A Temporal-Spectral Fused and Attention-Based Deep Model for Automatic Sleep Staging

Guidan Fu, Yueying Zhou, Peiliang Gong, Pengpai Wang, Wei Shao, and Daoqiang Zhang, *Senior Member, IEEE*

***Abstract*— Sleep staging is a vital process for evaluating sleep quality and diagnosing sleep-related diseases. Most of the existing automatic sleep staging methods focus on time-domain information and often ignore the transformation relationship between sleep stages. To deal with the above problems, we propose a Temporal-Spectral fused and Attention-based deep neural Network model (TSA-Net) for automatic sleep staging, using a single-channel electroencephalogram (EEG) signal. The TSA-Net is composed of a two-stream feature extractor, feature context learning, and conditional random field (CRF). Specifically, the two-stream feature extractor module can automatically extract and fuse EEG features from time and frequency domains, considering that both temporal and spectral features can provide abundant distinguishing information for sleep staging. Subsequently, the feature context learning module learns the dependencies between features using the multi-head self-attention mechanism and outputs a preliminary sleep stage. Finally, the CRF module further applies transition rules to improve classification performance. We evaluate our model on two public datasets, Sleep-EDF-20 and Sleep-EDF-78. In terms of accuracy, the TSA-Net achieves 86.64% and 82.21% on the Fpz-Cz channel, respectively. The experimental results illustrate that our TSA-Net can optimize the performance of sleep staging and achieve better staging performance than state-of-the-art methods.**

***Index Terms*— Sleep staging, EEG, feature fusion, multi-head attention, conditional random field.**

I. Introduction

GOOD sleep is a crucial human physiological activity intimately related to health. Studies have shown that sleep is inextricably linked to human metabolism [1], immune system [2], and memory [3]. Narcolepsy, sleep apnea syndrome, and insomnia are a few examples of sleep-related problems [4]. With the rapid development of modern civilization, people are having trouble in sleeping and are increasingly concerned about the quantity and quality of their sleep.

Sleep staging is essential to evaluating the quality of sleep. Typically, sleep specialists use acquisition equipment for sleep staging based on the entire night polysomnogram (PSG) [5] of the patient, which usually includes electroencephalogram (EEG), electromyogram (EMG), electrooculogram (EOG), and electrocardiogram (ECG). The manual sleep staging methods generally divide the PSG into 30-second segments (called epochs), and then the segments are manually classified by specialists based on relevant criteria. The existing criteria for evaluating the sleep stage include the American Academy of Sleep Medicine standard (AASM) [6], and the Rechtschaffe and Kales (R&K) rule [7]. The R&K rule divides the sleep process into six stages, including wake stage (W), rapid sleep eye movement stage (REM), S1, S2, S3, and S4 of non-rapid eye movement stage. The AASM rule combines S3 and S4 into N3, including W, REM, N1, N2, and N3 of non-rapid eye movement stage. According to the above rules, sleep specialists categorize the stages of sleep based on the distinctive waveforms and frequency characteristics that each stage of sleep exhibits. For instance, K-complex wave and sleep spindles wave are evident in the N2 stage, and the δ wave is more prevalent in the N3 stage of deep sleep. Similarly, the α wave is prominent during the W stage.

Generally, manual labeling and diagnosis in sleep staging need skilled and experienced specialists, which is time-consuming and laborious. Even if some specialists have received training, there are still certain biases in the manual sleep staging procedure. Besides, various specialists may have varying views on how the same PSG data should be evaluated. Considering these factors, it is important to construct objective and automatic sleep staging methods.

In recent years, some automatic sleep staging techniques have gained popularity with the growth of machine learning. Traditional machine learning of sleep staging requires hand-crafted feature extraction and classifier selection. Firstly, the data are preprocessed and filtered to obtain clean and non-impurity signals, then feature extraction is carried out on the preprocessed signals, and valuable features are selected and

Manuscript received 5 August 2022; revised 12 December 2022; accepted 16 January 2023. Date of publication 23 January 2023; date of current version 3 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62136004, Grant 61876082, and Grant 61732006; in part by the Research Fund for International Young Scientists under NSFC Grant 62050410348; and in part by the National Key Research and Development Program of China under Grant 2018YFC2001600 and Grant 2018YFC2001602. *(Guidan Fu and Yueying Zhou are co-first authors.) (Corresponding author: Daoqiang Zhang.)*

The authors are with the MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: fuguidan@nuaa.edu.cn; zhouyueying@nuaa.edu.cn; plgong@nuaa.edu.cn; pengpaiwang@nuaa.edu.cn; shaowei20022005@nuaa.edu.cn; dqzhang@nuaa.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3238852

input into the classifier to execute sleep staging. These handcrafted features include time-domain features [8], frequency-domain features [9], and time-frequency domain features [10]. In addition, various classifiers are used for sleep stages, such as support vector machines (SVM) [11], random forests (RF) [12], and adaptive boosting (AdaBoost) [13]. For instance, Guo et al. [11] proposed a sleep staging method using an SVM classifier based on the Hilbert-Huang transform and sample entropy features. In [14], the authors obtained the maximized feature based on multi-scale principal component analysis and discrete wavelet transform to realize the classification of sleep stages by integrating the RotSVM classifier. Based on the time domain and frequency domain, Timplalexis et al. [15] extracted the mixed features and tested multiple classifiers to generate the final classifier for sleep staging with the voting idea. Hassan et al. [16] used the tunable-Q wavelet transform to extract the features of sleep EEG signals and implemented a decision support system based on bootstrap aggregating to complete the classification of 2-state to 6-state sleep stages. In [17], the authors staged sleep by using several tiny classifiers and employed overlapping sampling to enhance the quality of the data samples.

Feature selection requires a certain amount of expertise and has limited performance. Deep learning can automatically extract meaningful features of polysomnography and further perform end-to-end sleep stage classification, gaining more and more interest and attempts. Existing methods can be categorized by model features into temporal domain feature-based, spectral domain feature-based, and other spatial feature-based or multimodal methods. Most of the methods are based on temporal domain features, and the model input is the raw temporal EEG sequence. For instance, based on the raw EEG signals, the convolutional neural network (CNN) was widely used in discriminative feature extraction to realize sleep stage classification [18], [19], [20]. To further learn the temporal correlation information, Supratak et al. [21] used CNN to gain shallow features and added two-layer bi-directional long short-term memory (Bi-LSTM) to mine the relationship of temporal sequence. In [22], Seo et al. proposed IITNet, which utilized Bi-LSTM to explore features inside and between the sleep stages. Drawing lessons from the transformer's idea of self-attention, the transform-based models [23], [24] can learn the features of sequence signals more effectively and have higher accuracy.

