Received September 7, 2020, accepted September 21, 2020, date of publication September 25, 2020, date of current version October 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3026814

Automatic Sleep Arousals Detection From Polysomnography Using Multi-Convolution Neural Network and Random Forest

YITIAN LIU, HONGXING LIU, AND BUFANG YANG
School of Electronic Science and Engineering, Nanjing University, Nanjing 210008, China

Corresponding author: Hongxing Liu (njhxliu@nju.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61271079.

ABSTRACT Sleep arousals is a type of sleep disorder, which refers to the phenomenon of waking up and falling asleep again. Monitoring the number and duration of sleep arousals is a crucial aspect of sleep quality assessment. The detection of sleep arousals caused by apnea is relatively easy, and existing methods have been able to give high quality results. However, sleep arousals caused by non-apnea remains an ongoing challenge, and this is also the subject of PhysioNet Computing in Cardiology Challenge 2018. We proposed a non-apnea sleep arousals automatic detection algorithm based on polysomnography (PSG) data. We took 8 most representative signals selected from the 13 channels of PSG signals as input, conducted preliminary classification through multiple convolutional neural networks, and then sent the initial results to the random forest module for ensemble voting, and obtained the final judgment. We carried out some experiments using the CinC 2018 database. We grouped the original dataset reasonably, and based on each group of data, we trained a corresponding CNN, ensuring the balance of positive and negative samples during the training. Our 4-fold cross validation results for the AUROC and AUPRC were 0.953 and 0.552, which were better than the results of the team which ranked first in the CinC 2018.

INDEX TERMS Sleep arousals, non-apnea, polysomnography, multiple convolutional neural networks, random forest.

I. INTRODUCTION

Sleep disorders refer to the abnormal phenomenon of sleep quality caused by various reasons, mainly including insomnia, sleep arousals, abnormal sleep behavior and so on. The phenomenon of arousal during sleep is the main reason affecting sleep, and it has a negative impact on the sleep/wake cycle [1]. Arousal is a brief intrusion of wakefulness into sleep, after which sleep resumes [2]. Although the correct detection of arousal can help to accurately assess sleep quality [3], and make it possible for patients to get well-directed treatment, it is difficult to be detected.

One of the most studied types of sleep arousal is Obstructive Sleep Apnea Hypopnea Syndrome (or simply, apnea) [4]. While apneas are well-known sleep disturbances, they are not the only cause of disturbance. In addition, some other factors, such as respiratory effort related arousals (RERAs), snoring, partial airway obstructions and teeth grinding may cause sleep arousals [5], collectively called non-apnea arousals. The region of a non-apnea arousal was defined as from 2 seconds before a RERA arousal begins, up to 10 seconds after it ends, or from 2 seconds before a non-RERA non-apnea arousal begins, up to 2 seconds after it ends [5].

The representative indicators to evaluate sleep arousals detection performance are area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC), which are often used in medical target detection, machine learning, data mining and other problems to judge the classification and test results [6].

The polysomnography (PSG) is the gold standard to diagnose sleep diseases [7], which contains physiological signals of 13 channels such as electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiogram (ECG) and oxygen saturation (SaO2). It is obtained by medical staff using a variety of recording devices on patients, and it typically contains 6 to 8 hours of sleep data, which are used to analyze sleep disorders.

Since it is a tedious and time-consuming work to manually annotate various physiological signals, automatic detection of sleep arousals becomes an essential research field. In the

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeswari Sundararajan.

previous researches, Cho S P, Lee J et al proposed an automatic detection method based on time-frequency analysis and support vector machine (SVM) classifier [8]. They used 9 patients single-channel EEG to do experiment. Alvarez Estevez and Moret-Bonillo proposed an automatic detection of arousals algorithm based on Fisher's linear and quadratic discriminants, support vector machines and artificial neural networks [9]. They used 2 EEG channels and EMG data from 20 patients. Kantar Tugce and Erdamar Aykut designed a decision support system algorithm for arousal by analyzing and obtaining features of EEG signals [10]. They used deep neural network to classification. However, all of these methods were developed on datasets containing relatively few patients, and may not generalize well across different populations.

However, the CinC 2018 [11] also interests this subject, and a lot of researches have been carried out. The winners of the CinC 2018 have come up with some models for arousal detection. Daniel Miller, Andrew Ward et al proposed a Convolution-Deconvolution arousal recognition network with cross connections [12], which consists of 8 convolution layers, 8 deconvolution layers and a fully connected layer. Matthew howe-patterson et al constructed a dense circular convolutional neural network (DRCNN) to detect sleep arousal [13]. This network has multiple dense convolution units (DCU) and residual skip connection bidirectional long short-term memory (BiLSTM) layer. Hei∂ar Már Þráinsson, Hanna Ragnarsdóttirand used recursive neural network (RNN) to classify between arousal regions and non-arousal regions [14]. Their network structure is divided into two layers, the first layer is a bidirectional recursive neural network (BRNN), the second layer is a dense neural network. From the achievements of CinC 2018, the large variance in performances across entrants (AUPRCs mean 0.28, and range from 0.07 to 0.55) indicates that arousal detection still requires further research [11].

