IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 13, 2020, accepted June 11, 2020, date of publication June 15, 2020, date of current version June 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3002548

Automatic Sleep Staging Based on a Hybrid Stacked LSTM Neural Network: Verification Using Large-Scale Dataset

CHIH-EN KUO AND GUAN-TING CHEN
Department of Automatic Control Engineering, Feng Chia University, Taichung 407, Taiwan
Corresponding author: Chih-En Kuo (cekuo@fcu.edu.tw)

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2221-E-035-064 and Grant 109-2634-F-006-013.

ABSTRACT Previously reported automatic sleep staging methods have usually been developed using healthy groups of fewer than 100 subjects. In this study, an automatic sleep staging method based on hybrid stacked long short-term memory (LSTM) was proposed and evaluated using a large-scale dataset of subjects with sleep disorders. Twenty-four features, including temporal and spectrum factors, were extracted from physiological signals and normalized after extracting the features. A variety of hybrid stacked LSTM structures and hidden units were used to determine the most suitable structure and parameters for the automatic sleep staging method. Finally, the proposed method was validated using a large-scale sleep disorder dataset from the PhysioNet Challenge 2018. To validate the robustness of the proposed system, half of the 994 subjects were randomly assigned to the training set, and the other half were assigned to the testing set. The best accuracy and kappa coefficient of the proposed method are 83.07% and 0.78, respectively. The best hybrid stacked structure was LSTM combined with bidirectional LSTM, which has 125 hidden units. In addition, four common sleep indices, including sleep efficiency, sleep onset time, wake after sleep onset, and total sleep time, were evaluated. The results, according to the intraclass correlation coefficient, indicated a moderate agreement with the results of the expert. The performance of the proposed method was compared with that of conventional machine learning, and it was noted that the hybrid stacked LSTM is a promising solution for automatic sleep staging. In future work, this method may assist clinical staff in reducing the time required for sleep staging.

INDEX TERMS Automatic sleep staging system, deep learning, hybrid stacked long short-term memory, large-scale sleep disorder dataset.

I. INTRODUCTION

Sleep is essential because it helps to restore the functions of the body and mind, such as the immune, nervous, skeletal, and muscular systems [1]. Sleep disorders, such as insomnia and sleep apnea, may cause daytime sleepiness, reduced cognitive function, weight gain, or even death. According to Philips' sleep survey, only half of the adults are satisfied with their sleep. In addition, 51% of adults report having sleep apnea. To diagnose sleep disorders, polysomnography (PSG) was used to record and analyze all-night sleep physiological signals from humans. The PSG data included electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), and electrocardiograph (ECG). After recording PSG data, the clinical staff uses manual sleep scoring to analyze human sleep architectures and evaluate their sleep quality. According to the American Academy of Sleep Medicine (AASM) [2], the sleep signals were segmented into many consecutive epochs with 30-s lengths, and then clinical staff scored each epoch as a specific sleep stage that contained wakefulness (Wake), nonrapid eye movement (Non-REM; stages 1-3), and rapid eye movement (REM).

However, diagnosing sleep disorders is time consuming and requires a considerable workload [3]. From the patient side, they need to wait at least two months to record sleep signals using PSG at a sleep center. In addition, the sensors and

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Moinul Hossain.

electrodes attached to the patient's body may cause mental stress and discomfort. From the clinical staff side, the scoring process is time intensive because the length of time it takes to record sleep data is approximately 6 to 8 hours, and they have to manually analyze patients' sleep data to conduct sleep scoring and annotate sleep-related events, which is a process that takes at least 1 hour.

Various automatic sleep staging methods have been proposed using all-night PSG recording data. These methods can be mainly divided into two steps: extracting different type of features from PSG data and training a classifier using the features. Different types of features, such as time-domain and frequency-domain, have been used to analyze PSG data [4], [5]. In addition, conventional machines, such as the support vector machine, are also used to help identify the sleep stages. The overall agreements of these methods were in the range of 80%–85%. Recently, deep learning, which has been promoted by strong computing power and massive datasets, has recently achieved good performance on complex medical pattern recognition tasks, such as pulmonary nodule and retinopathy screening. Therefore, sleep stages can also be classified using deep learning technology, such as deep belief nets (DBNs) [6], convolutional neural networks (CNNs) [7]–[9], or long short-term memory (LSTM) [9]–[11]. These networks learn the hierarchical representations or features from input data and classify them according to the learned features.

However, the previous automatic sleep classification methods generally use fewer than 100 PSG recordings from healthy individuals to develop and evaluate their methodologies. Although those methodologies have achieved high performance for healthy individuals, they are unlikely to have good generalizability. PSG signals vary widely due to individual differences, sleep conditions, and medicine effects. Therefore, the amount of PSG data is not enough. Furthermore, common clinical patients usually suffer from more sleep disorders than healthy individuals.

In this study, a massive sleep dataset from PhysioNet (nearly 1000), which was recorded from patients with sleep disorders, was used to train a hybrid stacked LSTM model and evaluate the model. Before classification, the 24 features were extracted from 30-s EEG, EOG, and EMG recordings and normalized to decrease the individual differences. The five hybrid stacked LSTM models with different numbers of hidden units were designed to find the model with the highest performance. The proposed method was validated using randomly selected subjects with independent training data. The performance evaluation used a confusion matrix to compute the overall agreement and kappa. In addition, the differences in the four common sleep indices between expert judgment and the proposed method were also compared using the Bland-Altman plot and intraclass correlation coefficient.

The significance of this study is the following. (1) The massive PSG sleep disorder data were used to train and evaluate the hybrid stacked LSTM models. (2) The five stacked LSTM models with the different numbers of hidden units were designed. (3) A suitable model with the appropriate number of hidden units to classify sleep stages was found. (4) We also used four sleep indices to compare the differences between the proposed method and expert judgment and the results are compared to those from previous studies.

II. MATERIALS AND METHODS

A. DATA DESCRIPTION

We used the PhysioNet2018 dataset in this study, which was taken from the PhysioNet Challenge 2018 [12]. The subjects in the dataset had all-night PSG recordings taken at an MGH sleep laboratory to diagnose sleep disorders. The dataset, which is a collection of 1893 all-night PSG recordings, was divided into two parts: a training set (n = 994) and a testing set (n = 989). Each PSG recording contained EEG (C3-M2, C4-M1, F3-M2, F4-M1, O1-M2, and O2-M1), left eye EOG (E1-M2), and chin EMG recordings with a sampling rate of 200 Hz. In addition, the all-night PSG recordings in the training set had their sleep stages and events annotated. According to the American Academy of Sleep Medicine (AASM) rule, each EEG with a 30-s interval in the training set was annotated with a corresponding sleep stage by the clinical staff. Therefore, we only used the training set to develop and validate the proposed method. The training set had 994 all-night PSG recordings that were obtained from patients with sleep disorders, and the mean ± standard deviation of the age and apnea-hypopnea index (AHI) of the patients was 55±14.2 years and 19±14.6 per hours, respectively. More information can be found on the official website [12].

B. METHODOLOGY

Fig. 1 illustrates the proposed automatic sleep stage classification method, including (1) preprocessing, (2) feature extraction, and (3) classification. The following figure presents each part in greater detail.

1) PREPROCESSING

The eight-order Butterworth bandpass filter with a cutoff frequency of 0.5-30 Hz was used to filter the EEG and EOG data, and the eight-order Butterworth bandpass filter with a cutoff frequency of 5-100 Hz was used to filter the EMG data. Next, the all-night PSG signals were segmented into consecutive epochs with a length of 30 s each.

According to the American Academy of Sleep Medicine (AASM) guidelines, multiple channel PSG recordings were taken because the physiological signals measured in the recording procedure may contain noise or artifacts. In actual clinical conditions, a channel with a better signal quality was selected to score by an expert. If the EEG channel selection method is not used, features with noise or artifacts may be extracted from the signal. This approach means that the classifier could not be successfully trained and would be unable to accurately classify sleep stages. An EEG with an amplitude greater than 250 μV during sleep generally means that it is an abnormal signal and will affect the classifier's performance.

