Received June 10, 2019, accepted June 19, 2019, date of publication June 26, 2019, date of current version July 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2924980

A Multi-Class Automatic Sleep Staging Method Based on Long Short-Term Memory Network Using Single-Lead Electrocardiogram Signals

YUHUI WEI[1,2], XIA QI[1,2], HUANING WANG[3], ZHIAN LIU[1,2], GANG WANG[1,2], (Member, IEEE), AND XIANGGUO YAN[1,2]

[1]The Key Laboratory of Biomedical Information Engineering of Ministry of Education, Key Laboratory of Neuro-informatics and Rehabilitation Engineering of Ministry of Civil Affairs, School of Life Science and Technology, Institute of Biomedical Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[2]National Engineering Research Center for Healthcare Devices, Guangzhou 510500, China
[3]Xijing Hospital, Fourth Military Medical University, Xi'an 710032, China

Corresponding authors: Gang Wang (ggwang@xjtu.edu.cn) and Xiangguo Yan (xgyan@mail.xjtu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grants 31571000 and Grant 61471291, and in part by the Fundamental Research Funds for the Central Universities of China under Grant xjj2017122.

ABSTRACT To overcome the disadvantage of clinical manual sleep staging, a convenient, economical, and efficient multi-class automatic sleep staging method is proposed based on long short-term memory network (LSTM) using single-lead electrocardiogram signals. From electrocardiogram signals, heart rate variability and respiratory signals were calculated, and, then, totally 25 features were extracted. Four different classifiers, including the two-class classifier to distinguish between wake and sleep, the three-class classifier to distinguish wake, non-rapid eye movement sleep, and rapid eye movement, the four-class classifier to distinguish wake, light sleep, slow wave sleep, and rapid eye movement, and the five-class classifier to distinguish wake, sleep stage N1, sleep stage N2, sleep stage N3, and rapid eye movement, were constructed using the LSTM. The single-lead electrocardiogram data from 238 patients with full sleep stages during sleep were used for the training set and the data from other 60 patients were regarded as a validation set. The rest of 75 patients have left aside for testing set. The accuracy of two-class, three-class, four-class, and five-class sleep staging was 89.84%, 84.07%, 77.76%, and 71.16% and the Cohen's kappa statistic k was 0.52, 0.58, 0.55, and 0.52, respectively, which realized the moderate agreement with clinical analysis. When expanding the dataset to extra 1068 patients with missing sleep stages, the accuracy has no obvious reduction but the Cohen's kappa statistic k dropped to 0.51, 0.52, 0.48, and 0.43, respectively. The proposed method, in this paper, is promising for low-cost, efficient, and convenient sleep staging in home care monitoring.

INDEX TERMS Electrocardiogram, heart rate variability, long short-term memory, sleep staging.

I. INTRODUCTION

Sleep is one of the most important physiological activities of human body, and sleep staging is one of the most efficient approaches to evaluate the equality of sleep. Nowadays, the most authoritative sleep staging standard is set by the American Academy of Sleep Medicine (AASM). Based on polysomnography (PSG) and the AASM Manual for the Scoring of Sleep and Associated Events Rules, sleep activities can be divided into five stages: wake (W), stage I (N1), stage II (N2), stage III (N3), and rapid eye movement (REM) [1]. The PSG consists of multi-channel biosignals including electroencephalogram (EEG), electromyogram (EMG), electrooculogram (EOG), electrocardiogram (ECG) and respiratory signals, and then experienced doctors examine the PSG signals of every 30-second frame to obtain the clinical classification results [2]. In non-clinical applications, there are also different standards of sleep staging, such as 2-class to distinguish W and sleep, and 3-class to distinguish W, non-rapid eye movement (NREM) and REM, 4-class to distinguish W, light sleep (LS), slow wave sleep (SWS), and REM [3]. The main purpose of this paper is to explore a general, convenient, and economical sleep staging method.

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

The sleep staging method based on PSG technology has obvious disadvantages in practical applications. All the signals are strictly measured in certain laboratory, which is inaccessible for most people. The measurement of multi-channel physiological signals may cause discomfort to sleep, making the measurement results deviate from the real situation. What's worse, expensive cost also limits the application of PSG. Therefore, it is of great significance for exploring an automatic sleep staging method. At present, the research of automatic sleep staging method can be divided into three aspects according to the used physiological signals. Firstly, the automatic sleep staging methods based on EEG signals have realized considerable performance since there is a direct link between EEG and electrophysiological activities of the brain [4]. The accuracy of 5-class sleep staging using single-lead EEG for healthy people can exceed 90% [5], [6]; 5-class sleep staging based on EEG, EMG, and EOG can achieve the accuracy of 92% or even higher for healthy people [7], [8], and 86% accuracy for patients with sleep disorders [9]. Secondly, the sleep staging methods based on cardiopulmonary coupling signals that mainly contain ECG and respiratory signals have attracted more and more attention. The reported study in 3-class sleep staging for patients with sleep disorder has achieved a maximum accuracy of 71.9% [10]. Thirdly, the sleep staging methods based on the acceleration signals during sleep is mainly to classify W and sleep, achieving the highest consistency of 91% with PSG system [11].

However, there are certain practical problems in the above studies. Although the staging methods based on EEG signals have high accuracy, there are current limitations in the recording technologies for measuring EEG activity in clinical and experimental applications [12]. EEG is very susceptible to various interference, thus the requirements for electrodes and measurement environment are strict, which usually lead to relatively high cost [13], [14]. Most of the sleep staging methods based on acceleration signals can only distinguish W and sleep, which is unreliable and lack of significance. However, ECG is one of the large-amplitude physiological signals which is relatively easy to obtain. Therefore, the sleep staging methods based on cardiopulmonary coupling is of great practical significance. At present, the most efficient approaches on sleep staging method based on cardiopulmonary coupling mainly focus on heart rate variability (HRV) and respiratory rate variability (RRV) which has obvious characteristics during different sleep states [15], [16]. Yucelbas, S., et al. compared the results of four different classification methods to classify 3-class sleep staging for healthy people, achieving the highest accuracy of 87.11%, but only 78.08% for patients with obstructive sleep apnea [17]. While another research using ECG and acceleration signals only achieved the accuracy of 74.5% [18]. Since the RRV based features can be extracted from ECG signals [19], the sleep staging methods involved HRV and RRV features can be simply implemented by using single-lead ECG signals. In general, the current research of sleep staging with ECG signals has two disadvantages. On the one hand, some methods just obtained good results on healthy subjects but performed poor on patients, lacking of universality and robustness. On the other hand, the state-of-art sleep staging accuracy was relatively low, far away from practical application. Thus, it is very important to further explore sleep staging methods using ECG signals.

The long short-term memory network (LSTM) has more advantages than other methods when dealing with pattern recognition problems for time series [20]. LSTM model adds a forgotten gate based on the traditional Recurrent Neural network (RNN) [21], making the neural network selectively forget the previously learned parameters. So LSTM can utilize the temporal correlation of time series and avoid the problem of long-term dependence [22]. Yulita et al. used Bi-directional LSTM for sleep staging using EEG, EOG, and EMG signals, achieving an accuracy of 86% for patients with sleep disorders [9], [23]. Radha et al. used the LSTM model to study sleep staging for healthy people and also yielded good results [24]. In this study, a novel method based on LSTM network was proposed for automatic sleep staging in home care monitoring for healthy people. Firstly, the HRV and RRV signals were extracted from only single-lead ECG signals. Then LSTM network was used for sleep staging on patients with mental disorders to respectively achieve 2-class, 3-class, 4-class and 5-class sleep staging task to meet the application needs on different occasions. The explored method in this paper has certain universality and can be easily transplanted to mobile devices to meet the sleep monitoring demands for more scenes like home care.

II. MATERIALS AND METHODS

A. DATA ACQUISITION

In order to verify the effectiveness of the method proposed in this paper, the PSG data is collected from the sleep disorders diagnosis center of Xijing Hospital, Fourth Military Medical University. The research was approved by the Ethics Committee of the First Affiliated Hospital of the Fourth Military Medical University. All the subjects were patients suffering either depression or schizophrenia without other mental disorders, and they were all given the informed consent before the experiments. Totally 1514 cases were collected. The sleep structures of those patients were much different from that of healthy subjects. The sleep stages of most patients missed REM or N3. Therefore, only valid cases were chosen for the research in this paper. The selecting criteria is that the patient must have complete sleep structures, containing all the five sleep stages (wake, N1, N2, N3, and REM), with no symptoms of sleep apnea. Therefore, 373 patients were chosen in the end. The demographic data and sleep parameters were presented in Table 1.