In addition to the original time domain EEG signals, some methods consider extracting features from the frequency domain. For instance, considering the frequency-domain information, Kuo et al. [25] employed the spectrogram obtained by wavelet transform as the model input and built the SNet model by CNN. Instead, Gupta et al. [26] applied a modified Fourier decomposition method to construct a new time-frequency image representation. Some methods used multi-channel EEG data to construct functional connectivity [27] representing the features of brain spatial domain or combined EOG and ECG data for sleep staging research [28], [29], [30]. Though the above deep learning models achieved acceptable sleep staging results, most of the existing methods only take into account the raw time-domain signal or the 2-D image of time-frequency representation. In this way, the feature representation power is limited, and applying image representation for training requires more resource consumption. Besides, these methods fail to consider the inner and continuous relationship between stages in sleep staging. Actually, for uncertain signals, the specialist would observe the adjacent epochs to determine the sleep stage of the current epoch.

To alleviate the above problems, we propose a Temporal-Spectral fused and Attention-based deep neural Network model (TSA-Net) for automatic sleep staging, using a single-channel EEG signal. The TSA-Net is a two-stream model that combines temporal and spectral features from two perspectives. In addition, we only use single-channel EEG data for our study, considering multimodal data can be informative, but single-channel EEG has lower requirements for equipment and is more portable. The TSA-Net starts with a two-stream feature extractor (TSFE) module, in which the raw time-domain signals and the converted frequency-domain signals are the inputs of temporal and spectral streams, respectively. In each stream of the TSFE, different feature extraction modules are designed. The feature context learning (FCL) module receives the extracted features and uses the multi-head self-attention [30] block to learn the correlation information between the features and the residual block to learn multi-level features. After that, we will get a preliminary result of sleep staging, similar to a sleep specialist's initial determination of the current epoch. Furthermore, we employ the conditional random field module (CRF) [31] to learn the transition rules between epochs and to obtain the final results.

Our contributions can be summarized as follows:

(1) We propose a fused dual-input model considering both temporal and spectral features from two perspectives to obtain salient features of epochs.

(2) We use the multi-head self-attention mechanism in the FCL module to learn the dependencies between features.

(3) We imitate the decision-making procedure of specialist in manual sleep staging and employ CRF module to build transition rules, which makes the model more reasonable.

(4) Extensive experiments are conducted on two public datasets, and the results indicate that the proposed TSA-Net model outperforms state-of-the-art models in terms of staging accuracy and overall performance.

The rest of the paper is arranged as follows. Section II describes the proposed TSA-Net model. Section III introduces the experimental design, including the dataset used in the experiment, the evaluation criteria, and specific model settings. Section IV presents the experimental results and performances of our model, discussion and limitations. Finally, Section V concludes the whole paper.

II. Method

In this section, we provide a thorough introduction to the proposed TSA-Net model, covering its overall structure and each module in depth.

A. Overview of Our Model

The time-domain information and frequency-domain information of different sleep stages have different characteristics,

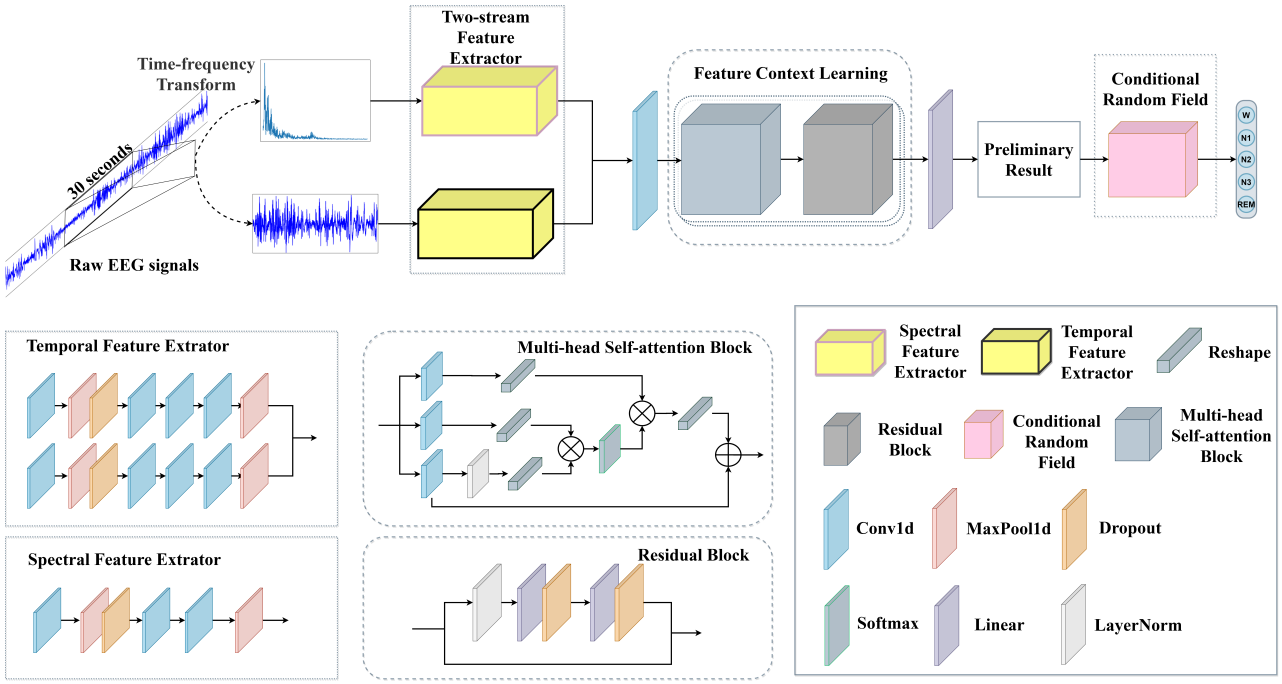


Fig. 1. The overall framework of the proposed model TSA-Net. Two branches of the raw EEG signals are input into the two-stream feature extractor; one branch into the temporal feature extractor, and the other branch into the spectral feature extractor after time-frequency transformation. The temporal and spectral fusion features obtained by the two branches were then entered into the feature context learning module to obtain the preliminary results of sleep staging. The conditional random field module will further optimize and gain the final sleep stage.