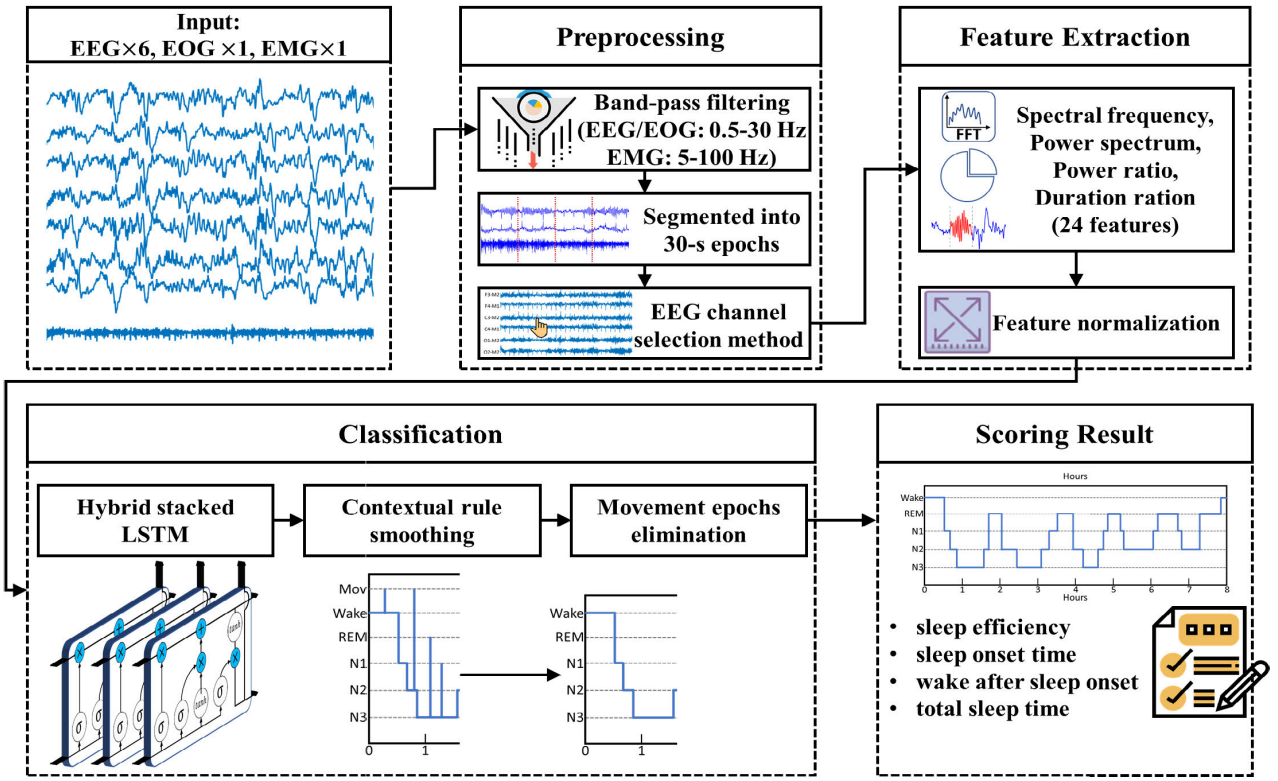

FIGURE 1. Proposed automatic sleep stage classification architecture.

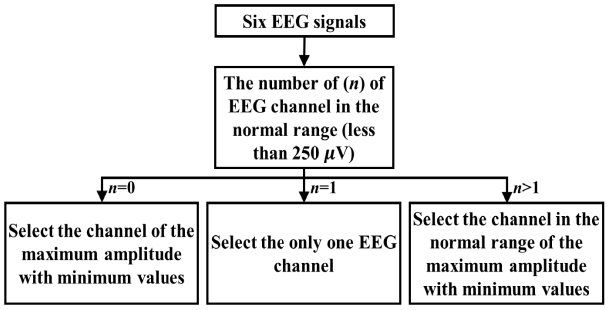

FIGURE 2. EEG selection method.

Therefore, we adopted an EEG channel selection method to choose a channel with less abnormal signals. Fig. 2 shows the flowchart of the EEG channel selection method. First, we computed the number of EEG channels (n) with amplitudes less than 250 μV (i.e., normal range) during each 30-s epoch. If $n \geq 1$, the EEG channel with a maximum amplitude that is at the minimum of the normal range was chosen. If $n = 0$, the 30-s epoch was considered to be body movement (Mov). Therefore, our classifier classified each 30-s epoch as either Wake, N1, N2, N3, REM, and Mov, and the Mov epochs were eliminated after smoothing according to the AASM rule.

TABLE 1. Sleep features considered in this study for automatic sleep staging.

No.	Type	Feature	Source	Label
1	PS	Total power of 0-30 Hz	EEG	0-30 E
2	PS	Total power of 0-30 Hz	EMG	0-30 M
3	PS	Total power of 0-30 Hz	EOG	0-30 O
4	PR	0-4 Hz/0-30 Hz	EEG	0-4 E
5	PR	4-8 Hz/0-30 Hz	EEG	4-8 E
6	PR	8-13 Hz/0-30 Hz	EEG	8-13 E
7	PR	13-22 Hz/0-30 Hz	EEG	13-22 E
8	PR	22-30 Hz/0-30 Hz	EEG	22-30 E
9	PR	0-4 Hz/4-8 Hz	EEG	0-4/4-8 E
10	PR	8-13 Hz/4-8 Hz	EEG	8-13/4-8 E
11	PR	0-4 Hz/0-30 Hz	EOG	0-4 O
12	SF	Mean frequency of 0-30 Hz	EEG	Mean(fre.) E
13	SF	Mean frequency of 0-30 Hz	EOG	Mean(fre.) O
14	SF	Mean frequency of 0-30 Hz	EMG	Mean(fre.) M
15	SF	Std. of frequency	EEG	Std(fre.) E
16	SF	Std. of frequency	EOG	Std(fre.) O
17	SF	Std. of frequency	EMG	Std(fre.) M
18	DR	Alpha ratio	EEG	Alpha E
19	DR	Spindle ratio	EEG	Spindle E
20	DR	Slow wave (i.e. N3) ratio	EEG	N3 E
21	DR	K-complex ratio	EEG	K-complex E
22	energy	Std. of amplitude	EOG	Std(AMP) O
23	energy	Mean amplitude	EMG	Amp M
24	energy	Std. of amplitude	EMG	Std(AMP) M

2) FEATURE EXTRACTION

The 24 features that were computed from EEG, EOG, and EMG signals were used in this study, as shown in Table 1. Each epoch, which is a consecutive signal with a length of 30 s, was segmented into 15 nonoverlapping subintervals (i.e., with each 2-s subinterval as a window) to avoid

losing the spectral characteristics, such as spindles and the K-complex. Then, the power of each subinterval was computed using the FFT. These features could be divided into five types: power spectrum (PS), power ratio (PR), spectral frequency (SF), duration ratio (DR), and energy. The PS is calculated by averaging the power of a specific frequency band. The PR is the power ratio of two frequency bands. The SF is the mean frequency of the spectral power. The DR is the ratio between the number of windows in which the energy of a specific frequency band is higher than a threshold to the total number of windows in an epoch (15). The energy represents the statistical features. More details regarding these features can be found in reference [4].

3) FEATURE NORMALIZATION

Feature normalization was applied to reduce the effects of the individual variability and was performed over values for each feature separately. This process can prevent extremely high or low values from influencing any conclusions. The procedure for feature normalization is summarized in the following steps.

Step 1: Calculate the means of the 10% lowest and highest values for the feature as the min and max values, respectively.

Step 2: Set the min and max values as 0 and 1, and then normalize the other values from 0 to 1.

Step 3: If the value is higher than 1, the value is specified as 1. If the value is lower than 0, the value is specified as 0.

Fig. 9 shows the distribution, means, and standard deviations of each feature corresponding to the five sleep stages in Table 1. It can be observed that some features have a clear distinction for a specific sleep stage. For example, the feature "0-30 E" could clearly distinguish between REM and N3, and the feature "22-30 E" could clearly distinguish between Wake an N3, as shown in Fig. 9 (a) and Fig. 9 (h), respectively.