The PSG data were measured by SOLAR3000B neurocentral monitoring analysis system developed by Beijing Solar Electronic Technologies Company Ltd. The SOLAR3000B could synchronously record eight-lead EEG, three-lead ECG,

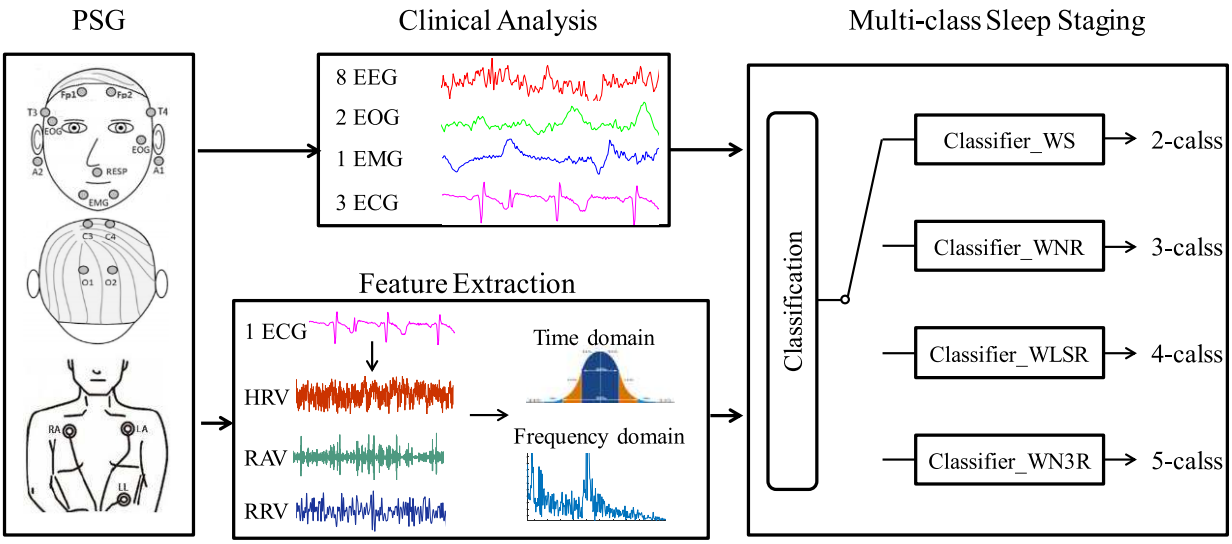


FIGURE 1. An overall description of the method proposed in this paper. The block diagram contains three parts: clinical analysis (The clinical technicians watched the signals from PSG to annotate the sleep stages), feature extraction (One single-lead ECG signals was used to extracted features for sleep staging) and multi-class classification sleep staging (Four different classifiers were constructed and trained independently to realize multi-class staging.).

TABLE 1. Demographic data and typical sleep parameters.

Demographic data	Age (years)	39±17
	Male	156
	Female	217
Sleep parameters	W	50384 (14%)
	N1	37755 (11%)
	N2	196070 (55%)
	N3	33245 (9%)
	REM	37860 (11%)
	Total	355314 (100%)

one-lead chin EMG, and two-lead EOG from the left and right eyes. The placement of eight-lead EEG signals were referenced to the international 10–20 system of electrode placement [25], including Fp1, Fp2, C3, C4, O1, O2, T3 and T4. Three-lead ECG included the lead-I (LA-RA), lead-II (RA-LL) and lead-III (LA-LL). The sampling rate of all signals was 100Hz [26]. The subjects were asked to sleep the whole night (approximately from 11. p.m. to 6:30 a.m.) as usual until they were awoken by the researchers in the morning. The standard sleep stages were determined by a sleep scoring technician according to the AASM rules.

B. METHODS

The proposed method used only one-lead ECG for automatic sleep staging. Firstly, HRV signals were calculated from de-noised ECG signals after pre-processing. Then HRV and respiratory amplitude variability (RAV) signals were extracted. Next, features were extracted from time and frequency domain. Finally, the multi-class sleep staging method was formed. Four different classifiers based on LSTM network were constructed, 2-class sleep staging for distinguishing W and sleep (Classifier_WS); 3-class sleep staging to distinguish between W, NREM and REM (Classifier_WNR); 4-class sleep staging to distinguish W, LS, SWS and REM (Classifer_WLSR); and 5-class sleep staging to distinguish between W, N1, N2, N3 and REM (Classifer_WN3R). An overall description of the method proposed in this paper is shown in Fig 1.

1) FEATURES EXTRACTION

An overview of feature extraction procedure is illustrated in Fig.2. Since it has been proved that body position has little impact on the measurement accuracy of respiratory signal estimation from ECG [27], any lead of ECG signals from PSG can be used to calculate HRV and RRV. In this paper, the lead-II ECG signals were chosen. A third-order bandpass Butterworth filter was used to filter out the valid components (0.5∼5Hz) from ECG signals. Then the peak point position of each R wave was identified using the maximum slope method [28]. The R-R interval was obtained from the difference between the adjacent R peak point positions and HRV signals were calculated from this R-R intervals signals using cubic spline interpolation. The process is used to convert the non-equidistantly sampled R-R interval time locations to an equidistantly sampled HRV signal that has amplitude equal to the R-R intervals signal at precisely that time location [29]. Finally, the HRV was down-sampled to 10 Hz by using a polyphase filter. Since the related frequency domain calculation used in this paper doesn't require signals with high frequency, 10 Hz is basically enough.

The original ECG signals contained the valid frequency components of respiratory information. Therefore, according to the theory of frequency modulation, the RAV was filtered by a third-order Butterworth bandpass filter with frequency band ranging from 0.15 Hz to 0.5 Hz from the original ECG signals. The RRV signal was calculated from the RAV signal with the same method as HRV extraction.

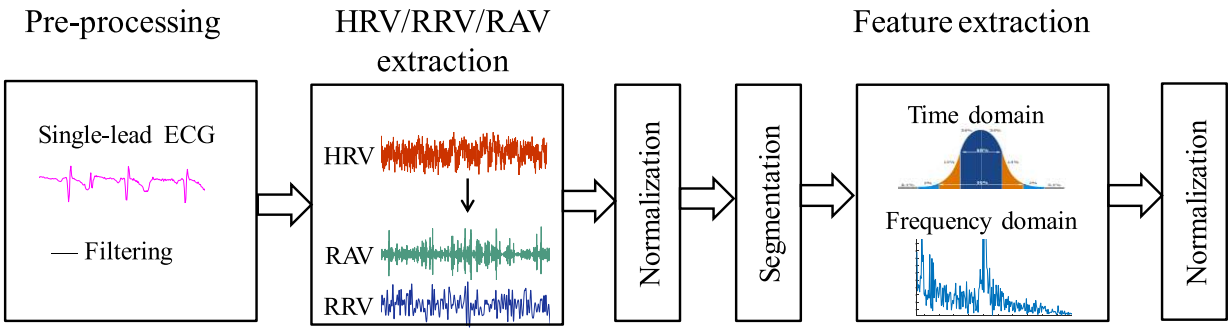


FIGURE 2. Overview of the feature extraction method proposed in this paper. One single-lead ECG signals were pre-processed by filtering. Then HRV, RAV and RRV were calculated. Features were extraction based on HRV, RAV and RRV after normalization and segmentation. Then all the extracted features were processed by normalization.

The time interval between adjacent respiratory peak positions was detected using the maximum slope, and then RRV was obtained from the respiratory time interval series with cubic spline interpolation. Finally, RRV was also down sampled to 10 Hz.

