Complex & Intelligent Systems (2024) 10:2689–2701
https://doi.org/10.1007/s40747-023-01290-2

ORIGINAL ARTICLE

Ssleepnet: a structured sleep network for sleep staging based on sleep apnea severity

Xingfeng Lv[1] · Jun Ma[1] · Jinbao Li[2] · Qianqian Ren[1]

Received: 7 April 2023 / Accepted: 4 November 2023 / Published online: 12 December 2023

Abstract
Sleep stage classification is essential in evaluating sleep quality. Sleep disorders disrupt the periodicity of sleep stages, especially the common obstructive sleep apnea (OSA). Many methods only consider how to effectively extract features from physiological signals to classify sleep stages, ignoring the impact of OSA on sleep staging. We propose a structured sleep staging network (SSleepNet) based on OSA to solve the above problem. This research focused on the effect of sleep apnea patients with different severity on sleep staging performance and how to reduce this effect. Considering that the transfer relationship between sleep stages of OSA subjects is different, SSleepNet learns comprehensive features and transfer relationships to improve the sleep staging performance. First, the network uses the multi-scale feature extraction (MSFE) module to learn rich features. Second, the network uses a structured learning module (SLM) to understand the transfer relationship between sleep stages, reducing the impact of OSA on sleep stages and making the network more universal. We validate the model on two datasets. The experimental results show that the detection accuracy can reach 84.6% on the Sleep-EDF-2013 dataset. The detection accuracy decreased slightly with the increase of OSA severity on the Sleep Heart Health Study (SHHS) dataset. The accuracy of healthy subjects to severe OSA subjects ranged from 79.8 to 78.4%, with a difference of only 1.4%. It shows that the SSleepNet can perform better sleep staging for healthy and OSA subjects.

Keywords Sleep stage · Obstructive sleep apnea · Deep learning · Structured learning

Introduction

Sleep is an essential primary physiological activity. Sleep can restore the spirit, relieve fatigue, improve immunity, and resist diseases. Poor sleep quality seriously affects human mental state and brain thinking ability and increases the incidence rate of hypertension, stroke, and heart disease [1].

✉ Jinbao Li
lijinb@sdas.org

Xingfeng Lv
lvxingfeng@hlju.edu.cn

Jun Ma
76159521@qq.com

Qianqian Ren
Renqianqian@hlju.edu.cn

[1] Department of Computer Science and Technology, Heilongjiang University, Harbin 150080, China

[2] Shandong Artificial Intelligence Institute , Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China

Sleep stage classification is primary research to evaluate sleep quality. The sleep stages are classified by analyzing the features of several polysomnographic (PSG) signals, such as electrooculographic (EOG), electroencephalographic (EEG), electromyographic (EMG) signals, and so on. Sleep experts divide the sleep records into several consecutive epochs, and each epoch marks the sleep stage according to the standard. There are two common standards: Rechtschaffen and Kales rules (R&K) and the American Academy of Sleep Medicine (AASM) [2]. R&K rules divide the sleep stages into wake, rapid eye movement (REM), and non-rapid eye movement (NREM). NREM includes stage 1, stage 2, stage 3, and stage 4. For the AASM rules, NREM is further divided into three stages, referred to as N1, N2, and N3. They merged stage 3 and stage 4 into stage N3. If sleep experts manually analyze these signals, this process is time-consuming, laborious, and expensive. Therefore, machine learning and deep learning methods are mainly used for automatic sleep staging.

Obstructive sleep apnea (OSA) is a common sleep disorder, mainly caused by intermittent sleep apnea caused by

upper airway obstruction [3]. OSA leads to an increase in transition stage N1 and changes the sleep structure [4]. During the whole night, healthy people experience the circulation sleep process. For OSA subjects, the transfer relationship between sleep stages also changes. The OSA severity has different effects on sleep structure. The OSA severity is mainly determined by the apnea-hypopnea index (AHI). AHI refers to the number of apnea per hour. If AHI<5, the subject is a healthy subject. If $5 \leq$AHI<15, the subject is judged as a mild OSA. If $15 \leq$AHI<30, the subject is considered a moderate OSA. If AHI≥ 30, the subject is regarded as a severe OSA. It is worth studying how OSA severity affects sleep staging performance and how to design models to reduce this effect.

Many researchers use predefined rules, machine learning, and deep learning methods for automatic sleep staging based on various physiological signals [5–11]. Combining various physiological signals can improve sleep staging performance and lead to poor sleep comfort. Therefore, many studies use single-channel EEG signals for sleep staging. In the predefined rules, the definition of threshold affects the classification performance. In the machine learning methods, classification performance depends on feature extraction and classifier selection. These methods require the prior knowledge to analyze physiological signals and manual feature extraction. So some researchers use deep learning models to learn features from physiological signals.

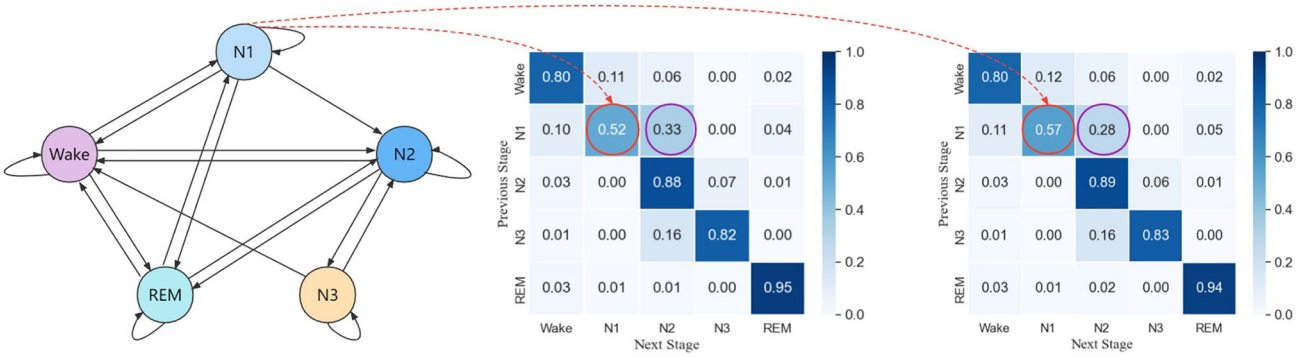
Some researchers use the convolutional neural network (CNN), recurrent neural network (RNN), or hybrid network models for automatic sleep staging. CNN automatically learns features from the raw physiological signal or time-frequency images. Tsinalis et al. [12] used the raw EEG signals to learn features by two-layer convolution and pooling operation. Phan et al. [13, 14] used the short-time Fourier transform to convert the EEG signal into time-frequency images. They added a multi-task classification layer for joint classification and prediction. Many studies use the RNN to learn the temporal features of sleep stages. Michielli et al. [15] used a cascaded RNN based on long short-term memory (LSTM) blocks to classify sleep stages. Phan et al. [16] proposed the dual RNN with attention to learn the temporal features within epochs and long epoch sequences. The EEG signal as input, RNN is complex and needs more training time. Therefore, many researchers build hybrid network models. Supratak et al. [17] proposed deepsleepnet, including two CNN branches and one LSTM layer, which belongs to two-stage training. To further simplify the model, these researchers designed an end-to-end tinysleepnet model [18]. Seo et al. [19] proposed the modified resnet-50 to learn the features within sleep stages and used bidirectional long short-term memory (Bi-LSTM) to obtain the temporal features, which can get an average accuracy of 83.9%. Mousavi et al. [20] used the dual-CNN to learn the internal features of each epoch and utilized the RNN with attention to discover the most relevant part of the input sequence.

