IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 17, 2019, accepted August 4, 2019, date of publication August 8, 2019, date of current version August 21, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2933814

A Two-Stage Neural Network for Sleep Stage Classification Based on Feature Learning, Sequence Learning, and Data Augmentation

CHENGLU SUN[1], JIAHAO FAN[1,2], CHEN CHEN[1], WEI LI[1], AND WEI CHEN[1,2], (Senior Member, IEEE)

[1]Center for Intelligent Medical Electronics, Department of Electronic Engineering, School of Information Science and Technology, Fudan University, Shanghai 200433, China
[2]Human Phenome Institute, Fudan University, Shanghai 200000, China

Corresponding author: Wei Chen (w_chen@fudan.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0112000, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2017SHZDZX01, and in part by the China Postdoctoral Science Foundation under Grant 2018T110346 and Grant 2018M632019.

ABSTRACT Sleep stage classification is a fundamental but cumbersome task in sleep analysis. To score the sleep stage automatically, this study presents a stage classification method based on a two-stage neural network. The feature learning stage as the first stage can fuse network trained features with traditional hand-crafted features. A recurrent neural network (RNN) in the second stage is fully utilized for learning temporal information between sleep epochs and obtaining classification results. To solve serious sample imbalance problem, a novel pre-training process combined with data augmentation was introduced. The proposed method was evaluated by two public databases, the Sleep-EDF and Sleep Apnea (SA). The proposed method can achieve the F1-score and Kappa coefficient of 0.806 and 0.80 for healthy subjects, respectively, and achieve 0.790 and 0.74 for the subjects with suspect sleep disorders, respectively. The results show that the method can achieve better performance compared to the state-of-the-art methods for the same databases. Model analysis displayed that the combination of the hand-crafted features and network trained features can improve the classification performance via the comparison experiments. In addition, the RNN is a good choice for learning temporal information in sleep epochs. Besides, the pre-training process with data augmentation is verified that can reduce the impact of sample imbalance. The proposed model has potential to exploit sleep information comprehensively.

INDEX TERMS Sleep stage classification, feature learning, sequence learning, EEG signal.

I. INTRODUCTION

Sleep is considered as an important state which exerts significant effects on human health. Analyzing people's sleep architectures and evaluating their sleep qualities are important. The sleep architecture and sleep qualities can be expressed by sleep stages. Sleep stages usually include nighttime wakefulness (Wake), rapid-eye movement (REM) stage and non-REM (NREM) stage. The NREM stage can be further divided into N1, N2 and N3 stage according to the American Academy of Sleep Medicine (AASM) rule that is a novel standard [1]. The Rechtschaffen and Kales (R&K) rule as an old standard divides the N3 stage into S3 and S4 [2].

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

Each sleep stage has its own characteristics, which can be reflected by electroencephalogram (EEG), electrooculogram (EOG) and electromyography (EMG). In sleep medicine, for simultaneously recording these physiological signals, a standard method called polysomnography (PSG) is utilized. The whole night sleep recording is divided into 30-second epochs, and each sleep epoch can be scored by sleep experts as different sleep stages through visually inspecting those multiple signals. PSG system often records more than 6 hours' data for the full night sleep, in order to exactly determine the sleep stage of an epoch, the sleep experts generally require checking at least EEG channels, EOG channels and chin EMG channels. One full night sleep recording commonly needs more than two hours to be reliably assessed by proficient experts. The sleep staging is the fundamental context in sleep medicine, so it's an unavoidable and cumbersome work to examine each epoch. Besides, those recorded signals are commonly collected in a specialized sleep laboratory, and it is inevitable to interfere with the natural sleep quality due to the complexity and inconvenience of deploying the PSG on the subject's body in the sleep laboratory. Indeed, the multi-channel sensors attached to the subject's body may cause mental stress and discomfort, which is not conducive to sleep monitoring.

To alleviate the subjects' pressure and save medical resources, a lot of studies attempt to develop machine learning methods to score sleep stage automatically based on several EEG channels [3]–[6]. Those machine learning methods are mainly classified into two categories: traditional machine learning methods and deep learning methods. The traditional machine learning methods used different classifiers to stage the sleep based on many designed hand-crafted features [7]–[9]. Those hand-crafted features commonly have obvious physical significance, like component of frequency spectrum. Those features are proved to have acceptable performance on classification, but they are hard to be understood by clinicians and not closer to the way of sleep staging. The sleep experts usually score the sleep stage by observing the morphology of signals. Besides, the feature extraction and selection process are complex and it is very difficult to extract new and effective features that exceed the existing features. If the characteristic of one specific stage is hard to be described by the provided hand-crafted features, this stage may be not easily identified based on the traditional methods. More importantly, classifiers in traditional methods are commonly not good at dealing with time series signals.

Recently, deep learning methods have attracted much attention in machine learning field. Deep learning methods have provided new ideas and achieved great success in many research directions [10]–[12]. For example, the convolutional neural network (CNN) and deep belief network (DBN) exhibit powerful effects on feature extraction [10], [13], the recurrent neural network (RNN) has been proved to have good capacity in time series signal processing [14], [15]. Thus, many researchers gradually turned to use those deep networks to process the physiological signals and achieve good performances [16], [17].

The sleep stage classification can be simply divided into two parts: feature extraction and sequential signal classification. Thus, the sleep data as the time series signals can be processed by various deep neural networks. For example, Tsinalis et al. [18] employed a two-layer CNN architecture to realize the sleep stage classification. Supratak et al. [19] proposed a deep learning model which utilized the CNN combined with bidirectional RNN for automatic sleep staging based EEG and EOG signals. Those studies have achieved promising performances and represented that deep learning method is competent for sleep stage classification. However, those studies utilized only the network to independently exploit the sleep information, rather than use the existing effective hand-crafted features. In addition, those studies cannot explain the effectiveness of the network. Hence, many studies attempt to use the deep learning methods to process the hand-crafted features for improving the model interpretability. Dong et al. [20] utilized a rectifier neural network to detect hierarchical features from hand-crafted features, and then use Long Short-Term Memory network (LSTM) to learn sequential information to improve the classification performance with EEG and EOG signals. Tsinalis et al. [21] used stacked sparse autoencoders to process the hand-crafted features for sleep staging. Similarly, Längkvist et al. [22] proposed a sleep stage classification method which used the DBN to process the hand-crafted features and hidden markov model (HMM) to capture sleep stage switching rules. Those deep learning structures only process the hand-crafted features without extracting the network trained features. A preferable method should fully utilize prior knowledge of sleep and network. Namely, the feature extraction process should both consider the hand-crafted features and network trained features.

Besides, most of those deep learning methods are tested on the healthy subjects, while the sleep architecture in the subjects with sleep disorders is more disordered than that of healthy people. Consequently, the classification work for people with sleep disorders is relatively difficult. Therefore, we should pay more attention to the improvement of the sleep staging algorithm for people with sleep disorders.

Hence, we introduce a novel two-stage network model for sleep stage classification based on single-channel EEG. The first stage aims to build an automatic process for extracting hand-crafted features and network trained features, following by fusing them. The second stage can realize the classification process by exploring the temporal information from the fused features obtained by previous stage. The main contributions of this work are as follows:

1) A new network architecture consisting of two stages is developed. In the first stage, a window-DBN (WDBN) is designed to learn filters for generating trained features which are combined with the hand-crafted features. The second stage uses bidirectional LSTM (BLSTM) to further exploit the temporal information for improving the classification performance.

2) To solve the serious imbalance of the samples and improve the generalization ability of the model, a novel pre-training process combined with data augmentation strategy was proposed.