In the above CinC 2018 researches, most of them use a single classifier to detect the arousal regions. Using a single classifier cannot consider the differences between different populations. Therefore, it cannot maximally learn the crucial features of physiological data for each group of people. In addition, most of them did not reorganize and recombine the experiment dataset, nor did they make in-depth analysis of which PSG channels are more meaningful for arousals detection, which could not guarantee that the most valuable information will be extracted from the subsequent networks.

In order to solve the above deficiencies, inspired by the thought of ensemble learning [15], in this paper we applied two classification schemes: multi-convolution neural network and random forest. We used many CNNs to extract signals features and make preliminary classification, and utilized random forest to determine the weight of these initial classifiers and give the final result.

The remainder of this paper is organized as follows. We show the database and related theories and concepts in section 2. Our methodology is presented in section 3. We perform extensive experiments and make a detailed explanation in section 4. We exhibit our results and evaluations in section 5. Then we analyze and discuss our methods in section 6. Finally, conclusions and future directions are outlined in section 7.

II. MATERIALS

A. DATABASE

We used the CinC 2018 training set to experiment, which consists of 994 PSG records of different subjects. The data were monitored at an MGH sleep laboratory for the diagnosis of sleep disorders. Each record contains approximately 7-8 hours of monitoring data, and it contains 13 channels physiological signals including: 6 channels of EEG, 1 channel of EOG, 3 channels of EMG, 1 channel of Airflow, 1 channel of SaO2, and 1 channel of ECG. All signals were sampled to 200 Hz and were measured in microvolts.

The dataset also counts the number of arousals and types, as shown in Table 1:

TABLE 1. The number of arousals and types in dataset.

The Arousals type	Number
Number of target arousals	
Bruxism	30
Cheyne-stokes breathing	3
Hypoventilation	4
Noise	1
Partial airway obstruction	11
Periodic leg movement	36
Respiratory effort-related arousals	43822
Snoring	28
Spontaneous	70
Number of non-target arousals	
Hypopnea	56936
Central apnea	22763
Mixed apnea	2641
Obstructive apnea	32547

In the training set, the target arousals labels had been given. By analyzing the whole dataset, we can know that the arousal regions marked with label '+1' account for 4 percent, the non-arousal regions labeled '0' for 80 percent, and the regions undefined marked with label '−1' account for 16 percent of the data.

The target arousals regions show that their minimum duration is about 30s, and the maximum duration is above 4 mins. Moreover, most of these regions of non-apnea arousals are located in stage 1, stage 2, and a few in stage 3 of the six sleep stages.

B. CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Network (CNN) is a kind of special deep Neural Network model. As long as a training schema is given, it can simulate the mathematical function relationship between input and output. In recent years, it has been widely used in classification and identification [16].

Its structure can be divided into three parts. The first part is the Convolutional Layer, which is generally placed at the beginning of the network as input to the system. It is often used for feature information extraction. In this layer, neurons are only connected to some adjacent neurons, and these neurons are connected together to form a feature plane. Neurons in this plane have the same weight, and the weight they share is called the convolution kernel. The advantage of this shared convolution kernel is that it can reduce the probability of network overfitting and reduce the complexity of the network by means of partial connection. The second part is the Pooling Layer. It exists between two successive convolutional layers. It compresses the data to reduce the risk of overfitting and it has characteristic invariance. In addition, it can reduce the dimension of the feature, by removing the redundant information. And the third part is the Full Connection Layer. Its structure is the same as the general neural network structure, each neuron is connected with each other. It connects the characteristic information of the parameter, and then send the output value to the next level structure.

C. RANDOM FOREST

As a newly emerging highly flexible machine learning algorithm, Random Forest (RF) has a wide application prospect. It is an algorithm that integrates multiple trees through the idea of ensemble learning. Its basic unit is decision tree. Each decision tree is a classifier, and it integrates the voting results of all categories, specifying the category with the most votes as the final output [17].

It can run efficiently on large datasets. It can also handle thousands of input variables without variable deletion. When a large proportion of the data are missing, it can also maintain accuracy.

III. METHODS

A. THE FRAME OF AROUSAL DETECTION MODEL

We built the overall architecture of the automatic detection system for non-apnea arousals as shown in Figure 1. We first preprocessed the raw data. The input of preprocessing module is the raw multi-channel PSG signals, and the output is many processed data segments whose lengths are all of 4s. Then, we respectively put these segmented data into the automatic recognizer module, which will give each data segment a classification result. After that, we conducted post-processing on these classification results, and finally output the corresponding arousal detection results of the whole PSG signals.

B. PRE-PROCESSING

We selected 3 channels of EEG (F4 - M1, C4 - M1, O2 - M1), 3 channels of EMG (Chin1 - Chin2, ABD), CHEST, Airflow, and SaO2, from 13 channels physiological signals, using these 8 channels to detect non-apnea arousals.