both of which can offer extensive details for sleep staging. Based on that, we propose the TSA-Net model. The pipeline of our dual-input single-channel based sleep staging method is shown in Fig. 1. The dual-input represents the raw temporal EEG sequence and the spectral sequence after the time-frequency transformation. Two-stream feature extractor includes temporal feature extractor and spectral feature extractor, and they vary according to their characteristics. The temporal feature extractor is referred to a previous work [23], where two branches of different convolution kernels are designed. The spectral feature extractor is simpler because the frequency domain is more conspicuous. These two feature extractors will deal with temporal and spectral sequences, respectively. After that, features from two streams will be concatenated, go through a convolutional layer mapping and enter the FCL module. The FCL consists of a multi-head self-attention block and a residual block. The multi-head self-attention block is used to learn the correlation and dependence information between the features learned from two streams. The residual structure can combine features from different levels. Sequentially, we get a preliminary sleep stage result after FCL, but this result does not consider the associations between epochs and the existing inappropriate transition. For uncertain epochs, we imitate the specialists who look at the adjacent epochs to determine the current epoch. Therefore, we propose the use of CRF to further modify the results and complete the sleep staging. The main modules are described below.

B. Dual-Input

Since the characteristics of EEG include the time domain and frequency domain, we mine relevant information from two perspectives, which is the dual input. For the original 30-second EEG with a sample rate of r Hz, the raw EEG time series is the temporal sequence, with the shape of $1 \times 30r$. We use Fast Fourier Transform for time-frequency transformation to obtain a spectral EEG sequence. Considering the low-frequency characteristics of EEG signals during sleep [32], we only select the frequency part around 0-25 Hz. These two parts of the input will be subjected to the corresponding feature learning part of the two-stream feature extractor.

C. Two-Stream Feature Extractor (TSFE)

The two-stream feature extractor includes two sub-modules, the temporal feature extractor, and the spectral feature extractor, which correspond to the temporal domain and spectral domain, respectively. The temporal domain sequence will be fed into the temporal feature extractor, and the spectral domain sequence will enter the spectral feature extractor. The time-frequency domain features obtained by the two feature extractors are integrated and flowed into the next module. These two sub-modules are described in detail next.

1) Temporal Feature Extractor: Since time-domain information is of the raw temporal sequence, inspired by works [21], [23], we implement two branches to capture valuable features of different bands. The convolutional layers of the two branches have different convolution kernel sizes to explore the feature information of different scales. The size setting of the convolution kernel is related to the sampling rate of the EEG signal. In the experiment of this study, with the sampling rate of 100 Hz, we set the convolution kernel size to 50 and 400, corresponding to time windows of 0.5 seconds and 4 seconds, respectively. Taking the 4-second time window as an example, it can capture sinusoidal signals as low as 0.25 Hz. The feature waveforms that can be captured for different size time windows are also different, so designing branches with

different convolution kernel sizes can obtain signal features on two scales with different sizes. Each branch consists of four convolution layers, two maximum pooling layers, and a dropout layer. Each convolution layer is followed by batch normalization and a Gaussian Error Linear Unit (GELU) activation function. Batch normalization normalizes the data so that the model can have a better generalization effect. To avoid gradient disappearance problems, we choose GELU as the activation function. The shape of the 1×3000 EEG temporal sequence is fed into the temporal feature extractor and outputs a shape of 128×80 feature vector.

2) Spectral Feature Extractor: The spectral feature extractor consists of a stack of one-dimensional convolution operations, including two convolutional modules for feature learning and a max-pooling module to reduce the dimension of features. Each convolutional layer in the convolutional module is followed by a batch normalization layer and GELU activation function, like the temporal feature extractor. In the max-pooling module, the max-pooling layer is followed by a dropout layer that drops with a certain probability to prevent the overfitting of the model. Since the feature of frequency domain sequence is more obvious, and a single branch can reduce unnecessary resource consumption, we just use one branch in the spectral feature extractor. In our model, the input of spectral feature extractors is 1×1000, and the output shape is 128×40.

D. Feature Context Learning Module (FCL)

Inspired by Transformer [33], the FCL module utilizes multi-head self-attention to encode and learn the extracted time-frequency domain features. The multi-head self-attention can process features in parallel, improving the parallel efficiency of the model over other models, like the recurrent neural network (RNN). The FCL module consists of a multi-head self-attention block and a residual block stacked twice. Multi-head self-attention block can learn dependence over a long period. Compared with the traditional self-attention method, multi-head self-attention divides the input features into subspaces composed of multiple heads, and each subspace will learn the attention weights in the space. The heads of different subspaces will interact with each other to convey attention information between different subspaces. After that, we design a residual structure to lessen information loss at different information levels, which makes up the second block. Therefore, the FCL module can improve the overall ability of the model and pay attention to different locations. The output of the TSFE module (denoted as $X \in R^{l \times d}$) will be input into this module, where l is the length of the feature, d is the dimension of the feature. After the two blocks, the final output of FCL is a preliminary sleep stage. Suppose we have H heads, the input features are evenly divided into H subspaces, each of which are represented as $x_n \in R^{l \times \frac{d}{H}}$, where $1 \leq n \leq H$. For each subspace n, we calculate its corresponding Q_n, K_n, V_n according to the learnable weight matrix W_n:

$$Q_n = X_n W_n^q \tag{1}$$

$$K_n = X_n W_n^k \tag{2}$$

$$V_n = X_n W_n^v \tag{3}$$

The self-attention A_n of each subspace n can be obtained by Q_n, K_n, V_n for dot product operation, and the specific operation formula is:

$$A_n = Attention\,(Q_n, K_n, V_n) = Softmax\left(\frac{Q_n K_n^T}{\sqrt{d}}\right) V_n \tag{4}$$

Multi-headed self-attention will concatenate the self-attention of each subspace A_n:

$$MHSA = Concat\,(A_1 \ldots A_n \ldots A_H) \tag{5}$$

The result of the multi-head self-attention calculation will be added with the input feature and then enter the residual block. In the residual block, the input M will first handle layer normalization and then enter two fully connected layers. Each fully connected layer is followed by a dropout layer with a probability of 0.1. The residual operation will be performed on the output of the multi-head self-attention block. The output of the residual block will be sent into the fully connected layer to output the preliminary predicted sleep staging results.