3) The proposed method was estimated on two different databases. The data in the first database were obtained from healthy people, and the data in the second database were obtained from the subjects with suspect sleep disorders. Through these two experiments, we can estimate the generalization of the model.

The main context of this study is organized as follows. In Section II, a detailed description of the methodology is introduced. In Section III, the experimental procedure and

TABLE 1. The employed hand-crafted feature of EEG.

Feature amounts	Feature description	Window size
5	Power spectral density (PSD) of δ (0.5–4 Hz), θ (4–8 Hz), α (8–13 Hz), β (13–20 Hz) and γ (> 20 Hz) band	5 s
3	Energy, Nonlinear energy and curve length [23]	5 s
4	Peak power and its corresponding frequency, mean, and median in PSD	5 s
1	Spectral entropy [23]	5 s
2	Hurst and fractal exponent [24], [25]	5 s
1	Kurtosis [26]	5 s
20	Multiscale sample entropy with length *m* of 1 and 2, coarse-graining scales of 1-10 [27]	30 s
4	Detrended fluctuation analysis (DFA) scaling exponent [28]	15 s, 20 s, 25 s, 30 s

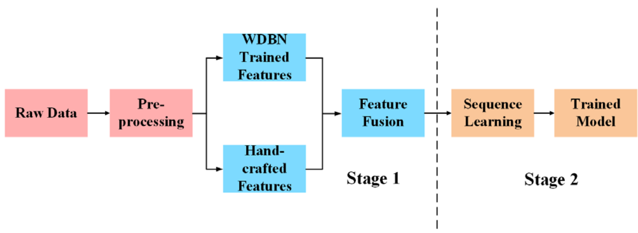


FIGURE 1. The architecture of the proposed network.

training parameters are described. The results that can reflect the performance of the model are presented in Section IV. In Section V, the discussion with regard to the experimental results and the model analysis are put forward. In the last section, we summarize this paper.

II. METHODOLOGY

A. ARCHITECTURE

The architecture of the proposed network consists of two main stages: the feature learning stage (Stage 1) and the sequence learning stage (Stage 2). The feature learning stage is used to provide fused feature matrix, and the sequence learning stage can learn the temporal information between successive epochs. Finally, the trained model can be obtained as shown in Fig. 1.

In Stage 1, the single-channel EEG as the raw data will be firstly pre-processed, then we extract features from the pre-processed signals by two different ways, one for extracting the hand-crafted features, and another is to obtain the network trained features by the WDBN. The two different feature matrices are fused to become the new feature matrix. Then the fused features from the Stage 1 is delivered to the Stage 2 with time order for capturing the temporal information. After the sequence learning process, the proposed model is trained, then the predicted sleep stages can be obtained by sending the testing data into the trained model.

B. FEATURE LEARNING STAGE

1) HAND-CRAFTED FEATURES

Each stage has its own characteristics for EEG signal. Those specific characteristics can be reflected by hand-crafted features to some extent. Hence extracting the specific features from EEG signal can express the differences between those sleep stages. In this study, a total of 40 EEG features are extracted after pre-processing as shown in Table 1. Those features were automatically extracted using sliding windows for each epoch with strides of 5 s, the feature amounts, feature descriptions and window sizes are provided. After obtaining the 40 features, a z-score normalization procedure is performed to reduce the impact of physiological differences and equipment-related variations from subject to subject. The feature vectors extracted from same sleep epoch would be concatenated for subsequent process. Although deep learning methods can extract useful features, completely abandoning those hand-crafted features is overcorrecting in the sleep medicine, because the design of those features respects the characteristics of the sleep stages in a certain degree.

2) NETWORK FEATURE LEARNING

The performance of classification methods based on the hand-crafted features is subject to the selection of features. However, finding novel and effective features is difficult. Besides, simply increasing number of features does not facilitate the classification process. So the deep learning approach for extracting and selecting useful features should be investigated.

DBN is a probabilistic generative model composed of multiple layers of stochastic, latent variables, and it's one of most popular networks used in feature learning. Each layer in DBN can be regard as a restricted Boltzmann machine (RBM) [29]. We can stack any number of RBMs to form the DBN, which is superior in feature extraction tasks [30].

Since the sampling rate of the EEG signals is commonly high, 30s-length signal contains thousands of points. However, a large proportion of downsampling may lose lots of information. Directly put such high-dimensional input into the DBN will increase computational complexity. Furthermore, many characteristics of the sleep stages can be reflected by a few second data, so we designed a WDBN to suit our scenario, as shown in Fig. 2. In this study, a 2-layer WDBN was adopted to handle the pre-processed EEG signal and obtain the network trained features, because the previous experiments found that two layers can take both accuracy and computational complexity into account.

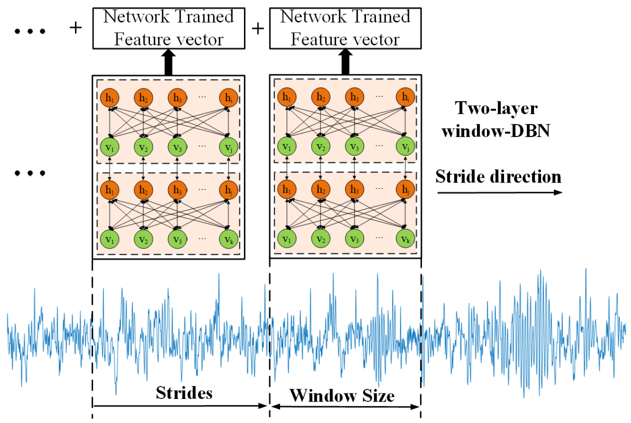


FIGURE 2. **The diagram of WDBN. The WDBN only process the signal with the length of window size, which is shorter than 30 s. The WDBN can go forward with a certain stride along the time direction, and each time it stays, DBN training will be performed. The obtained short network trained feature vector will be concatenated to form a 30 s epoch network trained feature.**

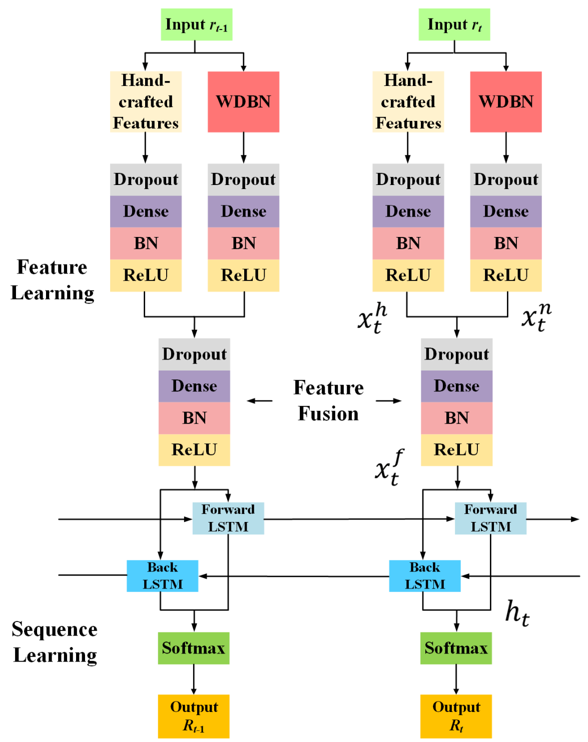


FIGURE 3. **The two-stage network structure which mainly contains the WDBN, hand-crafted feature extraction, and BLSTM. One dropout layer and dense layer are used for adjusting the hand-crafted features and network trained features before the fusion process. The fusion process firstly concatenate the two kinds of features, then utilizes one dropout layer and dense layer to select and fuse the concatenated features for obtaining the fused features. Then the fused features are transmitted to the sequence learning stage for training. The batch normalization (BN) operation and the rectified linear unit (ReLU) activation are added to each dense layer employed in this study.**

C. SEQUENCE LEARNING STAGE

The sleep stages in one full night have obvious temporal correlation and stage transition rules [31]. Sleep experts often use those rules to determine the current possible sleep stage based on the previous and next stages. For example, the REM stage typically follows epochs of stages N2 and, less commonly, stage Wake or N1 [31].

