Chapter 11
Biomedical Signal Processing for Automated Detection of Sleep Arousals Based on Multi-Physiological Signals with Ensemble Learning Methods

Navabeh Sadat Jalili Shani and Mohammad Hasan Moradi

Abstract Sleep-related breathing disorders such as sleep apnea and hypopnea are potentially serious disorders and can be the cause of a wide range of physical and mental health problems and also reduce the quality of life. Therefore, sleep studies are essential for identifying and treating these sleep disorders. This study aims to detect arousal regions caused by sleep non-apnea and non-hypopnea in polysomnography signals by using ensemble techniques. The dataset used in this study is related to Polysomnography measurement channels of 100 patients provided in the 2018 Physionet challenge database. The data was split into small epochs with 50% overlap. Several different features were extracted from each epoch in the time and frequency domain. Wilcoxon rank-sum test and Genetic Algorithm optimization algorithm were used to find a set of features with the most discriminative information. A technique for data augmentation was used to tackle the unbalanced data problem. For final classification, linear discriminant analysis, logistic regression, bagged tree from the bagging technique, and LightGBM from the boosting method were applied. Based on the Physionet Challenge indices, the area under the receiver operating characteristic curve (AUROC), and the area under the precision-recall curve (AUPRC), we compared the performance of classifiers on this dataset. The highest performance on 20 test subjects was 0.497 for AUPRC and 0.878 for AUROC.

N. S. Jalili Shani · M. H. Moradi (✉)
Biomedical Engineering Department, Amirkabir University of Technology, Tehran, Iran
e-mail: nava.jalili@aut.ac.ir; mhmoradi@aut.ac.ir

S. Mian Qaisar et al. (eds.), *Advances in Non-Invasive Biomedical Signal Sensing and Processing with Machine Learning*, https://doi.org/10.1007/978-3-031-23239-8_11

11.1 Introduction

The human body spends a third of its life sleeping, which is a complex process among living beings [1]. During sleep, fatigue is eliminated and mental and physical performance is restored [2]. During sleep, changes occur in many physiological functions of the body. There are also variations in these changes during different stages of sleep. The changes in physiological functions during sleep have led to associated variations in electrophysiological signals, which form the basis for research in sleep medicine [2].

Sleep disorders are a widespread problem in society today. Decreased consciousness due to sleep disorders has a negative impact on patients' lives [2]. Accurately diagnosing sleep disorders is difficult because of their clinical similarities. With the ability to record biosignals, the assessment of sleep quality became more accurate. To assess sleep quality, a variety of medical signals are used, including electroencephalogram (EEG), electrocardiogram (ECG), electromyogram (EMG), and electrooculogram (EOG), which the EEG signal is the most useful in evaluating sleep. These signals are called polysomnographic recordings [3, 4].

As defined by the American Sleep Disorders Association (ASDA) in 1992, arousal in sleep is a sudden increase in the frequency of the EEG signal that lasts at least 3 seconds and less than 15 seconds, and the person has been asleep for at least 10 seconds before it occurs [5]. A handbook of new sleep staging guidelines and respiratory, cardiac, and motor events was published in 2004 by the American Academy of Sleep Medicine (AASM). Accordingly, arousals can be defined as the transition between sleep and wakefulness, or the transition between REM and NREM stages. Brain arousals are defined as a sudden change in EEG frequency, including alpha, theta, or frequencies above 16 Hz, typically lasting between 3 and 15 seconds and at least 3 seconds. An important feature of arousal is unexpected changes in brainwave activity patterns. Arousals are usually observed in the second NREM stage or REM stage. Also, during the presence of arousals in the REM stage, the amplitude of the EMG signal increases further, which lasts at least 1 second [6, 7].

One of the leading causes of drowsiness during the day is the appearance of arousals that change the natural structure of sleep and cause sleep deprivation. This process is directly related to the development of sleep disorders in the individual [8]. As mentioned, these disorders can have significant effects on quality of life and daily cognitive function, including memory, learning, and concentration [9]. As a result, identifying EEG arousals in the polysomnographic recording is a good indicator for measuring sleep quality. The most important disorders that can cause arousals in sleep are sleep apnea, hypopnea, bruxism, snoring, and respiratory effort-related arousals (RERAs).

Visual examination of the signals recorded in polysomnography to diagnose the presence of arousals by physicians and specialists is a time-consuming and tedious task; in addition, Judging these events is also completely subjective and varies significantly from expert to expert [10]. Therefore, finding algorithms for automatic

detection of these disorders based on signals helps physicians in diagnosing them [11].

Diagnosis of arousal can also affect sleep staging. Diagnosis of arousal can also affect sleep staging. According to AASM rules, once an arousal region is detected, it may mean that it is best to rank the next 30-second sleep window as another (lighter) sleep stage unless the next 30-second window features maintain the current sleep stage (for example, the presence of rapid eye movements in the REM stage or the presence of sleep spindles and k complex in the N2 stage) [12].

Two types of factors can cause sleep arousal. The first cause is apnea and hypopnea disorders such as obstructive and mixed apneas, central apneas, and hypopneas. Therefore, many studies have been performed to diagnose these arousals [3, 13–15]. Another is arousal related to disorders other than apnea and hypopnea, which include arousal caused by RERA, bruxism, snoring, etc. These arousals, which are very important to diagnose, are relatively hidden in polysomnography and are difficult to detect [16]. RERA is a respiratory disorder characterized by an obstructive decrease in airflow due to narrowing the upper airways and lasts for more than 10 seconds. RERA does not reach the threshold for apnea and hypopnea, i.e., it is associated with less reduction in airflow and little or no hypoxia. In plethysmography or polysomnography, respiratory cycles are recorded from a decrease or increase in respiratory effort, which leads to the continuous appearance of arousals on physiological signals and ends before the criteria for apnea or hypopnea are met [17, 18]. This disorder can lead to serious consequences, such as loss of attention while driving [19]. Diagnosing sleep apnea with disorders other than apnea and hypopnea is a challenge. Therefore, our aim in this study focused on the automatic detection of arousal regions due to non-apnea disorders, especially RERA, which is more common.

There have been many studies in the field of arousal detection. There have been three methods used to develop an algorithm for detecting sleep disorders automatically in these studies. Most of these studies have proposed neural network-based algorithms. A group of studies [20–24] first considered the raw data as input of a neural network after making the necessary preprocessing on them, and another group first extracted a set of valuable features from the data after the necessary preprocessing to reduce data dimension and then trained their designed deep neural network with the extracted features [3, 25]. Another group of studies has used basic machine learning algorithms such as logistic regression, linear differential analysis [26], and support vector machine.

The remainder of this chapter is organized as follows: Section 11.2 discusses polysomnography. Section 11.3 is dedicated to sleep stages. Section 11.4 will cover the method. Results are reported in Sect. 11.5. Finally, the last section is the conclusion.

11.2 Polysomnography

Polysomnography is a nocturnal test during several hours of sleep in which different measurement techniques are used to simultaneously and continuously record neurophysiological, cardio-respiratory, and other physiological and physical parameters [27].

This method uses mentioned medical signals, pulse oximetry, airflow, and respiratory effort to provide data on physiological changes in many different systems of the body that are affected by sleep and therefore may not be present at waking time to assess the essential causes of sleep-related disorders. Polysomnographic recordings are divided into 30-second segments and are interpreted based on expert opinion and published sleep guidelines. Fig. 11.1 shows a segment of airflow and SaO2 signals. The patient has an apnea disorder in the green portion of the signal. When apnea occurs, the amplitude of the SaO2 signal and the airflow are reduced by more than 90%. Both signals are characterized by this reduction when apnea occurs. The part of the signal marked in blue is related to RERA arousal, and the decrease in airflow is less severe than in apnea.