E. Conditional Random Fields (CRF)

The preliminary results obtained in the FCL module only consider the characteristics of the current epoch. This is analogous to the fact that the specialist will have a pre-judgment of the sleep stage in each 30-second time window. However, when the EEG information within the time window cannot fully determine the sleep stage, the specialist will consider the epochs before and after the current window to determine the current sleep stage. Based on this ideology, we propose to use a CRF module to correct the transition rule of the sleep stage.

The CRF [31] is a discriminant probability model based on an undirected graph that considers the dynamic relationship between adjacent variables. Our method is based on a linear CRF, which defines two stochastic sequences. One is a state sequence $I = \{i_1, i_2, \ldots, i_T\}$ and the other is an observations sequence $O = \{o_1, o_2, \ldots, o_T\}$, $i_n, o_n \in \{W, N1, N2, N3, REM\}$, $(1 \leq n \leq T)$ Here, state sequence I is the ultimate label we want, and observed sequence O is the preliminary prediction in the FCL module. i_n represents the true sleep stage at n-time and o_n is the observed preliminary sleep staging result at n-time. Since our goal is to solve the problem of sleep staging, either the observation sequence or the state sequence can only be one of W, N1, N2, N3, and REM, a total of five possible states. We calculate the prediction result of the final sleep stage based on probability from the undirected graph, and its conditional probability distribution $P\,(i \mid o)$ is defined as:

$$P\,(i \mid o) = \frac{1}{Z\,(o)} \exp\left(\sum_{k=1}^{K} \omega_k f_k\,(i_n, i_{n-1}, o_n)\right) \tag{6}$$

$$z\,(o) = \sum \exp\left(\sum_{k=1}^{K} \omega k f_k\,(i_n, i_{n-1}, o_n)\right) \tag{7}$$

$$f_k(i_n, i_{n-1}, o_n = \begin{cases} t_k(i_n, i_{n-1}, o_n) \\ s_l(i_n, i_{n-1}, o_n) \end{cases} \tag{8}$$

$$\omega_k = \begin{cases} \lambda_k \\ \mu_l \end{cases} \tag{9}$$

where $f_k(i_n, i_{n-1}, o_n)$ is feature function, which is specifically divided into transition functions $t_k(i_n, i_{n-1}, o_n)$ and state functions $s_l(i_n, i_{n-1}, o_n)$. In Eq.(9), ω_k is the weight of the feature functions. λ_k is for transition functions, μ_l is for state functions correspondingly. K is the total number of feature functions.

For the conditional probability distribution constructed, we adopt a maximum likelihood estimate to calculate the conditional log-likelihood $L(\theta) = \sum_{j=1}^{N} logp(i^j|o^j)$, where N is the length of the predicted sequence, i^j and o^j represent the state value and observation value of the j-th sample, respectively. After obtaining the trained model, we use the Viterbi algorithm [33] to solve the predicted sleep stage. That is to say, the sequence of CRF optimizes the preliminary sleep staging result, and the final sleep staging result is obtained.

F. Loss Function

Due to the imbalance between categories in sleep stages, we use a weighted cross-entropy loss function [23]:

$$Loss = -\frac{1}{M}\sum_{m=1}^{M}\sum_{c=1}^{C}\omega_c y_m^c \log\left(\hat{y}_m^c\right) \tag{10}$$

where ω_c is a weight parameter that can be adjusted according to each category, M is the total number of samples, and C is the number of categories. y_m^c is the true label of the m-th sample, $\hat{y}_m^c$ is the predicted label of the m-th sample, which together constitute the training loss of the model.

III. Experiment

A. Datasets

Our experiments are evaluated on two public datasets from PhysioBank [34], called Sleep-EDF-20 and Sleep-EDF-78. Sleep-EDF-20 was released in 2013 and contained 20 healthy Caucasians, while Sleep-EDF-78 included 78 subjects without any sleep problems and was a later development version. All subjects in the experiment were aged between 25 and 101. They experimented for two consecutive day-night at home. For some reason, subjects 13, 36, and 52 were lost on one of the two nights of PSG records. Each PSG record contains Fpz-Cz and Pz-Oz two EEG channels, one horizontal EOG channel, one submental chin EMG channel, respiration, and rectal body temperature. Experienced specialists scored the sleep stage labeling of each PSG record according to R&K rules, which include W, S1, S2, S3, S4, REM, MOVEMENT, and UNKNOWN.

B. Data Preprocessing

In our research, following some previous studies [21], [23], we only use single-channel EEG Fpz-Cz, with a sampling rate of 100Hz. Both datasets go through the following preprocessing procedure: Firstly, we divide the EEG signal into epochs of 30 seconds, since the specialist marked the label by the epochs. Secondly, we follow previous studies and classify five stages of sleep according to the AASM standard, merging N3 and N4 into one, and removing MOVEMENT and UNKNOWN stages. Thirdly, since the non-sleep stages are useless for our research and we are only interested in sleep processes, we adopt 30 minutes before sleep and 30 minutes after sleep. Table I shows the sleep stage distribution for the two datasets after preprocessing.

TABLE I
THE DISTRIBUTION OF TWO OPEN DATASETS

Dataset	W	N1	N2	N3	REM	Total
Sleep-EDF-20	8285 19.6%	2804 6.6%	17799 42.1%	5703 13.5%	7717 18.2%	42308
Sleep-EDF-78	65951 33.7%	21522 11.0%	69132 35.4%	13039 6.7%	25835 13.2%	195479

C. Evaluation Metrics

We adopt five kinds of indicators to comprehensively evaluate the performance of the proposed model. They are accuracy (ACC), recall (RE), precision (PR), macro-averaged F1-score (MF1), and Cohen Kappa (K). MF1 and Kappa can provide a good evaluation of the unbalanced dataset. Suppose TP is True Positive, FP is False Positive, FN is False Negative, TN is True Negative, C is the kind of sleep stage, P_e is the hypothetical probability of agreement by chance. Their corresponding formulas are defined as follows.

$$ACC = \frac{\sum_{i=1}^{C} TP_i}{Total\ samples} \tag{11}$$

$$RE = \frac{TP}{TP + FN} \tag{12}$$

$$PR = \frac{TP}{TP + FP} \tag{13}$$

$$MF1 = \frac{1}{C}\sum_{i=1}^{C}\frac{2 * PR_i * RE_i}{PR_i + RE_i} \tag{14}$$

$$K = \frac{ACC - P_e}{1 - P_e} \tag{15}$$

D. Experimental Setup

We adopt k-fold cross-validation for sleep staging in this experiment, where k is set differently for two datasets. For Sleep-EDF-20, the k is set to 20, which is equivalent to leave-one-subject-out (LOSO) cross-validation. In each fold, 19 subjects are used for training, and the remaining one subject is used for testing. This operation is repeated 20 times until all the subjects are used. Due to the Sleep-EDF-78 dataset being large and k-fold cross-validation being time-consuming, we only use 10-fold cross-validation for this dataset like in other work [35].

