors during a patient's sleep (lasting 8.78h on average). The PSG recordings were collected and annotated during typical PSG evaluations in healthcare facilities, including labels of data segments corresponding to OSA events (see Methods for details).

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).

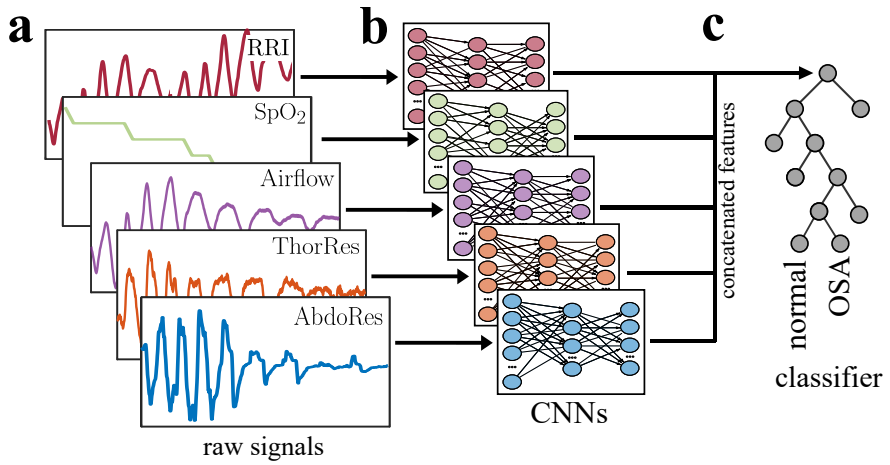

Figure 1: DRIVEN's pipeline. (a) Data are separated by channels (RRI, SpO_2, among others) and segmented into 60s windows. (b) For each channel, a distinct trained deep CNN extracts features (outputs) from the raw signal (input). (c) The extracted features are concatenated and fed to a LightGBM that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

PSG consists of a combination of sensors. However, some of these sensors are uncomfortable to wear. Hence, we considered 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. ECG data can also be hard to measure during sleep. Therefore, we replaced the ECG data with R-to-R interval (RRI) data, which can be readily obtained from common wearable devices such as smartwatches. The RRI data was 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 [25, 26]. Finally, data collected from various healthcare facilities had different sampling frequencies. These were normalized by resampling all channels to the maximum sampling frequency using the cubic spline as the interpolation method [27]; this method is simple to implement and, at the same time, it does not attenuate higher frequency components of the signal. Hence, all signals were standardized to the same length, which simplified the training of the neural networks.

PSG recordings were excluded from the study if they met at least one of the following criteria: (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.

DRIVEN: deep learning-model for the detection of OSA events. We developed a hybrid deep-learning model named DRIVEN for the classification of "normal" and "OSA" events from data. The inputs of DRIVEN 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. DRIVEN then learns whether a given input corresponds to a normal or an OSA event. Fig. 1 illustrates DRIVEN's pipeline. For each available physiological signal, a distinct deep convolutional neural network (CNN) is trained to 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) [28] 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 between OSA and normal samples. See Methods for details on data pre-processing and model training.

We trained and assessed DRIVEN on raw physiological signals. This avoids creating human-selected features, which can sometimes lead to bias. The performance was assessed via a 10-fold cross-validation strategy, separating data by patient record to avoid data leakage from samples taken from the same patient. The dataset is originally imbalanced with a ratio of 77 normal events per OSA event. To overcome the bias and inaccuracy associated with classification using imbalanced data between two classes, the minority and majority classes were oversampled and undersampled, respectively, to balance the number of OSA and normal samples from each record, which is a common strategy followed in the literature [29] . Labeled OSA 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 nearly 1 million balanced samples.

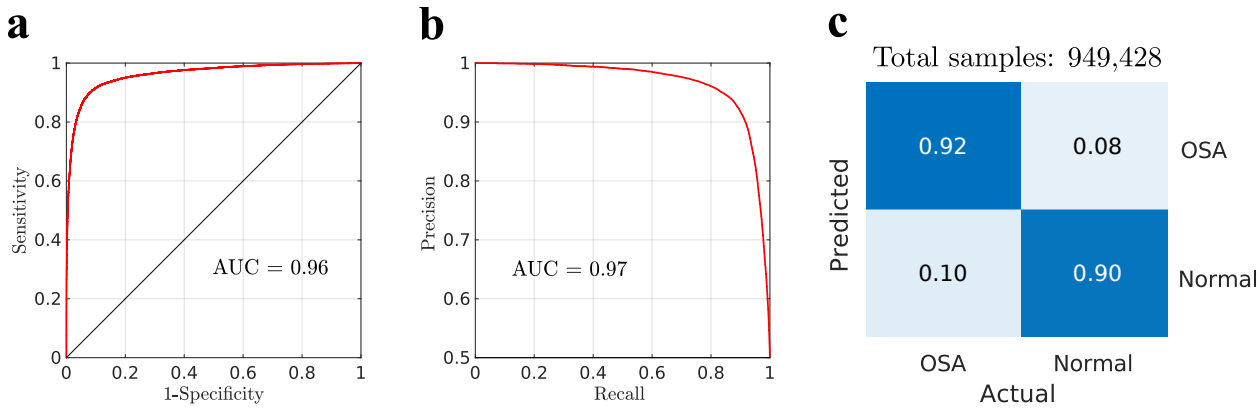

Figure 2: Performance of DRIVEN. (a) Receiver-operator characteristic curve. (b) Precision-recall curve. (c) Confusion matrix for a threshold of 0.5.