4) CLASSIFICATION

Long short-term memory (LSTM) [13] was used to classify each feature vector into one of the six sleep stages in this study. An LSTM network modifies the standard RNN to effectively overcome the problem that the standard RNN is not good at learning a time series with latent long-term dependencies. In addition, the bidirectional LSTM (BiLSTM) is also used. The basic idea of the BiLSTM is to present each training sequence forward and backward to two separate LSTM layers, both of which are connected to the same output layer. Therefore, for every point in a given sequence, the network has complete, sequential information regarding all the points before and after it. The sequence to sequence model is widely used and applied in the semantics field and can extract more rich semantic features. The purpose of this study is whether the stacked models using unidirectional and bidirectional LSTM layers are better than the same type of LSTM layer. Therefore, we designed hybrid stacked LSTM networks with double and triple layers to find the best

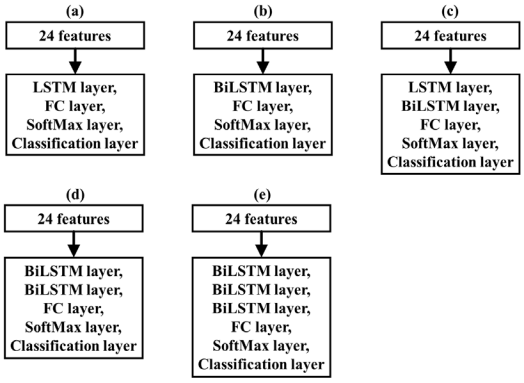

FIGURE 3. Five different structures of the hybrid stacked LSTM models for the proposed method. (a) Single LSTM, (b) single BiLSTM, (c) LSTM and BiLSTM, (d) double BiLSTM, and (e) triple BiLSTM (deep LSTM).

LSTM-based model for sleep staging and tested the various LSTM architectures and parameters of the network setting. In this study, five different structures of hybrid stacked LSTM models were tested, as shown in Fig. 3. The FC layer denotes the fully connected layer. To find the optimal hyperparameter of the network structure, the recurrent hidden unit was set from 10 to 250 with a step size of five. The maximum number of training epochs was 35. The adaptive moment estimation (ADAM) algorithm was adopted in this study for the backpropagation.

TABLE 2. Smoothing rules.

No.	Condition
1	Any REM epochs before the very first appearance of N2 are replaced with N1 epochs
2	Wake, REM, N2 → Wake, N1, N2
3	N1, REM, N2 → N1, N1, N2
4	N2, N1, N2 → N2, N2, N2
5	N2, N3, N2 → N2, N2, N2
6	N2, REM, N2 → N2, N2, N2
7	N3, N2, N3 → N3, N3, N3
8	REM, Wake, REM → REM, REM, REM
9	REM, N1, REM → REM, REM, REM
10	REM, N2, REM → REM, REM, REM
11	Mov, REM, N2 → Mov, N1, N2

5) SMOOTHING RULE

After scoring the sleep stages using the hybrid stacked LSTM model, a smoothing rule was used to increase the accuracy by considering the temporal contextual information since an expert may refer to the neighboring epochs in addition to the current epoch to make decisions. Therefore, we use the existing smoothing rules [4] to make sure that the automatic sleep staging results were similar to the expert's manual scoring. Table 2 presents the existing smoothing rules. For example, according to smoothing rule 1, any REM epochs before the first appearance of the N2 stage were replaced with N1 epochs. The three consecutive epochs of N2, REM, and N2 were replaced with the sequence N2, N2, and N2.

Similarly, the consecutive epochs of REM, N1, and REM were replaced with the sequence REM, REM, and REM.

6) MOVEMENT EPOCH ELIMINATION

After smoothing, a Mov elimination procedure using the AASM guidelines was used. If the Wake stage preceded or followed the Mov stage, Mov was replaced with Wake. If non-REM or REM preceded or followed Mov, Mov was replaced with the same stage as the epoch that followed it. The final result of the hypnogram was still characterized by the five stages.

C. EVALUATISON METRICS

A confusion matrix was used to compare the differences between our proposed method and the expert's manual scoring, and different metrics were employed to evaluate the performance of the proposed method, including the overall agreement (*overall*, i.e., accuracy), sensitivity (*Se*), and positive predictive value (*PPV*). These metrics are defined as follows:

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

$$Se = \frac{TP}{TP + FN}. \tag{2}$$

$$PPV = \frac{TP}{TP + FP}. \tag{3}$$

where *TP* and *TN* denote number of correct classifications, and *FP* and *FN* denote the number of incorrect classifications. *PPV* is the ratio of the true positives to the predicted positives. In addition, we also calculated Cohen's kappa coefficient (k) [14] to evaluate the agreement of the classification result between the expert and the proposed method. Cohen's kappa coefficient is a statistical measure of the interrater agreement among two or more raters.

To diagnose sleep issues, four common sleep indices can be used as a reference, including the sleep efficiency (SE), total sleep time (TST), sleep onset time (SOT), and wake after sleep onset (WASO). These indices are calculated from a hypnogram and defined as follows. The TST is defined as the amount of actual sleep time in a sleep episode. The SE is defined as the ratio of the TST to the time period from lights off to lights on. The SOT is defined as the amount of time it takes to go from being fully awake to sleep. The WASO is defined as the total minutes of wakefulness recorded after the SOT. In clinical diagnoses, a subject may have a poor night of sleep if their SE is lower than 85% [15]. The SOT is calculated to assess whether a subject can fall asleep promptly. The WASO is calculated to assess whether a subject has difficulty remaining asleep after the SOT.

We calculated the mean absolute errors (MAEs), the intraclass correlation coefficients (ICCs) [16], and the paired t-tests to evaluate the sleep indices. The MAE is the average magnitude of the absolute errors in a set of forecasts. The ICC represents the agreement between two or more raters or evaluation methods in the same dataset. The ICC form was set as ICC (2, 1) (i.e., two-way ANOVA, single measurement, and absolute agreement) in this study. The ICC can be interpreted as follows: an ICC less than 0.5 means poor reliability, an ICC from 0.5 to 0.75 means moderate reliability, an ICC from 0.75 to 0.9 means good reliability, an ICC greater than 0.9 means excellent reliability. The p-value, which was calculated using the paired t-test, was considered statistically significant when it was less than 0.05. In addition to the above metrics, we also used a Bland-Altman plot [17] and scatter plot to present the differences between the proposed method and the expert's manual scoring.

III. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETUP

Two-fold cross-validation was used to evaluate the performance of the proposed method. Specifically, half of the subjects were randomly grouped into the training set, and the others were used as the testing set. We repeated the random 2-fold cross-validation 64 times to test the robustness and generalization ability of the proposed model. In each 2-fold cross-validation program, the subjects that were part of the training data were independent of the subjects that were part of the testing data and were randomly selected. It means that the data from the same subjects do not simultaneously appear in both the training and testing sets; therefore, they fit the real situation.

Each run would compute their evaluation metrices, and the average evaluation metrics were used as the final results. According to Penzel *et al*. [18], our evidence grading for the performance evaluation studies is level one. The performance was evaluated based on the following respects: (1) the average performance of the method for five different hybrid stacked LSTM models with various numbers of hidden units, and (2) the sensitivity (*Se*) of each sleep stage obtained using the proposed method from the best hybrid stacked LSTM model.

B. AUTOMATIC SLEEP STAGING PERFORMANCE

Fig. 4 shows the average accuracy curves for the five hybrid stacked LSTM models with different numbers of hidden units. The following characteristics can be observed: (1) the accuracy may decrease when the number of hidden units is very few or many, (2) the accuracy is not better for the three-layer LSTM model than the two-layer LSTM model, and (3) the accuracy is better with BiLSTM than LSTM for sleep staging. Table 3 shows the highest average accuracy for the five hybrid stacked LSTM models with the number of hidden units. The LSTM+BiLSTM model with 125 hidden units obtained the highest average accuracy. In addition, each hybrid stacked model had a lower standard deviation, which confirmed the robustness and stability of the proposed model.

The 15,680 ($5 \times 49 \times 64$) hybrid stacked LSTM models with different network structures and different numbers of hidden units were trained in the experiment. The best hybrid stacked LSTM model that exhibited the highest accuracy throughout the experiment was recorded.