We apply the cross-entropy loss to train our model and assign different weights to different categories of losses to avoid the problem of unbalanced data. Considering the different severity caused by the uneven distribution of the number of the five categories, as used in [23], we set the weights of W, N1, N2, N3, and REM as 0.3, 0.4, 0.3, 0.2, and 0.3, respectively. During the model training, we use the Adam

TABLE II
THE CONFUSION MATRIX OF TSA-NET ON FPZ-CZ CHANNEL ON SLEEP-EDF-20 DATASET. THE NUMBERS IN BOLD ARE THE SAMPLE NUMBERS THAT PREDICTED CORRECTLY

	Predicted					Per-class metrics		
	W	N1	N2	N3	REM	PR	RE	MF1
W	**7746**	184	180	14	161	87.76	93.49	90.54
N1	423	**1029**	686	7	659	65	36.7	46.91
N2	440	243	**15870**	450	796	89.18	89.16	89.17
N3	35	2	601	**5064**	1	91.44	88.8	90.1
REM	182	125	459	3	**6948**	81.12	90.03	85.35

TABLE III
THE CONFUSION MATRIX OF TSA-NET ON FPZ-CZ CHANNEL ON SLEEP-EDF-78 DATABASE. THE NUMBERS IN BOLD ARE THE SAMPLE NUMBERS THAT PREDICTED CORRECTLY

	Predicted					Per-class metrics		
	W	N1	N2	N3	REM	PR	RE	MF1
W	**62294**	1843	1022	53	739	88.5	94.45	91.38
N1	5621	**6083**	7446	173	2199	48.4	28.26	35.69
N2	1155	3280	**60085**	2379	2233	81.92	86.91	84.34
N3	52	15	2601	**10360**	11	78.58	79.45	79.01
REM	1263	1348	2190	219	**20815**	80.07	80.57	80.57

optimizer with a learning rate of 0.001. The beta1 is set to 0.9, the beta2 is 0.999, and the batch size is 128. A total of 50 epochs are trained. The code for TSA-Net is available at https://github.com/Fuguidan/TSA-Net.

IV. RESULTS AND ANALYSIS

A. Classification Performance

We implement our temporal-spectral fused, attention-based and CRF-optimized method TSA-Net for validation on two public datasets Sleep-EDF-20 and Sleep-EDF-78. For the two datasets, we use the Fpz-Cz channel with 20-fold cross-validation and 10-fold cross-validation for the test, respectively. Table II and Table III show the confusion matrix of the two corresponding datasets. The confusion matrix is computed by the sum of the validation sets for each fold. Each row represents the actual sleep stages, and each column is the sleep stage predicted by our TSA-Net model. Take Table II, for example, the number 184 in row one, column two means 184 W stage samples are misclassified to the N1 stage. The right of the table indicates the five sleep stage performance calculated from the confusion matrix.

As shown in Table II and Table III, we find that the diagonal elements of the confusion matrix occupy the majority, proving that our method is effective. We can observe that the F1 score of W can reach more than 90% in two datasets, reaching 90.54% in Sleep-EDF-20 and 91.38% in Sleep-EDF-78, respectively. Compared with other stages, N1 obtain the lowest F1 score, and unlike other stages, its precision-score is higher

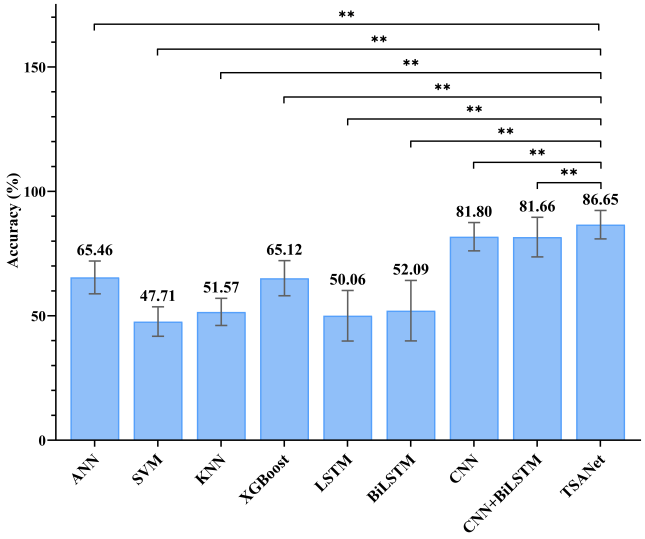


Fig. 2. Comparison between the TSA-Net and baseline classifiers. * indicates p-value < 0.1; ** indicates p-value < 0.05; paired samples t-test was used.

than the recall score. Looking at the confusion matrix, we also find that N1 is often misclassified as N2, REM, and W. In the case of misclassification, most stages are often misclassified as N2. This may be because N1 has a smaller sample size, while N2 has a sample size of about 40%. In contrast, the performance is better overall on the Sleep-EDF-20 dataset. Except for the W stage, all three indicators on the Sleep-EDF-78 dataset were lower than those in the Sleep-EDF-20 dataset.

B. Comparison With State-of-the-Art Models

We compare our TSANet with different classifiers on the Sleep-EDF-20 dataset using 20-fold cross-validation. Referring to the common features of EEG in previous works, we extract approximate entropy, sample entropy, fuzzy entropy [36], and power spectral density features [37] of five frequency bands. We adopt eight baseline models, including Artificial Neural Network (ANN), SVM, k-nearest neighbor (KNN), extreme gradient boosting (XGBoost), LSTM, BiLSTM, CNN, and CNN + BiLSTM. The first four methods are traditional classifiers with hand-crafted features, and the last four are deep models based on the raw sequence, among which CNN + BiLSTM was the DeepSleepNet model proposed by Supratak et al [21]. In the experiment, we also conduct paired samples t-test for statistical analysis on the classification accuracy of the 20-fold cross-validation.