Before feature extraction, HRV, RAV and RRV were all normalized into a coordinate scale. The way to normalize the given signal is to center it at zero mean and scale it to unit standard deviation. Then HRV, RAV and RRV were divided into 30-second epochs synchronizing in time with PSG classification results. To make representative features in context, the feature extraction procedure is carried out based on the sliding window technique by 90% overlapping frames from signal streams. Totally 25 features were computed in a 5-minute window centered on each 30-second epoch with 2-min extension to the left side and 2.5-minute extension to the right side. Feature calculation from HRV was the same with that from RRV. The detailed procedures were described in Table 2. The feature calculation from RAV was a little different, which depended on the maximum value and minimum value between the positions of two adjacent respiratory waves. In Fig 3, partial RAV signals from one subject were plotted, on which the maximum points of the absolute value of the slope of the respiratory wave (R_{POS}) were marked. The maximum values between two adjacent R_{POS} were referred to as $R_M(n)$. The time positions corresponding to the maximum values were referred to as $M(n)$. The minimum values between two adjacent R_{POS} were referred to as $R_N(n)$. The time positions corresponding to the minimum values were referred to as $N(n)$. Another 9 features were extracted from RAV signals. The detailed procedures of feature calculation from RAV signals were described in Table 3

Features were extracted within each epoch and then concatenated to form the final features series of each subject. Because of the differences of physiological signals between individuals, all the features were normalized to make them center at zero mean and scale to unit standard deviation.

2) MULTI-CLASS SLEEP STAGING

Multi-class sleep staging method in this paper was based on LSTM network. The layers in LSTM network use memory cells that can store long-term information from times series. The output of one LSTM layer is based on the current time step input, their last output (long-term recurrence) and the internal cell state. The cell state is adjusted through gating mechanisms. Furthermore, one of the great advantages of LSTM is its ability to handle variable length sequences. In this study, different subjects have different sleep time which resulted in the obtained variable length feature sequences. Therefore, LSTM network is the most appropriate choice. However, the implementation of LSTM network in Keras framework need to fix the length of input sequences [30]. In order to overcome the conflicts, the masking strategy was adopted. The length of all the sequence was set to 1036 30-second frames (8 hours and 38 minutes) which is the maximum sleep length of all the subjects. The sequence less than 1036 was padded with zeros. When constructing

TABLE 2. Feature calculation from HRV and RRV.

Feature calculation	Description
$Mean(S(n))$	The mean value of the segment;
$Std(S(n))$	The variance of the segment;
$\frac{Std(S(n))}{Mean(S(n))}$	The mean of the segment divided by the variance;
$\sum_{f=0.01}^{0.04} P(S(n))$	The sum of the power spectra of low frequency band (0.01 Hz-0.04 Hz) of the segment;
$\sum_{f=0.04}^{0.15} P(S(n))$	The sum of the power spectra of medium frequency bands (0.04 Hz-0.15 Hz) of the segment;
$\sum_{f=0.15}^{0.4} P(S(n))$	The sum of the power spectra of high frequency band (0.15 Hz-0.4 Hz) of the segment;
$\sum_{f=0.01}^{0.4} P(S(n))$	The sum of the total power spectra of the segment;
$\sum_{f=0.15}^{0.4} P(S(n)) \Big/ \sum_{f=0.04}^{0.15} P(S(n))$	The ratio of the sum of the power spectra of the high frequency band (0.15 Hz-0.4 Hz) to the sum of the power spectra of the mid-band (0.04 Hz - 0.15 Hz) of the segment.

PS: $S(n)$ presents HRV signals or RRV signals. $Mean(\,)$ indicates calculating the mean values; $Std(\,)$ indicates calculating the standard deviation. $P(\,)$ indicates calculating the power spectrum.

TABLE 3. Feature calculation from RAV.

Feature calculation	Description
$\frac{Median(R_N(n))}{F_{3/4}(R_N(n))-F_{1/4}(R_N(n))}$	The median value of the partial minimum values divided by the absolute difference between its upper and lower quarter quantiles.
$\frac{Median(R_M(n))}{F_{3/4}(R_M(n))-F_{1/4}(R_M(n))}$	The median value of the partial maximum values divided by the absolute difference between its upper and lower quarter quantiles.
$\frac{Mean(R_M(n)-R_N(n))}{Std(R_M(n)-R_N(n))}$	The mean value of relative values (maximum value minus minimum) divided by its variance;
$Median(\sum_{i=M(n)}^{M(n+1)}\lvert RAV(i)-RAV(i+1)\rvert)$	The median value of the absolute difference of the maximum values.
$Median(\frac{\sum_{i=M(n)}^{M(n+1)}\lvert RAV(i)-RAV(i+1)\rvert}{M(n+1)-M(n)})$	The median value of the absolute difference of the maximum values divided by its time interval.
$Median(\sum_{i=N(n)}^{N(n+1)}\lvert RAV(i)-RAV(i+1)\rvert)$	The median value of the differential sequence of the minimum values.
$Median(\frac{\sum_{i=N(n)}^{N(n+1)}\lvert RAV(i)-RAV(i+1)\rvert}{N(n+1)-N(n)})$	The median value of the absolute difference of the minimum values divided by its time interval.
$Median(\sum_{i=N(n)}^{M(n)}\lvert RAV(i)-RAV(i+1)\rvert)$	The median value of the absolute difference between the maximim values and minimum values.
$Median(\frac{\sum_{i=N(n)}^{M(n)}\lvert RAV(i)-RAV(i+1)\rvert}{M(n)-N(n)})$	The median value of the absolute difference between the maximum values and minimum values divided by its time interval.

PS: *Median*( ) means calculating the median value of the sequence; $F_{3/4}(\,)$ and $F_{1/4}(\,)$ means calculating the quarter quartile. *Mean*( ) means calculating the mean value of the signal; *Std*( ) means calculating the standard deviation.

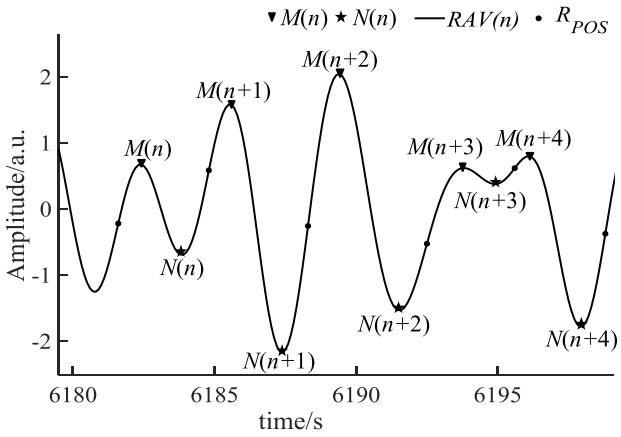


FIGURE 3. Partial RAV signals from one subject as an example. The position of R_{POS}, $M(n)$ and $N(n)$ were marked on RAV signals. Between two adjacent R_{POS}, only one maximum value and one minimum value were detected.

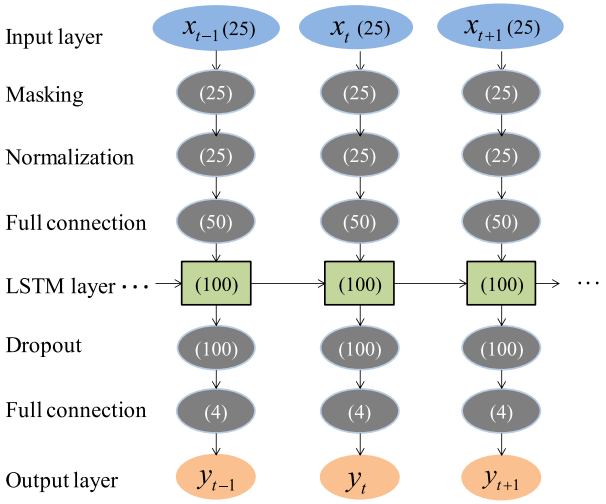


FIGURE 4. Detailed illustration of classifier structure. Each classifier contains eight layers. The numbers in brackets means the number of neurons of current layers.

classifiers, a masking layer must be added for right next to the input layer. As a result, the padding values in the sequence would all be filtered out.