11.2.1 EEG

The EEG signal is the main basis of the objective diagnosis of awakening, arousals, and sleep stages during the sleep study. In most cases, the brain waves are irregular, and there is no general pattern, but specific patterns appear in some cases, such as epilepsy. The EEG also contains waves with various frequency ranges, the presence of which in different brain regions corresponds to a specific activity or state. Rhythms in the EEG signal that include Delta, Theta, Alpha, Beta, and Gamma waves, as well as particular patterns in this signal that include Vertex sharp wave or V-waves, slow waves, K-complexes, and sleep spindles, are used to distinguish between waking, sleeping, classifying stages and disorders of sleep [6]. Table 11.1 shows the rhythms in the EEG signal. Figure 11.2 also shows these rhythms.

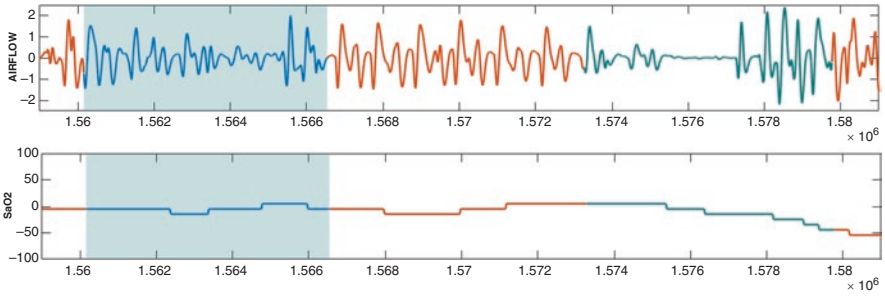

Fig. 11.1 A segment of airflow and SaO2 signals. The green parts are related to apnea arousal and the blue parts are related to RERA arousal

Table 11.1 The EEG rhythm specifications [28]

Delta	0.1–4 Hz	In adults, found in the frontal region
Theta	4–8 Hz	High amplitude waves in comparison to other EEG sub-bands
Alpha	8–12 Hz	Always seen in frontal and parietal
Beta	Above 12 Hz	During deep relaxation and while focusing

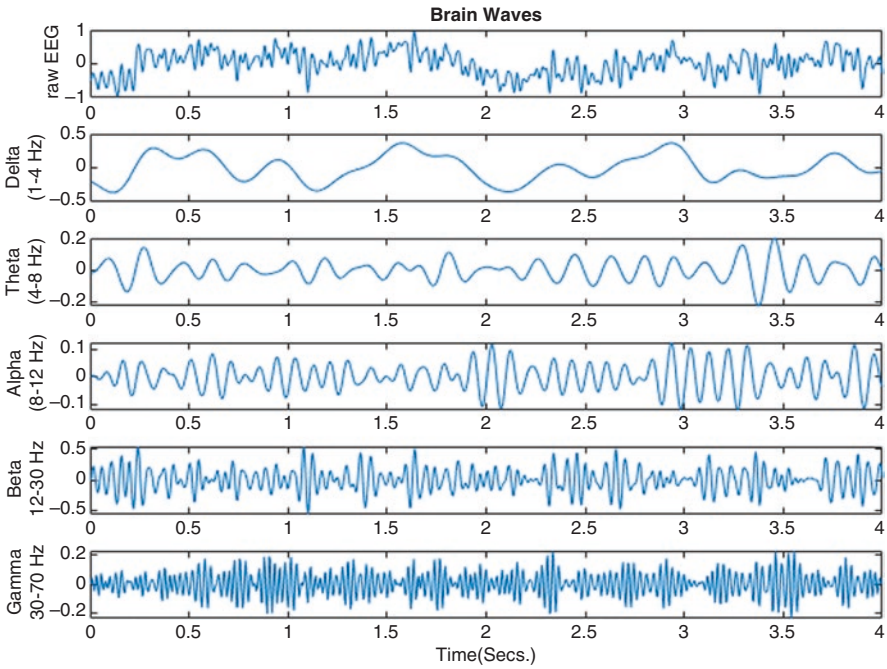

Fig. 11.2 Rhythms in EEG

11.2.1.1 Special Patterns in EEG

Identifying three specific patterns in sleep is very important in analyzing EEG recorded during sleep. Sharp vertex wave is one of the most important patterns of NREM sleep that are most often seen in stage N1 and sometimes in stage N2. They are seen in the EEG as quite distinctive, seven-shaped waves with peaks in the amplitude of about 100 to 200 microvolts [29]. Sleep spindles are a sequence of oscillations in the amplitude of the EEG signal, usually lasting between 0.5 and 3 seconds, and the frequency of these changes varies between 11 and 16 Hz [30]. K-complexes are complicated seven-shaped patterns of sleep waves that indicate the transition into a deeper stage of sleep. This pattern is seen as a sharp negative wave with a high amplitude followed by a low slope wave. K-complexes have a duration criterion that must at least last 0.5 seconds. K-complexes may include other events like periodic movements of the limbs in sleep, apnea, or appear with arousal (Alpha rhythm in EEG) [6].

11.3 Sleep Stage

REM and NREM are the two main stages of sleep. NREM sleep is also divided into four stages. Stages 1 and 2 are called the light stage of sleep, and stages 3 and 4 are called the deep stage or slow-wave sleep. There are four or five sleep cycles during each night, each consisting of one part of NREM sleep followed by REM sleep [31]. The first stage is wakefulness which may last 5 to 10 minutes. At this stage, the brain is between sleep and wake state. The predominant brain waves at this stage are Alpha waves. In the EOG signal at this stage, blinking and rapid movements are observed. Stage N1, or somnolence**,** is another stage of NREM sleep. At this stage, called light sleep, EEG activity is fast with low voltage. This stage is defined when Theta waves (4 to 7 Hz) form more than 15 seconds (more than half) of the epoch. The next stage is N2, also known as sigma, and the sleep spindle. This stage is in the middle of sleep and, at first, takes about 20 minutes. At this stage, sleep enters a deeper level, and heart rate and the body temperature decrease. Next is stage N3, which is also called deep sleep, sleep with slow waves, or delta. At this stage, muscle and brain activities decrease. The last stage is REM which is also called active or paradoxical sleep. At this stage, eyes are closed but have a rapid movement. In this state, the brain waves show an activity similar to the state of awakening. Dreaming is at this stage of sleep, and this stage occurs in adults between 90 and 120 minutes after the beginning of sleep [6, 32].

11.4 Methodology

In this section, we discussed the ensemble learning algorithms used in this study, the performance evaluation criteria, the dataset, their preprocessing process, extracted features, and methods for selecting the effective features.

11.4.1 Ensemble Learning

An ensemble learning algorithm is a machine learning method that incorporates a number of simple and weak algorithms to form a strong learning algorithm with higher accuracy. This type of learning is one of the best ways to maintain the balance between variance and bias. An effective model should be able to balance these two types of errors. Ensemble methods can be mainly divided into three groups: boosting, bagging, and stacking. In the present study, we use the random forest and XGboost algorithms, which apply “bagging” and “boosting” methods, respectively.

11.4.1.1 Bootstrap Aggregation (Bagging)

The bagging method creates algorithms parallel to one another. Several independent models are generated, and their predictions are averaged. The variance of the combined base models is reduced in proportion to the number of models or samples. In addition, the independence of the basic learning models significantly reduces the error through the application of averaging.

Bagging is used when the goal is to variance reduction of the decision tree classifier without impacting bias. The reduction of variance leads to increased accuracy and avoidance of overfitting, which is a major challenge in many predictive models. Multiple subsets of training instances are randomly created with replacement, and the data from each subset is used to learn the corresponding decision tree. In this way, a number of different models are eventually created. The average of all the predictions from the individual trees is the final result, which performs better than a single decision tree classifier.

11.4.1.1.1 Random Forest (RF)

RF is based on the ensemble learning algorithm and involves a large amount of classification and regression trees (CARTs). The core of the RF algorithm relies on the concept of 'bagging' and selecting features randomly. At the beginning of the RF algorithm, a bootstrap sample is taken from the training data. In each CART node, features are randomly selected, and feature subsets are searched to find the optimal split. An unpruned CART is formed when this process is repeated at each node. Additionally, for evaluating the model performance and estimating feature importance, bootstrap sample data - known as out-of-bag sample (OOB) – can be used as a test set. Repeating this process for each tree will result in a fully grown forest.