Performance of the hybrid model. This section evaluates the performance of DRIVEN in the classification task between normal and OSA events. Results are shown in Fig. 2 using all five physiological signals 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 (Fig. 2a,b), respectively, showing that the classification among OSA and normal classes is balanced. Fig. 2c shows the confusion matrix for the 10-folds on all 949,428 samples with a threshold of 0.5. Across all samples, the percentage of false negatives and false positives is under 10%.

The above results used all five measurement channels as inputs. However, it is likely that some physiological signals may contribute more than others in the classification task. Moreover, some physiological signals may not be available, or considered less comfortable to wear, in a homecare monitoring application of sleep apnea. Hence, we will next evaluate the performance of DRIVEN on different combinations of input channels (sensors). For each combination of input channels, the hybrid model is trained and tested as previously described.

Figure 3a shows the performance results for different sensor combinations. As expected, the best performance was obtained when all five channels were used in conjunction, and the performance slightly decreases in general with fewer input channels. This provides a good trade-off between the model's accuracy and the number and type of sensor devices worn by a patient, which are directly related to the patient's sleep comfort. Overall, models trained and tested with AbdoRes and/or ThorRes as inputs achieve higher performance. For instance, we note that with only two sensor devices (for AbdoRes and SpO_2 data), the model's accuracy and F1-score is around 88%, close to the top performance for all five sensors. On the other hand, performance seems to be less dependent on RRI data; the model trained on RRI data alone had the lowest performance, which may be related to the intrinsic time delay from the onset of an OSA event to the change of heart rhythm. Independently of the number and type of input channels, our results were always balanced in specificity and sensitivity.

Table 2 compares the performance of DRIVEN with previously published methods in OSA detection in the literature using the same databases and with at least 100 records. DRIVEN outperformed previous state-of-the-art work [30] when using similar sensors, with an accuracy of 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 S1). All other possible combinations of sensors are reported in table S2. 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 (Fig. 3a). This illustrates that the designed hybrid deep CNN classifier as well as the extensive dataset (consisting of three large databases) lead to a significant performance improvement while keeping its computational complexity and data pre-processing steps reasonably simple.

Independent test set performance. To evaluate how generalizable DRIVEN'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.

Figure 3b reports independent validation performances for each 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 databases, respectively. Moreover, for