We used ICA and double density DWT Algorithm [18] to remove the artifacts of 3 channels of EEG, and then applied a 50-order FIR filter (designed with Matlab FDA tool toolbox) to filter 8 selected channels signals.

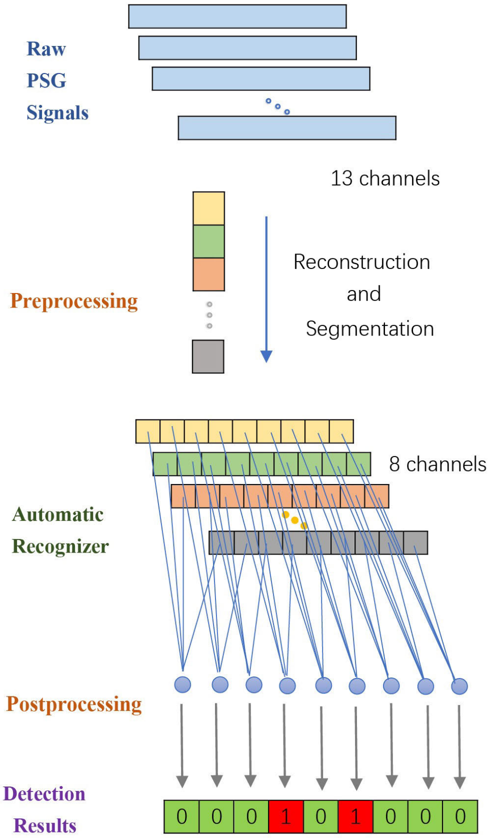

FIGURE 1. Frame diagram of the proposed method. It has 3 major parts: preprocessing, automatic recognizer, postprocessing. The system input is raw 13 channels PSG signals, output is the detection results containing a series of 0 and 1.

After that, we downsampled these signals to 50Hz, reducing the data size by 4 times. In order to eliminate the differences among these signals, we calculated and normalized the mean value and standard deviation of each signal. Finally, we divided the data into small segments of length 4s with non-overlap for the sake of facilitating subsequent network processing. The entire data preprocessing module is shown in Figure 2.

C. THE AUTOMATIC RECOGNIZER

This module is the core of the algorithm. We proposed a method that combines multiple CNN models with random forest. We input each data segment into n (the CNNs number) CNN models firstly, and each CNN model gives a classification result of this segment. Then we sent the multiple decision vector sets generated by CNNs to the random forest. We used the random forest to determine the weight of each CNN. The random forest gave the final verdict for each data segment. The structure of automatic recognizer is illustrated in Figure 3.

1) CNN NETWORK STRUCTURE

We used an original CNN model to extract features from small data segments. The proposed network structure is

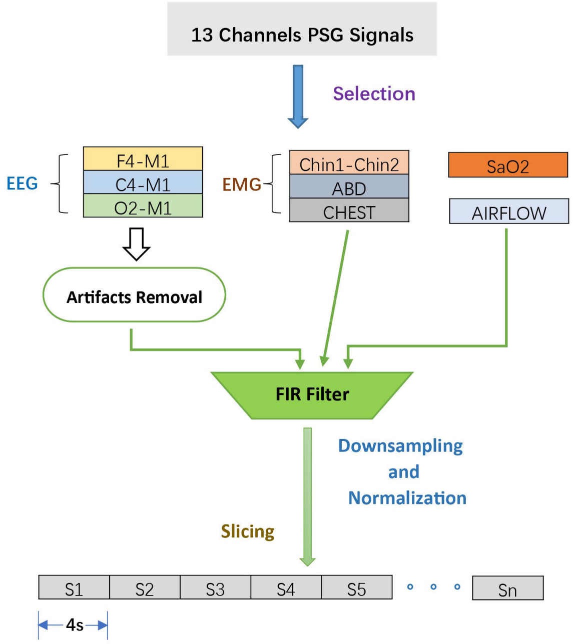

FIGURE 2. Flowchart diagram of pre-processing. S1, S2, S3, S4…Sn represent the divided data segments, and their length are all 4s.

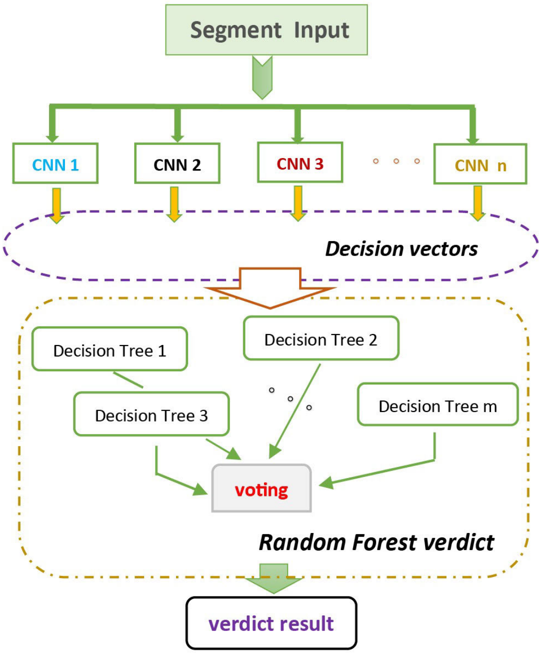

FIGURE 3. Structural diagram of the automatic recognizer. It has multiple CNN models and random forest part. n is the number of CNNs, and m is the number of decision trees in the random forest.