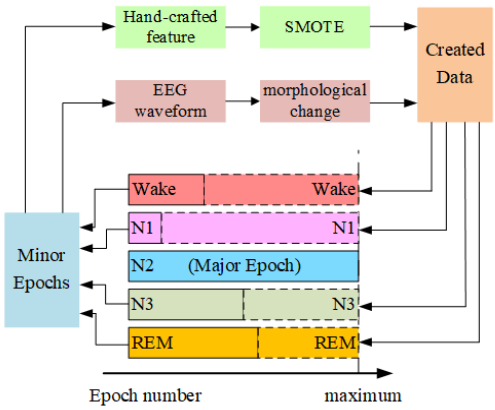


FIGURE 4. **The proposed oversampling approach combined with data augmentation.**

It is an important task to grasp those time series information and use them to adjust the results from the feature learning stage. LSTM is a good selection to complete this task, which is one kind of RNN module that has the advantage to explore dependencies between sequence inputs [32], [33]. It can selectively deliver the important information to next unit instead of all information which may contain useless message [34]. Besides, the LSTM has been already used for sleep staging [19], [20], whose effectiveness has been proved. But the LSTM architecture can only get information from the previous unit so that further improvements are introduced by BLSTM [35]. BLSTM is a two-direction LSTM structure, which means the current output can be simultaneously influenced by the forward and backward information. By this structure, the BLSTM can handle information both from the forward and backward direction, which makes BLSTM superior.

The two-stage model that combines the hand-crafted feature extraction, WDBN feature learning, and sequence learning is shown in Fig. 3. This procedure can be defined as following formulas:

$$x_t^n = WDBN(r_t) \tag{1}$$

$$x_t^h = Hand\_crafted(r_t) \tag{2}$$

$$x_t^f = Fusion(x_t^n \parallel x_t^h) \tag{3}$$

$$h_t = BLSTM_f(x_t^f) \tag{4}$$

$$R_t = softmax(h_t) \tag{5}$$

suppose a total of M 30-s epochs $\{r_1, ..., r_M\}$ with time order were generated after pre-processing, for $t = 1$ to M denotes the index of epochs. The *WDBN* and *Hand_crafted* represent the process of the WDBN and hand-crafted feature extraction, respectively. Thus x_t^n denotes the network trained features after the WDBN at tth epoch, and x_t^h means the hand-crafted features after network process at tth epoch. The *Fusion* and $\parallel$ express the feature fusion process and

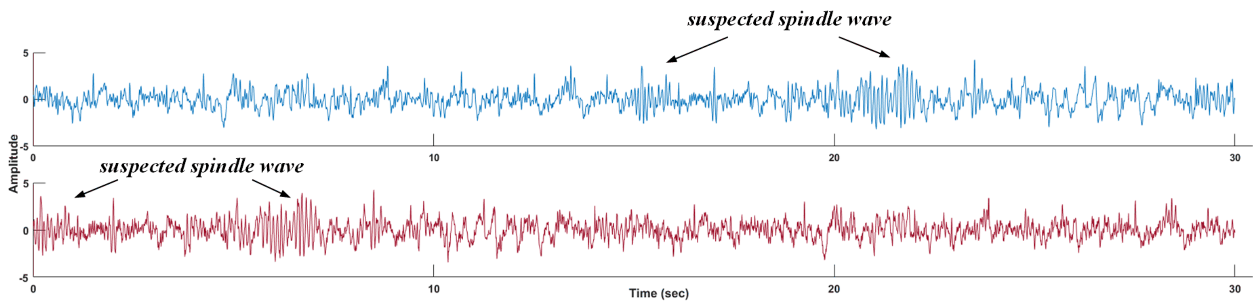


FIGURE 5. An example of morphological operation for one epoch EEG signal which contains two suspected spindle waves. The upper waveform is the original EEG signal, the lower one suffers one horizontal movements with displacement of 15s, and added the white noise with the SNR of 12 dB.

a concatenation operation, respectively. So the fused feature matrix x_t^f is the result obtained from the feature learning stage at tth epoch. The *BLSTM* represent the BLSTM process. The notation h_t express the output from BLSTM layer at tth epoch. The output of the BLSTM are then transmitted to the softmax layer, finally, the probability for five classes can be obtained.

D. PRE-TRAINING PROCESS WITH OVERSAMPLING

1) OVERSAMPLING

To prevent the class imbalance problem, the oversampling approach is commonly used to balance the samples in sleep staging. Traditional oversampling approach is to copy the samples from the minor classes in the training set until all the classes have the same number of samples [19]. Although this method can balance the weights in the network, it cannot let the network learn new patterns. So we designed a novel oversampling approach with data augmentation as shown in Fig. 4. This approach firstly identifies a sleep stage with maximum number of epochs, then create the same number of epochs from the minor sleep stages in the training set, such that all sleep stages have the same number of samples.

The data creation is realized by the SMOTE algorithm [36] and signal morphological change. Before transferring to the feature learning stage, the SMOTE algorithm is used to create the synthetic hand-crafted features for the minor epochs in the training dataset, simultaneously, the morphological operation which contains horizontal movement and noise addition is conducted on the EEG signal of the same epoch. The SMOTE algorithm can prevent the overfitting problem by improving the sample distribution. Horizontal movement represents an operation that the EEG signal is translated along the time axis with one random time period, in this study, we randomly select a time period between 5-25 s as the translation length. The noise addition means the signals randomly add a white noise with the signal to noise ratio (SNR) between 8-14 dB. An example of morphological operation for one epoch EEG signal that contains two suspected spindle waves is shown in Fig. 5. The lower EEG signal which suffers the morphological operation can be considered as a created data for augmentation.

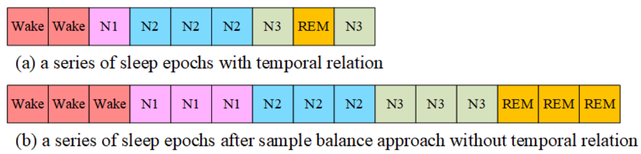


FIGURE 6. An example of a series of sleep epochs before and after the oversampling approach. (a) represents an original series of sleep epochs. (b) is the series (a) after the oversampling process whose temporal information is changed.

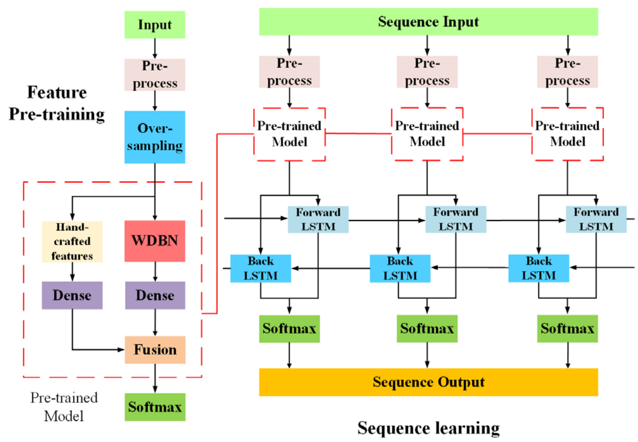


FIGURE 7. The whole training process which can both consider the sample balance and sequence information learning.