Perslev et al. [21] proposed the U-Time model based on the U-Net architecture. Jia et al. [22] proposed SalientSleepNet to detect the salient wave. The model is a temporal fully convolutional network based on the U2-Net architecture. Zhang et al. [23] used a "Dual-CNN" to process the temporal signals and time-frequency simultaneously. They combined CNN and RNN, and a Markov chain model fine-tuned the final results. Eldele et al. [24] proposed an attention-based deep learning architecture called AttnSleep to classify sleep stages using EEG signals. This architecture starts with the feature extraction module based on a multi-resolution convolutional neural network. Then an adaptive feature recalibration improves the quality of the extracted features. Yang et al. [25] combined a deep one-dimensional convolutional neural network and a hidden Markov model (HMM). They leveraged CNN to extract features for epoch-wise classification and HMM to get prior information on adjacent EEG epochs for subject-wise classification.

These deep learning models mainly study the sleep stages of healthy subjects. They classify sleep stages by learning the features of each epoch and the temporal features of multiple epochs, ignoring the transition relationship between sleep stages and the impact of OSA. There are transfer structures between sleep stages. The transfer structures are differences between healthy and OSA subjects on the SHHS dataset, as shown in Fig. 1. For healthy subjects, the transition probability of N1$\rightarrow$N1 is 0.52, and the transition probability of N1$\rightarrow$N2 is 0.33. When OSA occurs, the number of N1 stages increases, the transition probability of N1$\rightarrow$N1 increases to 0.57, and the transition probability of N1$\rightarrow$N2 is 0.28. For OSA subjects, the transition probability changes. We designed a structured sleep stage network (SSleepNet) to learn better the transition probability between sleep stages. The network uses the multi-scale convolution neural network module to learn the rich internal features of epochs and uses the structured learning module to learn the temporal features between epochs and the transfer structure between sleep stages. The whole model is an end-to-end deep learning network.

The main contributions are described as follows:

1. We propose a structured sleep staging network named SSleepNet for the sleep stages based on OSA subjects. We utilize a structured learning module (SLM) to learn the transfer structure between sleep stages and to reduce the impact of OSA on sleep stages.
2. We design a multi-scale feature extraction (MSFE) block for learning the high-frequency, low-frequency, and temporal features within an epoch. This block improves the sleep staging performance using rich features.



(a)Transition diagram of sleep stages (b)Transition matrix of healthy subjects (c)Transition matrix of OSA subjects

Fig. 1 Transition diagram and transition matrix of healthy and OSA subjects

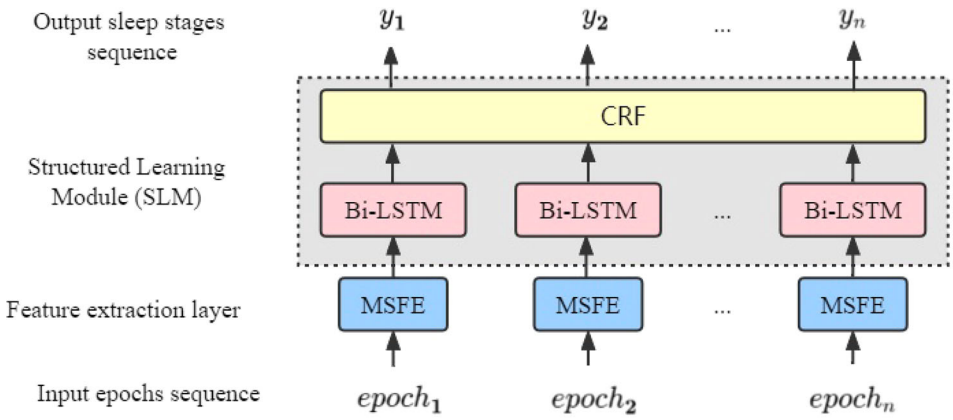


Fig. 2 The overall network framework of SSleepNet

3. We perform extensive experiments on two public datasets, and experimental results demonstrate that our SSleepNet model achieves good sleep staging performance for healthy and OSA subjects.

Materials and methods

OSA has a specific impact on the sleep stage structure, and the transfer structure between sleep stages is different. Therefore, we designed an SSleepNet based on OSA for sleep stage classification. The overall network framework is shown in Fig. 2. The model mainly consists of the multi-scale feature extractor (MSFE) and the structured learning module (SLM). First, the MSFE with three-branch CNN architectures is exploited to extract the features from a 30 s EEG signal (an epoch). In particular, it extracts low-frequency, high-frequency and temporal features by the different convolution branches. Second, we develop an SLM to capture subjects' transfer structure between sleep stages. The SLM mainly uses Bi-LSTM to obtain temporal features and conditional random field (CRF) to learn transfer structure between sleep stages labels. Third, the softmax activation function outputs the sleep stages.

Datasets

We used two public datasets to verify the effectiveness, namely, Sleep-EDF-2013 and Sleep Heart Health Study (SHHS). The Sleep-EDF-2013 was obtained from the PhysioBank [26, 27]. This dataset includes PSG records of 20 healthy subjects. Except that the 13th subject only contains the one-night data, the others have the two-night data. Only the Fpz-Cz channel EEG signal was used in our experiment, and the frequency was 100Hz. According to the R&K rule, the experts marked sleep with a duration of 30 s as wake, stage 1, stage 2, stage 3, stage 4, REM, movement time, and unscored. We excluded some movement and unscored epochs. We merged stage 3 and stage 4 into N3 according to the AASM rule. Each epoch is marked as Wake, N1, N2, N3, and REM. The statistical information of each sleep stage is shown in Table 1.

SHHS is a multi-center research cohort study implemented by the national heart, lung, and Blood Institute of the United States [28, 29]. It is mainly used to study the correlation between sleep disorders and high risk of cardiovascular diseases or other diseases. SHHS-1 contains 5793 PSG records from 6441 subjects over 40 years old. These

Table 1 Statistics of sleep stages on the Sleep-EDF-2013 dataset

Subjects	Wake	N1	N2	N3	REM	Total
20	8285	2804	17,799	5703	7717	42,308
	19.6%	6.6%	42.1%	13.5%	18.2%	100%

PSG records include two EEG channels (C4-A1 and C3-A2), two ophthalmic signal channels, an EMG signal channel, an ECG signal channel, and so on. In this experiment, we only use the EEG signal of the C4-A1 channel with a sampling rate of 125Hz. The subjects in the dataset had sleep-related diseases, such as lung disease, sleep apnea syndrome, cardiovascular disease, and coronary artery disease. These diseases may lead to deviations in the training model. We filtered the data using the following preprocessing steps to minimize this impact. First, according to the OSA severity, the subjects can be divided into four categories: healthy, mild OSA, moderate OSA, and severe OSA. Second, the sleep time at night is more than 7 h. The N3 sleep stage accounts for at least 5% of the whole sleep stage, REM accounts for at least 15%, and sleep efficiency is at least 75%. A total of 780 subjects are selected according to the above steps. The sleep stages statistics of OSA subjects is shown in Table 2. From the statistics, we can find that the proportion of sleep stages in severe OSA has changed. Among them, the percentage of Wake, N1, and N2 sleep stages increased, and the percentage of N3 and REM sleep stages decreased.

Feature extraction

Features are essential to sleep stage classification methods. CNN can capture the local correlation and spatial invariance of the information. So we developed MSFE module to learn rich features based on CNN. Figure 3 shows the MSFE module for feature extraction from raw single-channel EEG signals. The MSFE module consists of three-branch CNN architectures. Branch1 extracts high-frequency features by the small kernel convolutions. Branch2 extracts low-frequency features by the big kernel convolutions. Branch3 extracts temporal features by the temporal block with causal convolution. Because different sleep stages have their characteristic waves. For example, the wake stage mainly includes α (8–13 Hz) and β (14–30 Hz) waves. We can learn frequency information through different convolutional kernel sizes. Additionally, sleep is a temporal process, so the temporal features within an epoch also impact the classification performance. The features extracted by MSFE concatenate together and input into the SLM to the structural features of the sleep stages sequence.