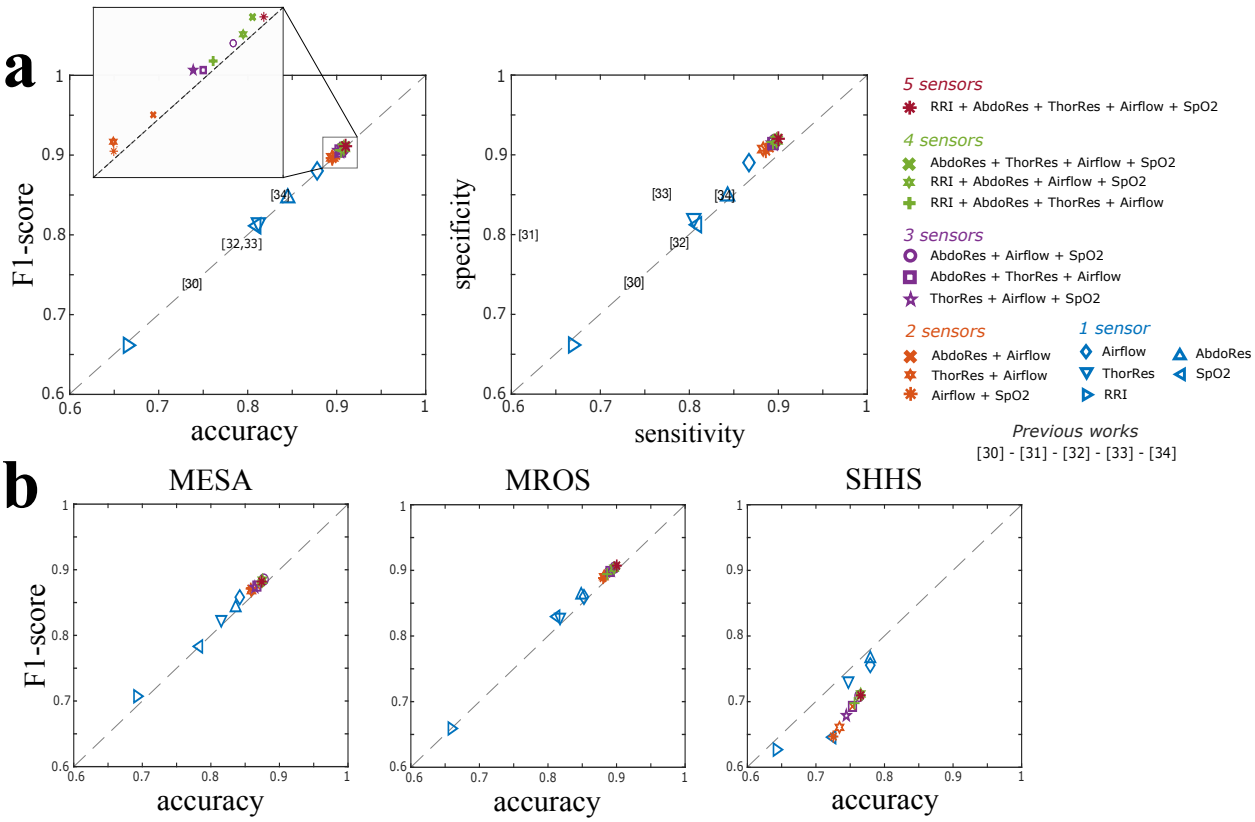

Figure 3: (a) Performance metrics of DRIVEN for different combinations of input channels (physiological signals). Performance metrics of previous works in the literature using one of the databases studied in this work are included for comparison purposes. (b) Performance of DRIVEN on independent test 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). Tables S2– S5 in Supplementary Material contain the specific values reported in these plots and all other possible sensor combinations.

Table 2: Comparison of DRIVEN 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	[31]	MESA	100	Airflow	CNN1D	0.74	0.74	0.74	0.74
2017	[32]	SHHS	2,100	AbdoRes+ThorRes	LSTM	0.71	0.62	0.80	0.48
2018	[33]	MESA	1,507	Airflow	CNN2D	0.79	0.79	0.79	0.79
2020	[34]	MROS	545	ECG	CNN1D-LSTM	0.79	0.77	0.85	0.79
2022	[30]	SHHS	8,444	ECG+AbdoRes+ThorRes+Airflow+SpO_2	Wavelet decomposition	0.84	0.84	0.85	0.85
2022	DRIVEN	All	10,880	RRI+AbdoRes+ThorRes+Airflow+SpO_2	Hybrid CNN classifier	0.91	0.90	0.92	0.91

most combinations of channel inputs, testing on MESA and MROS databases yielded accuracies and F1-scores ranging mostly between 80% and 90%, achieving better results than previous methods published in the literature that were trained and tested on the same database. Such results demonstrate the high generalizability of DRIVEN 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).

Removing the SHHS database from training resulted in a lower accuracy and F1-score (ranging between 60% and 85%). This difference in performance is likely due to the much lower number of available training samples and the fact that the sensors used in SHHS measure data at a lower rate; for instance, nasal airflow data was sampled at 10Hz in comparison to the sampling frequencies of 32Hz and 64Hz for the MESA and MROS databases, respectively.

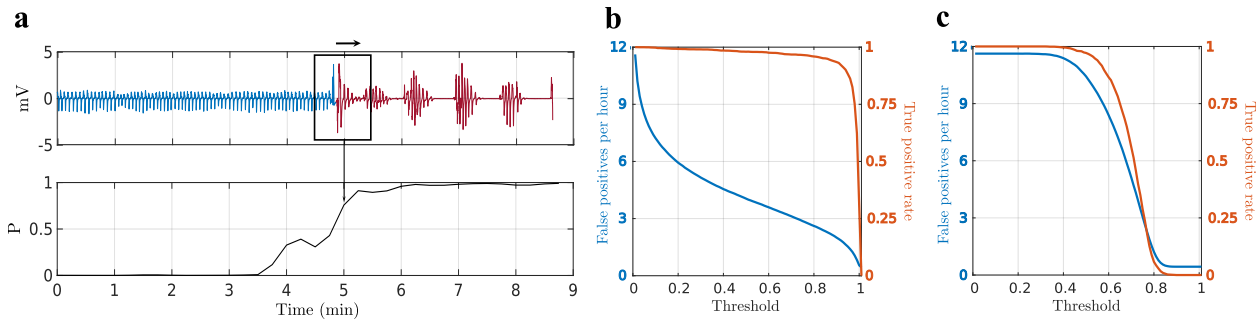

Figure 4: Automatic labeling of sleep apnea events from the MROS, MESA, and SHHS studies with an average Apnea-Hypopnea index (AHI) of 23.2 events per hour. (a) A moving window continuously samples short segments of time-series data (from all five channels) during a sleep study (top) and feeds them to 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 test set (b) using five channels and (c) only two sensors (AbdoRes and SpO_2).

In general, the more channels available as inputs to the model, the better the independent validation test performances of DRIVEN (Fig 3). As an exception, the best independent validation test performance on the SHHS database was obtained when leaving out airflow data, achieving an accuracy of 82% and an F1-Score of 80%. This could be due to the fact that around 60% of the total samples available for training are left out when using the SHHS database for testing, reducing the generalization capability of DRIVEN.

Automatic labeling of obstructive sleep apnea. The developed algorithm can also be used to automatically analyze whole sleep records to locate OSA events and provide statistics to patients and physicians. Moreover, with small modifications, it can also be used for real-time detection and alert of OSA events. To evaluate its performance, we conduct a retrospective analysis of all available full records from MESA, MROS, and SHHS studies, with an average duration of 8.8h. Fig. 4 shows the overall performance across all patients. A moving window is implemented to sample sequential segments of 60s, with time steps of 15s, 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 belonging to the OSA class, which changes over time depending on how far or close a patient is to an OSA event (Fig. 4a, bottom).