2) PRE-TRAINING PROCESS

Since we introduced the oversampling approach, the temporal information in series of sleep epochs was destroyed as shown in Fig. 6. Therefore, we need to introduce a pre-training process to protect the temporal relation and ensure the BLSTM can learn the sequence information under the premise of using the oversampling approach to learn comprehensive features.

The proposed model with designed pre-training process is shown in Fig. 7. It shows the structure from a temporal perspective. The feature pre-training is used to preliminarily train the feature learning stage with oversampling. After obtaining the pre-trained model, it would be reloaded and fine-tuned in the sequence learning stage, and the BLSTM layer is simultaneously trained to learn the sleep stage switch rules.

III. EXPERIMENT

A. DATA

A total of 63 full night recordings from two public databases: Sleep-EDF and Sleep Apnea (SA) Database were used to evaluate the proposed model [37]. The Sleep-EDF database has two subsets from two different studies: age effect in healthy subjects and Temazepam effects on sleep, which can be downloaded from PhysioNet [37]. We used all 20 healthy subjects from the first subset, which has total 39 full overnight recordings (one subject had only one night record and 19 subjects had two). Each recording contained 2 EEG channels (Fpz-Cz and Pz-Cz), 1 horizontal EOG and 1 chin EMG. Among them, the Fpz-Cz EEG was used in this study, whose sampling rate is 100 Hz. The SA database has been provided by St. Vincent's University Hospital and University College Dublin, which can also be downloaded from PhysioNet [37]. This database contains 25 full overnight recordings, from 25 adult subjects with suspected sleep disorder, for possible diagnosis of obstructive sleep apnea, central sleep apnea or primary snoring. Each record consists of 2 EEG channels (C3-A2 and C4-A1), 2 EOG channels, and 1 chin EMG channel. The C3-A2 EEG is used in this study, whose sampling rate is 128 Hz. One record has been removed due to the extremely abnormal annotation.

These recordings were manually scored into one of the six classes (Wake, S1, S2, S3, S4, and REM) by a sleep expert according to the R&K standard. Hence, the S3 and S4 stages were merged into the N3 stage to match the AASM standard. It's worth noting that if movement occurs in one sleep epoch and influences classification seriously, this epoch would be labeled as Movement or Artifact, one epoch would be labeled as Unknown if the sleep experts cannot estimate the stage of this epoch. We get rid of the epochs labeled without normal stages (Wake, N1, N2, N3, and REM) to improve the learning performance, because abnormal stages, such as Unknown, Artifacts, and Movement, do not contain learning objectives in the full night recording. After obtaining the raw data, it needs a de-noising process. A notch filter is applied to eliminate the 50 Hz or 60 Hz power frequency interference, and a band-pass filter of 0.3 to 35 Hz is used for the EEG signal.

The demographics, the percentages for five sleep stages and the total number of test epochs from the two databases are summarized in Table 2. It can be seen that the two databases have different sample distribution. The SA database has more Wake and N1 stages which indicates that the subjects with suspected sleep disorder have irregular sleep and bad sleep efficiency. The Leave One Subject Out (LOSO) approach was used to estimate the model performance for the two databases. The ratio of training subjects: validation subject: test subject is k-2: 1: 1, where the k is the number of subjects. The validation data is used to judge whether to store the trained model.

The experiments were performed on a Dell server with two Intel Xeon E5-2687W 3.0 GHz CPUs and four NVIDIA GeForce GTX1080Ti GPUs. The training process was implemented by utilizing Python3.6 with Tensorflow1.8 which is a deep learning library [38]. After the model is trained, it takes only tens of milliseconds to get one subject results.

TABLE 2. The percentage for five sleep stages from the two databases.

Databases	Sleep-EDF	Sleep Apnea
Subject Number (k)	39	24
Age (y)	29 ± 3	50 ± 10
Wake (%)	18.9	22.9
N1 (%)	6.7	15.9
N2 (%)	42.4	34.0
N3 (%)	13.6	12.9
REM (%)	18.4	14.3
Total Test Epochs	41950	20025

To assess the performance of the proposed model, we computed the accuracy, F1-score and Cohen's Kappa coefficient for the overall results and calculated the precision, recall and F1-score for each class. Thus, a total of five indices were used to assess the proposed model.

B. TRAINING PAREMETERS

During the feature pre-training process, a 2-layer WDBN was self-trained with the balanced samples via greedy layer-wise method, and then fine-turned by the labels. The weight matrix of hidden units was initialized to the normal distribution whose mean and variance is 0 and 1, respectively. During the WDBN training process, the stop mechanism was set to prevent overfitting, which would stop the WDBN training when the model cannot get better for more than continuous 10 training epochs. After the WDBN convergence, the network trained features are fused with the processed hand-crafted features. The cross-entropy loss was utilized to quantify the consistency between the predicted results and the groundtruth. The hyperparameters of network are shown in Table 3, which are chosen based on the previous experiments and literature reports [19], [22].

After the pre-training process, the input without the oversampling process, which preserves the time order information can be transmitted to the pre-trained model. One layer BLSTM is then utilized to learn the sequence information between tensors exported from the pre-trained model. We adopted 'sequence to sequence' LSTM structure, which can handle a certain number of input and obtain the same number outputs. This number is called sequence length which is set to 25 in this study. The cell and the hidden states in the BLSTM were reset in the beginning of each subject data. Ten sequences from the same subject data were simultaneously transferred to the BLSTM layer. The dropout technique was also employed in the BLSTM layer. The cross-entropy loss function was also utilized in the sequence learning stage. Besides, a heuristic gradient clipping technique is used to prevent the exploding gradients phenomenon. The model with best performance on the validation set is saved. After the

TABLE 3. The description of the network hyperparameters.

The description of hyperparameters	Value
Feature Learning Stage	
Batch size	100
Layer number of the WDBN	2
Hidden size of each layer in the WDBN	200
Number of training epochs of the WDBN	300
Window size of the WDBN	5 s
Strides of the WDBN	5 s
Initial biases of hidden units in the WDBN	0
Hidden size of the dense layers	400
Dropout probability	0.5
Learning rate of the dense layers	10^{-4}
Number of training epochs of the dense layers	200
Optimizer of the dense layers	Adam [39]
Sequence Learning Stage	
Sequence length	25
Batch size	10
Layer number of BLSTM	1
Hidden size of the BLSTM	500
Dropout probability	0.5
Learning rate for fine-tuning the pre-trained model	10^{-6}
Learning rate for sequence learning	10^{-4}
Clipping value of heuristic gradient clipping technique	5
Number of training epochs	200
Optimizer	Adam [39]

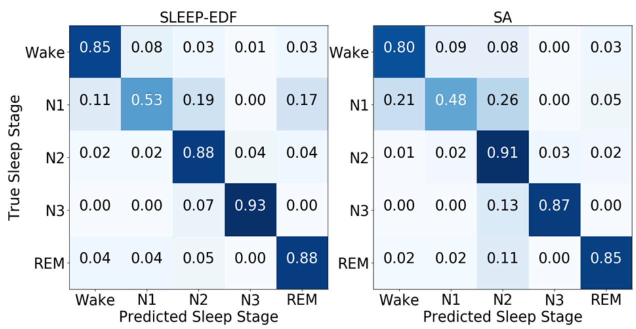


FIGURE 8. Normalized confusion matrices obtained by the proposed model from the Sleep-EDF database and SA database. The numbers in the main diagonal indicate the normalized epochs which were correctly classified in the corresponding stages. The darker the blue color in the blocks means the greater of the data.

whole training process, the results can be obtained by putting the test dataset into the trained model.

