Deep Learning-Based Approach for Sleep Apnea Detection Using Physiological Signals

A. R. Troncoso-García[1](✉), M. Martínez-Ballesteros[2], F. Martínez-Álvarez[1], and A. Troncoso[1]

[1] Data Science and Big Data Lab, Pablo de Olavide University, 41013 Seville, Spain
{artrogar,fmaralv,atrolor}@upo.es

[2] Department of Computer Science, University of Seville, 41012 Seville, Spain
mariamartinez@us.es

Abstract. This paper explores the use of deep learning techniques for detecting sleep apnea. Sleep apnea is a common sleep disorder characterized by abnormal breathing pauses or infrequent breathing during sleep. The current standard for diagnosing sleep apnea involves overnight polysomnography, which is expensive and requires specialized equipment and personnel. The proposed method utilizes a neural network to analyze physiological signals, such as heart rate and respiratory patterns, that are recorded during sleep to authomatic sleep apnea detection. The neural network is trained on a dataset of polysomnography recordings to identify patterns that are indicative of sleep apnea. The results compare the use of different physiological signals to detect sleep apnea. Nasal airflow seems to have the most accurate results and higher specificity, whereas EEG and ECG have higher levels of sensitivity. The best model concerning accuracy is compared to bias models previously applied to sleep apnea detection in literature, achieving greater results. This approach has the potential to provide automatic sleep apnea detection, being an accessible solution for diagnosing sleep apnea.

Keywords: Sleep apnea · Time series · Deep learning · classification · forecasting

1 Introduction

Sleep apnea is a sleep disorder that causes repeated pauses in breathing or shallow breathing during sleep. These pauses could last from a few seconds to several minutes and could occur many times throughout the night. These interruptions in breathing can cause a significant reduction in sleep quality and can lead to a variety of significant health consequences [12]. Obstructive sleep apnea (OSA) is the most common form of sleep apnea. Patients with OSA usually experience loud snoring, gasping, or choking during sleep, and fatigue during the day. Hypoapnea is a related term used to describe a partial reduction in airflow to the c

https://doi.org/10.1007/978-3-031-43085-5_50

lungs during sleep. Causes breathing to become shallower or slow for a period of time, typically lasting at least 10 s. Hypoapnea is often associated with OSA. Other symptoms of sleep apnea could include headaches in the morning, difficulty in concentrating, mood changes, and irritability. These symptoms directly affect the patient's daily life: sleep apnea contributes to motor vehicle accidents and reduced productivity, creating problems at work or school. Sleep apnea is closely related to obesity, smoking, alcohol consumption, and family history. Men are more likely to develop sleep apnea than women, which is more common in older adults. Sleep apnea also plays an increasing role in cardiovascular disease, particularly hypertension and congestive heart failure [12]. Treatment for sleep apnea can include lifestyle changes such as weight loss, exercise, and the elimination of alcohol and sedatives. There are also medical interventions, such as continuous positive airway pressure therapy [4]. Effective treatment can improve sleep quality, reduce symptoms, and improve overall health and well-being.

Polysomnography (PSG) is a widely used clinical test to diagnose sleep apnea and other sleep disorders. During a PSG test, a person is monitored while sleeping to measure various physiological parameters, such as heart rate or breathing rhythm. PSG is considered the standard for diagnosing sleep apnea because it provides a comprehensive assessment of the severity and frequency of breathing disturbances during sleep [1]. However, the PSG test has several disadvantages. PSG involves spending a night in a hospital or a sleep laboratory. The patient must be connected to various sensors, which is uncomfortable. In addition, PSG could be expensive. The costs may be high for public health systems and could be inaccessible for some patients. The test are also not accurate in certain situations. For example, the PSG test only provides information about the patient's sleep patterns during the stay in the hospital. It may not capture typical sleep patterns or account for the variability in sleep over time. PSG may also disrupt the patient's natural sleep patterns, as they are sleeping in an unfamiliar environment and connected to various sensors. Finally, the interpretation of PSG results requires specialized training and knowledge of relevant experts [8].