As shown in Fig. 2, we observe that TSANet significantly performs better than the baseline classifiers. In terms of accuracy, the TSA-Net improves by 21.19%, 38.94%, 35.08%, 21.53%, 36.59%, 34.56%, 4.85%, 4.99% compared to the ANN, SVM, KNN, XGBoost, LSTM, BiLSTM, CNN, CNN +BiLSTM, respectively. Among the baseline models, CNN and CNN +BiLSTM perform the best, both achieving an accuracy of more than 81%, which proves the superiority of CNN and the limitations of traditional classifiers. At the same time, among the traditional classifiers, ANN and XGBoost achieve the best accuracy. In contrast, LSTM-based methods

TABLE IV
COMPARISON AMONG TSA-NET AND STATE-OF-THE-ART MODELS. THE NUMBERS IN BOLD INDICATE THE BEST PERFORMANCE, AND THE NUMBERS UNDERLINED ARE THE SECOND BEST

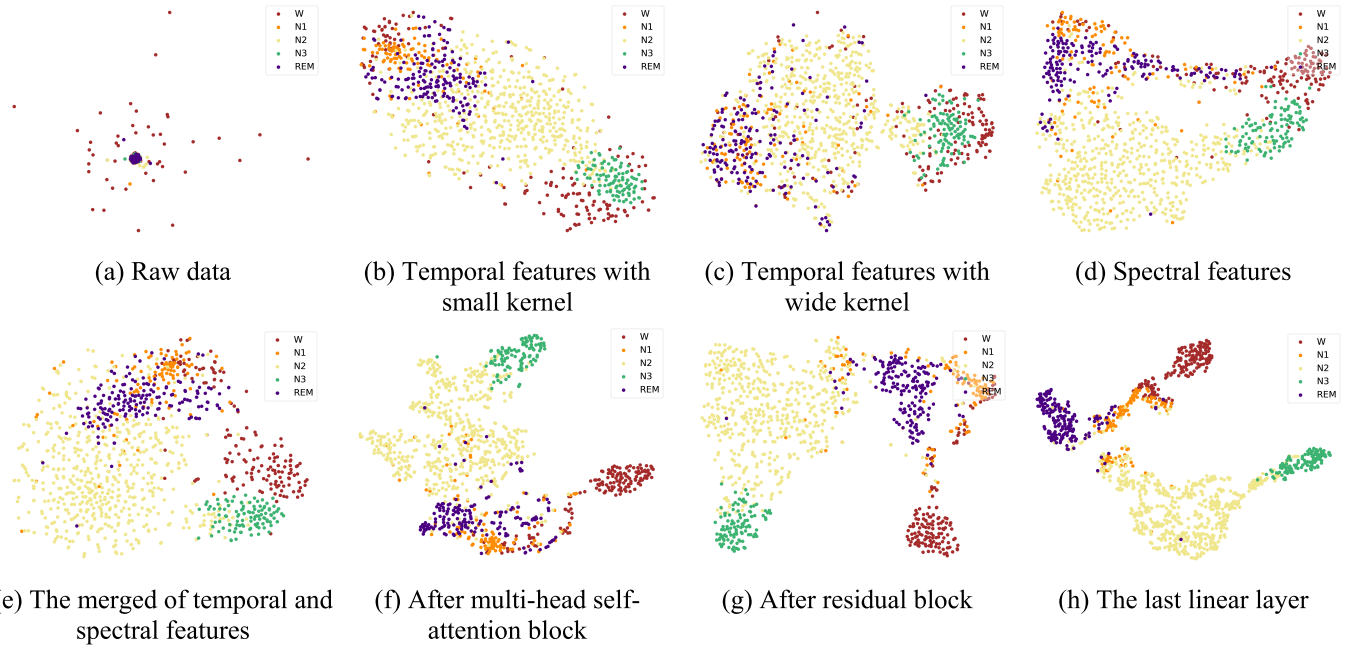
Model	Dataset	Fold	Channel	Per-class F1-score					Overall performances		
				W	N1	N2	N3	REM	ACC	Kappa	MF1
CNN-LSTM-CRF [39]	Sleep-EDF-20	20	Fpz-Cz	88.1	42.4	87.2	87.8	85.0	85.2	79	78.1
LightSleepNet [40]	Sleep-EDF-20	20	Fpz-Cz	90	31	88	89	78	83.8	78	75.3
CNN-HMM [41]	Sleep-EDF-20	20	Fpz-Cz	87.8	35.1	86.6	90.5	**86.8**	83.98	78	76.9
CCRRSleepNet [42]	Sleep-EDF-20	20	Fpz-Cz	89.01	**51.73**	87.5	88.2	82.86	84.29	78	79.81
RL-TCNN-CRF [43]	Sleep-EDF-20	20	Fpz-Cz	90.5	46.6	88.4	86.1	84.6	85.39	80	79.27
DeepSleepNet-Lite [36]	Sleep-EDF-20	20	Fpz-Cz	87.1	44.4	87.9	88.2	82.4	84	78	78
AttnSleep [23]	Sleep-EDF-20	20	Fpz-Cz	89.7	42.6	88.8	**90.2**	79	84.4	79	78.1
TSA-Net (ours)	Sleep-EDF-20	20	Fpz-Cz	**90.54**	46.91	**89.17**	90.1	85.35	**86.64**	**81.58**	**80.41**
DeepSleepNet-Lite [39]	Sleep-EDF-78	10	Fpz-Cz	91.5	**46**	82.9	79.2	76.4	80.3	73	**75.2**
AttnSleep [23]	Sleep-EDF-78	20	Fpz-Cz	82	42	**85**	**82.1**	74.2	81.3	74	75.1
TSA-Net (ours)	Sleep-EDF-78	10	Fpz-Cz	91.38	35.69	84.34	79.01	80.57	81.66	74.42	74.15
TSA-Net (ours)	Sleep-EDF-78	20	Fpz-Cz	**91.71**	30.11	84.94	80.02	**81.03**	**82.21**	**75.07**	73.57

do not perform well, probably because they are not suitable for 30-second-long sequences of sleep EEG and could not get effective information. Compared with hand-crafted features, CNN can effectively and automatically extract features, and abstract the features representation. The above results indicates the superiority of CNN-based deep learning models.