shown in Figure 4. The activation function of our network is RELU, and we used the MSRA (Micromechanical Silicon Resonant Accelerometer) method to initialize network weight, which has a better effect on the network initialization whose activation function is nonlinear [19]. Since the number of layers increases, its nonlinear fitting ability becomes better, we made a compromise between learning speed and accuracy. Our network consists of feature extraction part which is composed of 4 one-dimensional convolution layers and 2 pooling layers, and feature classification part which is composed of 2 fully connected layers and 1 softmax layer.

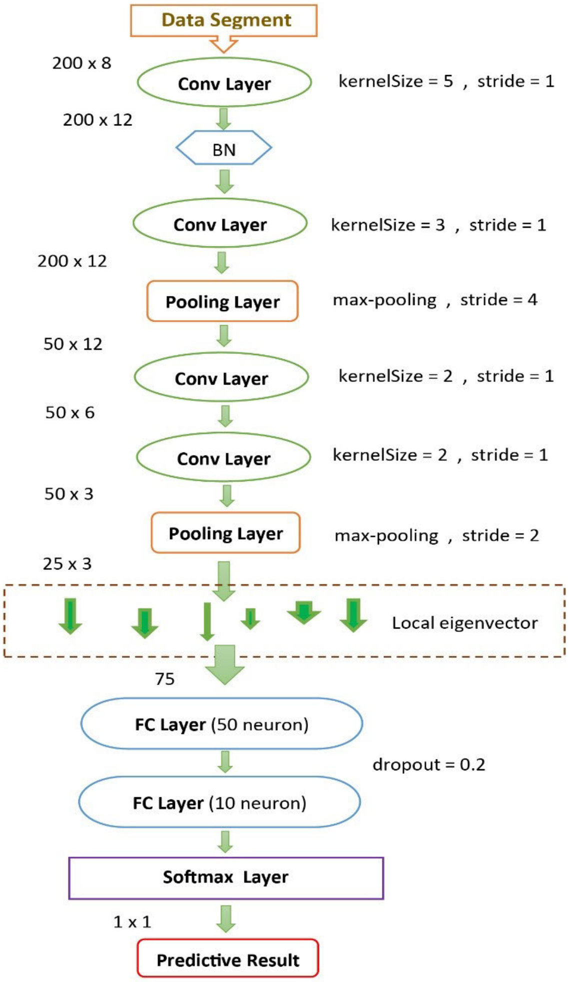

FIGURE 4. The structure of the proposed CNN. It has 4 1-D convolution layers and 2 pooling layers for feature extraction, 2 fully connected layers and 1 softmax layer for classification.

The input signal is the 8 channels data selected from the 13 channels of PSG signals. The data segment length is 4s and the sampling rate is 50Hz. Therefore, each channel has 200 data points and the input data size is 200*8.

Firstly, the input signals enter a one-dimensional convolution layer, where the number of convolution kernels is 12 and the size of the convolution kernel is 5. In order to ensure that the size of each channel data after passing through the convolution layer remains constant, we used zero-padding on the front and back ends of the data during convolution calculation. Then the batch normalization and ReLU activation function were performed. The output size after through the convolution layer is 200*12. Then we used a module with 3 convolutional layers and 2 pooling layers to extract the higher-order features of the signal. The number of convolution kernels in the convolution layer decreases by 2 times, and the size of the convolution kernel in the first convolution layer is 3, and the second is 2. The pooling method adopted in the pooling layer is maximum pooling. The first pooling layer stride is 4, and the second is 2. After passing this module, the size of the output data is 25*3. Then we stretched the data into vectors, mapped the feature space to a fully connected layer which is easier to classify, and regarded this layer as the final feature classification module. The first fully connected layer contains 50 units, and the second contains 5 units. The activation function adopted by the two full connection layers is ReLu, and the dropping rate is 0.2 to improve the generalization ability. Finally, we used the softmax layer to determine whether the area is an arousal region, and it gave a decision result.

2) RANDOM FOREST MODULE

Since each CNN model output a result, we need to use these results to make the final decision. Inspired by the idea of combining several weak classifiers into one strong classifier in ensemble learning [15], and considering that there is no dependency between each CNN model, we used the random forest which is an improved algorithm of Bagging algorithm [15] to realize this process. We utilized it so that the CNNs with better effect has a higher status, while diminish the impact of other CNNs judgment on the final result.