11.4.1.2 Boosting

Boosting is a sequential ensemble method in which the weight of each input is modified according to the latest classification. In boosting, when an input is misclassified by one hypothesis, its weight is increased, so the new hypothesis is more likely to be classified correctly. The goal at each step is to improve the prediction accuracy over the previous tree's performance. In boosting, like bagging, a single algorithm is used as the basic learning model, and in the process, weak learning models become a model with more robust performance. Boosting comes in many forms, including Adaboost, GBDT, XGBoost, and LightGBM. In this study, the LightGBM algorithm, the most recent one, is used, which is explained in the following.

11.4.1.2.1 Gradient Boosting Decision Tree (GBDT)

Another ensemble learning algorithm that uses several decision trees as base learners and a gradient descent optimization algorithm to calculate the error of its previous models is GBDT. Newly added decision trees increase the emphasis on samples misclassified by the previous decision tree.

In the GBDT algorithm, the residual of the previous decision tree is intended as input to the following decision trees. Then, the new decision tree adds to minimize the residual so that each time the iteration is repeated, the cost decreases as it progresses in the negative gradient direction. All decision tree results are then used to determine the prediction result.

Supposing the data samples are represented $\left(\mathrm{x_i},\mathrm{y_i}\right)_{i=1}^{m}$.

Which $\mathrm{x_i}$ is data value and $\mathrm{l_i}$ is the predicted label. The GBDT algorithm is explained as the follows.

First the initial constant value θ is obtained from the eq. (11.2).

$$F\left(x,\theta\right)=\sum_{m}^{n=1}L\left(\mathrm{y_n},\theta\right) \tag{11.1}$$

$$\mathrm{F_0}\left(\mathrm{x}\right)=\min_{\theta}F\left(x,\theta\right) \tag{11.2}$$

Where $\mathrm{L(y_i}, \theta)$ is the loss function.

Based on the gradient, the residual is calculated as follows:

$$\widehat{\mathrm{y}}_{\mathrm{i}}=-\left[\frac{\partial L\left(\mathrm{l_i},\mathrm{F}\left(\mathrm{x_i}\right)\right)}{\partial F\left(\mathrm{x_i}\right)}\right]_{\mathrm{f(x)=f_{n-1}(x)}} \tag{11.3}$$

Where n = 1, 2, …, N indicates the number of iterations.

The initial model T $(\mathrm{x_i}; \alpha_n)$ is determined by using the sample data, and α_n is computed based on the least square method as follows:

$$\mathrm{a_n}=\min_{\alpha,\theta}\sum_{i=1}^{m}\left(\widehat{\mathrm{y}}_{\mathrm{i}}-\theta T\left(\mathrm{x_i};\alpha_{\mathrm{n}}\right)\right)^2 \tag{11.4}$$

Based on the loss function minimization, the weight of the model is computed as follows:

$$\theta_{\mathrm{n}}=\min_{\theta}\sum_{\mathrm{m}}^{\mathrm{i=l}}\mathrm{L}\left(\mathrm{y_i},\mathrm{F_{n-l}}\left(\mathrm{x}\right)+\theta T\left(\mathrm{x_i};\alpha_{\mathrm{n}}\right)\right) \tag{11.5}$$

The model can be written as follows:

$$\mathrm{F_n}\left(\mathrm{x}\right)=\mathrm{F_{n-1}}\left(\mathrm{x}\right)+\theta_{\mathrm{n}}\mathrm{T}\left(\mathrm{x_i};\alpha_{\mathrm{n}}\right) \tag{11.6}$$

This loop continues until the number of iterations is defined or until the convergence is [33].

11.4.1.2.2 Light Gradient Boosting Machine (LightGBM)

LightGBM is another ensemble algorithm that is designed by Microsoft using the GBDT framework. It is intended to improve computation performance and thus provide a more accurate solution to the problems associated with handling big data in prediction. To calculate the information gain of all possible split points, conventional GBDT implementations must examine all sample data for each feature, which takes much time when dealing with big data (both for the number of features and data size). Thus, computational complexions would be proportional to the number of features and samples. In LightGBM, a histogram-based algorithm and growth strategy through the leaves of trees are adopted with maximum depth size to enhance training speed and optimize memory utilization.

In order to address these limitations and further improvement of model efficiency, the LightGBM algorithm presents *Gradient-based One-Side Sampling (GOSS)* and *Exclusive Feature Bundling (EFB)*.

(a) *Gradient-based one-side sampling*

In Adaboost, a sample's weight works as a good indicator of the importance of data samples. However, in GBDT, there is no local weight for the samples, so the proposed sampling methods for Adaboost cannot be applied directly. Gradient for each sample in GBDT provides practical information for data sampling. In other words, a sample with a small gradient had a small training error and had been well trained. The simple idea is to put aside data samples with small gradients. Nevertheless, it changes the data distribution and negatively affects the accuracy of the trained model. The Gradient-based One-Side Sampling (GOSS) method was developed to minimize this problem and maintain the accuracy of trained decision trees while reducing data samples.

GOSS maintains all samples with large gradients and takes randomly the ones with small gradients. GOSS selects the top a × 100% samples after sorting the data samples based on the absolute value of their gradients. Then it randomly selects b × 100% of samples from the remaining data. Then, for amplifying the sampled data with small gradients, GOSS uses a constant $\frac{1-a}{b}$ for computing the information gain to compensate for the effect of data distribution. In this way, it is possible to focus more on the data that has not been trained yet, without any changes to the original data distribution. More details about the theory of GOSS can be found in [34].

(b) *Exclusive Feature Bundling*

The purpose of the EFB method is to effectively reduce the number of features. In general, high-dimensional data are sparse. By considering the sparse nature of the feature space, we can devise a method of reducing the number of features almost without losing any information. Particularly, many features in sparse feature space are mutually exclusive; that is, they cannot take nonzero values simultaneously. Exclusive features can be grouped into a single feature (an exclusive feature bundle). By a carefully designed feature scanning algorithm, identical feature

histograms can be constructed from these feature bundles as the same histograms for single features. Thus, the histogram structure complexity from the number of features multiplied by the number of samples changes to the number of feature bundles multiplied by the number of samples, in which the number of feature bundles is much less than the number of features alone. In this way, the speed of GBDT training can be significantly increased without compromising accuracy. To do this by turning the ranking problem into a graph coloring process (adding edges for both non- exclusive features and features as vertices) and its solution is designed by another algorithm with a fixed approximation ratio of an effective algorithm.

Compared to the level-wise growth run in the XGBoost algorithm, the limited growth process through leaves used by the LightGBM algorithm is more efficient; because it is distributed only through a leaf with the highest information gain, and as a result, it achieves much better accuracy than other boosting algorithms, and the limited depth can effectively prevent overfitting. As mentioned, this algorithm is surprisingly fast; hence the word "light" is dedicated to it [34]. Studies have indicated that this algorithm can make the training process 20 times faster [35].

11.4.2 Evaluating Performance

For evaluating the performance of the implemented algorithms, we used two measurement criteria: the Area Under the Receiver Operating characteristic Curve (AUROC) and the Area Under the Precision-Recall Curves (AUPRC).

AUROC indicates the ability of a model to classify samples accurately. A ROC curve represents the exchange between true positive rate (TPR) and false positive rate (FPR) at a range of decision thresholds. In ROC curves, the decision threshold is implicit. There is no axis for decision thresholds, and AUROC is the area under the ROC curve. On the ROC curve, the x-axis represents the FPR, and the y-axis represents the TPR. AUROC for a specified curve is simply the area below it. The lowest AUROC value is 0.5, and the highest is 1.

The AUPRC criterion is used to evaluate model performance when classes are highly unbalanced and positive samples are essential. The precision-recall curve illustrates the relationship between precision and recall for different thresholds. Having a large area under the curve indicates high precision and recall of the model, where high precision is related to a low false positive rate (FPR) and high recall is related to a low false negative rate (FNR). A high value for both indicates that the classifier returns correct (high precision) and the most positive (high recall) results [36]. The area under the PR curve is known as the AUPRC. PR curves represent the tradeoff between accuracy and sensitivity over a range of decision thresholds. A recall is also known as a true positive rate (TPR). As a result, both AUPRC and AUROC use TPR.