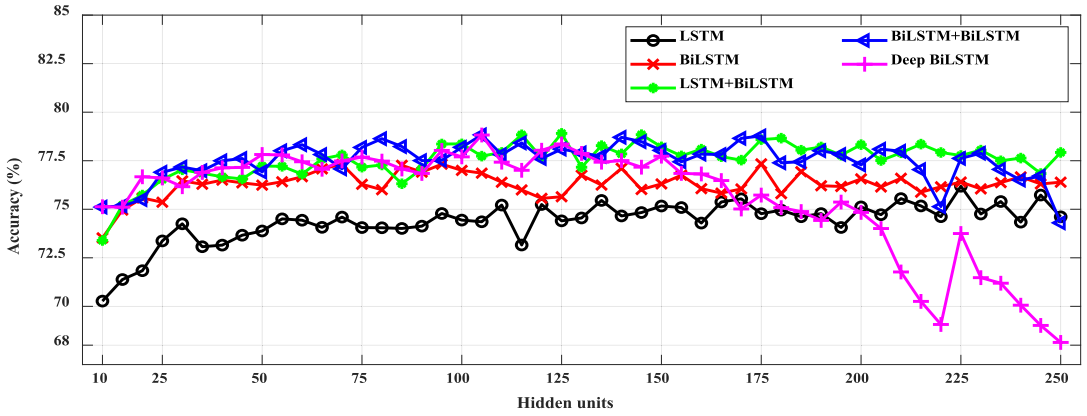

FIGURE 4. Average accuracy curve for five hybrid stacked LSTM models with different hidden units.

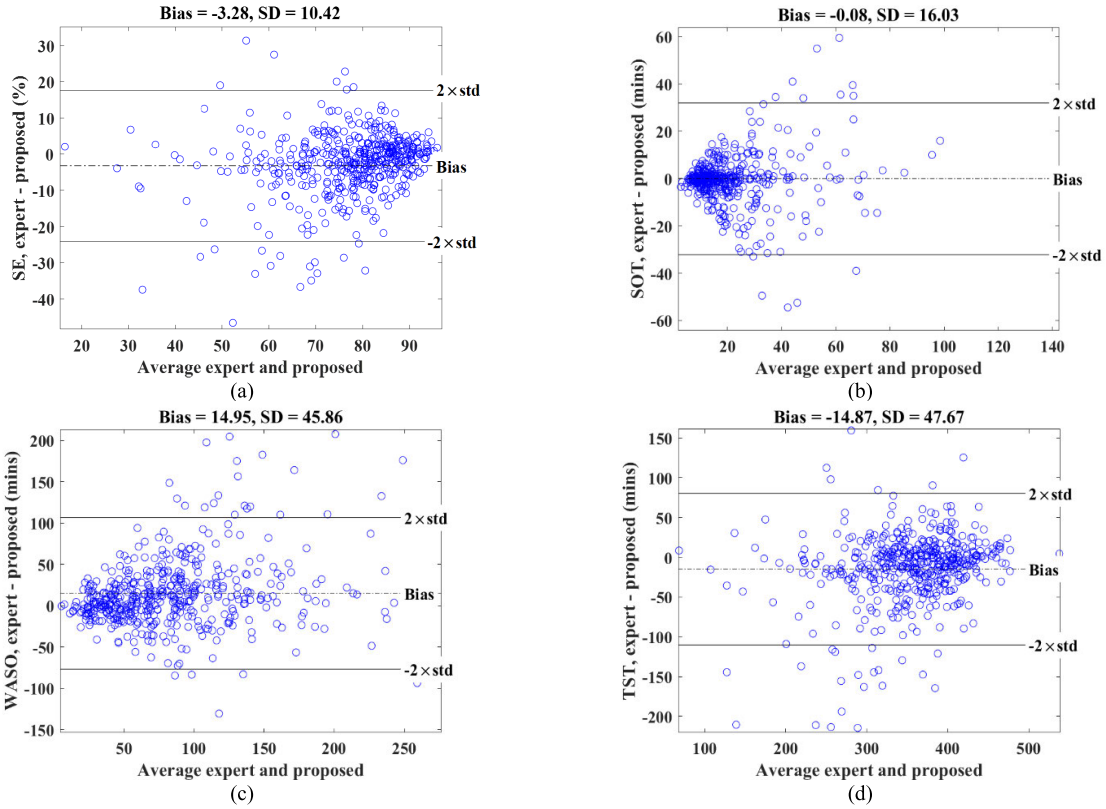

FIGURE 5. Bland-Altman graphs for each sleep index. (a) SE, (b) SOT, (c) WASO, and (d) TST. The X axis and Y axis represent the mean of the measurements of the experts and those of the proposed method and the difference between the results of the experts and the proposed method, respectively.

The confusion matrix was used to compare the best classification result from BiLSTM+LSTM to the expert result, as shown in Table 4. The testing set contained 497 PSG recordings with 476,750 30-s epochs, and the Mov epochs were not considered. The overall agreement and kappa were 83.07% and 77.52%, respectively. N1 has a low sensitivity due to class imbalance and sleep stage transition. N1 in PhysioNet2018 is only 19.6% of the dataset whereas N2 is 51%. In addition, the patients with sleep apnea may be aroused after falling asleep because they cannot breathe on their own. Therefore, the PhysioNet2018 dataset had more sleep stage transitions. Misclassification mostly occurred from one stage to another between the pairs Wake-N1, N1-REM, N2-N3, and REM-Wake [19]. Both the sleep onset time and total sleep time may not be accurately evaluated due to N1 having low sensitivity.

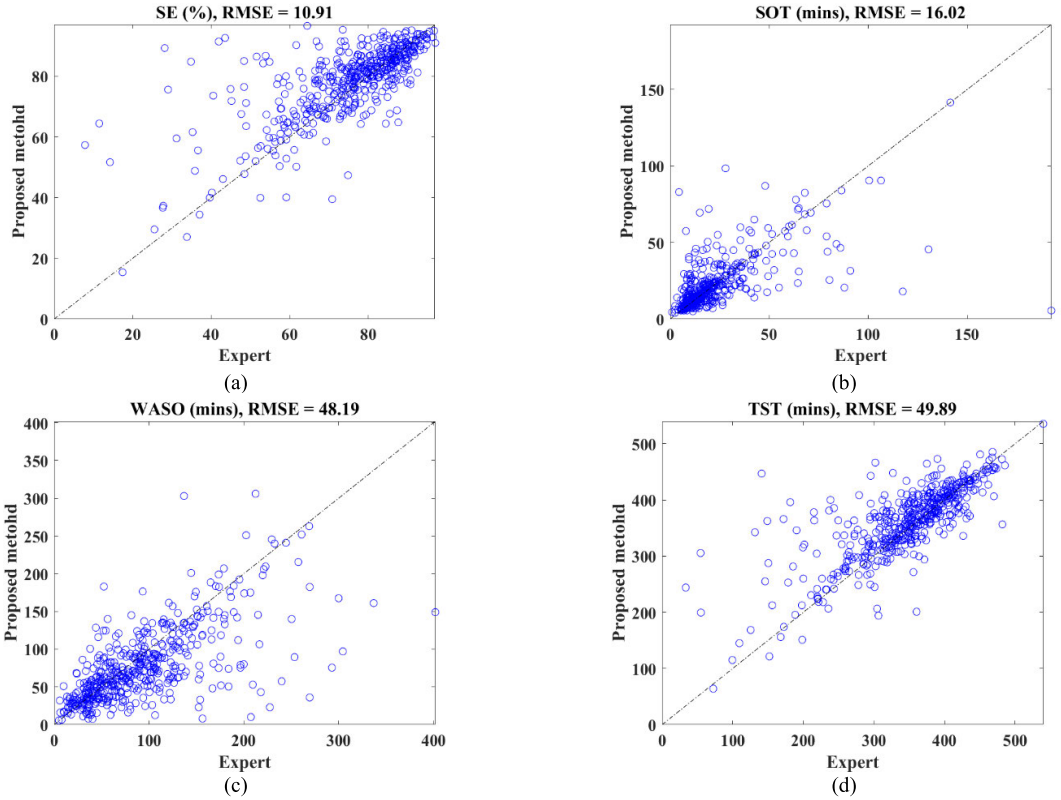

FIGURE 6. Scatter graphs for each sleep index. (a) SE, (b) SOT, (c) WASO, and (d) TST. The X axis and Y axis represent the measurements of the experts and the proposed method, respectively.