To realize 2-class, 3-class, 4-class or 5-class sleep staging tasks, four classifiers were designed. They have nearly the same structures but different outputs. As an example, the detailed illustration of the network structure of 4-class classifier is shown in Fig.4. Each classifier contains eight different layers, which acts different functions.

a) *Input layer*: input the data of 25 features (x_t).

b) *Masking*: the input sequence is ''masked'' with the given value to locate the time step that needs to be skipped. The masking value is set to zeros. As a result, the following layer will skip the time step with the value of zero.

c) *Normalization*: act as a regularizer, making the mean value of data in each batch close to zero and its deviation close to one. Batch normalization will accelerate network training by reducing internal covariate shift.

d) *Full connection*: consist of 50 parallel perceptions. Since there is no special treatment for the original features after normalization, all the parameters are set as default except that the units are set as 50.

e) *LSTM layer*: use one layer of LSTM. At each time step, the memory cell takes in the outputs of full connection layers and generates 100 new features. Then, those 100 new features, the output of last time step and the internal state of current memory cell, are all considered to calculate a result, which contains temporal information from the past and the present.

f) *Dropout*: the dropout layer is used to randomly forget some previous information, which is an effective strategy to avoid over-fitting.

g) *Full connection*: the full connection technique is carried out to map the output of dropout layers to a lower

dimensional space whose size depends on the classification of sleep staging tasks. The activation function is softmax. For example, if a 4-class sleep staging task is realized, the units are set as four. As a result, full connection layer will generate four-dimensional outputs representing the class probabilities at current time step.

h) *Output layers*: give the final sleep stage of current time step, which is the class label with the maximum probability.

3) MODEL TRAINING

The classifiers were implemented with Keras framework based on Tensorflow backend [30] in python environment with the interpreter of python 3.6. The rate of dropout layer is set as 0.5. The loss function of all classifiers is cross-entropy. The optimizer used in the classifier is ''adam''. The training epoch is set as 40. Too large batch size will decrease the number of iterations of one epoch, which may cause under-fitting. While too small batch size may make the training process not converge or lead the final convergence accuracy to fall into the local extrema. There has been researches suggesting that the best performance has been consistently obtained for mini-batch sizes between 2 and 32 for most deep learning problems [31]. Therefore, in this paper, the batch size is set to 5. Considering that the size of the data set used in this paper is not large enough, a relatively small batch size will not cause much calculation burden. Once the classifiers were trained well, they can be used for testing and predicting.

In this paper. The ratio of testing set is 0.2. Among the rest data, the ratio of training set to validation set is also 8:2. So the training set size was 238 subjects. The size of validation set was 60 subjects, and the size of testing set was 75 subjects. All the subjects from the validation and testing set were chosen randomly.

4) EVALUATION OF SLEEP STAGING AND SLEEP EQUALITY

To evaluate the sleep staging method proposed by this paper, sleep staging accuracy and Cohen's kappa statistic [32], [33] were used. Sleep staging accuracy p_o is calculated by the sum of correctly classified samples for each class divided by the total number of samples. Cohen's kappa statistic k is calculated as:

$$k = (p_o - p_e)/(1 - p_e) \tag{1}$$

where $p_e = (t_1 \times p_1 + t_2 \times p_2 + \ldots + t_n \times p_n) / (N \times N)$, and t is the number of true samples of each class, p is the number of predicted number of each class, n is the number of total classes, N is the total number of samples. Cohen suggested the kappa statistic k be interpreted as follows: values $\leq$ 0 as indicating no agreement between the proposed method and the standard criteria and 0.01$\sim$ 0.20 as none to slight, 0.21$\sim$ 0.40 as fair, 0.41$\sim$ 0.60 as moderate, 0.61$\sim$ 0.80 as substantial, and 0.81$\sim$ 1.00 as almost perfect agreement [34].

After the sleep staging was performed, the sleep equality was evaluated in the end. Therefore, the following averaged nightly summary sleep measures were also calculated for each classifier system [35]: total sleep time (TST), sleep onset latency to the first epoch of sleep (SOL), latency to persistent sleep of 10 continuous min (LPS), wakefulness after sleep onset (WASO), sleep efficiency (SE), the number of awakenings lasting at least 2 min (NA) and wakefulness time during sleep (WTDS) between the first and last sleep epoch since the recordings, time in REM sleep (TREM), time in light sleep (TLS), time in slow wave sleep (TSWS) and latency from the onset of the first epoch of sleep to the onset of the first epoch of REM sleep (REML). Statistical differences of sleep staging between the PSG and the proposed method were analyzed with SPSS.

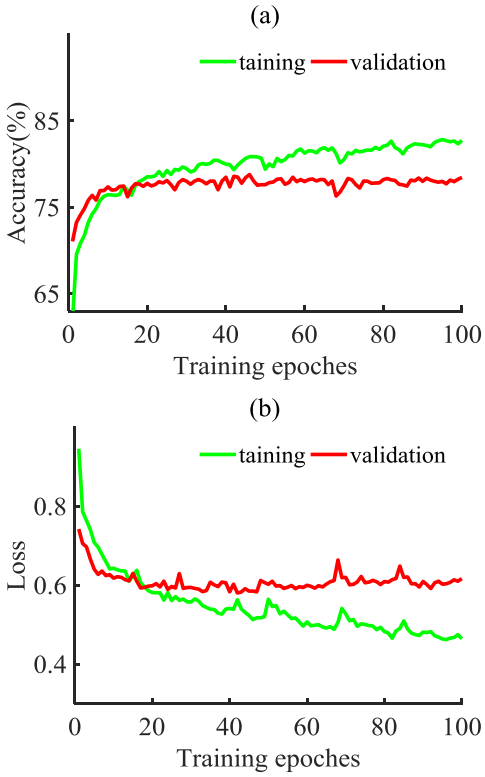


FIGURE 5. The training process of proposed method in this paper. (a) The change of accuracy with training epochs. (b) The change of loss with the training epochs.

III. RESULT

Taking the training process of 4-class sleep staging in dataset 1 as an example, the change process of accuracy and loss with training set and validation set were shown in Fig. 5. When the training epoch was set to 100, the training accuracy slowly increased with the increase of training epochs. At the same time the loss of training set slowly decreased. However, the loss of training set decreased while the loss of validation set stayed plateau once the epoch exceeded 40. This indicated that the model was fully trained and approached the tipping point of over-fitting. The same trend was also found in 2-class, 3-class and 5-class sleep staging. Thus it is suitable for using 40 epochs to train each sleep staging classifier.

As for the testing set, the confusion matrixes of 2-class, 3-class, 4-class and 5-class sleep staging were shown

TABLE 4. Confusion matrix of multi-class sleep staging.

2-class				
		Predicted result by proposed method		
		W	*Sleep*	*Total*
Clinical analysis result	*W*	5157	5366	10523
	Sleep	2128	58548	60676
	Total	7285	63914	71199
Accuracy:89.48%, Cohen's kappa statistic *k*:0.52				

3-class					
		Predicted result by proposed method			
		W	*NREM*	*REM*	*Total*
Clinical analysis result	*W*	5487	4443	593	10523
	NREM	1574	49729	1823	53126
	REM	385	2521	4644	7550
	Total	7446	56693	7060	71199
Accuracy:84.07%, Cohen's kappa statistic *k*:0.58					

4-class						
		Predicted result by proposed method				
		W	*LS*	*SWS*	*REM*	*Total*
Clinical analysis result	*W*	5880	3693	131	819	10523
	LS	1450	41728	1642	2217	47037
	SWS	53	2784	3229	23	6089
	REM	329	2774	0	4447	7550
	Total	7712	50979	5002	7506	71199
Accuracy:77.65%, Cohen's kappa statistic *k*:0.55						

5-class							
		Predicted result by proposed method					
		W	*N1*	*N2*	*N3*	*REM*	*Total*
Clinical analysis result	*W*	6936	552	2373	231	452	10523
	N1	1067	908	4141	53	1199	7368
	N2	1099	748	34622	1984	1216	39669
	N3	26	0	2659	3392	12	6089
	REM	759	439	1542	4	4806	7550
	Total	9887	2647	45337	5643	7685	71199
Accuracy:71.16%, Cohen's kappa statistic *k*:0.52							

in Table 4. The sleep staging accuracy and Cohen's kappa statistic k of each classification task were also listed in Table 2. From the testing results, 2-class sleep staging task achieved the best result among all sleep staging tasks. The accuracy is 89.48%, and Cohen's kappa statistic k is 0.52. For 3-class sleep staging problem, the accuracy is 84.07%, Cohen's kappa statistic k is 0.58, which indicates moderate agreement with the results of clinical analysis. The reason why sleep staging of 3-class achieves higher Cohen's kappa statistic k than 2-class is that the number of correctly classified W increased. The classification accuracy and Cohen's kappa statistic k of 4-class sleep staging was 77.65% and 0.55, which is able to meet the precision requirements of most sleep monitoring application. According to Table 4, it is very difficult to distinguish the sleep stage of N1 correctly, which is the main reason why the accuracy of 5-class sleep staging is relatively low. But the Cohen's kappa statistic k of 5-class sleep staging is 0.52, which is much higher than most of the current research.