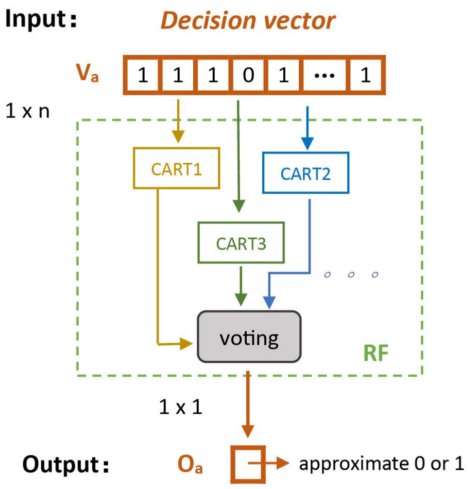

FIGURE 5. The schematic diagram of random forest work. n is the number of CNNs, Va is the decision vector of data segment a, and Oa is the classification result of data segment a.

The random forest we used is classification and regression trees (CART), which is commonly used in data mining. It selects the attribute with the smallest Gini coefficient as the optimal attribute partition method [20]. A schematic diagram of this process is shown in Figure 5. The input of the random forest is the decision vectors generated by the multiple CNN models, that is to say, we regarded as each CNN output as a character of a random forest, and the number of CNN structures is n, so the input data size is 1*n. Then each decision tree randomly selects these decision vectors for classification, and finally carries out a simple vote on the classification results of each decision tree to determine the final results. For each data segment, it gives a 1*1 result.

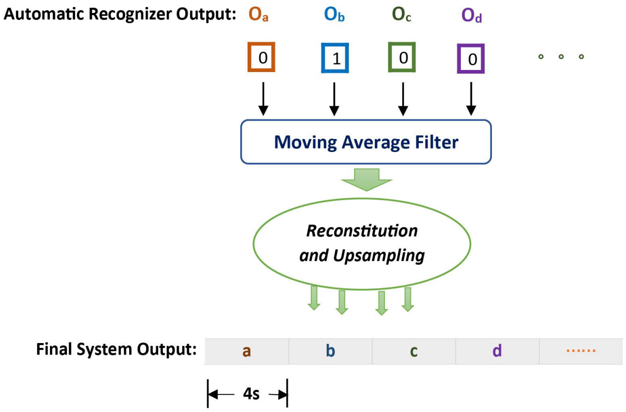

FIGURE 6. Flowchart diagram of post-processing. Oa, Ob, Oc, and Od are the automatic recognizer output for the different segments a, b, c, d, and their values are approximate to 0 or 1.

D. POST-PROCESSING

For each segment, the automatic recognizer gives a decision value. Since the time dependence of several successive data segments, we adopted some post-processing methods to improve the performance of the system, as shown in Figure 6. Firstly, we smoothed the output of the automatic recognizer. We used the Moving Average Filter [21] to process the judgment value of the consecutive data segments and correct the judgment error that may occur in a short time. Next, we regarded each segment output value as the label for each data point in the entire segment, and then we upsampled the data to 200Hz, in order to obtain the arousal detection results of the whole signals.

IV. EXPERIMENTS

A. INTRODUCTION OF THE EVALUATING INDICATORS FOR EXPERIMENT

We used the AUROC and AUPRC to evaluate the performance of the designed arousal detection system.

The ROC curve is based on a series of different ways of binary classification (boundary value or decision threshold), which is a curve drawn with sensitivity as vertical axis and specificity as horizontal axis. It can explore the relationship between the specificity and sensitivity of the algorithm and weigh the influence of fail to judge and miscarriage of justice. The area under this curve is AUROC. The larger the value (closer to 1), the higher the accuracy of the system will be. The system has a low accuracy when the AUROC is 0.5∼0.7, and has a certain degree of accuracy when it is 0.7∼0.9. When the AUROC is above 0.9, the system has a high accuracy [22].

The PRC curve is more able to reflect the real performance of classification when the ratio of positive and negative samples is large [23]. It reflects the relationship between precision and recall. Regarding them as vertical axis and horizontal axis, we can also draw a curve and get the area under it.

The larger its value (close to 1), the more comprehensive and accurate the recognition will be.

B. DATA PARTITION

We divided the records of 994 subjects into three parts, 696 records were used for training, 99 records were used for verification, and other 199 records were used for test. The entire dataset contains tr-03, tr-04, up to tr-14, a total of 12 series. We divided the three datasets according to the series, which is more universal and makes the final experimental results more convincing. Our data partition is shown in Table 2:

TABLE 2. The data partition for experiment.

Dataset	Training set	Validation set	Test set	total
tr03	97	14	28	139
tr04	78	11	23	112
tr05	116	17	33	116
tr06	67	10	19	96
tr07	77	11	22	110
tr08	24	3	7	34
tr09	34	5	10	49
tr10	38	5	11	54
tr11	43	6	12	61
tr12	62	9	17	88
tr13	45	6	13	64
tr14	15	2	4	21
total	696	99	199	994

C. BUILD TRAINING DATA

Since the ratio of the arousal regions to the non-arousal regions in the dataset is about 1:20, the detection of arousal regions can be regarded as a classification problem of unbalanced categories. Some Studies have shown that if 90% of the samples of the training set belong to the same category, the classifier will divide all samples into this class [24]. In this instance, the classifier is invalid, although the final classification accuracy is very high. Therefore, balancing dataset is crucial. In addition, when the data is unbalanced, using accuracy to evaluate has little reference significance. Under these circumstances, it is more likely to utilize precision and recall for evaluation [25].