TABLE 3. Best average accuracy, standard deviation, and number of hidden units for five hybrid stacked LSTM models (bold letters indicate the best results).

Model	Best averaged accuracy (%)	Standard deviation (%)	Number of hidden units
LSTM	76.19	± 2.23	225
BiLSTM	77.34	± 2.23	175
LSTM+BiLSTM	**78.90**	**± 2.29**	**125**
BiLSTM+BiLSTM	78.84	± 2.53	105
Deep BiLSTM	78.83	± 2.39	105

TABLE 4. Confusion matrix and evaluation metrices of the best hybrid stacked LSTM model.

		Computer					Metrics (%)			
		Wake	N1	N2	N3	REM	SE	PPV	overall	kappa
Expert	Wake	**51168**	4955	5521	0	840	81.89	74.67		
	N1	10282	**40344**	8199	70	12937	56.16	77.07		
	N2	3448	1334	**165953**	8929	411	92.16	85.81		
	N3	104	492	11084	**74784**	112	86.38	88.80		
	REM	3526	5221	2634	429	**62976**	84.21	81.49		
									83.07	77.52

TABLE 5. Comparison of the manual scoring and proposed method in terms of various objective sleep indices.

	SE (%)	SOT (mins)	WASO (mins)	TST (mins)
Expert	75.165±14.74	21.47±19.99	92.04±59.80	345.55±74.84
Proposed method	78.44±12.05	21.55±16.73	77.09±49.06	360.43±63.76
MAE	6.79	7.29	30.56	31.13
ICC	0.68	0.62	0.63	0.75
p-value	0.74	0.67	0.70	0.80

C. EVALUATION OF THE SLEEP INDICES

1) EVALUATION METRICS

The four common sleep indices were calculated using the proposed automatic sleep staging method and compared with the expert's manual scoring [20]. Table 5 shows the results that compared the expert's manual scoring and the proposed automatic sleep staging method. No significant differences were found among the sleep indices obtained from the manual scoring and the proposed automatic sleep staging method. The MAEs of the SE, SOT, WASO, and TST were 6.79%, 7.29 min, 30.56 min, and 31.13 min, respectively. The ICC indicated moderate agreement between the expert's manual scoring and the proposed automatic sleep staging method.

2) BLAND-ALTMAN PLOT ANALYSIS

Fig. 5 shows the Bland-Altman plots for each sleep index, and each subplot represented a sleep index. The Bland-Altman plot can present the mean and the standard deviation (std) of the differences between two methods. The Y-axis shows the difference between two methods (A-B), and the

X-axis shows the mean of the two methods ((A+B)/2). The positive or negative bias was represented as the difference between the mean difference and zero, and it also respectively represented the proposed method's overestimation or underestimation. In addition, the ±2 std represented the 95% confidence intervals, and we observed whether the data points lied within this interval. The biases of the SE, SOT, WASO, and TST were −0.03%, −0.08 min, 15.07 min, and −15.48 min, respectively.

The results of the Bland-Altman graph showed that the sleep indices obtained using the manual scoring and the proposed method were similar in terms of the SE and SOT but considerably different in terms of the WASO and TST. Because body movement indirectly affects the assessment of the sleep indices, if a subject's movement during sleep was unusual (e.g., with low sleep efficiency), the WASO and TST could not be accurately reported using the proposed method, and further examinations would be required.

3) SCATTER PLOT ANALYSIS

Fig. 6 shows the comparisons of the subject-by-subject sleep indices estimated using the proposed method and the results of the manual scoring. The root-mean-square errors (RMSEs) between the estimation obtained using the proposed method and the manual scoring for the various objective sleep indices were also calculated. The RMSE is a quadratic scoring rule that measures the average magnitude of the error. The RMSEs of the SE, SOT, WASO, and TST were 10.91%, 16.02 min, 48.19 min, and 49.89 min, respectively. The distributions of SE and SOT were extremely close to the diagonal, which indicated that the proposed method exhibited good performance when calculating the SE and SOT, although some large errors were observed in the WASO and TST. These errors may arise because the subject experienced many transitions in their sleep stages. This phenomenon was consistent with the findings shown in Fig. 5.

IV. DISCUSSION AND CONCLUSION

An automatic sleep staging method using the hybrid stacked LSTM was proposed in this study. The 24 features were extracted from the selected EEG, left-eye EOG, and chin EMG signals and normalized to reduce individual differences. Then, the five hybrid stacked LSTM models with different numbers of hidden units were designed to find the suitable number of hidden layers and their hidden units to score PSG data. In addition to the model evaluation, the four sleep indices that were computed using the proposed method were compared to those that were computed by experts. The hybrid stacked LSTM model, the LSTM+BiLSTM with 125 hidden units, had the highest mean accuracy of 78.90% and the highest accuracy of 83.07%. In addition, the std. of the best hybrid stacked LSTM models was only 2.29%. It proved the robustness and stability of the proposed method. For the assessment of sleep indices, the biases of SE, SOT, WASO, and TST were −3.28%, −0.08 min 14.95 min and −14.87 min, respectively. The results indicated a moderate agreement with the expert results for each sleep index according to the intraclass correlation coefficient.

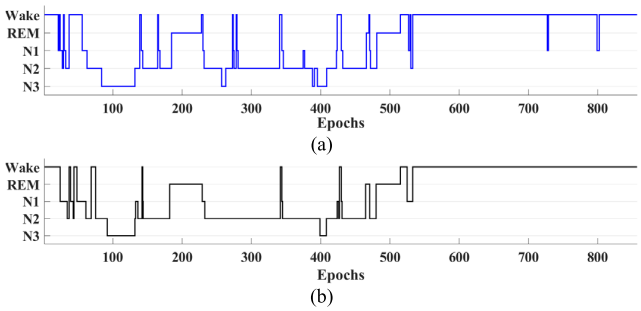

FIGURE 7. Hypnogram of the subject tr03-0005 (sleep efficiency: 54.38%). (a) The visual scoring. (b) The automatically scoring.

Our contributions in this study include four points. (1) The massive PSG data on sleep disorders (nearly 1000) was used to train and evaluate the hybrid stacked LSTM models. (2) Five hybrid stacked LSTM models with different numbers of hidden units were designed. (3) We find the suitable model and number of hidden units to classify sleep stages, and the suitable model was the LSTM+BiLSTM with 125 hidden units. (4) Similar to previous studies, we also use four sleep indices to compare the differences between the proposed method and experts.

Fig. 7 shows the hypnogram of Subject tr03-0005 (sleep efficiency: 54.38%). The manual scoring by the expert is shown in Fig. 7 (a), and the automatic scoring is shown in Fig. 7 (b). It can be observed that the hypnogram of the proposed automatic sleep staging system follows the changes in the sleep stages of the subject. The overall agreement for this subject is 88.50%. The sleep indices of Subject tr03-0982, that is, the SE, SOT, WASO, TST, show differences of 1.75%, −1.5 min, 9 min, and −7.5 min, respectively, when the results of the proposed method and those of the expert are compared.

Furthermore, we used six different classifier-based conventional machine learning techniques, namely, classification trees, linear discriminant analysis, the naive Bayes, the support vector machine, the K-nearest neighbors (KNN), and an ensemble of a subspace KNN, to compare conventional machine learning with the proposed method. The data are input into six different classifiers based on conventional machine learning techniques that have the same processing and validation method. Each epoch, a 30-s signal, had 24 features extracted as the input of six different classifiers, and these six different classifiers conducted training and testing according to the ground truth. The overall agreement is as shown in Table 6. The best overall accuracy was 69.50% for the conventional machine learning technique involving the ensemble of a subspace KNN. Compared with the LSTM-based neural network, these classifiers are not based on time series and they did not consider the temporal contextual information. However, sleep architecture has the temporal contextual information. Thus, they did not perform well.