IV. RESULTS

A. SLEEP STAGE CLASSIFICATION PERFORMANCE

Fig. 8 shows the normalized confusion matrix obtained by the proposed model via LOSO from the Sleep-EDF database and SA database. It can be seen that the classification performance for the Wake, N2, N3 and REM stages are great, except that a small number stages were incorrectly classified to the N2 stages. For the N1 stages, most of them were correctly predicted, but many N1 stages are misclassified as the Wake, N2 and REM stages.

The precision, recall, and F1-score for each stage and the overall accuracy, F1-score, Kappa coefficient were calculated

TABLE 4. Per-class and overall indices obtained from the Sleep-EDF and SA database by the proposed model.

	Sleep-EDF (healthy)			SA (patients)		
	Pre	Recall	F1	Pre	Recall	F1
Wake	0.881	0.846	0.863	0.825	0.800	0.812
N1	0.538	0.525	0.532	0.719	0.482	0.577
N2	0.909	0.878	0.893	0.770	0.911	0.835
N3	0.861	0.929	0.894	0.903	0.866	0.884
REM	0.819	0.876	0.846	0.836	0.852	0.844
Overall	Acc	F1	Kappa	Acc	F1	Kappa
	0.855	0.806	0.80	0.803	0.790	0.74

Pre=Precision, F1 = F1-score, Acc = Accuracy.

and provided in Table 4. From the per-class classification performance for the two databases, it represents that the proposed method has good ability to identify the Wake, N2, N3, and REM stages whose worst F1-score is 0.812 for classifying the Wake stages in the SA database. The poorest performance occurred in classifying the N1 stage, in which the best F1-score is 0.577 for the SA subset. With the small training samples and few characteristics of the N1 stage, the N1 stage is the most undistinguishable stage in the five classes. The overall accuracy, F1-score and Kappa coefficient for the Sleep-EDF database are 0.855, 0.806, and 0.80, respectively, while are 0.803, 0.790, and 0.74 for the SA database, respectively. The Kappa coefficients showed that our proposed method had a substantial agreement with the sleep expert, and the accuracies are promising and stable for the two databases with different characteristics. Besides, the F1-scores indicated that the proposed model can take into account both the precision and recall.

From the classification performance obtained from the two different databases, it can be seen that the Sleep-EDF database has better performance. This maybe because the data in the Sleep-EDF database is collected from healthy people, while the data in the SA database is collected from subjects with sleep disorders. The hypnogram of subjects with sleep disorders is more disordered than that of healthy people, consequently, the classification work for the SA database is relatively difficult. In addition, the differences of acquisition equipment and other environment would also lead to different classification results. Even though our model displayed different results on the two databases whose subject conditions, acquisition equipment and environment are various, the tests of two databases still show promising results.

An example of hypnogram for Subject-1 from the Sleep-EDF database for more than 800 epochs at one night is demonstrated in Fig. 9. We can see that most REM stages follow the epochs of stage N2 instead of other stages by the prediction of our proposed model, which is consistent with the stage switch rules as mentioned before. Besides, most misclassified stages appeared in the transitions between N1 and other stages. This phenomenon is similar with the information provided by Fig. 8 and Table 4 in a certain degree.

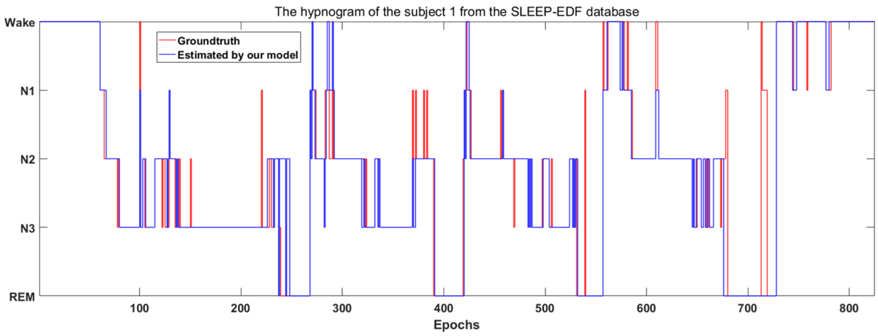


FIGURE 9. An example of hypnogram which is manually scored by a sleep expert and automatically scored by our model for Subject-1 from the Sleep-EDF database. The red solid line and the blue solid line represent the sleep stage interpreted by the sleep expert and our model, respectively, for more than 800 epochs at one night.

B. THE PERFORMANCE OF MODEL COMPONENTS

The proposed model has four important components: WDBN, hand-crafted features, BLSTM and novel pre-training process. To evaluate the benefits of different parts in the proposed model to the classification performance, we separated the proposed method and use other approaches to replace the components in our proposed method. To clarify the improvement brought by the BLSTM, support vector machine (SVM) and Hidden Markov Model (HMM) were used to replace the BLSTM to train the model. The SVM and HMM are widely used traditional machine learning methods which can be used as two comparison methods [40], [41]. Six methods were then derived from the proposed two-stage model as described below:

1) Method 1: Hand-crafted features + SVM. This method extracted 40 hand-crafted features, these features are then transmitted to the SVM with RBF-kernel for training. Finally, the trained model can obtain the classification results. The parameters of the SVM is selected automatically via the validation set.

2) Method 2: Hand-crafted features + BLSTM. This method transmitted 40 hand-crafted features to the BLSTM for training. The trained BLSTM can obtain the classification results. The architecture and the training parameters of BLSTM are the same as the proposed model.

3) Method 3: WDBN + BLSTM. This method utilized the EEG signal to train WDBN for obtaining the network trained features, then using the features to train the BLSTM. It's a two stage process without extracting the hand-crafted features. The architecture and the training parameters of the remaining parts are the same as the proposed model.

4) Method 4: WDBN + hand-crafted features + HMM. The WDBN combined hand-crafted features is the feature learning stage (Stage 1) in our proposed model, the fused features are then transferred to the HMM for training. Finally, the data is sent into the trained model for sleep stages classification. The architecture and the training parameters of Stage 1 are the same as the proposed model.

5) Method 5: WDBN + hand-crafted features + BLSTM. This method is our final proposed method without the novel pre-training process as shown in Fig. 3. The architecture and the training parameters of the remaining part are the same as the proposed model.

6) Method 6: WDBN + hand-crafted features + BLSTM + pre-training process. The method is actually our final proposed method as shown in Fig. 7.

The Sleep-EDF database was employed to test the six methods, the employed channel is Fpz-Cz EEG and the number of test epochs is 41950. The LOSO was performed to estimate the performance. Table 5 shows the comparison between the six methods across total accuracy, F1-score, Kappa coefficient and F1-score for each class. The confusion matrices obtained by Method 1 to Method 5 are demonstrated in Fig. 10. The confusion matrix and the results of Method 6 is already given in Fig. 8 and Table IV, respectively. When using the SVM to train the hand-crafted features (Method 1), it got the worst performance in the classification comparison. Only used the hand-crafted features (Method 2) or WDBN (Method 3) as the feature learning process, the classification performance cannot exceed Method 5, which combined the WDBN with the hand-crafted features. Almost all indices in Method 5 achieve the better performance than those indices in Method 2 or Method 3. It indicates that network trained features can provide effective information for the sleep staging. Besides, the hand-crafted features provide supplementary information for the network trained features. However, when using the HMM to replace the BLSTM (Method 4), the performance decreases sharply. The results demonstrated that the BLSTM in the sequence learning stage is superior to the HMM for sleep staging. The employment of the BLSTM for sequence learning would gain more than 5% increasing in total accuracy. The Method 6 as our final proposed model has the novel pre-training process, it achieved the best effects in most all indices which proved that the pre-training process combined with data augmentation can make the feature learning more comprehensive and robust.