The specific parameters of the three branches are shown in Table 3. The selection of these parameters is analyzed in the discussion section. The first convolution branch uses the convolution of (50, 1) to extract high-frequency features, with the stride of (6, 1) and 64 convolution kernels. The module uses max pooling to reduce dimension and remove redundant information. Then, it uses a dropout operation to prevent over-fitting. After three repeated convolution operations, the first branch outputs the feature F_{i1}. The second convolution branch uses a big convolution kernel with a size of (400, 1) to extract low-frequency features. Other operations are the same as the first branch and output the feature F_{i2}. The third convolution branch uses the 4-layer temporal block. Each temporal block layer includes two causal convolutions, two dropouts, and a residual connection. The convolution kernel size of each layer is (7, 1), and the dilated factor is 5. The third branch outputs the feature F_{i3}. MSFE uses the convolution structure of three branches to obtain features. These features concatenate to get more comprehensive features $F_i = Concat(F_{i1}, F_{i2}, F_{i3})$, which are input into SLM through dropout.

Structured learning module

We use SLM to learn the transfer relationship between different sleep stages based on MSFE. SLM consists of Bi-LSTM and CRF, where Bi-LSTM learns the temporal features of sequences, and CRF learns the transfer between labels. The input to SLM is a sequence of features F=(F_1, F_2,…, F_n) from the MSFE module. The specific structure is shown in

Table 2 Statistics of sleep stages on the SHHS dataset

OSA severity	Subjects	Wake	N1	N2	N3	REM	Total
Healthy	312	41,480	10,834	145,894	57,061	64,194	319,463
		13.0%	3.4%	45.7%	17.9%	20.1%	100%
Mild OSA	284	37,807	10,572	13,2328	51,745	58,718	291,170
		13.0%	3.6%	45.4%	17.8%	20.1%	100%
Moderate OSA	133	18,040	4898	64,061	22,749	26,281	136,029
		13.3%	3.6%	47.1%	16.7%	19.3%	100%
Severe OSA	51	7160	2267	24,309	8915	9598	52,249
		13.7%	4.3%	46.5%	17.1%	18.4%	100%

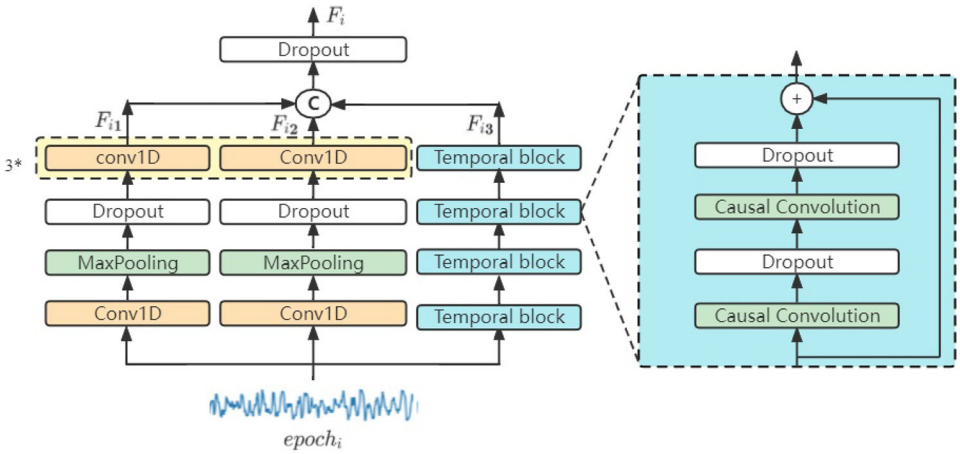


Fig. 3 Structure of multi-scale feature extraction (MSFE)

Table 3 Parameter table of MSFE module

Branch	Type	# filter	Size	Stride	Drop rate
Branch1	Conv1D	64	(50, 1)	(6, 1)	–
	Max pooling	–	(8, 1)	(8, 1)	–
	Dropout	–	–	–	0.5
	Conv1D*3	128	(4, 1)	(4, 1)	–
Branch2	Conv1D	64	(400, 1)	(50, 1)	–
	Max pooling	–	(4, 1)	(4, 1)	–
	Dropout	–	–	–	0.5
	Conv1D*3	128	(6, 1)	(1, 1)	–
Branch3	Temporal*4	64	(7, 1)	(1, 1)	–

Fig. 4. In Bi-LSTM, a jump connection is used to add MSFE and Bi-LSTM. Output features do not lose multi-scale features and learn the temporal features between epochs. h_i^f represents the features of the forward hidden layer, and h_i^b describes the features of the backward hidden layer. The fully connected (FC) in Bi-LSTM is used to ensure the consistency of dimensions. After the output of the features by Bi-LSTM is the second FC operation. The FC output performs two operations, respectively. One is to calculate the loss of the temporal features between sleep stages with the softmax function. The other is that CRF continues to learn the transfer relationship between labels.

Loss function

Inspired by multi-task learning, temporal feature and transfer relationship between sleep stages can be regarded as two tasks. Joint training of these two tasks can better learn the structured sleep staging results of different subjects and improve the performance of sleep staging. For the first task, the output of Bi-LSTM is used to generate the prediction label through softmax and calculate the cross-entropy loss as follows:

$$\ell_c = -\sum_{c=1}^{N} y_c \log(y_c{}') \tag{1}$$

where N represents the number of sleep stages, y_c represents the probability of actual classification, and $y_c{}'$ describes the probability of classification predicted by the model.

Each classification in the sleep stage is unbalanced. In the Sleep-EDF-2013 dataset, the wake accounts for 19.6%, while the N1 stage accounts for only 6.6%. N1 belongs to the minority classification. Unbalanced classification in deep learning may be more suitable for majority classification. There are three kinds of technologies to deal with unbalanced classification: sampling, threshold-moving, and adjusting cost or weight. Sampling eliminates or reduces data imbalance by changing training data distribution, such as over-sampling, clustering, under-sampling, and so on. Threshold-moving is to change the decision threshold to focus on the minority classification. Adjusting cost, also called cost-sensitive learning, biases the minority classification by adjusting the cost or weight of different categories to improve classification performance. In our SSleepNet, we use cost-sensitive learning to give weights to each classification. The definition of class cross-entropy with weight is as below:

$$\ell_{wc} = -w_c \sum_{c=1}^{n} y_c \log(y_c{}') \tag{2}$$

where w_c represents the weight of each classification. In our experiments, w_c is set as [1, 1.5, 1, 1, 1].