Unlike the ROC curve where the y-axis represents recall, the x-axis represents the FPR, the x-axis represents the PR-recall curve, and the y-axis represents precision.

One of the features of the PR curve is that it does not use TNs in any way. According to the definitions of precision and recall, they are calculated as follows:

$$Recall = TPR = \frac{True\ Positives\left(TP\right)}{True\ Positives\left(TP\right) + False\ Negatives\left(FN\right)} \tag{11.7}$$

The recall of a classifier can be defined as the ability to correctly identify all positive samples.

$$Precision = \frac{True\ Positives\left(TP\right)}{True\ Positives\left(TP\right) + False\ Positives\left(FP\right)} \tag{11.8}$$

A classifier's precision is its ability to avoid mislabeling a negative sample as positive.

In PR curves, there are no TNs, so AUPRC will not be affected by TNs. If AUPRC is used for datasets with 95% negative and 5% positive samples, it focuses on how 5% of the samples are treated. If the model performs well in detecting positive samples, the AUPRC will be high; if not, it will be low. The AUPRC is therefore most useful when its baseline is the lowest because it focuses on the smallest fraction of positive samples possible across large datasets with large numbers of TNs [37].

11.4.3 Data Description

As mentioned earlier, the current study aimed to provide a method for automatic detection of arousals related to non-apnea sleep disorders using polysomnographic signals with ensemble learning. In this study, we used Physionet Challenge 2018 database [16].

A total of 1983 polysomnographic data were provided, which included 994 subjects for training and 989 subjects for the test. Polysomnography in these data is recorded according to the AASM standard. A total of 13 signals are provided, including:

- six EEG channels; F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, and O2-M1, based on the international 10–20 system;
- one left eye EOG channel, referenced to the right ear EEG electrode (M2);
- Three EMG channels, one on the chin and two channels of breathing signal from the abdomen and chest;
- one airflow channel;
- one oxygen saturation channel (SaO2);
- And one ECG channel located below the right clavicle close to the sternum and across the left lateral chest wall

All signals were measured at a sampling frequency of 200 Hz except SaO2. SaO2 was upsampled to 200 Hz to synchronize samples using a sample and hold. Also, all signals are measured in microvolts.

The specialists labeled the training data in two different ways in the data source. In the first state, the samples were labeled into three classes −1, 0, and + 1. In these three classes:

- Label +1 is related to non-apnea and non-hypopnea arousals
- Label 0 is related to normal sleep
- Label −1 is related to apnea and hypopnea arousals

In the second state, the labeling has been done more precisely, and types of arousals related to each region are also determined. The sleep stages were labeled too.

According to the available data, arousals that were considered in this study are classified as spontaneous arousals, RERA, hypoventilation, bruxisms, hypopneas, snores, apneas (central, obstructive and mixed), vocalizations, periodic leg movements, Cheyne-Stokes breathing or partial airway obstructions.

The proposed method was evaluated using recordings belonging to 100 subjects from the training group. 80% of this data, which includes 80 subjects, have been considered for the training stage, and 20%, including 20 subjects, have been considered for the test stage [16].

11.4.4 Pre-Processing

Before extracting the features from the data, the signals were preprocessed to remove noise.

11.4.4.1 EEG and EOG Signals

Due to the non-invasive nature of EEG signals, the distance between the signal source and the recording electrode, as well as the low amplitude of this signal, adding unwanted signals can reduce the signal to noise ratio and cause many problems in the analysis of these signals; as a result, it is necessary to preprocess them before analyzing.

Both EEG and EOG signals are first preprocessed using a fourth-order low-pass Butterworth filter with a cut-off frequency of 42 Hz and then a fourth-order Butterworth high-pass filter with a cut-off frequency of 1 Hz.

Although EEGs are typically interpreted as recording brain activity, other electrical activities are also captured. Extra activities tend to be harmful artifacts, whether they are physiological, like EOG, EMG, or ECG, or extra artifacts like power line interference. Frontal electrodes are mainly responsible for recording EOG measurements, however they can also affect other electrodes [38]. Furthermore, due to the volume conduction effect, both effects of eye activity and EEG activity are diffused

to the surface of the head and recorded by electrodes. EOGs usually have a greater amplitude than EEGs. However, due to its similar frequency to EEG signals, removing this artifact is one of the most critical problems in studying brain activity [39].

The power spectrum of some of these artifacts created during recording is different from the power spectrum of the EEG, making its removal easy with an IIR filter. Even so, ECG and EOG, whose power overlaps significantly with EEG, are not easily removed and require specialized algorithms [40].

For this purpose, we used the adaptive filter, which is a successful method in the non-stationary signals processing field. This method uses raw EEG signals as inputs and EOG and ECG artifacts as reference signals. Pure EEG can be obtained by reducing the filtered output. The ability to adapt the parameters of an adaptive filter based on the minimization of mean square error (MSE) is between the output signal of the system and the desired signal. The recursive least squares (RLS) and the least mean square (LMS) are commonly used algorithms of adaptive filter algorithms. The advantages of the RLS algorithm over the LMS are its stability and fast convergence speed. Therefore, we used the RLS algorithm to adjust the filter coefficients to ensure that the reference signal has the most appropriate specifications for these artifacts.

To do so, we used the method proposed in [38]. In order to detect the additional presence of EOG and ECG, the filtered EEG, EOG, and ECG signals after being divided into segments with 2048 samples, the cross-correlation between each EEG segment with the respective EOG and ECG sections has been calculated. Adaptive RLS filters are used when the cross-correlation coefficients between EEG-EOG or EEG-ECG are greater than two thresholds of 0.5 and 0.2, respectively. If the threshold values for the cross-correlation coefficients are not met, the corresponding EEG part is transferred without changing the output. In Fig. 11.3 More details can be found in [38].

11.4.4.2 EMG Signal

After preprocessing with a Butterworth low-pass filter with a cut-off frequency of 70 Hz, the EMG signals were then passed through a Butterworth high-pass filter with a cut-off frequency of 10 Hz. Then, an adaptive filter was applied to remove ECG artifacts.

11.4.4.3 ECG Signal

This signal is also passed through a fourth-order Butterworth low-pass filter, with a cut-off frequency of 25 Hz, and a high-pass filter with a cut-off frequency of 0.1 Hz with the same specifications. Fig. 11.4 shows the ECG signal before and after the filtering.

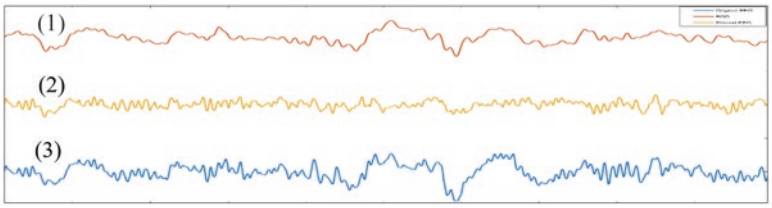

Fig. 11.3 One segment of (1) EOG signal, (2) EEG after the removal of EOG by RLS method and (3) EOG impregnated EEG

11.4.4.4 Airflow Signal

In order to preprocess this signal, in the first step, the DC value is removed from the signal. Thus, a window of 20,000 samples is considered. The median of the window is subtracted from the total in the length of the signal. In the next step, a fourth-order low-pass filter with a cut-off frequency of 5 Hz is applied to the signal. Fig. 11.5 shows one segment of the airflow signal before and after the filtering.

Any extra preprocessing step applied to SaO2 signal.

11.4.4.5 Signal Segmentation

One of the most critical early steps in biosignal processing is to segment the signals into smaller and relatively stationary segments. Each part must have nearly the same statistical characteristics, such as amplitude and frequency. In this study, first, the data are normalized using mean and standard deviation. Then, to divide the data into shorter epochs, we take some steps to choose the appropriate epoch size, and then labeling was done on each epoch.