Because 5-class sleep staging is recommended by the AASM criteria, this sleep staging problem is very important in intensive care unit or home care monitoring. Fig.6 showed the comparison results of predicted sleep stages by the proposed method with clinical analysis of one subject in the testing set. The shaded part marked the wrong classified frames, which often occurred between N2 and N3. But the accuracy of this subject is 84.78%, and Cohen's kappa statistic k is 0.76, which indicates substantial consistency with the clinical analysis.

TABLE 5. Summary sleep measures of PSG and LSTM network.

Summary sleep measures	PSG	proposed method	*P* value
TST(min)	404.51±58.33	408.75±53.58	0.85
SE(%)	0.84±0.12	0.85±0.11	0.85
SOL(min)	25.99±28.33	23.86±24.06	0.86
LPS(min)	474.85±51.95	467.90±49.89	0.59
WASO(min)	70.15±57.26	65.91±51.27	0.92
NA(#)	3.11±1.36	3.04±1.28	0.81
WTDS(min)	12.83±24.36	14.03±26.63	0.97
TREM(min)	50.33±32.86	51.23±30.82	0.68
TLS(min)	313.58±54.56	319.89±42.97	0.57
TSWS(min)	40.59±25.88	37.62±15.63	0.77
REML(min)	189.89±112.89	174.67±96.28	0.49

In order to evaluate the differences of the PSG and the proposed method in sleep quality assessment. Totally 11 summary sleep measures of the nightly averaged sleep stage and sleep/wakefulness measures were calculated for the 75 test subjects. Table 5 shows the results of the mean value and standard derivation of each measure. Analyzed by SPSS, all the summary sleep measures are not normal distributed, so the statistical differences between the PSG and the proposed method were tested by ANOVA. Individual pairs were compared by Mann-Whitney U significant difference. In this paper, the significance level was set at $P < 0.05$. As shown from Table 5, there were no significant differences between the PSG and the proposed method for all summary sleep measures. As the proposed method can almost realize the same results with PSG for the proposed 11 summary sleep measures, it is suggested that the proposed method is effective for assessing sleep quality.

IV. DISCUSSION

A. THE EFFECT OF DATASET

In this paper, only subjects with completed sleep structures were included in the training and testing of the model. However, for the original dataset obtained in this paper, missing sleep stages happened to nearly 75% of the subjects. To investigate a more practical sleep staging model, the whole dataset were fully used. Except for 73 cases (68 patients had symptoms of sleep apnea and lead-off problems happened to another 5 patients), totally 1441 patients (609 males, mean age 38±16 years) were involved. The rate of training set to testing set was also 8:2. As a result, the size of training set, validation set and testing set were 922, 231, 288 respectively. All the testing and validation set were chosen randomly. The model construction and training process were the same as the

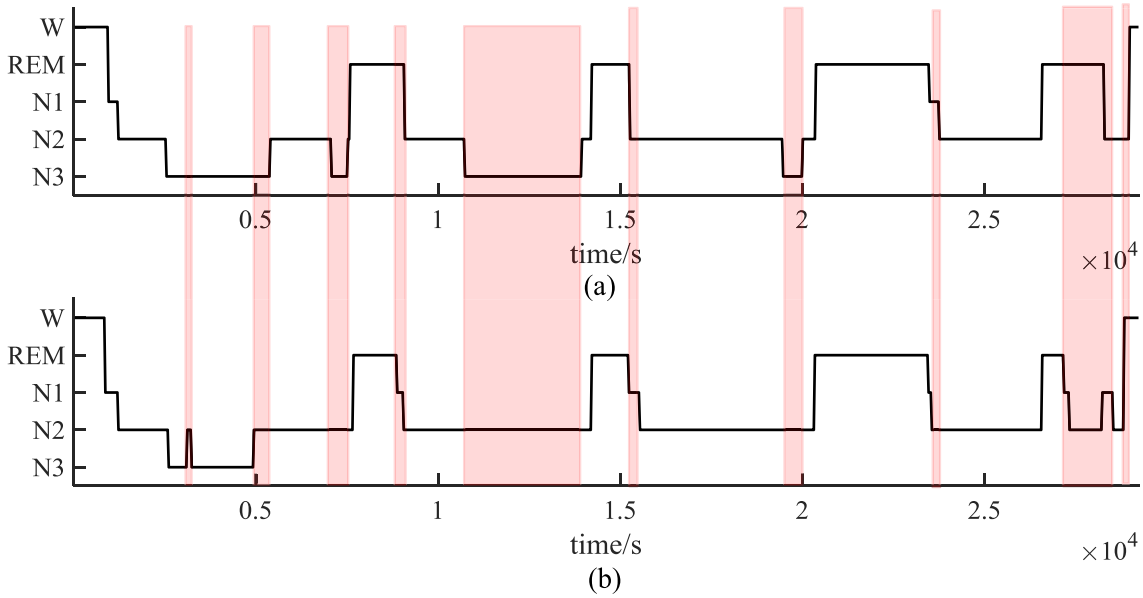


FIGURE 6. The comparison of predicted sleep stage by LSTM network with clinical technician of one subject from the test data. (a) Sleep stages by clinical analysis. (b) Sleep stages by LSTM network. The shaded parts indicate the wrong prediction of the model.

dataset with full sleep stages. The testing results were shown in Table 6.

Comparing the results in Table 6 with the results in Table 4, conclusion can be easily made that too much patients with missing sleep stages reduced the performance of LSTM network. In general, the accuracy of the 2-class, 3-class, 4-class and 5-class sleep staging has no obvious decrease, but the Cohen's kappa statistic k decreased a lot especially for 4-class and 5-class sleep staging. As for 5-class sleep staging, the Cohen's kappa statistic k is reduced by 17% from 0.52 to 0.43. Therefore, the dataset has relatively obvious effect on the performance of sleep staging model. Since too much samples with missing sleep stages would lead to great imbalance of the rate of different sleep stages, which may mislead the learning of LSTM network. Among all the 1441 valid patients, the sleep stage of N3 just covers around 2.71% of all the sleep stages. Most of the N3 stages were incorrectly recognized as N2, which resulted in the reduction of performance for 4-class and 5-class sleep staging. In both dataset, the number of W stages wrongly predicted as REM was much smaller than predicted as N2. The number of REM stages wrongly predicted as N2 was larger than predicted as W. As for the dataset with full sleep stages, the sensitivity of W and REM were both above 60%. Therefore, the proposed method had good capability to distinguish W and REM. Generally speaking, the features of HRV in W and REM stages were quite similar, which making it difficult to distinguish REM from W for most traditional methods. But LSTM network had its own advantages. The probability of transition between two sleep stages was quite different. For example, the transition probability from N3 to REM was larger than that from N3 to W. Such rules could be more easily learned by LSTM network rather than most traditional methods.

The aim of the study is for assessing the sleep quality of healthy people. Although missing sleep stages happens to healthy subjects in real life, it just happens occasionally. According to [36], [37] approximately 20% total sleep time is N3 and 25% is REM in adults and it may change with age. Instead, as for patients with depression or schizophrenia, their EEG signals are different from that of healthy people and most of their sleep time during the night is missing REM or N3 [38]–[41], which makes tremendous imbalance to the training data set. Therefore, the proposed method is recommended to apply to sleep monitoring for people with complete sleep structure. As for patients with serious sleep disorder, the proposed method may be incompetent.

B. THE EFFECT OF LSTM LAYER DEPTH

The method proposed in this paper just used one LSTM layer. To valid the impact of LSTM layer depth on sleep staging, the LSTM network was reconstructed using 2∼5 LSTM layers. The training and testing procedures of the new LSTM networks are the same as the proposed method. The sleep staging results from 1 to 5 LSTM layers were shown in Fig.7. The testing results indicate that the sleep staging accuracy has no evident improvement as the increase of LSTM layers. The Cohen's kappa statistic k seemed to increase a little when the LSTM layer rose to 3, but there was no more substantial growth when LSTM layer was larger than 3. But the Cohen's kappa statistic k of 2-class sleep staging increased a lot as LSTM layer increased from 1 to 2, but the growth tendency seemed to stop for the later layers. Eventually, the Cohen's kappa statistic k of 2-class and 3-class sleep staging simply oscillate around the value of 0.58. In recent study, the effectiveness of the layer depth has been proved in the model of convolutional neural networks for sleep staging [42], but

TABLE 6. Confusion matrix of multi-class sleep staging for the patients with missing sleep stages.