PSG records are considered as time series data because they involve the recording of physiological signals over time. During a PSG recording, various physiological signals such as brain waves (EEG), eye movements (EOG), muscle activity (EMG), heart rate (ECG), and breathing patterns are continuously monitored and recorded. These signals change over time and are typically sampled at a fixed frequency, resulting in a sequence of data points that can be analyzed as a time series. Analyzing PSG data as a time series reveals patterns and trends in physiological signals and provides insights into sleep disorders and other conditions that affect sleep. In this work, time series analysis techniques are used to study and interpret PSG data and extract relevant features for the diagnosis and treatment of sleep apnea. Artificial intelligence and deep learning (DL) techniques in particular are applied to detect sleep apnea. DL models analyze large amounts of data quickly and accurately. They represent a tool for doctors to help make decisions, leading to more reliable diagnoses and personalized treatment plans for patients. In addition, technology is applied to provide a noninvasive and cost-effective way to detect sleep apnea. Portable devices, such as wearable sensors and smartphone apps, could also be used at home, making sleep monitoring easy for patients. This paper proposes a DL approach to detect sleep apnea events. A neural network is applied to PSG data including ECG, EEG, blood pressure (BP), and nasal respiration. Sleep apnea detection methods using the four signals are compared. Then, the best method is compared to bias models using the same input data.

The remainder of the paper is structured as follows. First, in Sect. 2, the latest developments in deep learning regarding the detection of sleep apnea are discussed. Then, in Sect. 3, the experiments conducted are explained in detail, while Sect. 4 presents the results obtained. Lastly, the paper is concluded in Sect. 5.

2 Related Work

As previously introduced, computer-based sleep apnea detection is useful to help physicians diagnose the disease. Several examples of neural networks applied to the diagnosis of sleep apnea are found in the literature. The paper [5] revised existing algorithms that have been applied to the detection of obstructive sleep apnea using various sensors and the combination of different approaches. The paper presented 84 original research articles published between 2003 and 2017. The articles were selected to provide valuable information to researchers who want to implement potential signal-processing algorithms on hardware. The contributions of the article in [6] regarding automatic sleep apnea scoring processes are also discussed. Another review is found in [7]. The goal of the paper is to analyze the research published in the last decade, examining how different deep networks are implemented, what preprocessing or feature extraction is necessary, and the advantages and disadvantages of different types of networks. In the field of classifiers, neural networks are the most used models for the detection of sleep apnea. Namely, these models are deep vanilla neural network (DVNN), convolution neural network (CNN), and recurrent neural network (RNN).

The paper [9] discussed the usefulness of ML and DL models as a diagnosis-decision-support tool for the detection of sleep apnea. The article then focused on obstructive sleep apnea. ML models were applied to the analysis of the respiratory signal waveform to aid in its diagnosis. Local Interpretable Model-Agnostic (LIME) library was used to explain the results obtained from a PSG study for automatic detection of sleep apnea. The results obtained help humans to understand the importance of each feature. The study carried out in [2] proposed a CNN model to detect sleep apnea. The input data are four different types of sleep study that focus on portability and signal reduction. The CNN model used the level of oxygen saturation (SpO_2) as input. The results showed that it is a valid and cost-effective alternative to PSG. The study used 190000 samples from SPO2 sensors from 50 patients, and the overall accuracy of sleep apnea detection was 91.3%, with a loss rate of 2.3 using the cross-entropy cost function using the deep convolutional neural network.