Since arousals are the minority class, we first extracted the lengths of the signal intervals in which the arousal occurred continuously in those areas. First, the minimum and maximum arousal intervals were calculated for each individual to determine the most appropriate window size. About 95% of these intervals contain at least 1750 samples and in terms of time is 8.75 seconds. As a result of applying an adaptive filter after windowing, 149 samples were added to this range at the beginning and end to reduce distortion of the labeled signal. It will also be helpful for applying the hamming window to extract the frequency properties.

After analyzing the minimum arousal interval lengths of non-apnea among 994 patients, we realized that forty-nine patients had no non-apnea/hypopnea arousals, and no sample labeled +1 was observed in their data. We did not include these people in the training phase.

Therefore, each biosignal was segmented into an epoch with 2048 samples with 50% overlapping to eliminate sharp transitions between segments in such a way that 1750 samples in the middle of each segment must have the same label, and the whole window is considers to have that label.

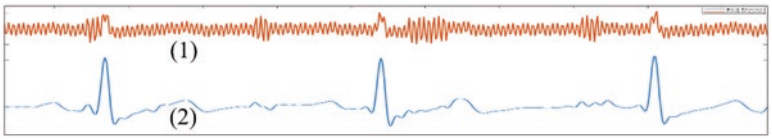

Fig. 11.4 A segment of (1) the noisy and (2) filtered ECG signal.

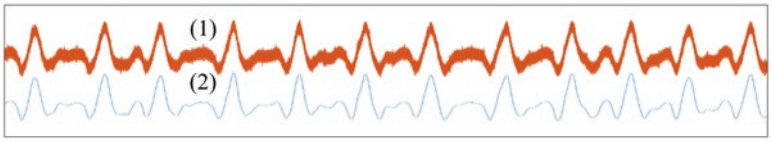

Fig. 11.5 A segment of (1) the noisy and (2) filtered airflow signal.

11.4.4.6 Labeling Epochs

The following were considered non-apnea arousals:

- In the period of two seconds before and ten seconds after a RERA arousal,
- A period of 2 seconds before a non-RERA, non-apnea arousal happens, up to 2 seconds after it ends [16].

In this study, we consider segmentation in the exact areas of arousal in the signals. In other words, the extra samples labeled at the beginning and end of arousal are not considered and are labeled 0 (Except for the 149 samples described above).

11.4.5 Feature Extraction

A detailed discussion of all the features that are extracted from signals is provided in this section.

11.4.5.1 Features Extracted from EEG Signals

EEG signals can provide helpful information about Sleep-related breathing disorders. This signal is, in fact, one of the most critical signals in diagnosing sleep-related breathing disorder [41]. Different features are extracted from each fragment of the preprocessed and segmented EEG signals, generally divided into the temporal, frequency, and time-scale domains.

Statistical features are some of the most common ones that can be extracted from the EEG. In this study we extracted median, mean, zero-crossing rate (ZCR), standard deviation (SD), variance, kurtosis, skewness, Root Mean Square (RMS), correlation between channels and three Hjorth parameters as time domain features.

The Activity parameter is the power of the signal (which is filtered wideband), the Mobility parameter is the mean frequency, and the Complexity parameter is the frequency change [42]. Studies show that changes in power along the EEG frequency bands provide information about arousal detection [43].

11.4.5.1.1 Frequency Features

As frequency features, the power spectral density estimate using the Welch Method with a hamming window with a length of about 2 seconds or 256 samples has been used to extract power from different EEG frequency bands. The relative and absolute power of the delta, theta, alpha, and beta bands are extracted from each segment. In addition, the power ratio of each sub-band relative to the others is also considered a feature.

11.4.5.1.2 Time-Frequency Features

Then the Deabeuchies order 8 Discrete Wavelet Transform (DWT) is used to decompose the signals into five stages. The coefficients obtained from this analysis include five detailed coefficients and an approximation coefficient. From these six coefficients, seven statistical features were extracted: Mean Absolute Value (MAV), variance, Average Power (AVP), standard deviation, entropy, mean, and skewness.

11.4.5.1.3 Nonlinear Features

Another category is nonlinear features. Because the EEG signal behaves erratically, entropy and fractal methods can provide a good indication of the irregularity of the signal. In other words, the more non-stationary the behavioral signal, the higher the entropy and fractal dimension. As entropy-based features, we extracted Renyi and relative spectral entropy that Renyi entropy provides information about the randomness of a system numerically and is the generalized Shannon entropy.

(a) ***Features based on fractal dimension***

A new method for measuring the dimension of irregular shapes was first proposed by Mandelbrot [44]. The fractal dimension is a non-linear measure of irregularity. As a measure of self-similarity, fractal dimension describes the behavior of random signals [45]. In this study, the methods of Katz and Higuchi for calculating the fractal dimension were used. Two algorithms, Higuchi fractal dimension (HFD) and, Katz fractal dimension (KFD) are used in this study.

- *Katz fractal dimension*

Katz fractal dimensions are shown below for a signal of length n:

$$D = \frac{\log(n)}{\log(n) + \log\left(\frac{d}{T}\right)} \tag{11.7}$$

Where T is the segment length (Euclidean distance between consecutive data points) and d indicates the maximum distance between the initial and other points [46].

- *Higuchi fractal dimension (HFD)*

The HFD is a fast, nonlinear method of measuring fractal dimension that gives more accurate results. Due to the shift in time series structure over a particular characteristic frequency, it is challenging to determine power law indices and a characteristic time scale based on the power spectrum. With very limited data points, the HFD method can still provide stable indices and time scales as a function of the characteristic frequency. Below are the equations of Higuchi's algorithm.

From a L-length one-dimensional time series x(1), x(2), …, x(l) with equal intervals, form a new time series X_n^k:

$$X_n^r = \left\{ x(n), x(n+r), x(n+2r), \ldots, x\left(n + \left|\frac{L-n}{r}\right|.r\right) \right\} \tag{11.10}$$

Where r represents the distance between two consecutive time series and n = 1, 2, …, r is the time value at the beginning of the sequence. For each x_n^r, the average length of each time series is calculated as follows:

$$L_n(r) = \sum_{[(n-L)/r]}^{i=1} \left| x(n+ir) - x(n+(i-1).r) \right| . \frac{(L-1)}{\left[(L-n)/r\right].r} \tag{11.11}$$

As a result, the average length of the whole series of discrete time signals is as follows:

$$L(r) = \frac{1}{r} \sum_{r}^{n=1} L_n(r). \tag{11.12}$$

The following result can be obtained using logarithmic transformation:

$$\ln(L(r)) \propto D_f . \ln\left(\frac{1}{r}\right). \tag{11.13}$$

The D_f slope is the Higuchi fractal dimension [47].

(b) ***Detrended Fluctuation Analysis (DFA)***

The other feature extracted from the EEG signal is the Features of *detrended fluctuation analysis (DFA)*. DFA is one of the nonlinear methods used to determine the fractal scalability properties of the EEG signal that is a modified method of root mean square for the random walk. A function of scale parameter is used to calculate the mean square distance of a signal from a trend line. Usually, there is a power-law dependency, and the main parameter is power. The terms related to it are explained below.

The time series $s(i)$ represents the EEG signal contribution of N_s samples. In the DFA algorithm, first, time series are corrected and integrated. Afterward, the integrated time series $s(\mathrm{m})$ is divided into separate sections of the same length. The length of this section in this study is considered as 1024 samples. The local trend is calculated for each section. By subtraction of the local trend $s_n(m)$ in each section, this trend is removed from the EEG time series $s(\mathrm{m})$. After determining the variation of the root square mean of this non-trended time series F(n), it can be expressed as follows:

$$\mathrm{F}(\mathrm{n}) = \sqrt{\frac{1}{N_s}\sum_{N_s}^{m=1}\left(S(m) - S_n(m)\right)^2} \tag{11.14}$$

As a function of section size, F(n) is calculated throughout all time scales to identify its dependence on average variations. On a log-log plot, linear relationships demonstrate fractal scaling. Therefore, the fluctuation is proportional to n^{α}. The parameter α, which indicates the variation rate of a complex signal, is the self-similarity or autocorrelation parameter. It can be defined as

$$\alpha \propto \frac{\log F(n)}{\log n} \tag{11.15}$$

By finding the slope of log F(n) - log n line, this can be calculated [48]. The value of this parameter is considered a feature.