2-class

		Predicted result by proposed method		
		W	*Sleep*	*Total*
Clinical analysis result	*W*	23794	24050	47844
	Sleep	10418	215524	225942
	Total	34212	239574	273786

Accuracy:87.41%, Cohen's kappa statistic *k*:0.51

3-class

		Predicted result by proposed method			
		W	*NREM*	*REM*	*Total*
Clinical analysis result	*W*	28325	18533	986	47844
	NREM	13588	189631	3686	206905
	REM	1746	9496	7795	19037
	Total	43695	217660	12467	273786

Accuracy:82.46%, Cohen's kappa statistic *k*:0.52

4-class

		Predicted result by proposed method				
		W	*LS*	*SWS*	*REM*	*Total*
Clinical analysis result	*W*	23641	22735	0	1468	47844
	LS	7031	186665	1007	4775	199478
	SWS	4	6284	1115	24	7427
	REM	1376	9338	0	8323	19037
	Total	32052	225022	2122	14590	273786

Accuracy:80.26%, Cohen's kappa statistic *k*:0.48

5-class

		Predicted result by proposed method					
		W	*N1*	*N2*	*N3*	*REM*	*Total*
Clinical analysis result	*W*	27858	2041	16615	0	1330	47844
	N1	6034	3483	21920	0	2824	34261
	N2	6515	2865	153843	27	1967	165217
	N3	21	0	7378	28	0	7427
	REM	2301	2090	5995	0	8651	19037
	Total	42729	10479	205751	55	14772	273786

Accuracy:70.81%, Cohen's kappa statistic *k*:0.43

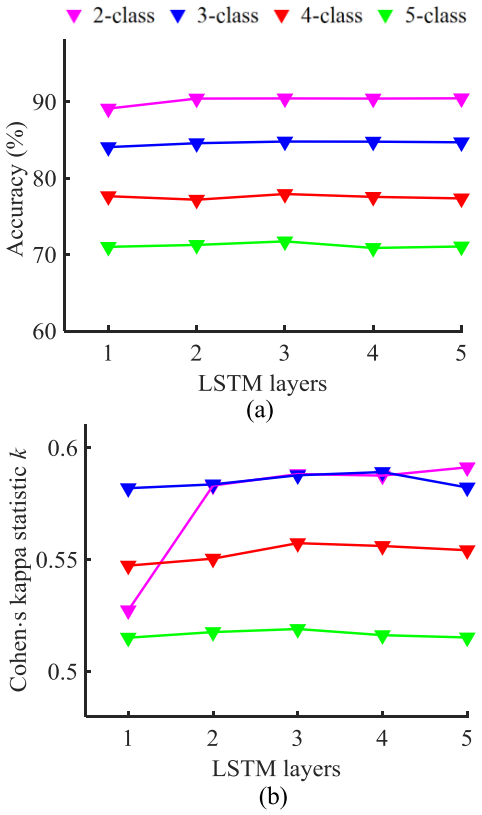


FIGURE 7. The performance of different LSTM layers (a) The change of accuracy with the depth of LSTM layers. (b) The change of Cohen's kappa statistic *k* with the depth of LSTM layers.

there is no previously report in LSTM networks. Through gating mechanisms, LSTM network can selectively store long-term information from time series, which could fully use the transition probability information between different sleep stages. Thus LSTM networks can achieve good performance in sleep staging problems. In fact, the situations in which performance improves with a second (or third, etc.) hidden LSTM layer are very few [43]. One hidden LSTM layer is sufficient for the large majority of sleep staging problems. However, sleep staging is a complex physiological process, which has various characteristics in different signals. In this paper, it has been proved that the depth of LSTM layers has no evident impacts on the results of sleep staging when using single-lead ECG signals. If multi-channel physiological signals were involved, things may be different. The impact of LSTM layer depth on sleep staging using various physiological signals should be further investigated.

C. THE IMPACT OF DROPOUT RATE ON OVER-FITTING

To avoid the problem of over-fitting, the dropout schema is usually recommended in the classification problem. The lower dropout rate would cause over-fitting and larger dropout rate may cause under-fitting. The dropout rate is usually set as 0.3 or 0.5. However, there is no report about the specific influence of dropout rate on LSTM network. Hence, the impact of dropout rate on sleep staging was investigated while the dropout rate was set from 0 to 0.9 with a step of 0.1. All the training and testing strategies are the same as those used by the proposed method. The sleep staging results of different dropout rates are shown in Fig. 8.

It can be observed from Fig. 8 that the change of Cohen's kappa statistic k was more sensitive to the change of dropout rate than accuracy. When the dropout rate increased from 0 to 0.9, the accuracy of all the four classifiers had little fluctuation. But the Cohen's kappa statistic k firstly increased and then significantly decreased once the dropout rate was larger than 0.7. When the dropout rate was set as 0, the LSTM cells would remember all the information from the former training process with no dropout. Too much redundant information would inevitably cause over-fitting. That's why the dropout rate at 0 got bad performance. By an overall view of Fig. 7, the dropout rates lower than 0.3 or larger than 0.7 would cause worse sleep staging results, while the dropout rates between 0.3 and 0.7 could seem to obtain good performance. As a result, the dropout rates between 0.3∼0.7 are suitable for the sleep staging using single-lead ECG signals based on the proposed method.

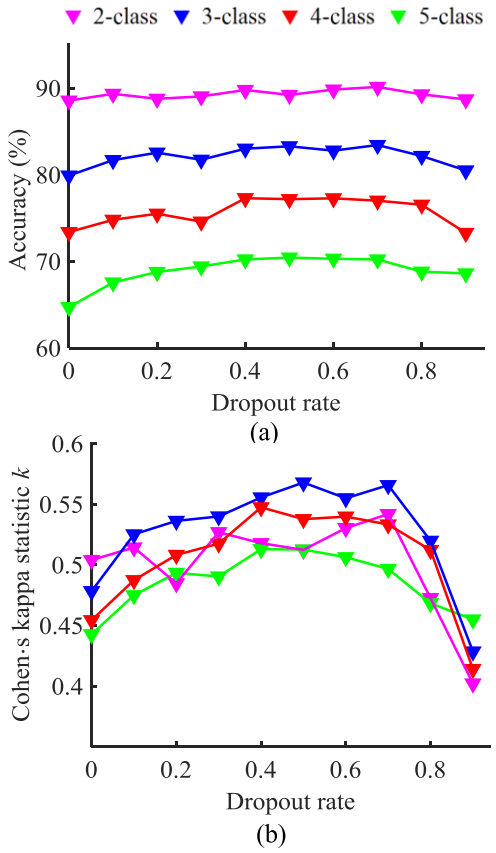


FIGURE 8. The change of accuracy (a) and Cohen's kappa statistic k (b) with different dropout rates.

D. THE VALIDATION OF PROPOSED METHOD

To validate the effectiveness for healthy people of proposed method, the public dataset Sleep Heart Rate and Stroke Volume Data Bank (SHRSV) [44] was used as prediction dataset in this study. The LSTM network was firstly trained by the dataset used in this study. Then, the sleep stages of 45 healthy subjects from SHRSV were predicted by the trained LSTM network. The accuracy and Cohen's kappa statistic k of 3-class were listed in Table 7.

Using single-lead ECG signals has already been proved valid for sleep staging recently [17]. Many conventional sleep staging methods based on HRV, including linear discriminant, random forest (RF), support vector machine (SVM), naive Bayes (NB), sleep stage transition (SST) model and time-dependent sleep stage transition (TSST) model, have been applied before [2], [17], [45], [46]. However, there were two main disadvantages of those conventional methods. On the one hand, the low accuracy of those models was unable to meet the clinical needs. On the other hand, the feature extraction process of those methods was somehow complicated. In the recent research [2], by using single-lead ECG signals, the RF method achieved the sleep staging accuracy of 72.58%. But totally 41 features were extracted when identifying the classification of W, NREM and REM. The results

TABLE 7. The comparison OF LSTM network with conventional method.

Method	Accuracy	Cohen's kappa statistic k
RF	72.58%	0.46
NB	67.6%	0.33
SVM	72.1%	0.24
SST	74.2%	0.36
TSST	76.0%	0.42
LSTM NETWORK	79.28%	0.50

of 3-class sleep staging by RF, NB, SVM, SST and TSST with the same dataset SHRSV were also shown in Table 7.