TABLE 5. Comparison of performance of different model components via loso from the sleep-edf database.

Methods	Overall results			F1-score for each class				
	Acc	F1	Kappa	Wake	N1	N2	N3	REM
Hand-crafted features + SVM	0.695	0.621	0.59	0.724	0.213	0.745	0.714	0.711
Hand-crafted features + BLSTM	0.795	0.740	0.72	0.843	0.458	0.823	0.800	0.776
WDBN + BLSTM	0.822	0.752	0.76	0.869	0.385	0.858	0.834	0.817
WDBN + hand-crafted features + HMM	0.790	0.738	0.72	0.828	0.422	0.834	0.836	0.769
WDBN + hand-crafted features + BLSTM	0.846	0.792	0.79	**0.886**	**0.511**	0.873	0.855	0.832
WDBN + hand-crafted features + BLSTM + pre-training process	**0.855**	**0.806**	**0.80**	0.863	**0.512**	**0.893**	**0.894**	**0.846**

Acc = Accuracy, F1 = F1-score.

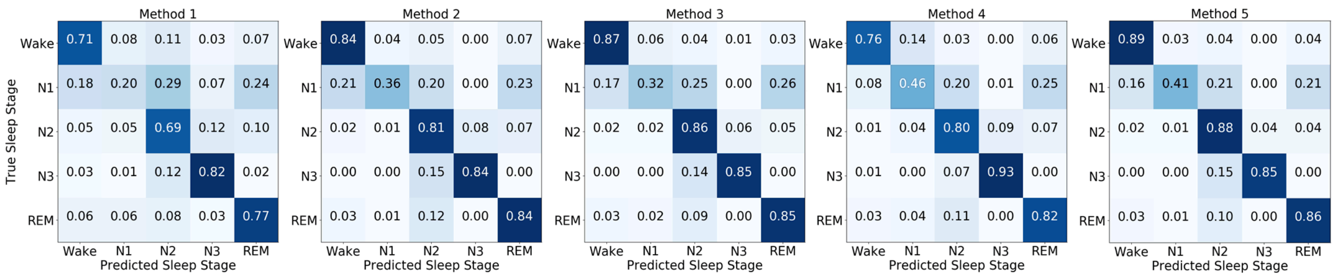


FIGURE 10. Normalized confusion matrix obtained by the Method 1-5 with different model components from the Sleep-EDF database.

TABLE 6. comparison of performance of different methods using the same signals from the same databases.

Databases	Methods	Employed Channels	Test epochs	Overall results			F1-score for each class				
				Acc	F1	Kappa	Wake	N1	N2	N3	REM
Sleep-EDF	Ref [21]	Fpz-Cz	37022	0.748	0.698	0.65	0.654	0.437	0.806	0.849	0.745
	Ref [18]	Fpz-Cz	37022	0.789	0.737	0.71	0.716	0.470	0.846	0.840	0.814
	Ref [19]	Fpz-Cz	41950	0.820	0.769	0.76	0.847	0.466	0.859	0.848	0.824
	Our model	**Fpz-Cz**	**41950**	**0.855**	**0.806**	**0.80**	**0.863**	**0.532**	**0.893**	**0.894**	**0.846**
SA	Ref [22]	C3-A2, EOG, EMG	20789	0.722	0.705	0.64	0.780	0.370	0.760	0.840	0.780
	Ref [9]	C3-A2, C4-A1	15000	0.810	-	-	-	-	-	-	-
	Our model	**C3-A2**	**20025**	**0.803**	**0.790**	**0.74**	**0.812**	**0.577**	**0.835**	**0.884**	**0.844**

Acc = Accuracy, F1 = F1-score.

C. COMPARISON WITH OTHER APPROACHES

To further understand the performance of the model, we compared it with other related studies which used the same databases for testing [9], [18], [19], [21], [22]. Those state-of-the-art methods are introduced in Section I. Table 6 shows a comparison between our method and other sleep stage classification methods across total accuracy, F1-score, Kappa coefficient and F1-score for each class. Those studies also conducted the subject independent verification. From Table 6, it can be seen that our method has advantages compared to those state-of-the-art methods. The [18], [19], [21] and our study used single-channel EEG (Fpz-Cz) in the Sleep-EDF database to train the models. The results represent that our proposed model generated higher performance. For the SA database, the [22] and [9] utilized multiple signals to train their model, while we employed only the C3-A2 EEG to train our model and get the results. For the comparison by using the SA database, our proposed model achieved higher accuracy, F1-score, and Kappa coefficient than the [22] which used C3-A2, EOG, and EMG channels, and obtained similar performance with the [9] which employed two EEG channels with highly screening the test data.

D. NETWORK TRAINED FEATURES

The hand-crafted features with obvious physical significance have clear interpretability, which can be proved to effective in theory. However, the hand-crafted features greatly compress the information of the signal, while the morphological characteristics of the signal might be ignored. So we use the WDBN to extract the network trained features which can preserve the morphological information, because the length of network trained features is equal to the window size, and the network trained features could be considered as many filters for match the specific patterns in the sleep medicine. The combination of the network trained features and hand-crafted features showed better effects on classification which indicated that these two kinds of features have complementary information. Hence, the network trained features are worth

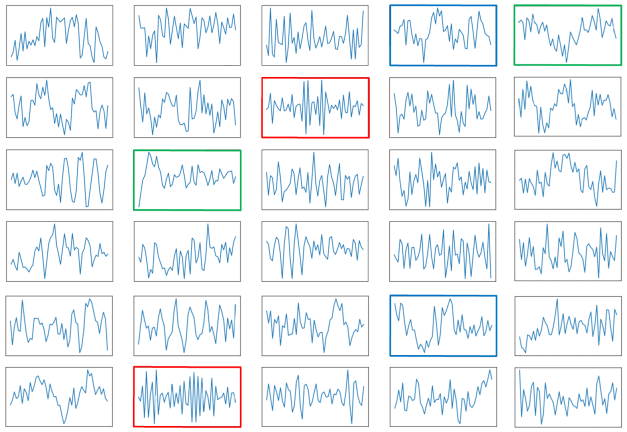


FIGURE 11. **Network trained features extracted by the WDBN. It can be observed that the learned features are of various forms and some waveforms are suspected of sleep events such as spindle wave, K-complex or slow wave. Only the first 30 features are shown here.**

exploring. We plot the network trained features extracted from the WDBN, which are shown in Fig. 11.

From Fig. 11, we can see that the network trained features reflect different forms of signal characteristics. To verify the diverse and complementary of these features, we validated them with the sleep experts. High similarities in the morphology between network trained features and the sleep events defined by AASM were found. To illustrate, the features denoted by red boxes and blue boxes have high coherence with sleep spindle and K-complex, respectively [31]. Many features with slow changes are similar to the slow waves that were denoted by greed boxes [31]. These different features are complemented to distinguish between different sleep stages. For example, in N2 stage, one of the typical sleep events is K-complex that can be presented as negative, sharp wave immediately followed by a positive component on the morphology of the EEG signal; The spindle wave and slow wave in the EEG signal are the characteristics of the N2 stage and N3 stage, respectively [31]. In the following layers, those features can be further selected and mapped to the most relevant stage. Namely, these features serve as significant characteristics of the morphology in distinguishing different stages, which is similar to the sleep medicine experts.