Our method of equalizing the dataset is to expand the arousal regions. We used a Synthetic Minority over-sampling Technique (SMOTE) algorithm [26] to oversampling, and used it to construct a new part of the data in arousal regions rather than the existing data. We eliminated those undefined data points, since they were not useful for the system to detect the arousal regions. We kept all the data points in non-arousal regions. Finally, we recombined these data into a new dataset and used it to train our network.

D. MODEL TRAINING

Our programming language was python 3.7, and we used the computer with a CORE i7 8700K CPU, a NVIDIA GTX 1080Ti GPU, and 32GB RAM to train this model.

We divided the records of 12 series in the reconstructed dataset into 12 groups. For each group of data, the total data length is about 5400, and we used a CNN model to train. There is no dependency between the training and calculations of each CNN model, so it can be done in parallel, which can save the model total training time. The batch size we set for training is 50. The loss function is the cross-entropy cost function, and Adam optimization method [27] was used to optimize our model. The learning rate is a parameter which affects the network convergence rate, and the optimal learning rate depends on the model structure and the experiment dataset. After several experimental adjustments, we set it to 0.0005. The maximum number of training epochs was set to 1000. If the value of AUPRC on the validation set growth rate does not exceed 0.0001 for 20 consecutive epochs, the training will be stopped. And the number of decision trees is 10 for training random forest. We randomly selected 8 properties as candidates from these 12 properties at a time, and we set the maximum number of iterations to 30.

V. RESULTS

We used the scoring program given by the CinC 2018 website [5] to test the performance of the system we designed. The AUROC on the whole test set was 0.953980±0.033123, and the AUPRC was 0.557513±0.113802. In addition, we calculated several other commonly used performance indicators (sensitivity of 88.92%, specificity of 96.56% and precision 87.6%), and the maximum and minimum scores of AUROC and AUPRC for each series, and the corresponding average score. The details are shown in Table 3 and Table 4.

TABLE 3. The AUROC performance.

Series	Maximum Score	Record Number	Minimum Score	Record Number	Average Score
tr03	0.9811	tr03-0179	0.9151	tr03-0241	0.9472
tr04	0.9783	tr04-0695	0.9326	tr04-0631	0.9613
tr05	0.9847	tr05-1190	0.9553	tr05-1097	0.9718
tr06	0.9609	tr06-0694	0.9049	tr06-0567	0.9487
tr07	0.9771	tr07-0230	0.9248	tr07-0509	0.9589
tr08	0.9774	tr08-0105	0.9520	tr08-0347	0.9678
tr09	0.9808	tr09-0331	0.9306	tr09-0478	0.9697
tr10	0.9539	tr10-0626	0.9247	tr10-0771	0.9324
tr11	0.9546	tr11-0459	0.9202	tr11-0354	0.9457
tr12	0.9613	tr12-0003	0.9478	tr12-0122	0.9538
tr13	0.9809	tr13-0475	0.9484	tr13-0080	0.9673
tr14	0.9690	tr14-0193	0.9221	tr14-0185	0.9499

It can be seen from the above table that the method we proposed can obtain a remarkable detection result for most patients. The results indicate that it will be a powerful tool for automatic arousal detection in clinic. Moreover, in order to further proving the generalization of our model, we performed a 4-fold cross-validation experiment. We reclassified the training set, verification set and test set so that each model can be evaluated on its own test set. The results are shown in Table 5.

Finally, we compared the models and results used by the CinC 2018 winners with ours, as shown in Table 6, which is sufficient to testify the superiority of our method.

TABLE 4. The AUPRC performance.

Series	Maximum Score	Record Number	Minimum Score	Record Number	Average Score
tr03	0.6953	tr03-0061	0.390131	tr03-0134	0.5334
tr04	0.8329	tr04-0695	0.423723	tr04-0404	0.5978
tr05	0.9274	tr05-1042	0.521648	tr05-1464	0.6758
tr06	0.8013	tr06-0317	0.357515	tr06-0709	0.6021
tr07	0.6855	tr07-0231	0.473156	tr07-0874	0.5418
tr08	0.6972	tr08-0105	0.555905	tr08-0113	0.6245
tr09	0.6286	tr09-0331	0.477302	tr09-0541	0.5623
tr10	0.7245	tr10-0626	0.292827	tr10-0768	0.5618
tr11	0.8519	tr11-0509	0.447249	tr11-0792	0.5743
tr12	0.4861	tr12-0515	0.322117	tr12-0122	0.4506
tr13	0.8428	tr13-0801	0.441147	tr13-0475	0.5689
tr14	0.5402	tr14-0193	0.518278	tr14-0185	0.5292

TABLE 5. The AUROC and AUPRC for cross-validation results.

	Model 1	Model 2	Model 3	Model 4	Average
AUROC	0.951	0.962	0.947	0.954	0.953
AUPRC	0.549	0.558	0.553	0.547	0.552

TABLE 6. The results of comparing to other models.