For performance evaluation, we randomly choose one of the trained models in the ten-fold cross-validation (Table S1), using only the AbdoRes + SpO_2 sensors since they have a reasonable trade-off between performance and patient comfort. Fig. 4b,c shows the true-positive rate and the number of false positives per hour as a function of the classification threshold when using five sensors and two sensors (AbdoRes and SpO_2), respectively. If we select the threshold 0.93 at the kink of the graph (from the validation set, Fig. S1) when using five sensors, we obtained an average number of false positives (per hour) of 1.71 and a true-positive rate of 0.91. When using only AbdoRes + SpO_2, with a threshold of 0.63, we obtained an average number of false positives (per hour) of 7.7 and a true-positive rate of 0.80 across all records in the testing set. In comparison, the average number of OSA events in these datasets is 23.2 events per hour. Fig. 4b,c 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 DRIVEN to specific patients, personalizing the choice of threshold according to patient requirements or OSA severity.

Discussion

This paper introduced a method, DRIVEN, for detecting OSA events using abdominal and thoracic pressure, nasal airflow, oxygen saturation, and R-to-R interval 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 a state-of-the-art performance of 91% accuracy and F1-score when using the five physiological signals. The best trade-off between model performance and patient comfort was achieved when combining data from only two sensors (abdominal movement and pulse oximetry), yielding an accuracy of 88%. In order to ensure the robustness of DRIVEN, we conducted a leave-one-out cross-validation on the MESA, MROS, and SHHS cohorts, yielding accuracies of 85%, 88%, and 81% respectively on abdominal movement and pulse oximetry data.

To test the automatic labeling capabilities of OSA events, we retrospectively analyze the sleep records from patients in the test set. Fig 4b,c shows that DRIVEN can correctly detect OSA events with a true positive rate of 0.91 and 1.71 false positives per hour when considering the five physiological signals, which is significantly less than the average of 23.2 OSA events. When using abdominal movement and pulse oximetry, DRIVEN yielded a true positive rate of 0.80 and 7.7 false positives per hour. The probability threshold can be tuned by doctors depending on the clinical application. Lower thresholds, for example, produce more sensitive models which might be employed for patients with severe OSA syndrome. Higher thresholds, on the other hand, produce more specific models with fewer false positives, making them ideal for monitoring healthy patients.

This result demonstrates the potential of DRIVEN for an automatic labeling application. Compared to a doctor's manual detection and diagnosis of OSA, computer-assisted signal analysis systems can reduce errors related to inter- and intra-operation variability and tiredness induced by the arduous annotating process [35]. Furthermore, most computer-based analyses can be conducted more cost-effectively and quickly [11]. Overall, DRIVEN could reduce the burden on doctors and costs to the healthcare system by making feasible the examination of patients through multiple nights from data that can be collected at home by wearable devices.

Methods

Data specification. The databases are available in European Data Format (EDF) [36] and XML files that have been annotated for each sleep stage using the Rechtschaffen & Kales (R&K) criteria [37]. 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 [38]. 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 [39]. It is a modified and optimized version of EfficientNet [40], a popular image classification algorithm that won the ImageNet 2019 competition [41]. 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), R-to-R intervals (RRI), thoracic movement (ThorRes), abdominal movement (AbdoRes), or airflow. Categorical cross-entropy was used as the loss function, ADAM as the optimizer [42], and stochastic gradient descent as the objective function optimizer [43]. 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 [44] and GBDT [45], 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 [28]. 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 [46].

Models comparison. We compared the performance of DRIVEN with two other CNN architectures that serve as a benchmark, the LSTM and Resnet [47], which are commonly used for classification tasks related to the detection of apnea and hypopnea events [10]. The Efficientnetv2 architecture outperformed Resnet and LSTM by 10% and 28%, respectively (Table S6). We also compared the performance of various classifiers on the automatically extracted features by Efficientnetv2. The LightGBM classifier achieved the best performance, with a 15% improvement compared to a logistic regression (Table S7). Furthermore, we compared the influence on the model performance for time-series windows with different lengths (ranging from 10s to 120s, see Table S8). A small window, such as 10 seconds, resulted in poor performance and could lead to information loss [48]. 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 [8–10]. 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 S9 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 are available in GitHub at `https://github.com/marinogb/osa_detection`.

Competing interests. The authors declare no competing financial or non-financial interests.

References

[1] Senaratna, C. V. *et al.* Prevalence of obstructive sleep apnea in the general population: a systematic review. *Sleep Medicine Reviews* **34**, 70–81 (2017).

[2] Riha, R. L. Defining obstructive sleep apnoea syndrome: a failure of semantic rules. *Breathe* **17** (2021).

[3] Mannarino, M. R., Di Filippo, F. & Pirro, M. Obstructive sleep apnea syndrome. *European Journal of Internal Medicine* **23**, 586–593 (2012).

[4] Punjabi, N. M. The epidemiology of adult obstructive sleep apnea. *Proceedings of the American Thoracic Society* **5**, 136–143 (2008).

[5] McNicholas, W. T. Diagnosis of obstructive sleep apnea in adults. *Proceedings of the American Thoracic Society* **5**, 154–160 (2008).