Moreover, we compare our model with the following recently published CNN-based deep learning models to better illustrate the overall classification performance of the TSA-Net:

- CNN-LSTM-CRF [38] used the splicing of the CNN module and LSTM module as pre-training, and then CRF was used for improvement.
- LightSleepNet [39] designed a residual lightweight model based on one-dimensional convolution.
- CNN-HMM [40] applied multi-core CNN for epoch classification, and HMM was used for optimization to obtain the final results.
- CCRRSleepNet [41] employed hybrid mixed relational inductive biases to learn different contributions between features from three levels of the frame, epoch, and sequence.
- RL-TCNN-CRF [42] was constructed utilizing representation learning combined with a temporal CNN.
- DeepSleepNet-Lite [35] first used the Monte Carlo dropout technique to enhance sleep scoring performance and detect uncertain instances.
- AttnSleep [23] proposed a temporal context encoder and deployed causal convolution to learn temporal correlation features.

Table IV shows the comparison of F1 scores in five stages and the three overall evaluations of Accuracy, Kappa, and MF1. For the sake of fairness, the datasets and experimental settings of the models we choose are as identical as possible. The sleep staging performances of the above methods are derived from their corresponding papers. We find that our method is the best at W, N2, and all three overall performances, achieving an accuracy rate of 86.64% on the Sleep-EDF-20 dataset. For the accuracy of Sleep-EDF-78 dataset, the result of the 10-fold cross-validation is 81.66% and the 20-fold cross-validation is 82.21%. As shown in Table IV, the TSA-Net outperforms the state-of-the-art models with improvements of 1.25% to 2.84% in terms of accuracy. In addition, the accuracies of TSA-Net improved by 0.91% (for 20-fold cross-validation) and 1.33% (for 10-fold cross-validation) on the Sleep-EDF-78. Compared with the CNN-HMM method, our model uses CRF for learning transition rules, and the structure of CRF undirected graphs is more effective. The extraction of dual features in the time domain and frequency domain can learn more comprehensive features in comparison to the CNN-LSTM-CRF model. Additionally, multi-head self-attention enhances parallelism capabilities, resulting in overall performance superior to LSTM. For the Sleep-EDF-78 dataset, the overall performance of accuracy and kappa in TSA-Net is also better than other models. And the kappa results prove that our model is not without class-to-category bias but is more scientific. Overall, our model performs better than the state-of-the-art models on both datasets under the same experimental settings. At the same time, we find that the improvement of TSA-Net on the two datasets is different. Firstly, we speculate that the amount and the data distribution differences is to cause. As shown in Table I, N1 accounts for only 6.6% in Sleep-EDF-20, while N2 accounts for 42.1%, which is much higher than in other stages. While the smallest proportion of Sleep-EDF-78 is N3, only 6.75%, N2 and W account for roughly the same proportion. The difference in the distribution of the two datasets should be responsible for the different performances. Secondly, leave-one-subject-out cross-validation is used on the Sleep-EDF-20, while leave-one-subject-out cross-validation used on the



(a) Raw data (b) Temporal features with small kernel (c) Temporal features with wide kernel (d) Spectral features

(e) The merged of temporal and spectral features (f) After multi-head self-attention block (g) After residual block (h) The last linear layer

Fig. 3. The visualization of the features of each module of TSA-Net.

Sleep-EDF-78 is not realistic considering resource constraints. The different experimental settings are also the influencing factors for the different effects. Finally, the differences between subjects are also an important factor. As the number of subjects increases, the model will be greatly affected by the differences between subjects. In this way, further optimization could be applied by reducing the differences between subjects using methods such as transfer learning.

C. Model Analysis

To further analyze the feature representation ability of each module in the TSA-Net, we use t-SNE [43] for visual analysis. The t-SNE is a nonlinear dimensionality reduction method that can map high-dimensional data to low-dimensional space to observe the characteristics of the data.

Fig. 3 visualizes the distribution of data processed by each module of the Sleep-EDF-20 dataset. Fig. 3(a) denotes the raw data without any preprocessing, and we find that the five sleep categories are all mixed up and difficult to distinguish from each other. Fig. 3 (b) and (c) show the characteristics of time-domain information processed by small convolution kernel and wide convolution kernel, respectively. There is a certain aggregation of categories in the same stage, but they cannot be clearly distinguished, especially between the W and N1 stages. Fig. 3 (d) represents the feature extraction in the frequency domain, and its inter-class aggregation is not clear.

The red W stage and orange N1 stage are scattered into the other three categories. In other words, this shows that both time domain and frequency domain features have certain feature representation abilities and can understand some of the traits of the various sleep stages, but it is not enough and needs further processing. Fig. 3 (e) exhibits the result of the integration of time and frequency domain feature extraction. From the comparison, we see that the integrated features have better

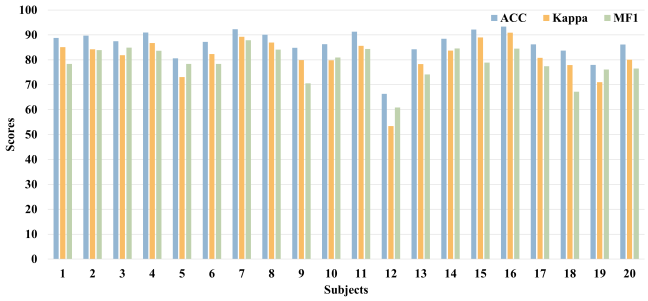


Fig. 4. Sleep staging performance of our TSA-Net for the 20 subjects on Fpz-Cz channel of Sleep-EDF-20 dataset. The y-axis represents the score of ACC, Kappa and MF1. The x-axis represents each subject.

expression ability than the separate time domain and frequency domain features, which further demonstrates the progress of temporal and spectral feature fusion. The data distribution after the multi-head self-attention block is shown in Fig. 3 (f), and we can see that the W, N2 and N3 stages have obvious boundaries, but the N1 stage is still confused with the REM stage. As Fig. 3 (g) displays, with the residual block combined with different levels of information, the classification effect is further improved. Fig. 3 (h) denotes the data distribution of the output of the last fully connected layer. The N1 stage has a closer aggregation, and the five stages are clearly separated. However, the overall performance of the N1 stage is not as good as other stages. We argue this may be due to the small sample size and feature learning is not as significant as in other stages.