Models	AUPRC Score	AUROC Score
CNNs-RF（this work）	**0.552**	**0.953**
DRCNN	0.543	0.931
BRNN-LSTM	0.452	0.901
CNN-RNN	0.430	0.927

TABLE 7. The results of different combinations of physiological signals.

Signals	Group Number							
	1	2	3	4	5	6	7	8
F3-M2	✓	×	✓	×	×	×	×	×
F4-M1	✓	✓	✓	✓	✓	✓	✓	✓
C3-M2	✓	×	✓	×	×	×	×	×
C4-M1	✓	✓	✓	✓	✓	✓	✓	✓
O1-M2	✓	×	✓	×	×	×	×	×
O2-M1	✓	✓	✓	✓	✓	✓	✓	✓
Chin1-Chin2	✓	✓	×	✓	✓	✓	✓	✓
ABD	✓	✓	×	✓	✓	✓	✓	×
CHEST	✓	✓	×	✓	✓	✓	×	×
AIRFLOW	✓	✓	×	×	✓	✓	✓	✓
SaO2	✓	✓	×	×	×	✓	✓	✓
AUPRC Score	0.518	0.519	0.374	0.531	0.546	**0.552**	0.451	0.417

VI. DISCUSSION

In the pre-processing part, we selected 8 physiological signals from 13 PSG channels and sent them to the automatic recognizer. This is the result of the experiments on different combinations of physiological signals we conducted, as shown in Table 7. We carried out one experiment for each combination. Among them, the sixth combination: 3 channels of EEG, 3 channels of EMG, Airflow, and SaO2 achieved the best results, and the AUPRC is 0.552.

These 8 channels data should be segmented before being sent to the automatic recognizer. If the data segment length is too small, the network could not extract some crucial features, and if the data segment length is too large, the accuracy of classification will be affected. We carried out several contrast experiments to find the optimal data segment length. The results are shown in Table 8. When the data segment length is 4s, the system gets the best effect.

TABLE 8. The results of using different data segment lengths.

Data segment length	AUPRC Score	AUROC Score
1s	0.374	0.886
2s	0.411	0.891
3s	0.505	0.917
4s	**0.552**	**0.953**
5s	0.549	0.942
6s	0.548	0.936

On account of the CNN can automatically extract the signal features and avoid the trouble of manual extraction, we utilized it in our model. And we used multiple CNNs to group training. The number of CNNs is consistent with the group of training data available.

Because of the individual differences of each subject, when they occur arousal, the different kind of physiological signals change is also distinct. In the CinC 2018 dataset, there are already 12 groups which are divided by population. In order to maximize the crucial characteristics of physiological data for each type of patient, we trained a CNN model with a data series respectively, which reduced the difficulty of CNN training and avoided the model underfitting due to the different group differences. We compared this group training method with the method of multiple CNN integration (training each CNN model with all the data). The results are shown in Table 9, and it demonstrates the superiority of group training.

TABLE 9. The comparison of different training methods.

	AUPRC Score	AUROC Score
Multiple CNNs with all training data	0.516	0.942
Multiple CNNs with training data for the corresponding series	**0.552**	**0.953**

In the post-processing part, we used the moving average filter to smooth the output of the automatic recognizer to correct the decision errors in a short time, and the AUPRC is 0.03 higher than that of the AUPRC without processing.

VII. CONCLUSION

In this study, we have proposed a method for identification of arousals by using multiple convolutional neural networks and random forest. The AUPRC score under 4-fold cross validation we get was 0.552, it is better than the best score of 0.54 in the CinC 2018, which exhibits that our method has achieved an outstanding effect. However, our model is not accurate enough to detect sleep arousals for few numbers of patients, and the detection results of different individuals have some dissimilarity. Therefore, we will intend to further generalize our method for a different dataset, and improve our model universality with feature selection and code optimization. It will become a powerful tool for clinical automatic sleep arousal detection.

REFERENCES

[1] K. Ryan, *Encyclopedia of Clinical Neuropsychology. Diss*. New York, NY, USA: Springer, 2013.

[2] R. B. Berry, "The AASM manual for the scoring of sleep and associated events: Rules terminology and technical specifications," *Amer. Acad. Sleep Med.*, to be published.

[3] G. L. Sorensen, P. Jennum, J. Kempfner, M. Zoetmulder, and H. B. D. Sorensen, "A computerized algorithm for arousal detection in healthy adults and patients with parkinson disease," *J. Clin. Neurophysiol.*, vol. 29, no. 1, pp. 58–64, Feb. 2012.

[4] G. Parati, C. Lombardi, and K. Narkiewicz, "Sleep apnea: Epidemiology, pathophysiology, and relation to cardiovascular risk," *Amer. J. Physiol.-Regulatory, Integrative Comparative Physiol.*, vol. 293, no. 4, pp. R1671–R1683, 2007.

[5] A. L. Goldberger, L. A. N. Amaral, L. Glass, and J. M. Hausdorff, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2003.