[6] Hutchison, K. N., Song, Y., Wang, L. & Malow, B. A. Analysis of sleep parameters in patients with obstructive sleep apnea studied in a hospital vs. a hotel-based sleep center. *Journal of Clinical Sleep Medicine* **4**, 119–122 (2008).

[7] Yumino, D. *et al.* Differing effects of obstructive and central sleep apneas on stroke volume in patients with heart failure. *American Journal of Respiratory And Critical Care Medicine* **187**, 433–438 (2013).

[8] Ramachandran, A. & Karuppiah, A. A survey on recent advances in machine learning based sleep apnea detection systems. In *Healthcare*, vol. 9, 914 (MDPI, 2021).

[9] Mendonca, F., Mostafa, S. S., Ravelo-Garcia, A. G., Morgado-Dias, F. & Penzel, T. A review of obstructive sleep apnea detection approaches. *IEEE Journal of Biomedical And Health Informatics* **23**, 825–837 (2018).

[10] Mostafa, S. S., Mendonça, F., G. Ravelo-García, A. & Morgado-Dias, F. A systematic review of detecting sleep apnea using deep learning. *Sensors* **19**, 4934 (2019).

[11] Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S. & Acharya, U. R. Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods And Programs In Biomedicine* **161**, 1–13 (2018).

[12] Tang, J., Alelyani, S. & Liu, H. Feature selection for classification: A review. *Data Classification: Algorithms And Applications* 37 (2014).

[13] Ganapathy, N., Swaminathan, R. & Deserno, T. M. Deep learning on 1-d biosignals: a taxonomy-based survey. *Yearbook of Medical Informatics* **27**, 098–109 (2018).

[14] Bury, T. M. *et al.* Deep learning for early warning signals of tipping points. *Proceedings of the National Academy of Sciences* **118**, e2106140118 (2021).

[15] Angra, S. & Ahuja, S. Machine learning and its applications: A review. In *2017 International Conference On Big Data Analytics And Computational Intelligence (icbdac)*, 57–60 (IEEE, 2017).

[16] Chalapathy, R. & Chawla, S. Deep learning for anomaly detection: A survey. *Arxiv Preprint Arxiv:1901.03407* (2019).

[17] Ribeiro, A. H. *et al.* Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications* **11**, 1760 (2020).

[18] Lee, H., Shin, S.-Y., Seo, M., Nam, G.-B. & Joo, S. Prediction of ventricular tachycardia one hour before occurrence using artificial neural networks. *Scientific Reports* **6**, 1–7 (2016).

[19] Nikkonen, S., Afara, I. O., Leppänen, T. & Töyräs, J. Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea. *Scientific Reports* **9**, 1–9 (2019).

[20] Radha, M. *et al.* Sleep stage classification from heart-rate variability using long short-term memory neural networks. *Scientific Reports* **9**, 1–11 (2019).

[21] Chen, X. *et al.* Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (mesa). *Sleep* **38**, 877–888 (2015).

[22] Blackwell, T. *et al.* Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study. *Journal of The American Geriatrics Society* **59**, 2217–2225 (2011).

[23] Quan, S. F. *et al.* The sleep heart health study: design, rationale, and methods. *Sleep* **20**, 1077–1085 (1997).

[24] Zhang, G.-Q. *et al.* The national sleep research resource: towards a sleep data commons. *Journal of The American Medical Informatics Association* **25**, 1351–1358 (2018).

[25] Pan, J. & Tompkins, W. J. A real-time qrs detection algorithm. *IEEE Transactions On Biomedical Engineering* 230–236 (1985).

[26] Sedghamiz, H. Matlab implementation of pan tompkins ecg qrs detector. *Code Available at at File Exchange Site of Mathworks* (2014).

[27] Keys, R. Cubic convolution interpolation for digital image processing. *IEEE Transactions On Acoustics, Speech, And Signal Processing* **29**, 1153–1160 (1981).

[28] Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Advances In Neural Information Processing Systems* **30** (2017).

[29] Mohammed, R., Rawashdeh, J. & Abdullah, M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th International Conference On Information And Communication Systems (icics)*, 243–248 (IEEE, 2020).

[30] Sharma, M., Kumbhani, D., Tiwari, J., Kumar, T. S. & Acharya, U. R. Automated detection of obstructive sleep apnea in more than 8000 subjects using frequency optimized orthogonal wavelet filter bank with respiratory and oximetry signals. *Computers In Biology And Medicine* **144**, 105364 (2022).

[31] Haidar, R., Koprinska, I. & Jeffries, B. Sleep apnea event detection from nasal airflow using convolutional neural networks. In *International Conference On Neural Information Processing*, 819–827 (Springer, 2017).

[32] Van Steenkiste, T., Groenendaal, W., Deschrijver, D. & Dhaene, T. Automated sleep apnea detection in raw respiratory signals using long short-term memory neural networks. *IEEE Journal of Biomedical And Health Informatics* **23**, 2354–2364 (2018).

[33] McCloskey, S., Haidar, R., Koprinska, I. & Jeffries, B. 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*, 361–372 (Springer, 2018).