To understand how LSTM learns the features of different sleep stages in the network model, we visualized the output

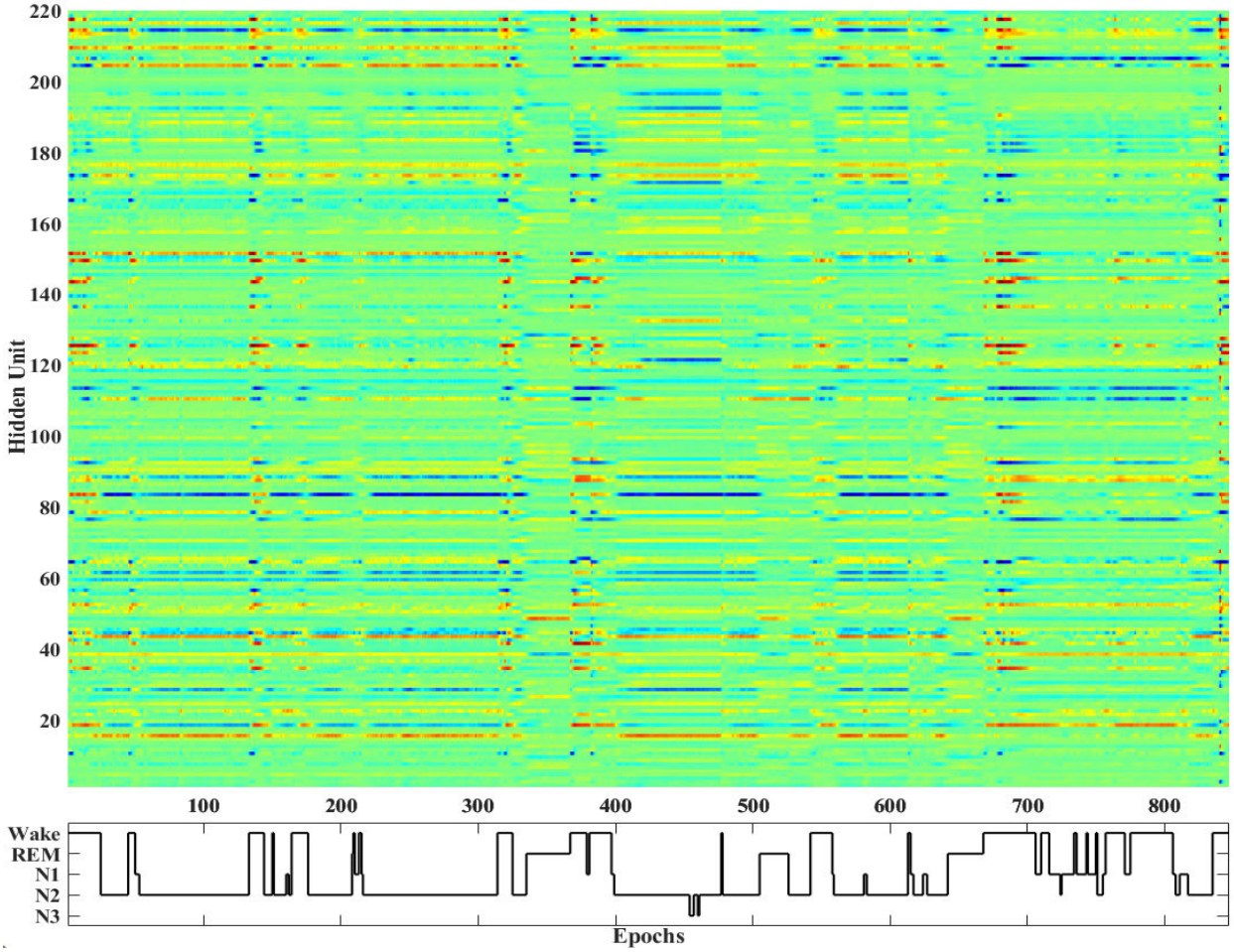

FIGURE 8. Activation of the LSTM hidden units from the automatic sleep staging for the subject tr03-0982. The top image shows the heat map, and the bottom image shows the hypnogram from the automatic sleep staging. The X axis represents the number of epochs, and Y axis denotes the number of hidden units.

TABLE 6. Overall agreement of the classifier based conventional machine learning techniques.

Shallow Neural Networks	Overall agreement (%)
Classification Trees	62.00
Linear Discriminant Analysis	62.80
Naive Bayes	54.50
Support Vector Machine	69.10
K-nearest Neighbors (KNN)	69.10
Ensemble of a Subspace KNN	69.50
Our Proposed Method	**78.90**

of LSTM in a heat map for one subject (tr03-0982), as shown in Fig. 8. The X-axis is the number of epochs, and the Y-axis is the number of hidden units. It was observed through the heat map that the output intensity of the hidden neurons of the LSTM changed with time for the different sleep stages. In addition, it can be observed from Fig. 8 that a hidden unit learns the features of a single sleep stage, and thus the output intensity of the hidden unit changes with the sleep stage. For example, the 20th hidden unit has a higher output intensity in the Wake stage and a lower output intensity in the non-Wake stages, the 115th hidden unit has a lower output intensity in the Wake stages and a higher output intensity in the non-Wake stages, and the 45th hidden unit has a higher output intensity in the REM stages (epochs 335-366) and a lower output intensity in the non-REM and Wake stages.

Table 7 shows a comparison of our method and other sleep stage scoring methods based on the deep learning and conventional machine learning classifier in terms of the kappa, overall agreement, and sensitivity. Although reference [21] has the highest performance, they only used eight healthy subjects and tested the method by subject dependent. This did not meet the real situation in the clinical application. In terms of the used dataset, we used PhysioNet2018 in the comparison since it has the greatest number of subjects (994). On the other hand, both the Sleep-EDF and Montreal Archive of Sleep Studies (MASS) were made up of data from healthy individuals, but the PhysioNet2018 was made up of data from patients with sleep apnea. It means the PhysioNet2018 dataset