From Table g, it can be observed that the performance of proposed method is much higher than the conventional methods both in accuracy and the Cohen's kappa statistic k. The dataset for training LSTM network model were collected from hospitalized patients who suffered from either depression or schizophrenia. Although the sleep structures of those patients were different from healthy subjects, the proposed method can still achieve high accuracy and Cohen's kappa statistic k for healthy subjects. This demonstrated that the proposed method had wider applicability and broader effectiveness than conventional methods and it can be used for healthy people. If the training data could contain a certain number of healthy subjects, the sleep staging results would be much better.

E. ADVANTAGES AND LIMITATIONS

In this paper, the features extracted from single channel ECG signals and the LSTM network were proposed for automatic sleep staging, including 2-class, 3-class, 4-class, and 5-class sleep staging. Especially for 5-class sleep staging, the sleep staging accuracy exceeded 70%, and the Cohen's kappa statistic k was 0.52. Although there was a little difference with the gold standard of PSG, these results had achieved much improvement than those of other methods using single-lead ECG signals. The advantage of the sleep staging method explored in this paper is that this approach can be easily transplanted to portable mobile devices or other monitoring devices. Because only single-lead of ECG signals was used as the input of sleep staging, these input signals could be easily obtained by many medical devices. Once the ECG signals were used as the input of proposed method, detailed sleep stages could be obtained conveniently. In family monitoring, the proposed method can greatly improve the fineness and accuracy of sleep staging compared with the existing monitoring methods such as wrist sensors. Since many wrist sensors using actigraph [39] would only distinguish between sleep and wake, the sleep staging of bracelets is invalid when classifying SWS and LS. Therefore, the proposed method is considered as an effective approach of sleep staging in family health care.

On the other hand, the proposed method in this paper also has certain limitations. Firstly, the used dataset was collected from hospitalized patients. They were suffering from mental

illness like depression or schizophrenia. The sleep structures of those patients were different from healthy subjects, mainly in the decrease of N3 and REM sleep, and the increase of N2 sleep [38], [47]. If dataset from healthy subjects was added when training the model of LSTM network, the sleep staging results would be much better. Secondly, although the dataset of 373 subjects used in this paper was relatively larger than the previous investigations, it is still too small for deep learning. A large dataset is expected to achieve better performance. What's more, the results of full dataset containing 1441 patients were not good, which was mainly caused by the imbalance of the data. So the new model for patients with incomplete sleep structure should be explored in the future. In order to better fit the full dataset, a slightly simplified model based on LSTM network should be redesigned. Last but not least, the subjects involved in this paper might be receiving medication. Medicine for treating mental illness usually promotes sleep by stimulating the nerve activity, which also has certain impact on heart rate. Therefore, the heart rate features used in this paper may be different from that of healthy subjects. The method proposed in this paper may not be fully applicable to healthy subject for sleep staging. In the future, when the method is applied to healthy subjects, it needs to be re-evaluated more rigorously. Another improvement direction is to measure ECG signals from front head to involve certain information about eye movements. Thus, the detection of REM may be more precise.

V. CONCLUSION

A multi-class automatic sleep staging method based on single-lead ECG signals was proposed in this paper. This method can achieve different staging tasks of 2-class, 3-class, 4-class and 5-class sleep staging, and has a high consistency with the clinical results of PSG. At the same time, the proposed method, which was trained by the dataset based on mental illness patients, can also achieve quite good performance on ECG data of healthy subjects. So the method proposed in this paper has certain universality and stability, which can be used for sleep monitoring in care units, family monitoring, and mobile medical treatment.

REFERENCES

[1] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, and B. Vaughn, "The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications," Amer. Acad. Sleep Med., Darien, IL, USA, 2012.

[2] M. Xiao, "Sleep stages classification based on heart rate variability and random fores," *Biomed. Signal Process. Control*, vol. 8, no. 6, pp. 624–633, Nov. 2013.

[3] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, J., "Sleep stage classification with ECG and respiratory effort," *Physiol. Meas.*, vol. 36, no. 10, p. 2027, 2015.

[4] Z. Cao and C.-T. Lin, "Inherent fuzzy entropy for the improvement of EEG complexity evaluation," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 1032–1035, Apr. 2018.

[5] E. Alickovic and A. Subasi, "Ensemble SVM method for automatic sleep stage classification," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 6, pp. 1258–1265, Jun. 2018.

[6] P. Memar and F. Faradji, "A novel multi-class EEG-based sleep stage classification system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 1, pp. 84–95, Jan. 2018.

[7] S. I. Dimitriadis, C. Salis, and D. Linden, "A novel, fast and efficient single-sensor automatic sleep-stage classification based on complementary cross-frequency coupling estimates," *Clin. Neurophysiol.*, vol. 129, no. 4, pp. 815–828, 2018.

[8] G. Wang, C. Teng, K. Li, Z. Zhang, and X. Yan, "The removal of EOG artifacts from EEG signals using independent component analysis and multivariate empirical mode decomposition," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 5, pp. 1301–1308, Sep. 2016.

[9] I. N. Yulita, M. I. Fanany, and A. M. Arymuthy, "Bi-directional long short-term memory using quantized data of deep belief networks for sleep stage classification," *Procedia Comput. Sci.*, vol. 116, pp. 530–538, 2017.

[10] H. Wenhan, L. Anishchenko, L. Korostovtseva, B. J. Kooij, M. Bochkare, and Y. Sviryaev, "The study of sleep stage classification based on ECG and respiratory signal," *Intell. Comput. Appl.*, vol. 1, pp. 49–54, Jul. 2018.

[11] G. Jean-Louis, D. F. Kripke, R. J. Cole, J. D. Assmus, and R. D. Langer, "Sleep detection with an accelerometer actigraph: comparisons with polysomnography," *Physiol. Behav.*, vol. 72, nos. 1–2, pp. 21–28, Feb. 2001.

[12] C. T. Lin, C.-H. Chuang, Z. Cao, A. K. Sing, C.-S. Hung, Z. Cao, A. K. Sing, C.-S. Hung, Y.-H. Yu, M. Nascimben, Y.-T. Liu, J.-T. King, T.-P. Su, and S.-J. Wang, "Forehead EEG in support of future feasible personal healthcare solutions: Sleep management, headache prevention, and depression treatment," *IEEE Access*, vol. 5, pp. 10612–10621, 2017.

[13] D. Wang, R. Ren, K. Li, Y. Feng, D. Ma, X. Yan, and G. Wang, "Epileptic seizure detection in long-term EEG recordings by using wavelet-based directed transfer function," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 11, pp. 2591–2599, Nov. 2018.

[14] G. Wang, D. Ren, K. Li, D. Wong, M. Wang, and X. Yan, "EEG-based detection of epileptic seizures through the use of a directed transfer function method," *IEEE Access*, vol. 6, pp. 47189–47198, 2018.

[15] S. Liu, J. Teng, X. Qi, S. Wei, and C. Liu, "Comparison between heart rate variability and pulse rate variability during different sleep stages for sleep apnea patients," *Technol. Health Care*, vol. 25, no. 3, pp. 435–445, 2017.

[16] G. Gutierrez, J. Williams, G. A. Alrehaili, A. McLean, R. Pirouz, R. Amdur, V. Jain, J. Ahari, A. Bawa, and S. Kimbro, "Respiratory rate variability in sleeping adults without obstructive sleep apnea," *Physiol. Rep.*, vol. 4, no. 17, 2016, Art. no. e12949.

[17] . Yücelba , C. Yücelba , G. Tezal, S. Öz en, and . Yosunkaya, "Automatic sleep staging based on SVD, VMD, HHT and morphological features of single-lead ECG signal," *Expert Syst. Appl.*, vol. 102, pp. 193–206, Jul. 2018.

[18] E. Yuda, Y. Yoshida, R. Sasanabe, H. Tanaka, T. Shiom, and J. Hayano, "Sleep stage classification by a combination of actigraphic and heart rate signals," *J. Low Power Electron. Appl.*. vol. 7, no. 4, p. 28, 2017.

[19] H. Sharma, K. K. Sharma, and O. L. Bhagat, "Respiratory rate extraction from single-lead ECG using homomorphic filtering," *Comput. Biol. Med.*, vol. 59, pp. 80–86, Apr. 2015.