[34] Banluesombatkul, N., Rakthanmanon, T. & Wilaiprasitporn, T. Single channel ecg for obstructive sleep apnea severity detection using a deep learning approach. In *Tencon 2018-2018 IEEE Region 10 Conference*, 2011–2016 (IEEE, 2018).

[35] Alvarez-Estevez, D. & Moret-Bonillo, V. Computer-assisted diagnosis of the sleep apnea-hypopnea syndrome: a review. *Sleep Disorders* **2015** (2015).

[36] Kemp, B. & Olivan, J. European data format 'plus'(edf+), an edf alike standard format for the exchange of physiological data. *Clinical Neurophysiology* **114**, 1755–1761 (2003).

[37] Rechtschaffen, A. A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects. *Brain Information Service* (1968).

[38] Huang, L. *et al.* Normalization techniques in training dnns: Methodology, analysis and application. *Arxiv Preprint Arxiv:2009.12836* (2020).

[39] Tan, M. & Le, Q. Efficientnetv2: Smaller models and faster training. In *International Conference On Machine Learning*, 10096–10106 (PMLR, 2021).

[40] Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference On Machine Learning*, 6105–6114 (PMLR, 2019).

[41] He, K., Girshick, R. & Dollár, P. Rethinking imagenet pre-training. In *Proceedings of the IEEE/cvf International Conference On Computer Vision*, 4918–4927 (2019).

[42] Jais, I. K. M., Ismail, A. R. & Nisa, S. Q. Adam optimization algorithm for wide and deep neural network. *Knowledge Engineering And Data Science* **2**, 41–46 (2019).

[43] Montavon, G., Orr, G. & Müller, K.-R. *Neural networks: tricks of the trade*, vol. 7700 (springer, 2012).

[44] Chen, T. *et al.* Xgboost: extreme gradient boosting. *R Package Version 0.4-2* **1**, 1–4 (2015).

[45] Ye, J., Chow, J.-H., Chen, J. & Zheng, Z. Stochastic gradient boosted distributed decision trees. In *Proceedings of the 18th Acm Conference On Information And Knowledge Management*, 2061–2064 (2009).

[46] Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* **13** (2012).

[47] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition*, 770–778 (2016).

[48] Kraemer, K. H., Donner, R. V., Heitzig, J. & Marwan, N. Recurrence threshold selection for obtaining robust recurrence characteristics in different embedding dimensions. *Chaos* **28**, 085720 (2018).

SUPPLEMENTARY MATERIAL

A PREPRINT

Marino E. Gavidia[1,*], Arthur N. Montanari[1,2], and Jorge Goncalves[1,3,*]

[1]Luxembourg Center for Systems Biomedicine, University of Luxembourg, Belvaux L-4367, Luxembourg
[2]Department of Physics and Astronomy, Northwestern University, Evanston, IL 60208, USA
[3]Department of Plant Sciences, Cambridge University, Cambridge CB2 3EA, United Kingdom
[*]Corresponding authors: marino.gavidia@uni.lu; jmg77@cam.ac.uk

February 15, 2023

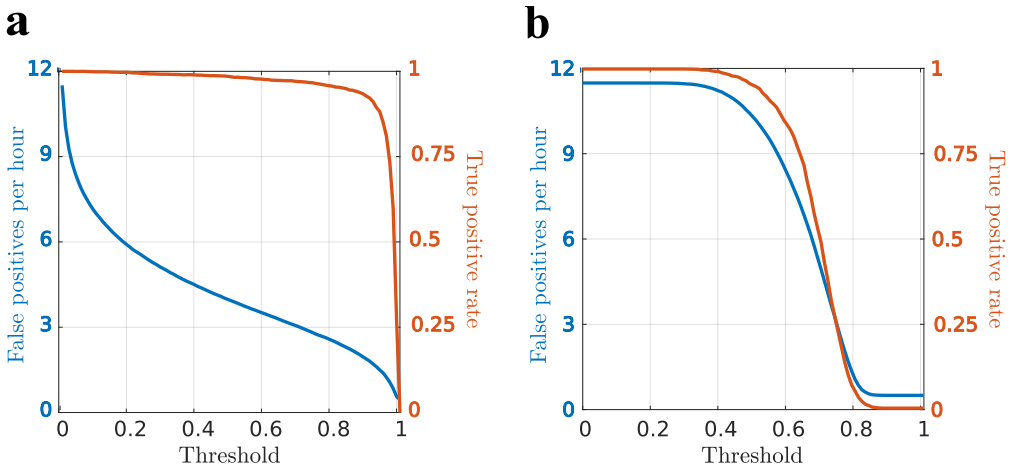

Figure S1: True-positive rate and number of false positives per hour as a function of the classification threshold on the validation set (b) five sensor channels and (c) only two sensors (AbdoRes and SpO_2).

Table S1: 10-fold cross-validation accuracy on the testing sets after training the EfficientNetV2 and the Light Gradient Boosting Machine using all five 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 S2: Performance of DRIVEN on all databases.