[6] B. Sahiner, W. Chen, and A. Pezeshk, "Comparison of two classifiers when the data sets are imbalanced: The power of the area under the precision-recall curve as the figure of merit versus the area under the ROC curve," *Proc. SPIE*, vol. 10136, Mar. 2017, Art. no. 101360G.

[7] A. L. Chesson, Jr., R. A. Ferber, J. M. Fry, and M. Grigg-Damberger, "The indications for polysomnography and related procedures," *Sleep*, vol. 20, no. 6, p. 423, 1997.

[8] S. P. Cho, J. Lee, H. D. Park, and K. J. Lee, "Detection of arousals in patients with respiratory sleep disorders using a single channel EEG," in *Proc. IEEE Eng. Med. Biol. 27th Annu. Conf.*, Jun. 2005, pp. 2733–2735.

[9] D. Álvarez-Estévez and V. Moret-Bonillo, "Identification of electroencephalographic arousals in multichannel sleep recordings," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 54–63, Jan. 2011.

[10] T. Kantar and A. Erdamar, "Continuous wavelet transform based method for detection of arousal," in *Proc. 25th Signal Process. Commun. Appl. Conf. (SIU)*, May 2017, pp. 1–4.

[11] M. M. Ghassemi, B. E. Moody, H. L. Li-wei, C. Song, Q. Li, H. Sun, R. G. Mark, M. B. Westover, and G. D. Clifford, "You snooze, you win: The physionet/computing in cardiology challenge 2018," *Hypertension*, vol. 40, no. 41, pp. 6–40, 2018.

[12] D. Miller, A. Ward, and N. Bambos, "Automatic sleep arousal identification from physiological waveforms using deep learning," in *Proc. Comput. Cardiol. Conf. (CinC)*, Sep. 2018, pp. 1–4.

[13] M. Howe-Patterson, B. Pourbabaee, and F. Benard, "Automated detection of sleep arousals from polysomnography data using a dense convolutional neural network," in *Proc. Comput. Cardiol. Conf. (CinC)*, Sep. 2018, pp. 1–4.

[14] H. M. Práinsson, H. Ragnarsdóttir, G. F. Kristjánsson, B. Marinósson, E. Finnsson, E. Gunnlaugsson, S. Æ. Jónsson, J. S. Ágústsson, and H. Helgadóttir, "Automatic detection of target regions of respiratory effort-related arousals using recurrent neural networks," in *Proc. Comput. Cardiol. Conf. (CinC)*, Sep. 2018, pp. 1–4.

[15] Hastie, Trevor, R. Tibshirani, and J. Friedman, "'Ensemble learning," in *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2009.

[16] N. Chumerin, "Convolutional neural network," Tech. Rep., 2017.

[17] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: A classification and regression tool for compound classification and QSAR modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947–1958, Nov. 2003.

[18] V. Roy and S. Shukla, "A NLMS based approach for artifacts removal in multichannel EEG signals with ICA and double density wavelet transform," in *Proc. 5th Int. Conf. Commun. Syst. Netw. Technol.*, Apr. 2015, pp. 461–466.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[20] D. Dension, "Bayesian CART algorithm," *Biometrika*, vol. 85, no. 2, pp. 363–377, 1999.

[21] S. Golestan, M. Ramezani, J. M. Guerrero, F. D. Freijedo, and M. Monfared, "Moving average filter based phase-locked loops: Performance analysis and design guidelines," *IEEE Trans. Power Electron.*, vol. 29, no. 6, pp. 2750–2763, Jun. 2014.

[22] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[23] [Online]. Available: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

[24] J. Q. Li, "Statistical learning with imbalanced training set in a machine vision application: Improve the false alarm rate and sensitivity simultaneously," *Proc. SPIE*, vol. 6070, Feb. 2006, Art. no. 607002.

[25] K. M. Ting, "Precision and recall," in *Encyclopedia of Machine Learning*. New York, NY, USA: Springer, 2011.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2011.

[27] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer*, to be published.

YITIAN LIU received the B.S. degree in communication engineering from Northeastern University, China, in 2019. He is currently pursuing the M.S. degree in electronic information engineering with Nanjing University, China.

HONGXING LIU received the Ph.D. degree from Xi'an Jiao Tong University, China, in 1997. He is currently a Professor of information electronics with the School of Electronic Science and Engineering, Nanjing University, China, and a Master's Tutor for signal and information processing. He is also a Member with the Biomedical Electronics Branch of the Chinese Institute of Electronics. In recent years, he has mainly carried out research in the field of intelligent information processing and biomedical electronics. He has focused on fetal electrocardiogram technology and achieved remarkable results. He has presided over three projects of the National Natural Science Foundation, one Social Development Project of Jiangsu Science and Technology Support Program, more than ten horizontal cooperation projects, and published nearly 100 articles by the first author or correspondent author, including more than 20 articles by SCI. The first inventor applied for 33 invention patents, of which 17 were authorized.

BUFANG YANG, photograph and biography not available at the time of publication.