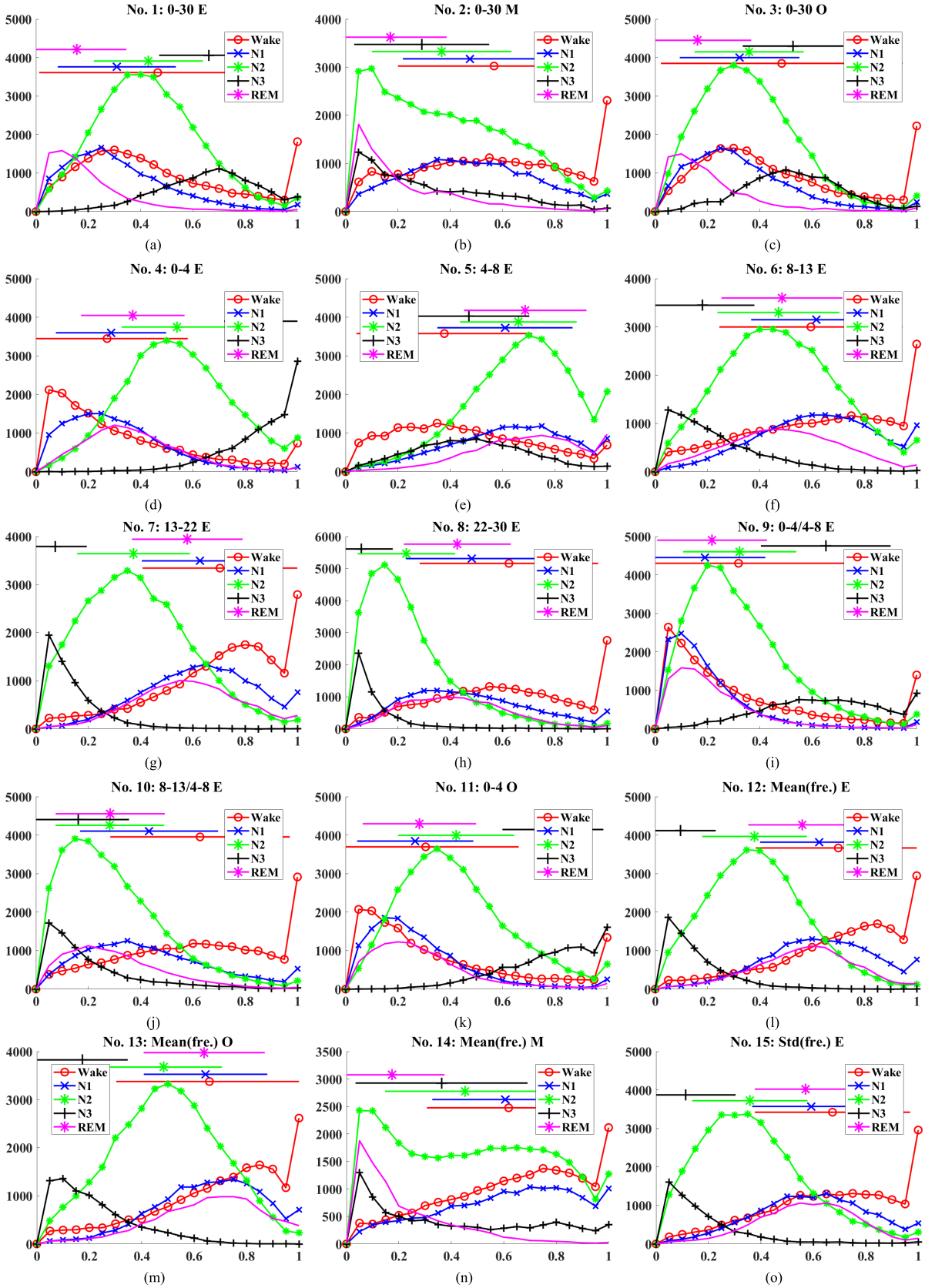

FIGURE 9. The distributions of the epoch numbers and feature values in Table 1 for the Wake, N1, N2, N3, and REM stages, respectively. The X axis represents the distribution of the magnitude of the normalized feature values; its range is 0 to 1, and the bins are 0.05. The Y axis represents the total number of each corresponding stage.

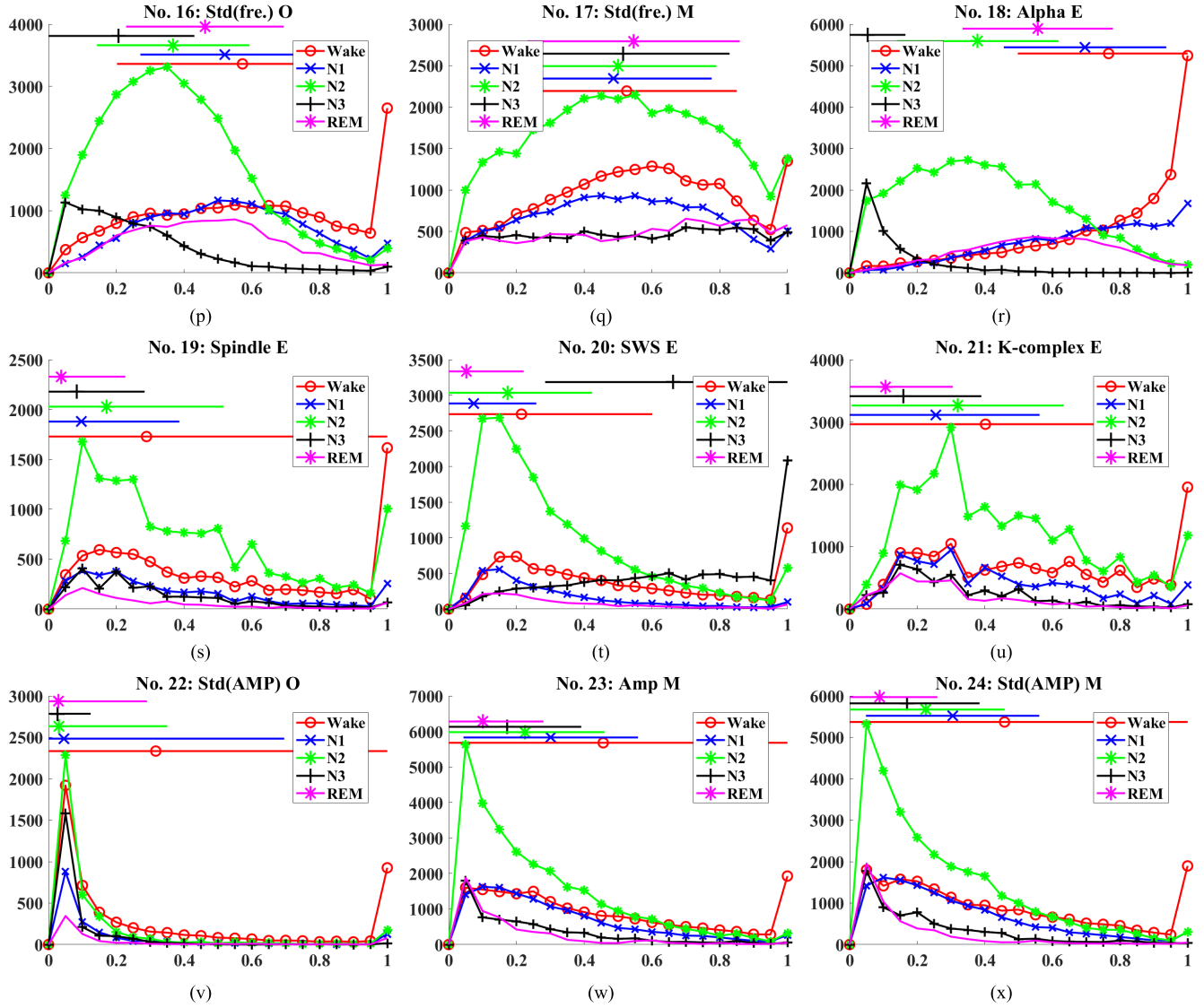

FIGURE 9. ***(Continued.)*** **The distributions of the epoch numbers and feature values in Table 1 for the Wake, N1, N2, N3, and REM stages, respectively. The X axis represents the distribution of the magnitude of the normalized feature values; its range is 0 to 1, and the bins are 0.05. The Y axis represents the total number of each corresponding stage.**

has a higher frequency of sleep stage changes and more complexity than both the Sleep-EDF and MASS dataset.

In addition, most patients who have PSG recordings taken in sleep center or hospital are more likely to have sleep disorders than healthy individuals. This means that PhysioNet2018 is closer to the actual clinical situation. In terms of the validation procedure, the 2-fold cross-validation strategy has a greater number of testing subjects than the other methods. It showed that the proposed method was more robust than the other methods. In addition, we reported our iterations to prove the generalization ability. The kappa also showed that the proposed model has substantial agreement.

Perslev *et al.* [22] proposed a U-Net-based automatic sleep scoring method and evaluated the method using the PhysioNet2018 dataset. Their method was based on time series segmentation for sleep scoring. Compared to their results, the agreement and the sensitivities of Wake, N2, and N3 for our proposed method were better.

In summary, the model was trained using the data of patients with sleep disorders, which could guarantee the clinical applicability. On the other hand, the data of patients with sleep disorders contained significant noise; thus, the robustness of the model may increase more than when using the data from healthy individuals.

In future work, this method can assist clinical staff in reducing the time required for sleep staging. We can try to use the same architecture to analyze at healthy group or other group with sleep-disorders to ensure the transferability of the system. In addition, the existing research proposed some features that could distinguish between high and low sleep efficiency [5], [23]. These features can be combined with the proposed method to improve the system robustness.

TABLE 7. Comparison between the proposed method and other sleep staging methods.