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO_2	0.91	0.9	0.92	0.922	0.911
AbdoRes+ThorRes+Airflow+SpO_2	0.909	0.899	0.92	0.922	0.911
RRI+AbdoRes+Airflow+SpO_2	0.908	0.898	0.918	0.92	0.909
RRI+AbdoRes+ThorRes+Airflow	0.905	0.894	0.916	0.918	0.906
RRI+ThorRes+Airflow+SpO_2	0.904	0.894	0.915	0.917	0.905
RRI+AbdoRes+ThorRes+SpO_2	0.885	0.875	0.895	0.898	0.886
AbdoRes+Airflow+SpO_2	0.907	0.896	0.917	0.919	0.908
AbdoRes+ThorRes+Airflow	0.904	0.894	0.914	0.916	0.905
ThorRes+Airflow+SpO_2	0.903	0.894	0.914	0.916	0.905
RRI+AbdoRes+Airflow	0.902	0.892	0.912	0.914	0.903
RRI+Airflow+SpO_2	0.897	0.886	0.907	0.91	0.898
RRI+ThorRes+Airflow	0.896	0.884	0.909	0.912	0.898
AbdoRes+ThorRes+SpO_2	0.884	0.874	0.894	0.896	0.885
RRI+AbdoRes+SpO_2	0.879	0.87	0.888	0.891	0.88
RRI+AbdoRes+ThorRes	0.871	0.861	0.881	0.884	0.872
RRI+ThorRes+SpO_2	0.861	0.853	0.87	0.873	0.863
AbdoRes+Airflow	0.899	0.89	0.909	0.911	0.9
ThorRes+Airflow	0.895	0.883	0.907	0.91	0.897
Airflow+SpO_2	0.895	0.886	0.905	0.907	0.896
RRI+Airflow	0.881	0.869	0.894	0.897	0.883
AbdoRes+SpO_2	0.877	0.869	0.886	0.888	0.879
AbdoRes+ThorRes	0.865	0.856	0.874	0.877	0.866
ThorRes+SpO_2	0.859	0.852	0.866	0.869	0.86
RRI+AbdoRes	0.855	0.85	0.861	0.863	0.856
RRI+ThorRes	0.826	0.817	0.835	0.84	0.828
RRI+SpO_2	0.824	0.821	0.827	0.828	0.825
Airflow	0.878	0.867	0.89	0.893	0.88
AbdoRes	0.845	0.843	0.848	0.849	0.846
ThorRes	0.812	0.805	0.82	0.825	0.815
SpO_2	0.81	0.809	0.812	0.812	0.811
RRI	0.665	0.669	0.661	0.653	0.661

Table S3: Independent validation test performances of DRIVEN on the MESA database.

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO_2	0.874	0.831	0.93	0.939	0.882
RRI+AbdoRes+Airflow+SpO_2	0.875	0.826	0.942	0.951	0.884
RRI+AbdoRes+ThorRes+Airflow	0.874	0.834	0.924	0.933	0.881
AbdoRes+ThorRes+Airflow+SpO_2	0.873	0.833	0.923	0.933	0.88
RRI+ThorRes+Airflow+SpO_2	0.866	0.819	0.928	0.939	0.875
RRI+AbdoRes+ThorRes+SpO_2	0.85	0.815	0.894	0.906	0.858
AbdoRes+Airflow+SpO_2	0.877	0.831	0.938	0.946	0.885
RRI+AbdoRes+Airflow	0.876	0.83	0.937	0.946	0.884
RRI+Airflow+SpO_2	0.876	0.823	0.949	0.958	0.885
AbdoRes+ThorRes+Airflow	0.867	0.828	0.917	0.927	0.875
RRI+ThorRes+Airflow	0.865	0.822	0.922	0.932	0.874
RRI+AbdoRes+SpO_2	0.864	0.822	0.917	0.928	0.872
ThorRes+Airflow+SpO_2	0.864	0.819	0.923	0.934	0.873
RRI+AbdoRes+ThorRes	0.854	0.837	0.871	0.878	0.857
AbdoRes+ThorRes+SpO_2	0.844	0.808	0.888	0.901	0.852
RRI+ThorRes+SpO_2	0.829	0.796	0.869	0.883	0.838
AbdoRes+Airflow	0.867	0.82	0.929	0.939	0.876
ThorRes+Airflow	0.859	0.816	0.916	0.927	0.868
Airflow+SpO_2	0.858	0.801	0.943	0.954	0.871
RRI+Airflow	0.856	0.799	0.94	0.951	0.868
AbdoRes+ThorRes	0.849	0.84	0.859	0.863	0.851
AbdoRes+SpO_2	0.845	0.8	0.906	0.92	0.856
RRI+AbdoRes	0.844	0.81	0.886	0.899	0.852
ThorRes+SpO_2	0.82	0.79	0.858	0.873	0.829
RRI+SpO_2	0.815	0.778	0.865	0.883	0.827
RRI+ThorRes	0.807	0.802	0.813	0.817	0.809
Airflow	0.842	0.779	0.94	0.953	0.858
AbdoRes	0.836	0.813	0.864	0.874	0.842
SpO_2	0.815	0.786	0.851	0.866	0.824
ThorRes	0.783	0.784	0.783	0.783	0.783
RRI	0.692	0.674	0.714	0.743	0.707

Table S4: Independent validation test performances of DRIVEN on the MROS database.