The study in [10] presented a new approach for automatically detecting sleep-disordered breathing events, such as sleep apnea. RNN was used to analyze nocturnal electrocardiogram (ECG) recordings. The proposed RNN model included recurrent layers with LSTM and a gated-recurrent unit (GRU). The model was trained and tested on ECG recordings from 92 patients, resulting in a F1-score of 98.0% for LSTM and 99.0% for GRU. These results showed that the proposed method outperformed conventional methods and could be used as a screening and diagnostic tool for patients with sleep breathing disorders. Furthermore, the study in [11] proposed an algorithm based on DL models to automatically detect sleep apnea events in respiratory signals. The algorithm improved the scoring per patient when assigned to the apnea-hypopnea index. The proposed algorithm was proved to be a useful tool for trained staff to quickly diagnose sleep apnea. Finally, the study in [13] explored an alternative to PSG for detecting sleep apnea and hypopnea syndrome using ECG and SpO_2 signals. The paper proposed a combination of classifiers to improve classification performance by using complementary information from individual classifiers.

3 Methodology

This paper proposes the application of a neural network to time series classification in the problem of detecting sleep apnea. Figure 1 shows the methodology carried out in this article. As introduced, several signals are used to characterize sleep apnea in the clinical scope. Here, the focus is on comparing nasal respiration, BP, EEG, and ECG signals to see which time series is better for the detection task of sleep apnea. A grid search is carried out to optimize the hyperparameters of the neural network model. Then, the signal obtaining best results is selected, and this model is compared to bias models using the same input data.

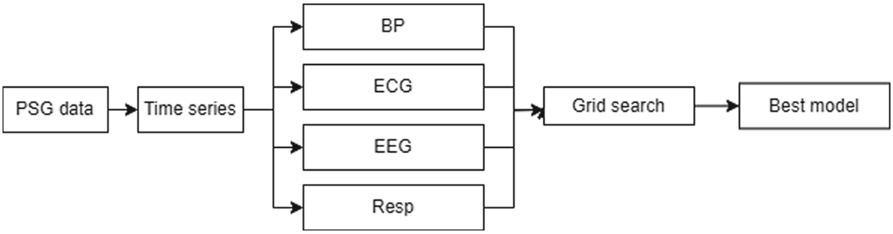


Fig. 1. Purposed methodology.

3.1 Data Preprocessing

The PSG data are in a standardized format that is commonly used in sleep studies. PSG data includes a variety of signals, such as EEG, ECG, respiration, and body movement, among others. The data also include annotations that provide information on sleep stages, arousal, and other events that occur during the sleep study. These annotations are created by experienced physicians who visually inspect signals and label them according to established guidelines. The Waveform Database (WFDB) is an open source software package developed by Physionet that provides tools for reading, writing, and processing physiological signals. WFDB allows users to easily access and manipulate large databases of physiological signals, including PSG data. The WFDB Python package is a Python interface to the WFDB software, providing a convenient way to access and analyze PSG data using Python [3]. In this paper, WFDB has been used to covert waveform and signal data to time series data. The signals have been processed and filtered to detect errors and outliers. PSG data is divided into four different dataset, each of them containing one physiological signal such as EEG or ECG.

3.2 Bias Models

The results of the purposed DL methodology are compared to the benchamark algorithms. Bias models have been used with default configuration and have been applied to the same input data. The bias models are presented in a previous work for the International Conference KES 2022 [9]. The models are detailed as follows:

1. Logistic Regression (LR) is a basic model used for binary classification that predicts targets using a linear approximation.
2. K Nearest Neighbors (KNN) classifier. The model uses a k-nearest neighbor vote for classification. The parameter **k** is set to 5 in this case.
3. Decision Tree (DT) classifier. The model is a nonparametric supervised learning method that creates a model by learning simple decision rules from the data features to predict the target variable.
4. Gradient Boosting Classifier (GBC). GBC builds an additive model in a forward stage-wise manner and optimizes differentiating loss functions. In each stage, regression trees are fitted on the negative gradient of the binomial or multinomial deviance loss function. For binary classification, only a single regression tree is induced.