The second task in the multi-task model is to use CRF to learn the transfer relation between labels and optimize the output sequence. The loss function of CRF consists of two parts: the score of the actual path and the total score of all paths. The score of the actual path should be the highest of all paths, and there is only one path. CRF uses the transmission

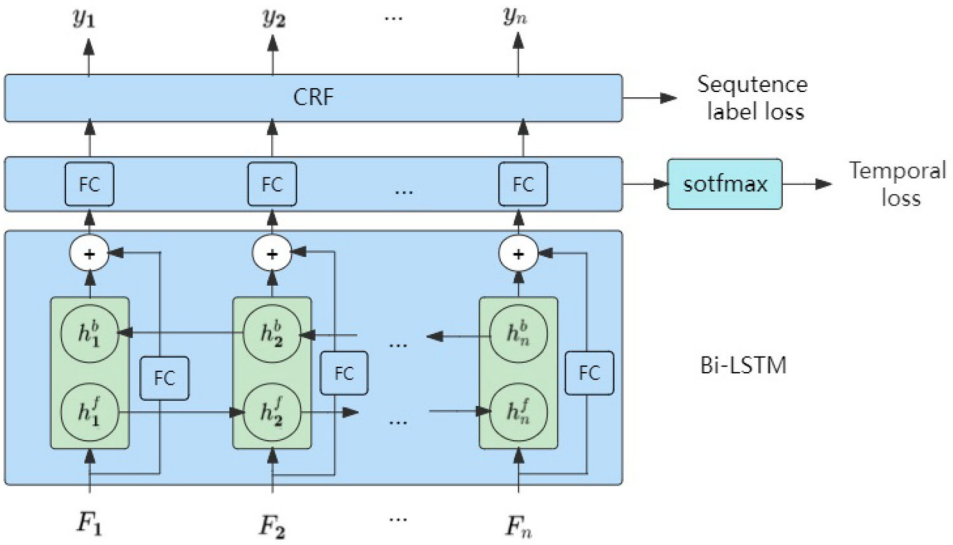


Fig. 4 Framework of structured learning module (SLM)

and transfer scores to calculate the score. These two scores are the parameters of CRF. CRF calculates the loss function by comparing the actual path score with all path scores expressed by ℓ_e. To make the loss from temporal feature learning and transfer relation learning part at the same scale, we use a λ, which is a hyper-parameter. In our experiment, this parameter is set to 10. The network adds the L2-norm regularization loss ℓ_2 to prevent over-fitting. Therefore, the multi-task loss function is defined as follows:

$$\ell_{mtl} = \lambda \ell_{wc} + \ell_e + \ell_2 \tag{3}$$

Performance evaluation

We evaluate and compare the performance of different methods using classification Accuracy, Recall, Precision, Kappa coefficient, and F1 score. They are defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \times 100\% \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{6}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \tag{7}$$

where TP, FP, TN, and FN represent the number of true positive, false positive, true negative, and false negative epochs. The proportion of the correctly identified epochs is measured by sensitivity. Specificity reflects the detection effect of negative samples.

In the performance evaluation, we also adopted macro-averaged F1-score (MF1), Kappa coefficient, and confusion matrix to evaluate the general performance of all classes. MF1 is a common metric to evaluate the performance of the models and can be defined as below:

$$MF1 = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{8}$$

where N is the number of sleep stage classification.

Kappa coefficient is to characterize the interrater agreement level and can be calculated as below:

$$Kappa = \frac{p_c - p_e}{p_t - p_e} \tag{9}$$

where p_c represents the number of correctly scored stages, p_t represents the total number of stages, and p_e represents the expected number of agreements for each sleep stage.

The confusion matrix also is used. Each row of the confusion matrix represents the epoch in actual labels, while each column represents the epoch in the predicted labels. We also standardized the confusion matrix by rows to obtain different probabilities. Use colors with different shades to represent the probability. The darker the color, the greater the probability. On the contrary, the smaller the probability.

Results

Experimental setup

To verify the effectiveness of the SSleepNet, we design the three groups of experiments. First, using the Sleep-EDF-2013 dataset to demonstrate the efficacy for healthy subjects. Second, in the SHHS dataset, 50 subjects from healthy subjects, mild, moderate, and severe OSA subjects are selected for training to explore the performance of OSA subjects.

Finally, ablation experiments verify the role of each module.

The experiment divides the dataset into training, verification, and test set. The training set trains the parameters in the model, the verification set selects the optimal model, and the test set evaluates the model's performance. Sleep-EDF-2013 dataset uses 20-fold cross-validation. For 20 subjects, we adopted the leave one subject out method. We selected one subject as the test set for each fold, four subjects as the validation set, and 15 subjects as the test set. The prediction results of all 20 people are combined and compared with the expert labels to obtain the evaluation performance of the network. In the SHHS dataset, we choose 200 subjects. The 50 subjects of each type of OSA patient are selected to train, verify, and test. We adopted the tenfold cross-validation. 10% of subjects are chosen as the test set, 10% of the remaining 90% are selected as the verification set, and the rest are the training set.

In the experiment, the details of the model implementation, the network model adopts the Adam optimization, the learning rate is 1e−4, the epoch of training is 10, the batch size is 16, and the sequence length is 10. The number of layers in the time convolution model is 4, the convolution kernel size is (7, 1), and the dilated factor is 5. The number of hidden units in Bi-LSTM is 128. The experimental implementation uses the Tensorflow framework, and the hardware adopts NVIDIA GTX 2080 Ti GPU.

Experimental results

The results on the Sleep-EDF-2013 dataset

We used the Sleep-EDF-2013 dataset to verify that the network can get good sleep staging performance in healthy subjects. The experimental results are shown in Table 4. The table shows the confusion matrix and the performance of each sleep stage score. The average accuracy on the test set is 84.6%, the MF1 score is 79.5%, and the kappa coefficient is 0.79. For the classification results of each sleep stage, precision, recall, and F1 scores are used for evaluation. The precision and F1 scores of the Wake are 90.5% and 89.5%, and the recall of the N3 is 91.4%. N1 belongs to the minority classification, with the lowest classification performance, and other stages show good performance.

To visually observe the classification results of SSleepNet, Fig. 5 shows the sleep stage histogram of the SC415 subject. The horizontal axis represents the number of 30 s epoch. The vertical axis represents five sleep stages. In the figure, (a) represents the sleep stage histogram labeled by human experts for each epoch, while (b) represents the histogram predicted by the SSleepNet. The classification accuracy of the SSleepNet is 92.7%, the Kappa coefficient is 0.90, and the F1 score is 83.9%. Many N1 are misclassified as REM, and there are a few misclassifications between N3 and N2. Most of the other classification results are very close to the results marked by human experts. The experimental results show that the proposed model performs better in sleep stage classification.

The results on the SHHS dataset

The model was trained on the SHHS dataset to analyze the impact of OSA on sleep staging. Figure 6 shows the confusion matrix of OSA subjects. For healthy subjects, the average accuracy is 79.8%, and the accuracy of Wake, N2, and REM can reach more than 79%. As a minority classification, N1 has the lowest accuracy of 40%, and 25% is misclassified as N2. For mild OSA subjects, the accuracy of REM increased, but the accuracy of N1 decreased. The overall average accuracy is the same as that of healthy subjects, which is 79.8%. For the moderate OSA subjects, compared with healthy subjects, the accuracy of the N1 stage and N3 stage decreased by 7% and 4%, the N2 stage decreased by 1%, REM increased by 5%, and the accuracy of the Wake stage remained unchanged. The sleep staging accuracy of the model in this set reached 79.1%, and the average accuracy decreased by 0.7% compared with healthy subjects. The main reason is that with the increase in the OSA severity, the number of N1 classifications increases, and the transfer of sleep stages is more complex. The accuracy is 78.4% for severe OSA subjects, which decreased by 0.7% compared with moderate OSA. The accuracy of Wake in moderate OSA is 79%, but in severe OSA patients, the number of Wake classifications increased, and the accuracy increased by 3%. The accuracy of N1 and N3 decreased by 2% and 1%, respectively.