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO_2	0.9	0.851	0.965	0.97	0.907
AbdoRes+ThorRes+Airflow+SpO_2	0.9	0.852	0.965	0.97	0.907
RRI+AbdoRes+Airflow+SpO_2	0.896	0.847	0.962	0.968	0.903
RRI+ThorRes+Airflow+SpO_2	0.893	0.844	0.957	0.963	0.9
RRI+AbdoRes+ThorRes+Airflow	0.892	0.84	0.963	0.969	0.9
RRI+AbdoRes+ThorRes+SpO_2	0.886	0.836	0.953	0.96	0.894
AbdoRes+Airflow+SpO_2	0.897	0.848	0.962	0.968	0.904
ThorRes+Airflow+SpO_2	0.892	0.844	0.956	0.962	0.899
AbdoRes+ThorRes+Airflow	0.89	0.837	0.964	0.97	0.898
RRI+AbdoRes+Airflow	0.888	0.836	0.96	0.967	0.896
AbdoRes+ThorRes+SpO_2	0.885	0.833	0.954	0.961	0.893
RRI+ThorRes+Airflow	0.881	0.83	0.95	0.958	0.889
RRI+Airflow+SpO_2	0.879	0.832	0.943	0.952	0.888
RRI+AbdoRes+SpO_2	0.878	0.827	0.948	0.956	0.887
RRI+AbdoRes+ThorRes	0.867	0.812	0.946	0.956	0.878
RRI+ThorRes+SpO_2	0.866	0.819	0.929	0.939	0.875
AbdoRes+Airflow	0.886	0.832	0.961	0.968	0.895
ThorRes+Airflow	0.881	0.83	0.951	0.959	0.89
Airflow+SpO_2	0.88	0.835	0.94	0.948	0.888
AbdoRes+SpO_2	0.877	0.825	0.948	0.956	0.886
AbdoRes+ThorRes	0.863	0.805	0.948	0.958	0.875
ThorRes+SpO_2	0.862	0.815	0.926	0.937	0.872
RRI+Airflow	0.854	0.815	0.904	0.916	0.863
RRI+AbdoRes	0.852	0.799	0.927	0.94	0.864
RRI+SpO_2	0.827	0.785	0.882	0.899	0.838
RRI+ThorRes	0.824	0.778	0.889	0.908	0.838
Airflow	0.852	0.818	0.893	0.905	0.859
AbdoRes	0.848	0.791	0.934	0.947	0.862
SpO_2	0.817	0.782	0.862	0.879	0.828
ThorRes	0.812	0.76	0.889	0.911	0.829
RRI	0.658	0.657	0.658	0.659	0.658

Table S5: Independent validation test performances of DRIVEN on the SHHS database.

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO_2	0.765	0.928	0.692	0.574	0.709
RRI+AbdoRes+ThorRes+SpO_2	0.818	0.882	0.773	0.735	0.802
RRI+AbdoRes+Airflow+SpO_2	0.765	0.925	0.693	0.578	0.711
AbdoRes+ThorRes+Airflow+SpO_2	0.763	0.924	0.691	0.573	0.707
RRI+AbdoRes+ThorRes+Airflow	0.756	0.92	0.684	0.561	0.697
RRI+ThorRes+Airflow+SpO_2	0.746	0.92	0.674	0.539	0.68
AbdoRes+ThorRes+SpO_2	0.816	0.878	0.771	0.734	0.8
RRI+AbdoRes+SpO_2	0.811	0.883	0.762	0.717	0.792
RRI+ThorRes+SpO_2	0.805	0.848	0.771	0.743	0.792
RRI+AbdoRes+ThorRes	0.801	0.856	0.76	0.723	0.784
AbdoRes+Airflow+SpO_2	0.763	0.92	0.691	0.575	0.708
RRI+AbdoRes+Airflow	0.754	0.919	0.683	0.558	0.694
AbdoRes+ThorRes+Airflow	0.753	0.918	0.681	0.555	0.692
RRI+Airflow+SpO_2	0.746	0.913	0.675	0.544	0.682
ThorRes+Airflow+SpO_2	0.744	0.917	0.673	0.537	0.678
RRI+ThorRes+Airflow	0.732	0.917	0.661	0.511	0.656
AbdoRes+SpO_2	0.808	0.877	0.76	0.715	0.788
ThorRes+SpO_2	0.801	0.842	0.769	0.741	0.788
AbdoRes+ThorRes	0.796	0.849	0.757	0.72	0.78
RRI+SpO_2	0.787	0.826	0.756	0.727	0.773
RRI+AbdoRes	0.785	0.849	0.741	0.693	0.763
RRI+ThorRes	0.756	0.786	0.731	0.703	0.742
AbdoRes+Airflow	0.753	0.915	0.681	0.557	0.692
Airflow+SpO_2	0.746	0.91	0.676	0.547	0.683
ThorRes+Airflow	0.734	0.913	0.663	0.517	0.66
RRI+Airflow	0.725	0.907	0.656	0.502	0.646
AbdoRes	0.779	0.844	0.734	0.683	0.755
SpO_2	0.779	0.817	0.749	0.719	0.765
ThorRes	0.747	0.782	0.72	0.686	0.731
Airflow	0.724	0.901	0.655	0.503	0.645
RRI	0.642	0.656	0.631	0.598	0.626

Table S6: Performance for different network architectures.

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

Table S7: 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

Table S8: 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 S9: 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