Similar features as EEG signals are also extracted from the EOG signal.

11.4.5.2 Features Extracted from EMG

From the EMG signal, features such as MAV, AVP, standard deviation, variance, 95th percentile, skewness, kurtosis, frequency band energy of each four Hz window (five frequency bands), waveform length (WL), ZCR, the derivative zero-crossing rate, amplitude, and root mean square, are extracted.

11.4.5.3 Features Extracted from SaO2

From the SaO2 signal, thirteen features are extracted, including mean, standard deviation, coefficient of variation, kurtosis, mean, variance, and range. Since arousals occur in a few seconds, power spectrum density is calculated using Yule-Walker autocorrelation estimation of order 5, and an average frequency range of 0.016–5 Hz is also considered a feature. In addition, the percentage of time when $SaO2 \geq 96$, $90 \leq SaO2 \leq 96$, $80 \leq SaO2 \leq 90$ and $SaO2 < 80$ in each window used to determine the duration of normal, mild, moderate, and severe oxygen saturation are also extracted as features.

11.4.5.4 Features Extracted from Airflow

Features extracted from airflow, similar to other signals, are in the time and frequency domain. These features include statistical features such as respiration rate, frequency band energy per 0.25 Hz, average, minimum, maximum, range, variance, coefficient of variation, skewness, and kurtosis, and also other features like integral of absolute value (IAV), MAV, ZCR, slope sign changes (SSC), WL, RMS. Since the airflow is also nonstationary, Welch's method is used to calculate power spectrum density as another feature.

11.4.5.5 Features Extracted from ECG

Features such as heart rate, heart rate variability, low-frequency band-power, and the ratio of low-frequency band-power (0.003 to 0.4 Hz) to high-frequency power (0.04 to 0.15 Hz) are extracted from ECG signals.

This study considers another set of features associated with signal variation over time. Due to the time intervals between the selected segments and their effect, it's necessary to effectively use the features of each segment adjacent. However, considering the purpose of this study of detecting non-apnea arousals and ineffective regions related to apnea, not all selected segments have adjacent necessarily. For this reason, we used the ratio between the first half of the segment and its second half as new features extracted from each segment in order to consider the course of change in each segment. In this way, each segment first is divided into two equal parts, and the features considered for each part are extracted separately. Then, the ratio of these extracted features in two parts is considered the new feature.

Also, the autocorrelation of the segments is considered. In this way, autocorrelation is calculated from all segments, and then time features are extracted from them.

A total of 1741 features were extracted from 13 channels. Table 11.2 shows the number of features extracted from each channel.

Table 11.2 The number of features that extracted

	EEG + EOG	Corr of EEGs	EMG	SaO2	Airflow	ECG	All
Initial features	1160	15	325	26	114	101	1741

11.4.6 Data Balancing

In order to improve the classifier learning performance and also to apply statistical analysis, the data in the two classes need to balance. Due to the significant imbalance, the class needs to choose a method for balancing the data. In this method, the classifier is trained each time by a random subset of data. The second approach is increasing the minority class by multiplying a little random noise of about 0.1% by the feature values of each part of the class labeled 1.

11.4.7 Feature Selection

There are two general types of feature selection methods: individual evaluation and subset evaluation. Feature ranking also known as individual evaluation measures the relevance of individual features by giving them a score. However, using this approach provides a subset of selected features according to a specific method of searching. This method compares each selected subset to the previous best by evaluating it based on a specific evaluation criterion.

Besides this category, three general feature selection approaches are filter, wrapper, and embedded. In this study, the first two approaches are used.

The filter method does the feature selection process as a preprocessing step without considering any specific classification algorithm and only by ranking features based on their importance in the dataset. The advantage of this method is its low computational cost and good generalization ability. In this study, we used two methods of correlation-based feature selection (CFS) and the Wilcoxon rank-sum test among the filtering methods [49, 50].

To determine the usefulness of a subset of features, the wrapper method uses a learning algorithm as a black box. In order to find practical features to improve the performance and reduce the speed of classification, the significance of the extracted features at the significance level of 0.01 was evaluated using Wilcoxon statistical test. This test was used as the first step in selecting better features, and features with were removed from the feature vector. Thus, 320 features out of a total of 1741 features did not reject the Wilcoxon null hypothesis of having samples of continuous distributions with different medians and so were removed from the feature vector.

Then, the 711 features with higher scores are selected by ranking the features using the CFS method, and then by using the genetic algorithm with KNN fitness function, which is one of the wrapper methods, the features have increased to 249. Table 11.3 shows the number of features selected in each step.

Based on this table, we can see which of these 13 signals contributes most to classifying target class or non-apnea arousals. Among the selected features, the relative power of different EEG frequency bands and their autocorrelation signals, the ratio of the first half spectrum power to the second half of the epoch related to EEG and EMG signals, 95th percentile autocorrelation of EEG signals can be named.

11.5 Classification Result

In this section, the results of the classifiers are reported in several steps. First, using the features obtained after the CFS, i.e., 711 features, the performance results of the two approaches of data reduction and data augmentation in order to balance. Tables 11.4, 11.5, and 11.6 shows the results of the RF algorithm with 50 trees in both steps. RF parameters are selected experimentally. According to the definition of the AUPRC criterion, it is necessary to calculate the baseline to compare the value of this criterion. The baseline in each step is equal to the total number of epochs with label 1 to the total available data. In Table 11.4, the baseline for the test data used is 0.114.

Data augmentation improves classification results because all zero-class data samples are included every time that model is trained, so the whole dataset is taken into account when used to train the model. For this reason, this method has been used to balance the data in the following.

In this section, the random forest algorithm is implemented separately on the extracted initial features of each of the 13 channels. The results are presented in Table 11.5 N is the number of features of each channel.

As a result of measuring the importance of each feature group for the detection of non-apnea arousal with the AUPRC criteria, it can be concluded that airflow with a score equal to 0.367 is the second most important feature after the EMG signal. The best result is related to the EMG signal with AUPRC equal to 0.375. This is understandable since more than 90% of the non-apnea arousal regions identified in the data include RERA, related to respiration. The best result is related to the EMG signal with AUPRC equal to 0.375. The reason for this is understandable since more than 90% of the non-apnea arousal regions identified include RERA, which is a breath-related disorder. Among the EEG signals, the features of the C3-M2 central channel were more important than the other channels, and this channel alone can be used to identify arousals.

Table 11.3 The number of features that selected in each stage of feature selection

Selected by	EEG + EOG	Corr of EEGs	EMG	SaO2	Airflow	ECG	Total
Wilcoxon	952	10	286	22	61	90	1421
CFS	496	2	163	4	22	24	711
GA	166	1	60	2	12	8	249

Table 11.4 Comparing the performance of two approaches of data reduction and data augmentation for balancing

Random Forest					
Data Reduction			Data Augmentation		
	Recall/ TPR	F1-score		Recall/ TPR	F1-score
Class 0	0.814	0.882	Class 0	0.834	0.894
Class 1	0.750	0.457	Class 1	0.793	0.487
Macro average	0.782	0.669	Macro average	0.801	0.685
Weighted average	0.807	0.834	Weighted average	0.829	0.848
Accuracy	0.806		Accuracy	0.811	
Baseline	0.114		Baseline	0.114	
AUPRC	0.378		AUPRC	0.436	
AUROC	0.834		AUROC	0.860	

Table 11.5 The result of applying random forest on each signal

AUROC	AUPRC	N		AUROC	AUPRC	N	
0.789	0.329	99	Chin EMG	0.751	0.277	136	F3-M2
0.740	0.265	92	Abdominal EMG	0.724	0.310	136	F4-M1
0.722	0.298	95	Chest EMG	0.766	0.321	136	C3-M2
0.815	0.375	286	All channels of EMG	0.748	0.309	136	C4-M1
0.802	0.367	61	Airflow	0.742	0.293	136	O1-M2
0.770	0.298	22	SaO2	0.752	0.305	136	O2-M1
0.791	0.252	90	ECG	0.814	0.332	816	All channels of EEG