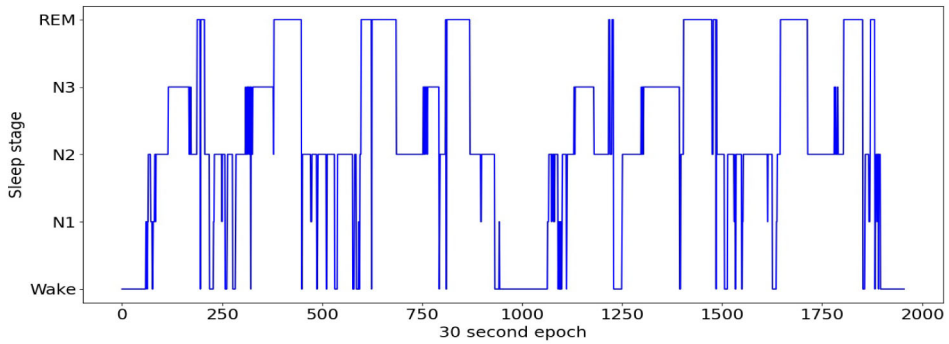
To observe the effect of OSA severity on sleep staging, Fig. 7 shows the performance of per-class on OSA severity. And it also shows the average accuracy, F1 score, and kappa coefficient of these subjects. From the figure, we can find that with the increase in OSA severity, the performance of the N2 is unchanged. The Recall and F1 scores of N3 show a downward trend. The changes in Wake and N1 are more complex. The Recall and F1 scores of the Wake increased slightly, while the Precision and F1 scores of N1 decreased. The F1 score of REM changes little with the increase in OSA severity. We can find that OSA with different severity affects each sleep stage. The average accuracy, F1 score, and kappa coefficient of all sleep stage classifications decreased with the increase in OSA severity.

Ablation experiments

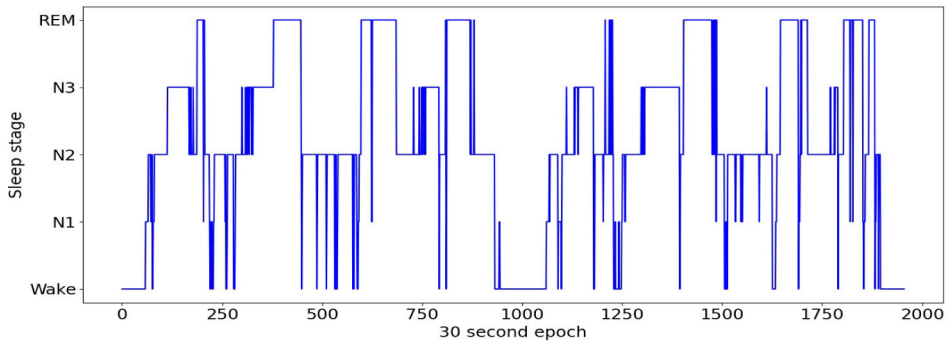
The SSleepNet is based on MSFE, CRF, and loss function. We perform the ablation experiments on the Sleep-EDF-2013 dataset and SHHS subset to analyze the effectiveness of each part. The ablation experiment forms five variables by elimi-

Table 4 The confusion matrix and per-class results on the Sleep-EDF-2013 dataset

	SSleepNet output					Per-class results (%)		
	Wake	N1	N2	N3	REM	Precision	Recall	F1
Wake	7327	522	164	82	190	90.5	88.4	89.5
N1	442	1407	516	13	426	50.3	50.2	50.2
N2	163	517	15266	962	891	89.5	85.8	87.6
N3	24	5	445	5214	15	83.0	91.4	87.0
REM	137	347	665	9	6559	81.2	85.0	83.0



(a) Hypnogram of Human experts



(b) Hypnogram of SSleepNet

Fig. 5 Human experts and SSleepNet hypnograms of SC415 subject

nating different positions, and the first four variable modules do not consider the class-sensitive loss function.

Variant 1: (MSFE_1): contains only two-branch CNN and Bi-LSTM, without the third branch convolution (Branch3).

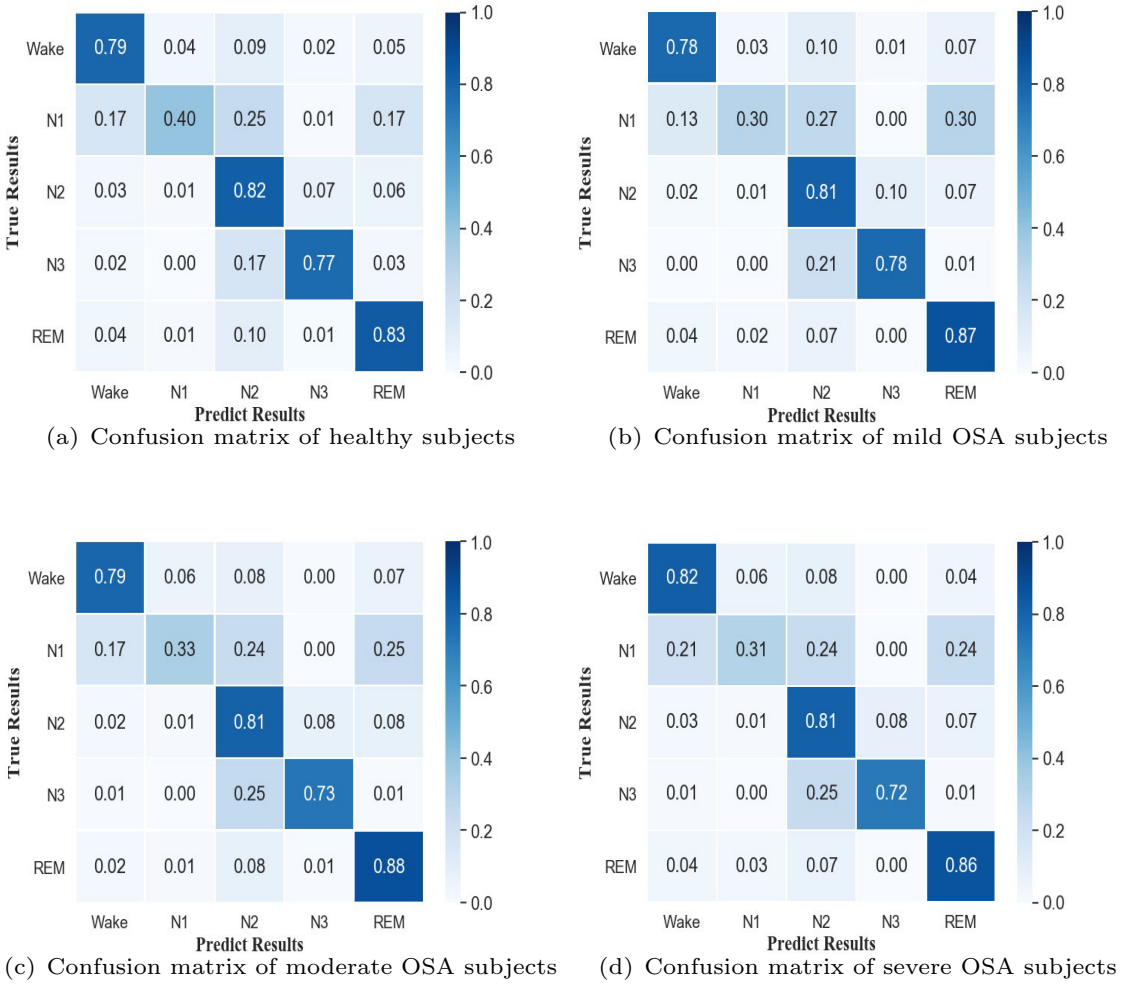
Variant 2: (MSFE_2): based on Variant 1, an epoch internal temporal feature extension (Branch3) is added to form a three-branch multi-scale feature learning module.

Variant 3: (MSFE_1 + CRF): based on Variant 1, add CRF to learn the transfer relationship between sleep stages.

Variant 4: (MSFE_2 + CRF): add CRF based on Variant 2, learn the temporal features within the epoch and the transfer relationship between labels simultaneously.

SSleepNet: (structured sleep network): based on Variant 4, add a class-sensitive loss function to strengthen the learning of the minority classification.