[20] F. A. Gers, E. Douglas, and J. Schmidhuber, "Applying LSTM to time series predictable through time-window approaches," in *Proc. Int. Conf. Artif. Neural Netw.*, 2001, pp. 669–676.

[21] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, Sep. 2000.

[22] B. Bakker, "Reinforcement learning by backpropagation through an LSTM model/critic," in *Proc. IEEE Int. Symp. Approx. Dyn. Program. Reinforcement Learn.*, Apr. 2007, pp. 127–134.

[23] G. Wang, Z. Sun, R. Tao, K. Li, G. Bao, and X. Yan, "Epileptic seizure detection based on partial directed coherence analysis," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 3, pp. 873–879, May 2016.

[24] M. Radha, P. Fonseca, M. Ross, A. Cerny, P. Anderer, and R. M. Aarts, "LSTM knowledge transfer for HRV-based sleep staging," 2018, *arXiv:1809.06221*. [Online]. Available: https://arxiv.org/abs/1809.06221

[25] Z. Cao, C.-H. Chuang, J.-K. King, and C.-T. Lin, "Multi-channel EEG recordings during a sustained-attention driving task," *Sci. data*, vol. 6, Apr. 2019, Art. no. 19.

[26] G. Wang and D. Ren, "Effect of brain-to-skull conductivity ratio on EEG source localization accuracy," *BioMed Res. Int.*, vol. 2013, Mar. 2013, Art. no. 459346.

[27] C. O'Brien and C. Heneghan, "A comparison of algorithms for estimation of a respiratory signal from the surface electrocardiogram," *Comput. Biol. Med.*, vol. 37, no. 3, pp. 305–314, Mar. 2007.

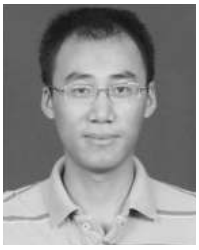
[28] R. A. Álvarez, A. J. M. Penín, and X. A. V. Sobrino, "A comparison of three QRS detection algorithms over a public database," *Procedia Technol.*, vol. 9, pp. 1159–1165, Jul. 2013.
[29] S. Chatlapalli, H. Nazeran, V. Melarkod, R. Krishnam, E. Estrada, Y. Pamula, and S. Cabrera, "Accurate derivation of heart rate variability signal for detection of sleep disordered breathing in children," in *Proc. 26th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 1, Apr. 2004, pp. 538–541.
[30] L. Z. Bigmoyan and Z. E. Zhourunlai. (2018). *Keras Development Document*. https://keras-cn.readthedocs.io/en/latest/
[31] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," 2018, *arXiv:1804.07612*. [Online]. Available: https://arxiv.org/abs/1804.07612
[32] A. B. Cantor, "Sample-size calculations for Cohen's kappa," *Psychol. Methods*, vol. 1, no. 2, pp. 150–153, 1996.
[33] M. L. Mchugh, "Interrater reliability: The kappa statistic," *BiochemiaMedica*, vol. 22, no. 3, pp. 276–282, 2012.
[34] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, Aug. 1977.
[35] J. R. Shambroom, S. E. Fábregas, and J. Johnstone, "Validation of an automated wireless system to monitor sleep in healthy adults," *J.Sleep Res.*, vol. 21, no. 2, pp. 30–221, 2012.
[36] D. Shrivastava, S. Jung, M. Saadat, R. Saadat, and K. Crewson, "How to interpret the results of a sleep study," *J. Community Hospital Internal Med. Perspect.*, vol. 4, no. 5, 2014, Art. no. 24983.
[37] J. R. D. Espiritu, "Aging-related sleep changes," *Clinics Geriatric Med.*, vol. 24, no. 1, pp. 1–14, Feb. 2008.
[38] M. Thase, A. S. Fasiczka, and A. Simons, "Electroencephalographic sleep profiles before and after cognitive behavior therapy of depression," *Arch. Gen. Psychiatry*, vol. 55, no. 2, pp. 138–144, Feb. 1998.
[39] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. P. Pollak, "The role of actigraphy in the study of sleep and circadian rhythms," *Sleep*, vol. 26, no. 3, pp. 342–392, 2003.
[40] Z. Cao, W. Ding, Y.-K. Wang, F. K. Hussain, A. Al-Jumaily, and C.-T. Lin, "Effects of repetitive SSVEPs on EEG complexity using multiscale inherent fuzzy entropy," *Neurocomputing*, to be published. doi: 10.1016/j.neucom.2018.08.091.
[41] Z. Cao, C.-T. Lin, K.-L. Lai, L.-W. Ko, J.-T. King, K.-K. Liao, Fuh, and S.-J. Wang, "Extraction of SSVEPs-based inherent fuzzy entropy using a wearable headband EEG in migraine patients," *IEEE Trans. Fuzzy Syst.*, to be published.
[42] S. Paisarnsrisomsuk, M. Sokolovsky, F. Guerrero, C. Ruiz, and S. A. Alvarez, "Deep sleep: Convolutional neural networks for predictive modeling of human sleep time-signals," in *Proc. KDD Deep Learn. Day*, London, U.K., Aug. 2018.
[43] K. Yao, "Depth-gated recurrent neural networks," 2015, *arXiv:1508.03790*. [Online]. Available: https://arxiv.org/abs/1508.03790
[44] *Institute of Psychophysiology & Rehabilitation, Kaunas University of Medicine*. Accessed: Oct. 12, 2011. [Online]. Available: http://www.pri.kmu.lt/datbank/index.php
[45] F. Ebrahimi, S. K. Setarehdan, J. Ayala-Moyeda, and H. Nazeran, "Automatic sleep staging using empirical mode decomposition, discrete wavelet transform, time-domain, and nonlinear dynamics features of heart rate variability signals," *Comput. Methods Programs Biomed.*, vol. 112, no. 1, pp. 47–57, 2013.
[46] T. Takeda, O. Mizuno, and T. Tanaka, "Time-dependent sleep stage transition model based on heart rate variability," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2015, pp. 2343–2346.
[47] R. M. Benca, "Sleep and psychiatric disorders: A meta-analysis," *Arch. Gen. Psychiatry*, vol. 49, no. 8, p. 651, 1992.

YUHUI WEI received the B.S. degree in biomedical engineering from Xi'an Jiao Tong University, Xi'an, China, in 2016, where she is currently pursuing the M.S. degree in biomedical engineering. Her research interests include medical signal processing, sleep staging analysis, and machine learning.

XIA QI received the B.S. degree in biomedical engineering from Xi'an Jiao Tong University, Xi'an, China, in 2017, where she is currently pursuing the M.S. degree in biomedical engineering. Her research interests include health care, sleep staging analysis and application, and deep learning.

HUANING WANG received the Ph.D. degree in psychiatry from Fourth Military Medical University, China, in 2009, where he is currently an Associate Chief Physician and an Associate Professor of psychiatry with the Xijing Hospital. He has authored more than 20 international journal articles. His main research interest includes physiotherapy for mental illness, such as PTSD, depression, and bipolar disorders.

ZHIAN LIU received the B.E. and M.E. degrees in automobile engineering from Shandong University, Jinan, China, in 2012 and 2018, respectively. He is currently pursuing the Ph.D. degree in biomedical engineering with Xi'an Jiaotong University, Xi'an, China. His research interests include biomedical functional neuroimaging, biomedical signal processing, and deep learning.



GANG WANG (M'12) received the Ph.D. degree in biomedical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2008.

He was a Postdoctoral Associate with the Department of Biomedical Engineering, University of Minnesota, Twin Cities. Since 2011, he has been with the Faculty of Xi'an Jiaotong University as an Associate Professor with the School of Life Science and Technology, Institute of Biomedical Engineering. He has authored more than 40 international journal articles, including the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING and the IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS. He holds a number of patents and copyrights. His research interests include biomedical functional neuroimaging, biomedical signal processing, neural engineering, and bioelectromagnetism.

XIANGGUO YAN received the B.S. degree in industrial automation from the Zhengzhou University of Light Industry, Zhengzhou, China, in 1983, the M.S. degree in automatic control, and the Ph.D. degree in biomedical engineering from Xi'an Jiaotong University, Xi'an, China, in 1990 and 1995, respectively.

From 1996 to 1998, he was a Visiting Scientist with the Juelich research center, Germany. He is currently a Professor with the School of Life Science and Technology, Xi'an Jiaotong University. His main research interests include biomedical signal and image processing, and the development of medical instrumentation.