V. DISCUSSION AND FUTURE WORK

This study provides a novel sleep stage classification method which combined WDBN, hand-crafted features, BLSTM and novel pre-training process. Those important components compose the two-stage network, which has become the popular structure in dealing with physiological signals [19], [20]. The experimental results show that the proposed method has promising performance for the sleep stage classification with single-channel EEG.

As shown in the Fig. 8 and Table 4, the error rate of stage N1 is obviously higher than other stages, and the error mainly occurs when the real N1 epochs are misclassified to other major stages such as Wake, N2, and REM epochs. This may be caused by the features for distinguishing the stage N1 are relatively less than other stages, and the number of stage N1 epochs is also less than other stages which leads to poor learning. From a technical point of view, the lack of powerful features and temporal information to identify the N1 stage could be the reason for this phenomenon. Hence, dig out novel and valid features and networks, further explore the temporal information of N1 stage, to improve the classification model will be the next mission. Transfer learning [42] may solve the problem of insufficient learning in the N1 stage. For example, we can extract the network layer related to the N1 stage from the trained networks of other databases, then put those layers into our network to assist the N1 classification.

It can be speculated that if we change the parameters of WDBN, for example, adopt a variety of window sizes and strides, the obtained network trained features can contain more information in theory, but this will greatly increase calculation complexity, so novel network structure which can extract more information while reducing the computational complexity is worth exploring. In addition, the hand-crafted features extracted from single-channel EEG expressed limited information for sleep analysis, the performance of classification maybe improved by exploiting new and effective features which can characterize additional discriminating information.

In Section IV-B, Method 1 (Hand-crafted features + SVM) and Method 2 (Hand-crafted features + BLSTM) sent only the hand-crafted features into the SVM model and BLSTM for classification, respectively. The results show that only the hand-crafted features cannot provide enough information for the classification. In addition, the performance of Method 2 is better than Method 1, which indicated that the sequence learning is effective. If we only use the WDBN for feature learning as applied in Method 3 (WDBN + BLSTM), the results demonstrated that the WDBN is superior to the hand-crafted features for sleep staging. When combining the WDBN with the hand-crafted features (Method 5), the classification performance would be further improved, the phenomenon indicates that the information obtained from the hand-crafted features and the network trained features are mutual complementation. Besides, the BLSTM outperforms the traditional approach for the sequence learning in this study.

For comparison with traditional methods, the deep learning approaches have advantages in sleep staging. From the perspective of feature extraction of the proposed model, the WDBN successfully extracted different morphological features, which is consistent with the way of sleep experts for judging the sleep stage according to observation. Besides, the classification is conducted by the BLSTM based on the time series information in the successive sleep epochs. While the traditional methods are usually weak in exploring the relations between sleep epochs. For comparison with other deep learning approaches. The traditional feature extraction approaches (hand-crafted) are combined with our proposed model, to extract the comprehensive characteristics of the employed signals. Besides, we employed the pre-training strategy to make the model have better generalization.

Finally, we need to highlight the limitations and the future directions of this study. Although we analyzed the network trained features in a certain degree and found that this approach based on the deep learning is similar to the sleep expert, the deep learning framework has relatively poor interpretability compared with the traditional machine learning methods. Besides, extract the hand-crafted features requires prior knowledge of sleep medicine to estimate the effect of the features. Furthermore, the sleep staging performance for the subjects with sleep disorders is worse than that of healthy people. Therefore, we should pay more attention to the improvement of the sleep classification algorithm for patients with sleep disorders, and the detection of events related to sleep diseases in the future. In addition, the proposed data augmentation strategy may just increase the receptive field of the WDBN instead of creating new patterns for the minor sleep epochs. Hence, explore novel data augmentation strategy to solve the serious sample unbalance problem is one of our research directions. For example, generative adversarial network (GAN) [43] is currently the most popular data generation tool which can be used to artificially synthesize sleep epochs of minor stages for balancing the samples. In addition, it's significant to explore which is the most relevant hand-crafted and network trained features to different stages by decoding the mapping of the network. But it's still extraordinarily difficult in deep learning field due to the complex network structure and large number of parameters. Maybe we can solve this problem by the study of model interpretability in the future.

VI. CONCLUSION

This paper presents a novel sleep stage classification method based on a two-stage neural network. In the first stage, the proposed model used WDBN to extract network trained features from single-channel EEG, and the network trained features can be fused with hand-crafted features by the model. The fused features are then transferred to the second stage for obtaining the sequential results. The evaluation of the proposed model was conducted on two public databases, and the performance is promising. The combination of the hand-crafted features and the network trained features can improve the classification performance via the model comparison experiments. The BLSTM can achieve good performance for learning the temporal information. In addition, the novel pre-training process combined with data augmentation strategy was verified that it can improve the accuracy and generalization of the model. We believe that our proposed model can be a good choice for sleep stage classification.

REFERENCES

[1] C. Iber, S. Ancoli-Israel, and A. L. Chesson, *The AASM Manual for the Scoring of Sleep and Associated Events : Rules, Terminology and Technical Specifications*. Westchester, IL, USA: American Academy of Sleep Medicine, 2007.

[2] J. A. Hobson, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," *Electroencephalogr. Clin. Neurophysiol.*, vol. 26, no. 6, p. 644, 1969.

[3] A. R. Hassan and M. I. H. Bhuiyan, "A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features," *J. Neurosci. Methods*, vol. 271, pp. 107–118, Sep. 2016.

[4] C.-S. Huang, C.-L. Lin, L.-W. Ko, S.-Y. Liu, T.-P. Su, and C.-T. Lin, "Knowledge-based identification of sleep stages based on two forehead electroencephalogram channels," *Frontiers Neurosci.*, vol. 8, p. 263, Sep. 2014.

[5] R. Sharma, R. B. Pachori, and A. Upadhyay, "Automatic sleep stages classification based on iterative filtering of electroencephalogram signals," *Neural Comput. Appl.*, vol. 28, no. 10, pp. 2959–2978, 2017.

[6] A. R. Hassan and A. Subasi, "A decision support system for automated identification of sleep stages from single-channel EEG signals," *Knowl. Based Syst.*, vol. 128, pp. 115–124, Jul. 2017.

[7] T. Lajnef, S. Chaibi, P. Ruby, P.-E. Aguera, J.-B. Eichenlaub, M. Samet, A. Kachouri, and K. Jerbi, "Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines," *J. Neurosci. Methods*, vol. 250, pp. 94–105, Jul. 2015.

[8] R. Boostani, F. Karimzadeh, and M. Nami, "A comparative review on sleep stage classification methods in patients and healthy individuals," *Comput. Methods Programs Biomed.*, vol. 140, pp. 77–91, Mar. 2017.

[9] J. Shi, X. Liu, Y. Li, Q. Zhang, Y. Li, and S. Ying, "Multi-channel EEG-based sleep stage classification with joint collaborative representation and multiple kernel learning," *J. Neurosci. Methods*, vol. 254, pp. 94–101, Oct. 2015.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[11] T.-H. Wen, M. Gašić, and N. Mrkšić, "Semantically conditioned LSTM-based natural language generation for spoken dialogue systems," in *Proc. Conf. Empirical Methods Natural Lang.*, Lisbon, Portugal, Sep. 2015, pp. 1711–1721.

[12] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, Fort Lauderdale, FL, USA. Apr. 2011, pp. 315–323.