Table 11.6 Comparison of results for two ensemble algorithms under the same condition

Bagged Tree (RF)				LightGBM			
		Predicted				Predicted	
		Class 0	Class 1			Class 0	Class 1
Actual	Class 0	85%	15%	Actual	Class 0	86%	14%
	Class 1	27%	73%		Class 1	26%	74%
		Recall TPR/	F1-score			Recall TPR/	F1-score
Class 0		0.847	0.900	Class 0		0.861	0.909
Class 1		0.724	0.497	Class 1		0.734	0.521
Accuracy		0.832		Accuracy		0.846	
Macro average		0.785	0.698	Macro average		0.797	0.715
Weighted average		0.833	0.854	Weighted average		0.847	0.865
AUPRC		0.466		AUPRC		0.496	
AUROC		0.865		AUROC		0.877	

Figure 11.6 shows the performance of the four classifiers using the AUPRC criteria on each feature category compared to each other. As it turns out, the respiratory properties performed best.

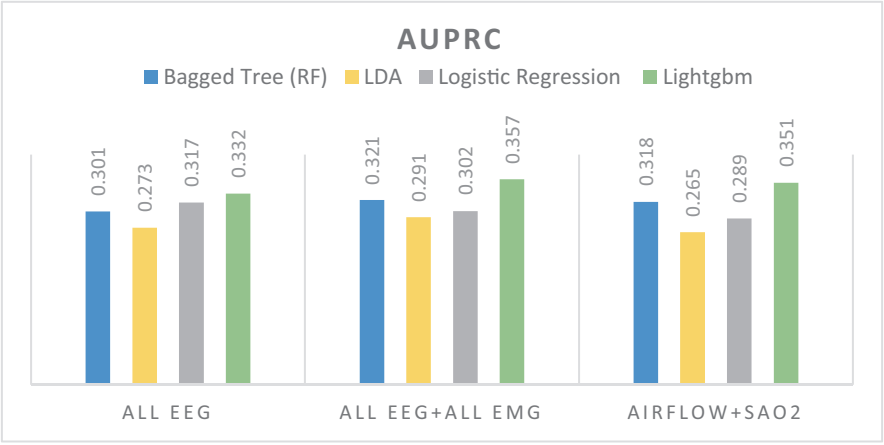

Fig. 11.6 The performance of classifier on different feature set

The mentioned algorithms are trained using 249 features obtained after applying the genetic algorithm with the five-fold cross-validation method, and the test results are given in Tables 11.6 and 11.7. The baseline here is about 0.18.

Figure 11.7 shows the diagrams for two main criteria for measuring performance in this study, AUPRC, and AUROC. As can be seen, in these diagrams and the results above, the ensemble algorithms, LightGBM of boosting category, and random forest of bagging category, in contrast to the individual algorithms, have a significant ability to identify arousal regions. The LightGBM algorithm performed better with much higher speeds than the other algorithms.

11.6 Conclusion

Studies show that detecting non-apnea arousal regions has less accuracy than identifying apnea arousals, and most of the studies have been done to identify apnea. The goal of the challenge held by Physionet in 2018, is to identify non-apnea arousal regions. In more than 90% of the models presented in this challenge, neural networks were used. The best results were obtained using deep neural networks with the area under the precision-recall and the area under the precision-recall scores of 0.93 and 0.54, respectively. However, intelligent feature extraction can still be promising, as only four papers in this challenge have applied pattern recognition methods using manual feature extraction and reached to the area under the precision-recall in the range of 0.21 to 0.29.

In this study, different features of all available signals have been extracted, and using ensemble algorithms that have received much attention today, results have been obtained that are relatively close to neural network-based methods.

Table 11.7 Comparison of results for two basic machine learning algorithms under the same condition

Linear discriminative Analysis				Logistic Regression			
Predicted		Predicted				Predicted	
		Class 0	Class 1			Class 0	Class 1
Actual	Class 0	89%	11%	Actual	Class 0	90%	10%
	Class 1	53%	47%		Class 1	49%	51%
		Recall TPR/	F1-score			Recall TPR/	F1-score
Class 0		0.890	0.908	Class 0		0.903	0.891
Class 1		0.465	0.401	Class 1		0.506	0.448
Accuracy		0.841		Accuracy		0.857	
Macro average		0.678	0.654	Macro average		0.704	0.683
Weighted average		0.841	0.850	Weighted average		0.857	0.840
AUPRC		0.310		AUPRC		0.345	
AUROC		0.806		AUROC		0.824	

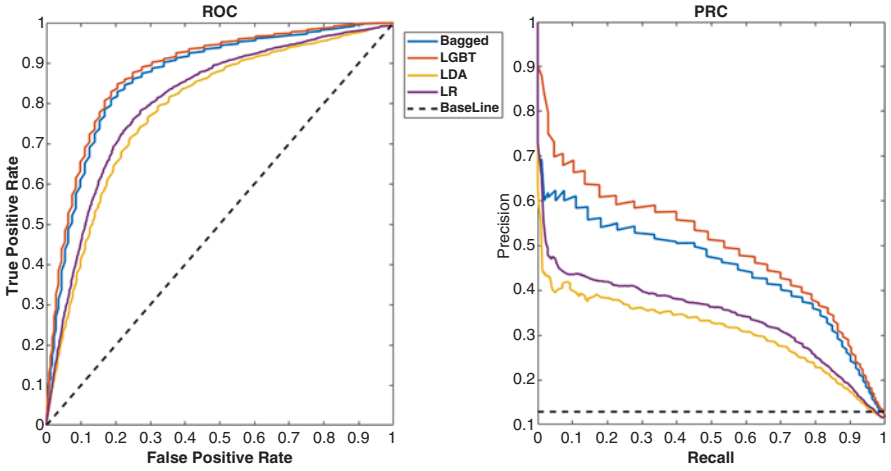

Fig. 11.7 Comparison of PRC and ROC curves

References

1. M.K. Pavlova, V. Latreille, Sleep disorders. Am. J. Med. **132**(3), 292–299 (2019)
2. W. Wen, Sleep quality detection based on EEG signals using transfer support vector machine algorithm, Front. Neurosci. 15, (2021)
3. H. Ragnarsdóttir, B. Marinósson, E. Finnsson, E. Gunnlaugsson, J. S. Ágústsson, H. Helgadóttir, Automatic detection of target regions of respiratory effort-related arousals using recurrent neural networks, in *2018 Computing in Cardiology Conference (CinC)*, (2018), vol. 45: IEEE, pp. 1–4
4. M. Diykh, Y. Li, Complex networks approach for EEG signal sleep stages classification. Expert Syst. Appl. **63**, 241–248 (2016)