3.3 Deep Learning Model

The neural network used in this study is a dense neural network with several layers with a certain number of neurons. To avoid overfitting, a dropout layer has been added to the network. Dropout is a regularization technique that randomly removes some of the neurons in a layer during training, helping to prevent the network from relying too heavily on any one neuron or feature.

The neural network architecture has been implemented using grid search. In particular, the following hyperparameters have been tuned: number of layers, number of neurons per layer, and dropout. The range of the hyperparameters is presented in Table 1. Dropout with a value equal to 0 means that there is no dropout layer.

Table 1. Hyperparameters range.

Hyperparameter	Range	Optimal value
Number of layers	2, 3, 4, 5	4
Neurons	100, 200, 300, 400, 500	[100, 200, 300, 400]
Dropout	0, 0.1, 0.2, 0.3, 0.4, 0.5	0.2

The best model obtained includes four dense layers and a dropout layer after each of them. Other hyperparameters such as batch size, learning rate, or training epochs have not been tuned. The batch size determines how many samples are processed at once during each training iteration, and the learning rate controls how much the network weights are updated during each training iteration. The network was trained with a batch size of 64, a learning rate of 0.1 and during 10 epochs. In general, the architecture of the neural network and the training parameters have been carefully selected to optimize the performance of the model for the specific task of detecting apnea events using PSG data.

4 Results

The results of sleep apnea detection using the neural network model are presented in this section. First, the input data is characterized. Then, the performance of the model is evaluated using several quality measures, such as accuracy, sensitivity, specificity, and F1 score. Finally, the best model is compared to bias model previously used in sleep apnea detection.

4.1 Input Data

This study uses data from the MIT-BIH Polysomnographic Database [3], which contains recordings from 16 patients. Different physiological signals are presented, including ECG, invasive BP, EEG, and nasal respiration airflow. Health professionals carefully study the signals and annotate them based on the existence of apnea events and the stage of sleep. The PSG data are processed to create a dataset for classification tasks, where each instance is a 30-second window of the four datasets of pyhsiological signal labeled as either an apnea or hypoapnea event (1), or normal breathing (0). The final dataset has 7500 attributes per instance due to the 250 Hz sampling rate. Therefore, the dataset is treated as a time series where each instance represents a 30-second interval measurement. Figure 2 illustrates an example of a complete PSG record with four signals and their measure unit. Namely, the signals are: ECG (mV), blood pressure (BP mmHg), EEG (mV) and nasal respiration (l). Data must be normalized because of the different scales in the four signals.

The classification task is performed using this dataset. The four different signals recorded during 30 s are every single instance of the input data. The classification model achieves a binary classification task. One of the biggest problems

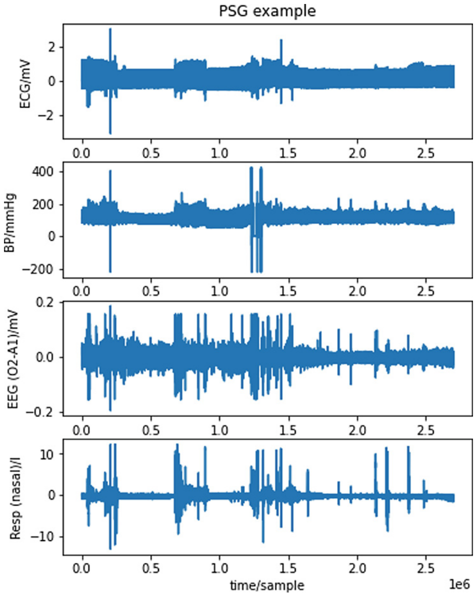


Fig. 2. PSG example record.

when performing sleep apnea detection is that the data is clearly unbalanced. Figure 3 shows the distribution of apnea events: normal (meaning that there is no apnea event, normal breathing) and anomalous. Anomalous events mean hypoapnea (partial breathing interruption), obstructive apnea (total obstruction), and central apnea. In this paper, the target value has been summarized as "apnea" (1) and "no apnea" (0).