Table 5 shows the classification performance of different variable modules. The results of ablation experiments verify the role of each variable. First, the third branch convolution can improve classification performance. The classification accuracy of Variant 1 is 83.4%, and that of Variant 2 is 83.9%. The classification performance of Variant 2 is better than that of Variant 1. Variant 2 adds the Branch3 to learn the temporal features inside the epoch, improving accuracy by 0.5% and MF1 score by 0.5%. This result shows that the temporal features within the epoch can also learn valuable features for classification. Observing Variant 3 and Variant 4 can obtain the same conclusion. Secondly, by comparing the performance of Variant 1 and Variant 3, or Variant 2 and Variant 4, we find that CRF is also important for sleep staging, which improves the accuracy by 0.5%. CRF can learn the transfer relationship between different sleep stages.



(a) Confusion matrix of healthy subjects (b) Confusion matrix of mild OSA subjects

(c) Confusion matrix of moderate OSA subjects (d) Confusion matrix of severe OSA subjects

Fig. 6 Sleep stage confusion matrix of subjects with OSA severity

Finally, SSleepNet added the class-sensitive loss function to alleviate the problem of unbalanced classification. The average accuracy is improved by 0.2% and MF1 by 0.7%.

To further validate the role of each variant module, we also perform the ablation experiments on SHHS subsets. SHHS dataset are divied into four subsets, including the health subjects, mild OSA subjects, moderate OSA subjects and severe OSA subjects. By using this subset to verify the sleep staging performance of subjects with different OSA severity. Table 6 shows the classification performance of different variable modules on the first subset (healthy subjects). Through this ablation experiment, we can find that in the SHHS subset, each variable module has varying degrees of improvement in performance. Especially Variant 2 to Variant 1 improved accuracy by 0.4%, indicating that Branch3 improves performance by learning temporal features. Additionally, Variant 4 to Variant 2 improved accuracy by 0.2%, indicating that CRF improves performance.

Discussion

Sensitivity analysis in MSFE

MSFE module is one key component of SSleepNet, it is important to study the size of convolution kernels of each branch. We fix the other parameters and test different size of convolution kernels on Sleep-EDF-2013 dataset. In the branch1, we design our model using (25, 1), (50, 1), (75, 1) and (100, 1). In the branch2, we design our model using (200, 1), (400, 1), (600, 1) and (800, 1). In the branch3, we design temporal block with small convolution kernels. The convolutional kernel of this branch has a weak impact on performance. Figure 8a shows the model performance in terms of accuracy and F1 score in the Branch1. Figure 8b shows the model performance in the Branch2. We can observe that the size of the convolutional kernel affects the performance of the model, as using smaller convolutional kernels can extract

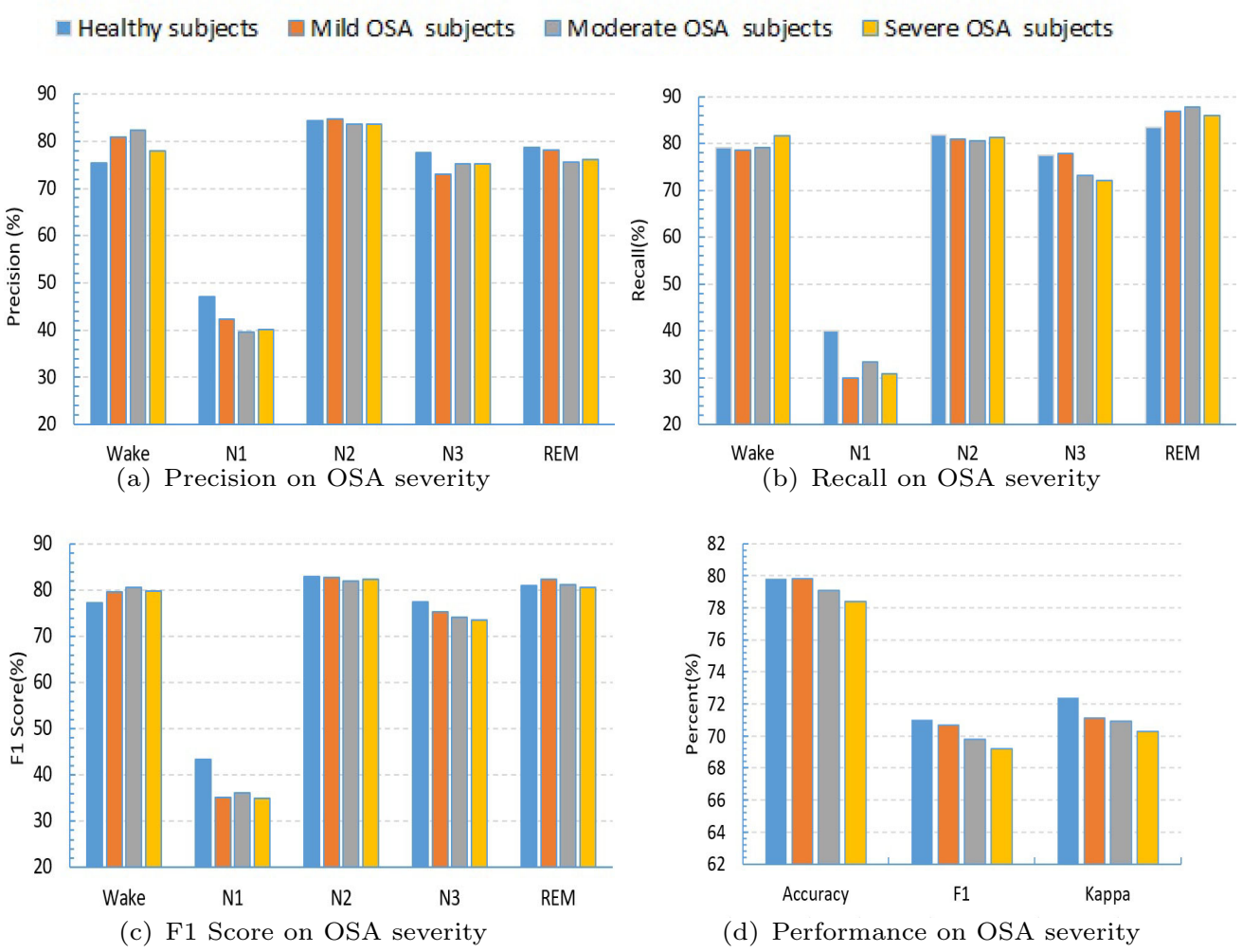


Fig. 7 Per-class performance on OSA severity on the SHHS dataset

Table 5 Results of the ablation experiments on Sleep-EDF-2013 dataset

	MSFE_1	MSFE_2	CRF	Loss	Accuracy (%)	MF1 (%)	Kappa
Variant 1	✓				83.4	78.4	0.77
Variant 2		✓			83.9	78.9	0.78
Variant 3	✓		✓		83.9	78.4	0.78
Variant 4		✓	✓		84.4	78.8	0.79
SSleepNet		✓	✓	✓	84.6	79.5	0.79

Table 6 Results of the ablation experiments on SHHS subset

	MSFE_1	MSFE_2	CRF	Loss	Accuracy (%)	MF1 (%)	Kappa
Variant 1	✓				80.7	71.4	0.72
Variant 2		✓			81.1	72.0	0.73
Variant 3	✓		✓		81.1	72.3	0.73
Variant 4		✓	✓		81.3	72.4	0.73
SSleepNet		✓	✓	✓	81.5	72.8	0.73