5. E. Scoring, EEG arousals: Scoring rules and examples: A preliminary report from the sleep disorders atlas task force of the American sleep disorders association. Sleep **15**(2), 174–184 (1992)
6. R.K. Malhotra, A.Y. Avidan, Sleep stages and scoring technique. Atlas Sleep Med., 77–99 (2013)
7. T. Penzel, S. Canisius, Polysomnography, in *Sleep Apnea*, vol. 35: (Karger Publishers, 2006), pp. 51–60
8. D. Álvarez-Estévez, V. Moret-Bonillo, Identification of electroencephalographic arousals in multichannel sleep recordings. IEEE Trans. Biomed. Eng. **58**(1), 54–63 (2010)
9. S. Mariani, S. M. Purcell, S. Redline, Automated processing of big data in sleep medicine, in *Signal Processing and Machine Learning for Biomedical Big Data*, (CRC Press, 2018), pp. 443–463
10. R. Heinzer et al., Prevalence of sleep-disordered breathing in the general population: The HypnoLaus study. Lancet Respir. Med. **3**(4), 310–318 (2015)
11. S. Cho, J. Lee, H. Park, K. Lee, Detection of arousals in patients with respiratory sleep disorders using a single channel EEG, in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, (2006) IEEE, pp. 2733–2735
12. D. Chylinski et al., Validation of an automatic arousal detection algorithm for whole-night sleep EEG recordings. Clocks & sleep **2**(3), 258–272 (2020)
13. A.H. Khandoker, J. Gubbi, M. Palaniswami, Automated scoring of obstructive sleep apnea and hypopnea events using short-term electrocardiogram recordings. IEEE Trans. Inf. Technol. Biomed. **13**(6), 1057–1067 (2009)
14. R. Lazazzera et al., Detection and classification of sleep apnea and hypopnea using PPG and SpO $ _2 $ signals. IEEE Trans. Biomed. Eng. **68**(5), 1496–1506 (2020)
15. T. Sugi, F. Kawana, M. Nakamura, Automatic EEG arousal detection for sleep apnea syndrome. Biomed. Signal Proc. Control **4**(4), 329–337 (2009)
16. M.M. Ghassemi et al., You snooze, you win: the physionet/computing in cardiology challenge 2018," in *2018 Computing in Cardiology Conference (CinC)*, (2018), vol. 45: IEEE, pp. 1–4
17. V. Tsara, A. Amfilochiou, M. Papagrigorakis, D. Georgopoulos, E. Liolios, Guidelines for diagnosis and treatment of sleep-related breathing disorders in adults and children. Definition and classification of sleep related breathing disorders in adults: Different types and indications for sleep studies (part 1). Hippokratia **13**(3), 187–191 (2009)
18. E. J. Olson, W. R. Moore, T. I. Morgenthaler, P. C. Gay, B. A. Staats, Obstructive sleep apnea-hypopnea syndrome, in *Mayo Clinic Proceedings*, vol. 78, no. 12: Elsevier, 2003, pp. 1545–1552
19. J.F.M. Jiménez, M.R. González, L.J. Findley, Sleepy drivers have a high frequency of traffic accidents related to respiratory effort-related arousals. Archivos de bronconeumologia **39**(4), 153–158 (2003)
20. M. Howe-Patterson, B. Pourbabaee, F. Benard, Automated detection of sleep arousals from polysomnography data using a dense convolutional neural network, in *2018 Computing in Cardiology Conference (CinC)*, (2018), vol. 45: IEEE, pp. 1–4
21. R. He et al., Identification of arousals with deep neural networks (DNNs) using different physiological signals, in *2018 Computing in Cardiology Conference (CinC)* (2018), vol. 45: IEEE, pp. 1–4
22. A. Patane, S. Ghiasi, E. P. Scilingo, M. Kwiatkowska, Automated recognition of sleep arousal using multimodal and personalized deep ensembles of neural networks, in *2018 Computing in Cardiology Conference (CinC)*, (2018), vol. 45: IEEE, pp. 1–4
23. P. Warrick and M. N. Homsi, "Sleep arousal detection from polysomnography using the scattering transform and recurrent neural networks," in *2018 Computing in Cardiology Conference (CinC)*, 2018, vol. 45: IEEE, pp. 1–4
24. H. Li, Q. Cao, Y. Zhong, Y. Pan, Sleep arousal detection using end-to-end deep learning method based on multi-physiological signals, in *2018 Computing in Cardiology Conference (CinC)*, (2018), vol. 45: IEEE, pp. 1–4
25. B. Varga, M. Görög, P. Hajas, Using auxiliary loss to improve sleep arousal detection with neural network, in *2018 Computing in Cardiology Conference (CinC)*, (2018), vol. 45: IEEE, pp. 1–4

26. N. Sadr, P. de Chazal, Automatic scoring of non-apnoea arousals using the polysomnogram, in *2018 Computing in Cardiology Conference (CinC)*, (2018), vol. 45: IEEE, pp. 1–4
27. K.E. Bloch, Polysomnography: A systematic review. Technol. Health Care **5**(4), 285–305 (1997)
28. V.R. Badrakalimuthu, R. Swamiraju, H. de Waal, EEG in psychiatric practice: To do or not to do? Adv. Psychiatr. Treat. **17**(2), 114–121 (2011)
29. C.L. Drake, K.M. Mason, S.M. Bowyer, T. Roth, G.L. Barkley, N. Tepley, Vertex sharp waves during sleep localized by 2DII. Cortex **1**, 8 (2002)
30. M. Schönauer, D. Pöhlchen, Sleep spindles. Curr. Biol. **28**(19), R1129–R1130 (2018)
31. S. Tong and N. V. Thankor, Quantitative EEG Analysis Methods and Clinical Applications.. Artech House, 2009
32. M. Sharma, D. Goyal, P. Achuth, U.R. Acharya, An accurate sleep stages classification system using a new class of optimally time-frequency localized three-band wavelet filter bank. Comput. Biol. Med. **98**, 58–75 (2018)
33. H. Rao et al., Feature selection based on artificial bee colony and gradient boosting decision tree. Appl. Soft Comput. **74**, 634–642 (2019)
34. G. Ke et al., Lightgbm: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Proces. Syst. **30**, 3146–3154 (2017)
35. J. Brownlee, how to Develop a Light Gradient Boosted Machine (LightGBM) Ensemble, ed (2020)
36. M. Thoma, Wikimedia Commons, the free media repository, Roc-draft-xkcd-style.svg, Ed., ed (June 2018)
37. Draelos. Measuring Performance: AUPRC and Average Precision. https://glassboxmedicine.com/2019/03/02/measuring-performance-auprc/. accessed
38. A. A. Gharbali, J. M. Fonseca, S. Najdi, and T. Y. Rezaii, "Automatic eog and emg artifact removal method for sleep stage classification," in *Doctoral Conference on Computing, Electrical and Industrial Systems*, 2016: Springer, pp. 142–150
39. P. Shooshtari, G. Mohammadi, B. Molaee Ardekani, M. B. Shamsollahi, Removing ocular artifacts from EEG signals using adaptive filtering and ARMAX modeling, in *Proceeding of World Academy of Science, Engineering and Technology*, vol. 11, no. CONF, (2006) pp. 277–280
40. X. Jiang, G.-B. Bian, Z. Tian, Removal of artifacts from EEG signals: a review. Sensors **19**(5), 987 (2019)
41. X. Li, S.H. Ling, S. Su, A hybrid feature selection and extraction methods for sleep apnea detection using bio-signals. Sensors **20**(15), 4323 (2020)
42. C. Vidaurre, N. Krämer, B. Blankertz, A. Schlögl, Time domain parameters as a feature for EEG-based brain–computer interfaces. Neural Netw. **22**(9), 1313–1319 (2009)
43. J.V. Liu, H.K. Yaggi, Characterization of Arousals in Polysomnography Using the Statistical Significance of Power Change, in *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (2018): IEEE, pp. 1–6
44. B. Mandelbrot, How long is the coast of Britain? Statistical self-similarity and fractional dimension. Science **156**(3775), 636–638 (1967)
45. E.B. Sadeghian, M.H. Moradi, Fractal dimension for detection of ERD/ERS patterns in asynchronous brain computer interface, in *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*, (2008): IEEE, pp. 560–563
46. C.-T. Shi, Signal pattern recognition based on fractal features and machine learning. Appl. Sci. **8**(8), 1327 (2018)
47. A. Yilmaz, G. Unal, Multiscale Higuchi's fractal dimension method. Nonlinear Dynamics **101**(2), 1441–1455 (2020)
48. A. Adda, H. Benoudnine, Detrended fluctuation analysis of EEG recordings for epileptic seizure detection, in *2016 International Conference on Bio-engineering for Smart Technologies (BioSMART)*, (2016): IEEE, pp. 1–4
49. V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, *Feature Selection for High-Dimensional Data* (Springer, 2015)
50. C.-j. Tian, J. Lv, X.-f. Xu, Evaluation of feature selection methods for mammographic breast cancer diagnosis in a unified framework. BioMed Res. Int. **2021**, 1–9 (2021)