4.2 Quality Measures

Classification refers to the task of automatically assigning input data to one of several predefined categories or classes based on a set of characteristics. In this paper we work in the field of binary classification, meaning that there are two classes: "No apnea" (0) and "Apnea" (1). The LSTM algorithm learns to identify patterns in the input data that are characteristic of each class and then uses these patterns to classify new unlabeled data. Classification aims to create a model that can accurately predict the class of new data based on the features provided.

The quality measures used to evaluate the proposed methodology are presented as follows.

- Accuracy. It measures the percentage of correct predictions made by the model out of all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

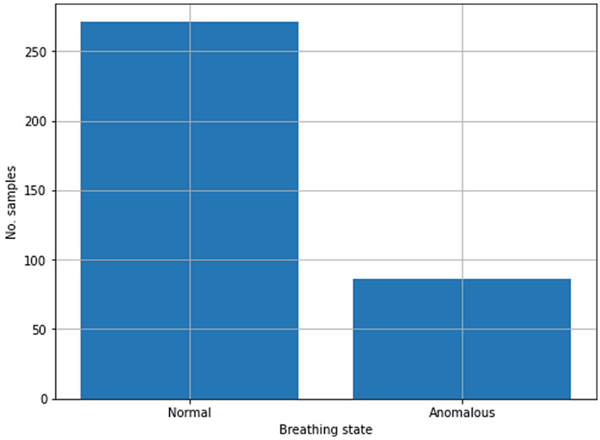


Fig. 3. Types of apnea distribution.

- Precision. It is the fraction of true positive predictions (correctly classified as positive instances) out of all positive predictions (instances classified as positive). It is calculated as the ratio of the number of true positives (TP) to the sum of true positives and false positives (FP).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

- Sensitivity. This measure refers to the fraction of true positive predictions out of all actual positive instances. Calculated as the ratio of the number of true positives to the sum of true positives and false negatives. It could also be called recall.

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{3}$$

- Specificity. It is a metric that measures the ability of a classification model to correctly identify negative instances as negative. Specifically, it is the fraction of TN predictions (correctly classified as negative instances) out of all negative predictions (instances classified as negative). It is calculated as the ratio of the number of true negatives to the sum of true negatives and false positives. High specificity indicates that the model is good at avoiding false positives.

$$Specificity = \frac{TN}{FP + TN} \tag{4}$$

- F1 Score. This metric is known as the harmonic mean of precision and recall. It is a weighted average of precision and recall, where the weights are equal and ranges from 0 to 1, with 1 being the best possible score. The F1 score is calculated as 2 times the product of precision and recall, divided by the sum of precision and recall.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{5}$$

In the clinical scope, priority is often given to metrics related to the detection of TP cases. The consequences of FN, such as missing a positive case, could be severe and potentially life-threatening. In this way, sensitivity can be a more important metric than precision or accuracy, as it measures the proportion of actual positive cases that were correctly identified by the model. For example, recall can be crucial in medical diagnosis or disease screening to ensure that all positive cases are identified, even if that means sacrificing some precision or accuracy.

4.3 Sleep Apnea Classification

The LSTM model is used to analyze time series of sleep apnea using different signals as input. Signals include BP, ECG, EEG, and respiration (Resp). The quality measures have been evaluated for models using each signal as input data including accuracy, sensitivity, specificity, and F1 score that have been introduced in the previous Sect. 4.2. Table 2 presents the quality measures for the four physiological signals.