Ref.	Dataset	Number of subjects	Validation strategy	Method	Agreement (%)	kappa (%)	Sensitivity (%)				
							Wake	N1	N2	N3	REM
[24]	DREAMS	20 (healthy)	2-fold (20 times) subject dependent	MCFS+SVM [a]	83.31	77.09	93.7	29.7	88.4	83.9	82.6
[21]	Sleep-EDF	8 (healthy)	10-fold subject dependent	TQWT+AdaBoost [b]	91.36	86.4	98.9	39.7	90.2	82.3	83.0
[25]	Sleep-EDF	8 (healthy)	10-fold subject dependent	EEMD+RUSBoost [c]	83.49	84.1	95.2	42.0	79.5	77.38	80.5
[26]	Sleep-EDF	20 (healthy)	leave-one-out	HMM [d]	78.1	70.0	86.9	18.3	87.6	79.5	86.4
[27]	MASS	62 (healthy)	60%-20%-20% (training-validating-testing)	MLP+LSTM [e]	85.92	79.09	84.55	56.31	90.73	84.76	86.12
[28]	Sleep-EDF	20 (healthy)	95%-5% (training-testing)	CNN+LSTM	82.00	78.70	84.7	46.60	85.90	84.80	82.40
[29]	MASS	200 (healthy & patients)	90%-5%-5%	Two-layers LSTM	**87.10**	**81.5**	**89.40**	**59.70**	90.90	80.20	**93.50**
[30]	Sleep-EDF	20 (healthy)	20-fold	Two-layers BiLSTM	82.0	76.0	83.4	50.1	81.7	94.2	83.9
[22]	PhysioNet Challenge 2018	994 (patients)	5-fold	U-Net	78.76	71.4	80.43	57.41	85.94	76.36	83.43
proposed method			2-fold (64 times)	Hybrid stacked LSTM	83.07	77.5	81.89	56.16	**92.16**	**86.38**	81.49

[a]MCFS+SVM: multi-cluster/class feature selection (MCFS) + support vector machine (SVM)
[b]TQWT+AdaBoost: tunable-Q factor wavelet transform (TQWT) + adaptive boosting (AdaBoost)
[c]EEMD+RUSBoost: ensemble empirical mode decomposition (EEMD) + random under sampling boosting (RUSBoost)
[d]HMM: hidden Markov model
[e]MLP: multilayer perceptron

First, these features can be used to distinguish between satisfactory and unsatisfactory sleep, and a specific model trained by different groups can later be used to score subject sleep stages. In addition, for clinical application, we can develop a semiautomatic human machine interface to assistant a sleep expert in scoring the sleep stage. The function of the human machine interface should include manual and automatical scoring and calculation of the sleep indices. In automatical scoring, the system's uncertain epoch can be marked and 24 features of that epoch can be provided for the experts to analyze.

APPENDIX

Fig. 9 shows the distributions of the epoch numbers and feature values in Table 1 for the Wake, N1, N2, N3, and REM stages, respectively. The X axis represents the distribution of the magnitude of the normalized feature values; its range is 0 to 1, and the bins are 0.05. The Y axis represents the total number of each corresponding stage.

REFERENCES

[1] F. S. Luyster, P. J. Strollo, P. C. Zee, and J. K. Walsh, "Sleep: A health imperative," *Sleep*, vol. 35, no. 6, pp. 727–734, Jun. 2012.
[2] C. Iber, and C. Iber, *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.
[3] A. Roebuck, V. Monasterio, E. Gederi, M. Osipov, J. Behar, A. Malhotra, T. Penzel, and G. D. Clifford, "A review of signals used in sleep analysis," *Physiol. Meas.*, vol. 35, no. 1, pp. R1–R57, Jan. 2014.
[4] S.-F. Liang, C.-E. Kuo, Y.-H. Hu, and Y.-S. Cheng, "A rule-based automatic sleep staging method," *J. Neurosci. Methods*, vol. 205, no. 1, pp. 169–176, Mar. 2012.
[5] S.-F. Liang, C.-E. Kuo, F.-Z. Shaw, Y.-H. Chen, C.-H. Hsu, and J.-Y. Chen, "Combination of expert knowledge and a genetic fuzzy inference system for automatic sleep staging," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 10, pp. 2108–2118, Oct. 2016.
[6] M. Längkvist, L. Karlsson, and A. Loutfi, "Sleep stage classification using unsupervised feature learning," *Adv. Artif. Neural Syst.*, vol. 2012, Jul. 2012, Art. no. 107046.
[7] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J.-F. Payen, "A convolutional neural network for sleep stage scoring from raw single-channel EEG," *Biomed. Signal Process. Control*, vol. 42, pp. 107–114, Apr. 2018.
[8] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic sleep stage scoring with single-channel EEG using convolutional neural networks," 2016, *arXiv:1610.01683*. [Online]. Available: http://arxiv.org/abs/1610.01683
[9] Y. Jeon, S. Kim, H.-S. Choi, Y. G. Chung, S. A. Choi, H. Kim, S. Yoon, H. Hwang, and K. J. Kim, "Pediatric sleep stage classification using multi-domain hybrid neural networks," *IEEE Access*, vol. 7, pp. 96495–96505, 2019.
[10] N. Michielli, U. R. Acharya, and F. Molinari, "Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals," *Comput. Biol. Med.*, vol. 106, pp. 71–81, Mar. 2019.
[11] Y. Wei, X. Qi, H. Wang, Z. Liu, G. Wang, and X. Yan, "A multi-class automatic sleep staging method based on long short-term memory network using single-lead electrocardiogram signals," *IEEE Access*, vol. 7, pp. 85959–85970, 2019.
[12] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp.Z e215–e220, 2000.
[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[14] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.
[15] H. Danker-Hopfe, D. Kunz, G. Gruber, G. Klösch, J. L. Lorenzo, S. L. Himanen, B. Kemp, T. Penzel, J. Röschke, H. Dorn, A. Schlögl, E. Trenker, and G. Dorffner, "Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders," *J. Sleep Res.*, vol. 13, no. 1, pp. 63–69, Mar. 2004.
[16] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychol. Bull.*, vol. 86, no. 2, p. 420, Mar. 1979.
[17] J. M. Bland and D. G. Altman, "Comparing methods of measurement: Why plotting difference against standard method is misleading," *Lancet*, vol. 346, no. 8982, pp. 1085–1087, Oct. 1995.

[18] T. Penzel, M. Hirshkowitz, J. Harsh, R. D. Chervin, N. Butkov, M. Kryger, B. Malow, M. V. Vitiello, M. H. Silber, and C. A. Kushida, "Digital analysis and technical specifications," *J. Clin. Sleep Med.*, vol. 3, no. 2, pp. 109–120, 2007.
[19] R. S. Rosenberg and S. Van Hout, "The American Academy of sleep medicine inter-scorer reliability program: Sleep stage scoring," *J. Clin. Sleep Med.*, vol. 09, no. 01, pp. 81–87, Jan. 2013.
[20] F. Mendonca, S. S. Mostafa, F. Morgado-Dias, A. G. Ravelo-Garcia, and T. Penzel, "A review of approaches for sleep quality analysis," *IEEE Access*, vol. 7, pp. 24527–24546, 2019.
[21] A. R. Hassan and M. I. H. Bhuiyan, "An automated method for sleep staging from EEG signals using normal inverse Gaussian parameters and adaptive boosting," *Neurocomputing*, vol. 219, pp. 76–87, Jan. 2017.
[22] M. Perslev, M. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-Time: A fully convolutional network for time series segmentation applied to sleep staging," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4417–4428.
[23] S.-F. Liang, C.-E. Kuo, Y.-H. Hu, Y.-H. Pan, and Y.-H. Wang, "Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 6, pp. 1649–1657, Jun. 2012.
[24] S. Seifpour, H. Niknazar, M. Mikaeili, and A. M. Nasrabadi, "A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal," *Expert Syst. Appl.*, vol. 104, pp. 277–293, Aug. 2018.
[25] A. R. Hassan and M. I. H. Bhuiyan, "Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting," *Comput. Methods Programs Biomed.*, vol. 140, pp. 201–210, Mar. 2017.
[26] H. Ghimatgar, K. Kazemi, M. S. Helfroush, and A. Aarabi, "An automatic single-channel EEG-based sleep stage scoring method based on hidden Markov model," *J. Neurosci. Methods*, vol. 324, Aug. 2019, Art. no. 108320.
[27] 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.
[28] S. Mousavi, F. Afghah, and U. R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0216456.
[29] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, "SeqSleepNet: End-to-End hierarchical recurrent neural network for Sequence-to-Sequence automatic sleep staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 400–410, Mar. 2019.
[30] 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.

CHIH-EN KUO received the B.S. degree in mathematics from National Cheng Kung University (NCKU), Tainan, Taiwan, in 2005, the M.S. degree in information management from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2009, and the Ph.D. degree from the Department of Computer Science and Information Engineering, NCKU, in 2013. He is currently an Assistant Professor with the Department of Automatic Control Engineering, Feng Chia University, Taichung, Taiwan. His current research interests are biomedical signal/image processing, machine learning, deep learning, computer-aided diagnosis systems, mathematical modeling, and human sleep EEG analysis.

GUAN-TING CHEN received the B.S. degree in automatic control engineering from Feng Chia University, Taichung, Taiwan, in 2018, where he is currently pursuing the M.S. degree in automatic control engineering. His research interests are machine learning and automatic sleep scoring.