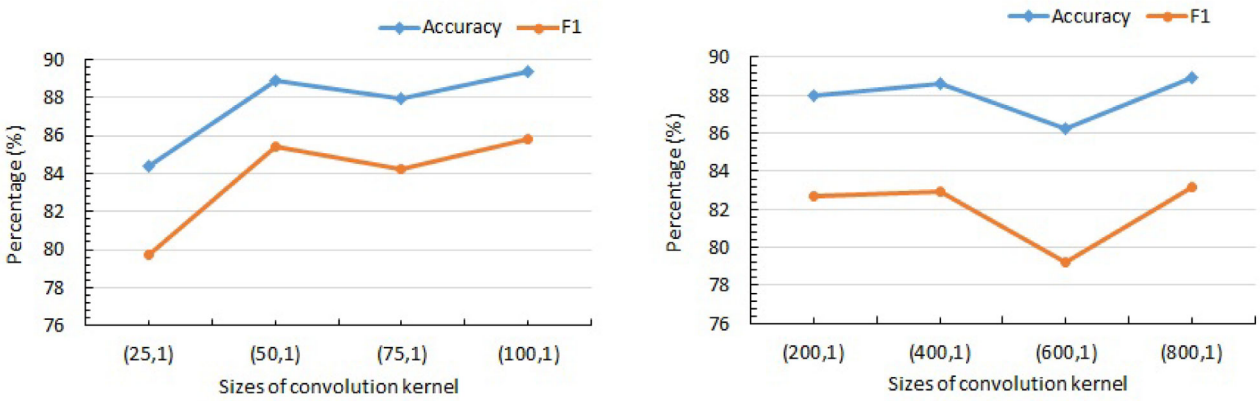


Fig. 8 Sensitivity analysis for the sizes of convolution kernel in MSFE

Table 7 Performance comparison between SSleepNet and other literature on Sleep-EDF-2013 dataset

Method	Classifier	Results (%)		Per-class F1-score (%)				
		Acc	MF1	Wake	N1	N2	N3	REM
Tsinalis et al. [12]	CNN	78.9	73.7	65.4	43.7	80.6	84.9	74.5
Supratak et al. [17]	DeepSleepNet	82.0	76.9	84.7	46.6	85.9	84.8	82.4
Phan et al. [13]	1-max CNN	79.8	72.0	77.0	33.3	86.8	86.3	76.4
Seo et al. [19]	IITNet	83.9	77.6	87.7	43.4	87.7	86.7	82.5
Zhu et al. [30]	CNN+Attention	82.8	77.8	90.3	47.1	86.0	82.1	83.2
Yang et al. [25]	1D-CNN-HMM	83.9	76.9	87.8	35.1	86.6	90.5	86.8
Li et al. [31]	CAttSleepNet	84.1	78.2	89.6	47.1	87.2	85.0	82.1
Qu et al. [32]	Res+Attention	84.3	79.0	90.2	48.3	87.8	85.6	83.0
SSleepNet	MSFE+SLM	84.6	79.5	89.5	50.2	87.6	87.0	83.0

high-frequency features, while larger convolutional kernels can extract low-frequency features. In our experiments, we eventually set (50, 1) and (400, 1) in the branch1 and branch2.

Comparision with other literature

The SSleepNet model uses MSFE to learn comprehensive features and SLM to learn the temporal features and the transfer relationship between sleep stages. This network performs well on the dataset containing healthy subjects and analyzes its impact on sleep staging for OSA subjects with different severity. To illustrate the progressive nature of the network performance, we compare the model with other literature regarding average accuracy (Acc), MF1, and F1 scores of each sleep stage on two datasets. The performance comparison on Sleep-EDF-2013 dataset are shown in Table 7. For a fair comparison, all networks are trained on the same EEG signal of the Fpz-Cz channel. These networks use 20-fold cross-validation and independent subjects as the test set. We also use the same SHHS dataset as other literature and conduct training and testing. The performance comparison on SHHS dataset are shown in Table 8.

Tsinalis et al. [12] used CNN to learn features from the raw EEG signal directly, and the model can obtain an average accuracy of 78.9%. Supratak et al. [17] proposed a deep sleep net (DeepSleepNet) to learn features from the raw EEG signal, and the model adopts two-stage training. In the first stage, the dual branch CNN training model is used to learn the features of different frequencies. In the second stage, the learned features are input into Bi-LSTM to continue training to obtain the sleep staging results. The accuracy can reach 82.0%. Phan et al. [13] preprocessed the raw EEG by short-time Fourier transform. They got the power spectrum and extracted the features by 1-max CNN, with an average accuracy of 79.8%. Considering that sleep experts pay more attention to individual features in analyzing signals, Zhu et al. [30] proposed a CNN model integrating attention mechanisms, with an accuracy of 82.8%. The F1 scores of the Wake and N1 stages are also higher than other methods. Seo et al. [19] proposed an intra- and inter-epoch temporal context network (IITNet). IITNet extracts representative features at a sub-epoch level by a residual neural network and captures intra- and inter-epoch temporal contexts from the sequence of the features via Bi-LSTM. This model can achieve an accuracy of 83.9%. Yang

Table 8 Performance comparison between SSleepNet and other literature on SHHS dataset

Method	Classifier	Results (%)		Per-class F1-score (%)				
		Acc	MF1	Wake	N1	N2	N3	REM
Supratak et al. [17]	DeepSleepNet	81.0	73.9	85.4	40.5	82.5	79.3	81.9
Eldele et al. [23]	Attnsleep	84.2	75.3	86.7	33.2	87.1	87.1	82.1
SSleepNet	MSFE+SLM	84.3	77.5	86.0	47.8	85.6	82.0	86.3

et al. [25] used CNN and the HMM to learn the temporal features of long epochs and obtained 83.98% accuracy. Li et al. [31] exploited CNN with attention mechanism and Bi-LSTM to learn contextual information intra-epoch and continuous epochs, respectively. However, they did not fully consider the use of large convolutional kernels to extract low-frequency features intra-epochs. Qu et al. [32] proposed a multi-scale deep architecture and utilized the multi-head self-attention module of the transformer model to obtain global temporal context. But they ignored the temporal features within an epoch and the transition relationship between sleep stages. Eldele et al. [23] utilized a multi-head attention mechanism to capture the temporal correlation between features extracted in 30 s epoch. However, it did not exploit the sleep transition rules between sleep stages.

In our experiment, the SSleepNet is a helpful sleep stage classification model. The accuracy on the Sleep-EDF-2013 dataset is 84.6% and MF1 of 79.5%. On the SHHS dataset, accuracy is 84.3% and MF1 is 77.5%. The model's input is only the raw EEG signal and no special preprocessing, which makes the sleep stage classification more concise. The performance is superior to that of other literature, mainly for some reasons. The model learns rich features through MSFE, including low-frequency features, high-frequency features, and temporal features within an epoch. The model learns the temporal sequence features between epochs through module Bi-LSTM and uses CRF to learn the transfer relationship between sleep stages. The model can also perform better for OSA subjects. Cost-sensitive learning is used for minority classification to improve the classification performance of N1. The F1 score of N1 is higher than other literature.

Our study has some limitations. Firstly, we only used physiological signals in PSG for sleep stage classification. The sleep stage is also related to other features of the human body. Some diseases also affect sleep stage classification. To accurately classify and comprehensively evaluate the sleep stage, we will mine information related to sleep quality from the data in electronic medical records in the future. Secondly, for a minority classification in the sleep stage, we will consider using the data augmentation strategy to generate new sleep stage sequences.

Conclusion

We propose an SSleepNet based on OSA for sleep stage classification. The network can learn the transfer relationship between sleep stages and perform well for healthy subjects and OSA patients. The SSleepNet uses the MSFE to learn rich features, the Bi-LSTM to obtain the temporal features between epochs, and CRF to learn the transfer relationship between sleep stages. The network performs better than other methods on the Sleep-EDF-2013 dataset for healthy subjects. Using a single-channel EEG signal, the accuracy can reach 84.6%. Moreover, the ablation experiments verify the effectiveness of each module. On the SHHS dataset, healthy, mild OSA, moderate OSA, and severe OSA subjects were tested, respectively, and the accuracy is 79.8%, 79.8%, 79.1%, and 78.4%, respectively. The experimental results show that the SSleepNet can perform well in healthy people and patients with OSA. With the increase in OSA severity, the classification accuracy decreases slightly. The SSleepNet can perform sleep staging for patients with different severity OSA, making the model application more universal.