[13] D. F. Wulsin, J. R. Gupta, R. Mani, J. A. Blanco, and B. Litt, "Modeling electroencephalography waveforms with semi-supervised deep belief nets: Fast classification and anomaly measurement," *J. Neural Eng.*, vol. 8, no. 3, 2011, Art. no. 036015.

[14] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, Chiba, Japan, Sep. 2010, pp. 1045–1048.

[15] Y. Du, W. Wang, and L, Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1110–1118.

[16] Z. Zhou, Y. Wang, J. Yu, Y. Guo, W. Guo, and Y. Qi, "High spatial–temporal resolution reconstruction of plane-wave ultrasound images with a multichannel multiscale convolutional neural network," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 65, no. 11, pp. 1983–1996, Nov. 2018.

[17] J. Liu, F. Chen, C. Pan, M. Zhu, X. Zhang, L. Zhang, and H. Liao, "A cascaded deep convolutional neural network for joint segmentation and genotype prediction of brainstem gliomas," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1943–1952, Sep. 2018.

[18] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic sleep stage scoring with single-channel EEG using convolutional neural networks," *stat*, vol. 5, p. 1050, Oct. 2016.

[19] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017.

[20] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, Feb. 2018.

[21] O. Tsinalis, P. M. Matthews, and Y. Guo, "Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders," *Ann. Biomed. Eng.*, vol. 44, no. 5, pp. 1587–1597, May 2016.

[22] M. Längkvist, L. Karlsson, and A. Loutfi, "Sleep stage classification using unsupervised feature learning," *Adv. Artif. Neural Syst.*, vol. 2012, May 2012, Art. no. 107046.

[23] A. Noviyanto, S. M. Isa, I. Wasito, and A. M. Arymurthy, "Selecting features of single lead ECG signal for automatic sleep stages classification using correlation-based feature subset selection," *Int. J. Comput. Sci. Issues*, vol. 8, no. 5, pp. 139–148, 2011.

[24] U. R. Acharya, E. C.-P. Chua, O. Faust, T.-C. Lim, and L. F. B. Lim, "Automated detection of sleep apnea from electrocardiogram signals using nonlinear parameters," *Physiol. Meas.*, vol. 32, no. 3, pp. 287–303, Feb. 2011.

[25] E. Pereda, A. Gamundi, R. Rial, and J. González, "Non-linear behaviour of human EEG: Fractal exponent versus correlation dimension in awake and sleep stages," *Neurosci. Lett.*, vol. 250, no. 2, pp. 91–94, Jun. 1998.

[26] T. Gasser, P. Bäecher, and J. Möecks, "Transformations towards the normal distribution of broad band spectral parameters of the EEG," *Electroencephalogr. Clin. Neurophysiol.*, vol. 53, no. 1, pp. 119–124, Jan. 1982.

[27] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Phys. Rev. Lett.*, vol. 89, Jul. 2002, Art. no. 068102.

[28] X. Long, J. Foussier, P. Fonseca, R. Haakma, and R. M. Aarts, "Analyzing respiratory effort amplitude for automated sleep stage classification," *Biomed. Signal Process. Control*, vol. 14, pp. 197–205, Nov. 2014.

[29] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio, "Learning algorithms for the classification restricted Boltzmann machine," *J. Mach. Learn. Res.*, vol. 13, pp. 643–669, Mar. 2012.

[30] B. Taji, A. D. C. Chan, and S. Shirmohammadi, "Classifying measured electrocardiogram signal quality using deep belief networks," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, Turin, Italy, May 2017, pp. 1–6.

[31] R. B. Berry, *Fundamentals of Sleep Medicine*. Amsterdam, The Netherlands: Elsevier, 2011.

[32] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 6645–6649.

[33] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, Lille, France, Jul. 2015, pp. 2342–2350.

[34] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[35] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[36] P. Jeatrakul, K. W. Wong, and C. C. Fung, "Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm," in *Proc. Int. Conf. Neural Inf. Process.* Berlin, Germany: Springer, 2010, pp. 152–159.

[37] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, C. K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. 215–220, 2000.

[38] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement.*, Savannah, GA, USA, Nov. 2016, pp. 265–283.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*.m [Online]. Available: https://arxiv.org/abs/1412.6980

[40] T. Willemen, D. Van Deun, V. Verhaert, M. Vandekerckhove, V. Exadaktylos, J. Verbraecken, S. Van Huffel, B. Haex, and J. V. Sloten, "An evaluation of cardiorespiratory and movement features with respect to sleep-stage classification," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 2, pp. 661–669, Mar. 2014.

[41] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "An unsupervised approach for automatic activity recognition based on hidden Markov model regression," *IEEE Trans. Autom. Sci. Eng.*, vol. 10, no. 3, pp. 829–835, Jul. 2013.

[42] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

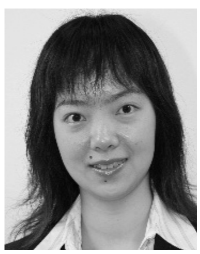
[43] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018.

CHENGLU SUN received the B.S. degree in mechatronic engineering from the Shanghai Normal University, Shanghai, in 2013, and the M.S. degree in mechanical manufacture and automation from the University of Shanghai for Science and Technology, Shanghai, in 2016. He is currently pursuing the Ph.D. degree with the Center for Intelligent Medical Electronics, Department of Electronic Engineering, School of Information Science and Technology, Fudan University, Shanghai, China. His research interests include machine learning, signal processing, and computer vision with an emphasis on unobtrusive physiological signal monitoring.

JIAHAO FAN received the B.S. degree in communication engineering in 2013, and the M.S. degree in electronic and communication engineering from Jilin University, in 2016. He is currently pursuing the Ph.D. degree with the Center for Intelligent Medical Electronics, Department of Electronic Engineering, School of Information Science and Technology, Fudan University, Shanghai, China. His research interest focuses on sleep-related algorithms, which mainly include automatic stage scoring and sleep disease detection. He is also interested in signal processing, machine learning, and artificial neural networks.

CHEN CHEN received the M.S. degree in embedded system from the Institut Supérieur d'Electronique de Paris (ISEP), in 2013, and the Ph.D. degree in computer science from the Université Pierre et Marie Curie (UPMC), in 2016. She is currently a Postdoctoral Researcher with the Centre for Intelligent Medical Electronics, School of Information Science and Technology, Fudan University. Her research interest includes biomedical engineering, focusing on biomedical signal processing, wearable sensor systems, sleep analysis, and personalized health monitoring.

WEI LI received the B.S. degree from Fudan University, Shanghai, China. He is currently pursuing the M.S. degree with the Center for Intelligent Medical Electronics, Department of Electronic Engineering, Fudan University, where he studied with the Department of Electronic Engineering, from 2012 to 2016. His research focuses on physiological signal measurement and sleep monitoring. His research interests include biomedical signal processing, sensor technology, as well as unobtrusive physiological measurement techniques.



WEI CHEN received the B.S. and M.S. degrees from Xian Jiaotong University, in 1999 and 2002, respectively, and the Ph.D. degree from The University of Melbourne, in 2007. From 2008 to 2015, she was an Assistant Professor with the Department of Industrial Design, Technical University of Eindhoven. She is currently a Full Professor with the Department of Electronic Engineering, Fudan University. Her research interests include wearable sensor system, biomedical signal processing, and health monitoring. She is also an Associate Editor of the IEEE Journal of Biomedical Health Informatics (JBHI), a regional representative of the IEEE EMBS Technical Committee on Wearable Biomedical Sensors and Systems, and the Vice Chair of the IEEE Sensors and Systems Council.