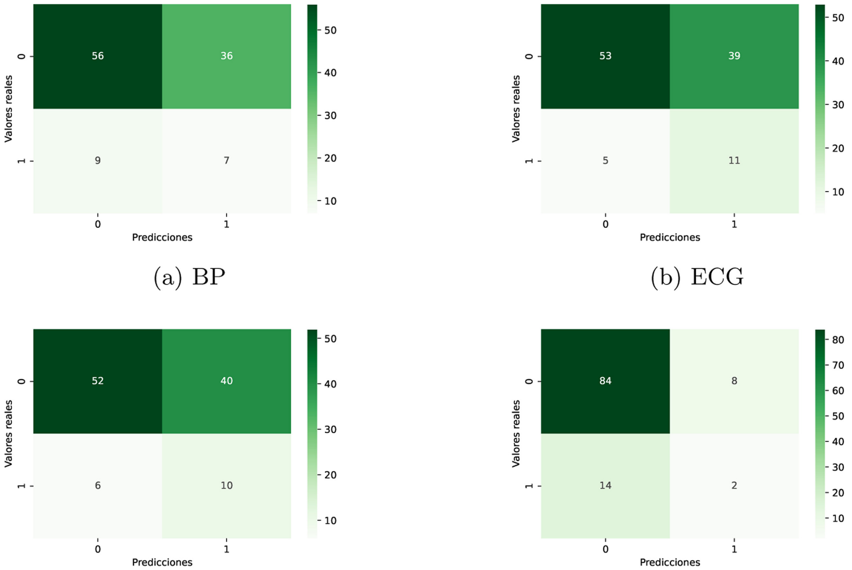
Table 2. Quality measures of the LSTM model using different signals as input.

Signal	Accuracy	Sensitivity	Specificity	F1
BP	0.583	0.438	0.641	0.238
ECG	0.593	**0.688**	0.576	0.334
EEG	0.574	0.625	0.565	0.303
Resp	**0.712**	0.125	**0.913**	**0.370**

Table 2 shows that the model achieved the highest accuracy (0.712) when using respiration as input. However, the sensitivity values was lower than the other signals, indicating that the model tended to classify more instances as negative. On the contrary, the model that used EEG as input achieved higher sensitivity values to detect apnea events (positive class). Similar results are obtained with ECG. Regarding specificity, the best results are again obtained with Resp signal, with greater differences with the others. Finally, concerning F1 score the four signals have similar values although Resp has a little better result.

The confusion matrix for each experiment is presented in Fig. 4d. The confusion matrix compares the predicted values with the actual values. It summarizes the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions made by the model.

In this case, positive means the presence of an apnea event, whereas negative means that there is not apnea. The rows of the matrix represent the actual class labels, while the columns represent the predicted class labels. The diagonal of the matrix shows the number of correct predictions, while the off-diagonal elements represent the incorrect predictions. On the one hand, concerning the predictions



(a) BP

(b) ECG

(c) EEG

(d) Resp

Fig. 4. Confusion matrix of the four experiments.

obtained using nasal respiration, the highest accuracy is reached. However, this results are not ideal, as the model tends to classify every instance as a 'no apnea' (0) event, as shown in Fig. 4d. This can be attributed to the imbalanced classes. On the other hand, the models that use ECG and EEG perform better in detecting apnea events (1) with a sensitivity of 0.688 and 0.625, respectively. However, these models also tend to classify several cases as positive (1) that are not apnea events. This is shown in the lower levels of specificity. The results of the experiments highlight the importance of choosing the appropriate input signal for the model to achieve the best performance. It also shows that while accuracy is important, other measures, such as sensitivity, should also be considered when evaluating the performance of the model, specially in clinical scope.

4.4 Comparation with Bias Model

The better model which is trained with nasal respiration signal which obtains the best results, is compared to the bias algorithms previously detailed in Sect. 3. Quality measures, namely accuracy, ROC-AUC score and F1 score for both bias models and the purposed model are presented in Table 3. The neural network outcomes clearly LR and DT models. Concerning GBC, results are quite similar.

Table 3. Performance of bias ML models.