Funding This paper is supported by the National Natural Science Foundation of China (62172143), the National Key R&D Program of China under Grant 2020YFB1710200, and the Natural Science Foundation of Heilongjiang Province under Grant LH2019F028, and the Harbin science and technology bureau innovation under Grant 2017RAQXJ131.

Data availability We evaluated our model with the Sleep-EDF-2013 and SHHS datasets. The datasets are publicly available on the Internet http://www.physionet.org/physiobank/database/sleep-edfx/ and http://sleepdata.org/datasets/shhs.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Vallat R, Shah VD, Redline S, Attia P, Walker MP (2020) Broken sleep predicts hardened blood vessels. PLoS Biol 18(6):3000726
2. Berry RB, Budhiraja R, Gottlieb DJ, Gozal D, Iber C, Kapur VK, Marcus CL, Mehra R, Parthasarathy S, Quan SF, Redline S, Strohl KP, Davidson Ward SL, Tangredi MM (2012) Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events. deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. J Clin Sleep Med 8(5): 597–619
3. Benjafield A, Ayas N, Eastwood P, Heinzer R, Ip M, Morrell M, Nunez C, Patel S, Penzel T, Pepin J (2019) Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med 7(8):687–698
4. Korkalainen H, Aakko J, Nikkonen S, Kainulainen S, Leino A, Duce B, Afara IO, Myllymaa S, Töyräs J, Leppänen T (2020) Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea. IEEE J Biomed Health Inform 24(7):2073–2081
5. Chriskos P, Kaitalidou DS, Karakasis G, Frantzidis C, Gkivogkli PT, Bamidis P, Kourtidou-Papadeli C (2017) Automatic sleep stage classification applying machine learning algorithms on EEG recordings. In: 2017 IEEE 30th international symposium on computer-based medical systems (CBMS), pp 435–439
6. Li X, Cui L, Tao S, Chen J, Zhang X, Zhang G (2018) Hyclasss: a hybrid classifier for automatic sleep stage scoring. IEEE J Biomed Health Inform 22(2):375–385
7. Yetton BD, Mcdevitt EA, Cellini N, Shelton C, Mednick SC (2018) Quantifying sleep architecture dynamics and individual differences using big data and Bayesian networks. PLoS ONE 13(4):0194604
8. Mosheyur M, RahmanHassan MI, Bhuiyan RA (2018) Hassan: sleep stage classification using single-channel EOG. Comput Biol Med 102(2):211–220
9. Klok AB, Edin J, Cesari M, Olesen AN, Sorensen HBD (2018) A new fully automated random-forest algorithm for sleep staging. In: 2018 40th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp 4920–4923
10. Chen C, Liu X, Ugon A, Zhang X, Amara A, Garda P, Ganascia JG, Philippe C, Pinna A (2019) Symbolic fusion: a novel decision support algorithm for sleep staging application symbolic fusion: a novel decision support algorithm for sleep staging application. EAI Endorsed Trans Pervasive Health Technol 16(8):4
11. Frantzidis CA, Nday CM, Chriskos P, Polyxeni G, Papadeli C (2020) A review on current trends in automatic sleep staging through bio-signal recordings and future challenges. Sleep Med Rev 55(5):1–34
12. Tsinalis O, Matthews PM, Guo Y, Zafeiriou S (2016) Automatic sleep stage scoring with single_channel EEG using convolutional neural networks. CoRR arXiv:1610.01683
13. Phan H, Andreotti F, Cooray N, Chén OY, De Vos M (2018) DNN filter bank improves 1-max pooling CNN for single-channel EEG automatic sleep stage classification. In: 2018 40th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp 453–456
14. Phan H, Andreotti F, Cooray N, Chén OY, De Vos M (2019) Joint classification and prediction CNN framework for automatic sleep stage classification. IEEE Trans Biomed Eng 66(5):1285–1296
15. Michielli N, Acharya UR, Molinari F (2019) Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals. Comput Biol Med 106:71–81
16. Phan H, Andreotti F, Cooray N, Chén OY, De Vos M (2019) Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. IEEE Trans Neural Syst Rehabil Eng 27(3):400–410
17. Supratak A, Hao D, Chao W, Guo Y (2017) Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Trans Neural Syst Rehabil Eng 25(11):1998–2008
18. Supratak A, Guo Y (2020) Tinysleepnet: an efficient deep learning model for sleep stage scoring based on raw single-channel EEG. In: 2020 42nd annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp 641–644
19. Seo H, Back S, Lee S, Park D, Kim T, Lee K (2020) Intra- and inter-epoch temporal context network (IITNET) using sub-epoch features for automatic sleep scoring on raw single-channel EEG. Biomed Signal Process Control 61:102037
20. Mousavi S, Afghah F, Acharya UR (2019) Sleepeegnet: automated sleep stage scoring with sequence to sequence deep learning approach. PLoS ONE 14(5):0216456
21. Perslev M, Jensen MH, Darkner S, Jennum PJ, Igel C (2019) U-time: a fully convolutional network for time series segmentation applied to sleep staging. In: Advances in neural information processing systems, vol 32. Annual conference on neural information processing systems 2019, pp 4417–4428
22. Jia Z, Lin Y, Wang J, Wang X, Xie P, Zhang Y (2021) Salientsleepnet: multimodal salient wave detection network for sleep staging. In: 2021 international joint conference on artificial intelligence (IJCAI), pp 1–10
23. Zhang L, Chen D, Chen P, Li W, Li X (2021) Dual-CNN based multi-modal sleep scoring with temporal correlation driven fine-tuning. Neurocomputing 420:317–328
24. Eldele E, Chen Z, Liu C, Wu M, Guan C (2021) An attention-based deep learning approach for sleep stage classification with single-channel EEG. IEEE Trans Neural Syst Rehabil Eng 29:809–818
25. Yang B, Zhu X, Liu Y, Liu H (2021) A single-channel EEG based automatic sleep stage classification method leveraging deep one-dimensional convolutional neural network and hidden Markov model. Biomed Signal Process Control 68(2):102581
26. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation 101(23):215–220
27. Rechtshaffen A (1968) A manual of standardized terminology, techniques and scoring systems for sleep stages of human subjects, vol 204. National Institute of Health, New York
28. Zhang GQ, Cui L, Remo M, Tao S, Matthew K, Michael R, Sara M, Daniel M, Susan R (2018) The national sleep research resource: towards a sleep data commons. J Am Med Inform Assoc 25(10):1351–1358
29. Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, Rapoport DM, Redline S, Robbins J, Samet JM, Wahl PW (1997) The sleep heart health study: design, rationale, and methods. Sleep 20(12):1077–1085
30. Zhu T, Luo W, Yu F (2020) Convolution- and attention- based neural network for automated sleep stage classification. Int J Environ Res Public Health 17(11):1–13
31. Li T, Zhang B, Lv H, Hu S, Xu Z, Tuergong Y (2022) Cattsleepnet: automatic end-to-end sleep staging using attention-based deep neural networks on single-channel EEG. Int J Environ Res Public Healthy 19:5199
32. Qu W, Wang Z, Hong H, Chi Z, Feng DD, Grunstein R, Gordon C (2020) A residual based attention model for EEG based sleep staging. IEEE J Biomed Health Inform 24(10):2833–2843