Model	Accuracy	ROC-AUC	F1
NN - Resp	0.712	0.675	0.370
LR	0.698	0.514	0.363
DT	0.750	0.567	0.292
GBC	0.758	0.686	0.333

5 Conclusions and Future Work

In summary, the LSTM model was evaluated for the classification of time series of sleep apnea using different physiological signals as input, including blood pressure, electrocardiogram, electroencephalogram, and nasal airflow from respiration. The results showed that the best accuracy and specificity were obtained when respiration was used as input. However, the models that used ECG and EEG achieved higher sensitivity values to detect apnea events. In clinical scope is essential identifying the positive instances, meaning the patients with a certain disease. In general, the four models tended to classify more instances as negative (no apnea) than positive (apnea). The fact that the classes are imbalanced may have contributed to this bias. Nasal respiration has been proved as the most useful signal to detect sleep apnea.

Therefore, the results in the paper are useful in choosing the appropriate input signal to evaluate the model performance, particularly in clinical applications. An accurate automatic detection system is capable of improving sleep apnea management by providing a more objective measure of treatment efficacy and an objective feedback on treatment efficacy. The existence of these systems could help both patients and clinicians. Automatic detection systems are used as a diagnostic support tool for doctors. These systems could also monitor treatment progress and adjust therapy as needed, moving towards personalized medicine.

Future works could investigate the performance of ensemble DL models. These models could be applied to improve the accuracy and reliability of classification results. Another approach is to combine multiple physiological signals to improve the robustness of the classification model.

Acknowledgements. The authors would like to thank the Spanish Ministry of Science and Innovation the support within the projects PID2020-117954RB-C21 and TED2021-131311B-C22.

References

1. Brockmann, P.E., Schaefer, C., Poets, A., Poets, C.F., Urschitz, M.S.: Diagnosis of obstructive sleep apnea in children: a systematic review. Sleep Med. Rev. **17**(5), 331–340 (2013)

2. Chaw, H.T., Kamolphiwong, S., Wongsritrang, K.: Sleep apnea detection using deep learning. Tehnički glasnik **13**(4), 261–266 (2019)
3. Goldberger, A.L., et al.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Biomedicallation **101**(23), e215–e220 (2000)
4. Kakkar, R.K., Berry, R.B.: Positive airway pressure treatment for obstructive sleep apnea. Chest **132**(3), 1057–1072 (2007)
5. Mendonca, F., Mostafa, S.S., Ravelo-Garcia, A.G., Morgado-Dias, F., Penzel, T.: A review of obstructive sleep apnea detection approaches. IEEE J. Biomed. Health Inform. **23**(2), 825–837 (2018)
6. Mostafa, S.S., Mendonça, F., Ravelo-García, A.G., Morgado-Dias, F.: A systematic review of detecting sleep apnea using deep learning. Sensors **19**(22), 4934 (2019)
7. Sherstinsky, A.: Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D **404**, 132306 (2020)
8. Tan, H.L., Kheirandish-Gozal, L., Gozal, D.: Pediatric home sleep apnea testing: slowly getting there! Chest **148**(6), 1382–1395 (2015)
9. Troncoso-García, A., Martínez-Ballesteros, M., Martínez-Álvarez, F., Troncoso, A.: Explainable machine learning for sleep apnea prediction. Procedia Comput. Sci. **207**, 2930–2939 (2022)
10. Urtnasan, E., Park, J.U., Lee, K.J.: Automatic detection of sleep-disordered breathing events using recurrent neural networks from an electrocardiogram signal. Neural Comput. Appl. **32**, 4733–4742 (2020)
11. 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 J. Biomed. Health Inform. **23**(6), 2354–2364 (2018)
12. White, D.P.: Sleep apnea. Proc. Am. Thorac. Soc. **3**(1), 124–128 (2006)
13. Xie, B., Minn, H.: Real-time sleep apnea detection by classifier combination. IEEE Trans. Inf. Technol. Biomed. **16**(3), 469–477 (2012)