Taking Sleep-EDF-20 as an example, we experiment with a leave-one-subject-out cross-validation. Fig. 4 shows the overall performance of each subject in the 20-fold validation. We can see that the accuracy of most subjects is above 80%, of which the overall performance of subject 7 is the best and

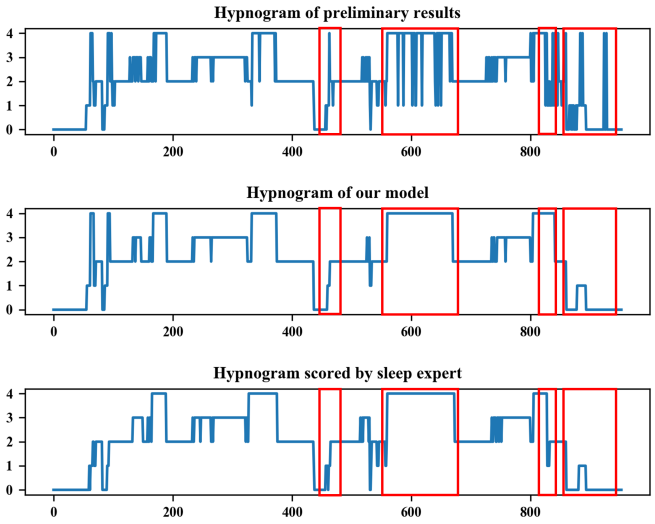


Fig. 5. Examples of hypnogram on Fpz-Oz channel for a random subject (i.e. subject 15) in Sleep-EDF-20 dataset, including the hypnogram of preliminary results (top), finally hypnogram of our model TSA-Net (center), corresponding hypnogram scored by sleep specialists (bottom). The mis-classification epochs are marked in the red box. 0-4 of the y-axis represent W, N1, N2, N3, and REM, respectively. The x-axis represents epochs of sleep.

the three indicators of subject 12 are not ideal. The possible reason is that in the cross-validation of fold 12, the sample size of test subjects is smaller than that of other subjects, and the data in the N1 stage is even less, only 31 samples, which may be the reason for the poor performance of fold 12.

D. CRF Refinement

In our model, we use CRF to learn transition rules for further correcting the preliminary predictions. To observe the improvement effect of CRF, we exhibit the whole-night sleep stage of the preliminary results of TSA-Net, the final results of TSA-Net with CRF, and corresponding labels scored by sleep specialists for visualization.

As shown in Fig. 5, there are some abnormal transitions in the preliminary results, which manifest as frequent jumps in the stages. The results of CRF correction are presented in the center hypnogram of Fig. 5. We can observe that the jump during the stages is significantly reduced, which is highlighted in the red box in the figure. The first red box indicated that N2 can be corrected when it is misclassified as REM and REM frequently misclassified as N1 in the second red box and can also be corrected. Similarly, W is wrongly classified as REM. The optimized stage conversion has been significantly improved and is closer to the sleep stage labeled by specialists.

We further explore the transition of the CRF module and plot the transition probability based on the preliminary results of the Sleep-EDF-20 dataset, as shown in Fig. 6. Additionally, Fig. 7 shows the sample size statistics of the preliminary results and the final results in each stage when the correction occurred. We observe that the most frequently occurring sleep transition rule correction is optimizing N1 to REM, followed by N1 to W. We conjecture that the most likely explanation is that N1, the stage between REM and N2, has the smallest

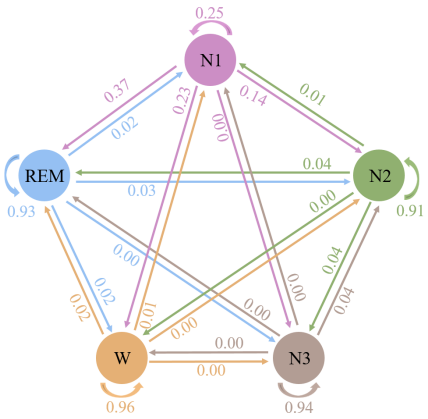


Fig. 6. Transition probability of CRF correction module on the Sleep-EDF-20 dataset.

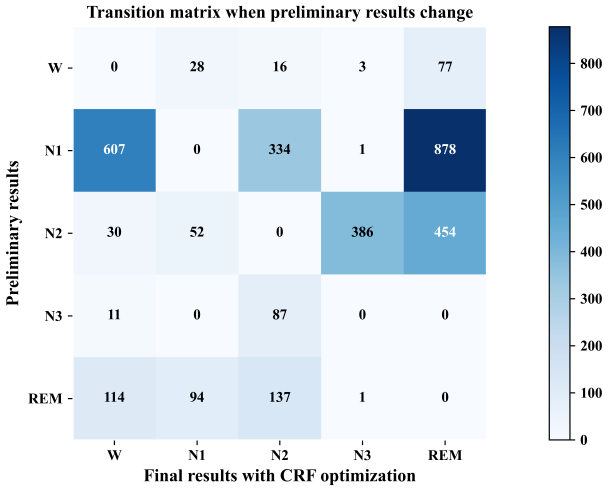


Fig. 7. Transition matrix of sleep stages calculated on the Sleep-EDF-20 dataset when preliminary results and final results differ.

sample size and only makes up 5%–10% of the entire sleep cycle. Besides, the preliminary results are easy to be misclassified as its adjacent stages, while the CRF could modify this phenomenon according to the unreasonable jitter of its adjacent stages. In addition, we could see that the mutual transition probability between N3 and N1, as well as between N3 and REM are both 0. This may be because N3, which is a stage of deep sleep, exhibits a significant slow wave (0.5-2Hz), whereas REM and N1, which are stages of preparation for sleep and light sleep, commonly exhibit the alpha wave (8-13Hz) and the theta wave (4-8Hz). The above two-paired stages can be well separated in the preliminary staging results based on the combination of temporal features and spectral features so that no further correction is required for the CRF module, which leads to its low transition probability. At the same time, the differences in correction between different stages also verify the rationality and necessity of the CRF module.

E. Ablation Study

To evaluate the effect of each module of the model, we perform ablation analysis on the Sleep-EDF-20 dataset based on the modules of the flowchart in Fig.1, and the specific variant model designs are as follows:

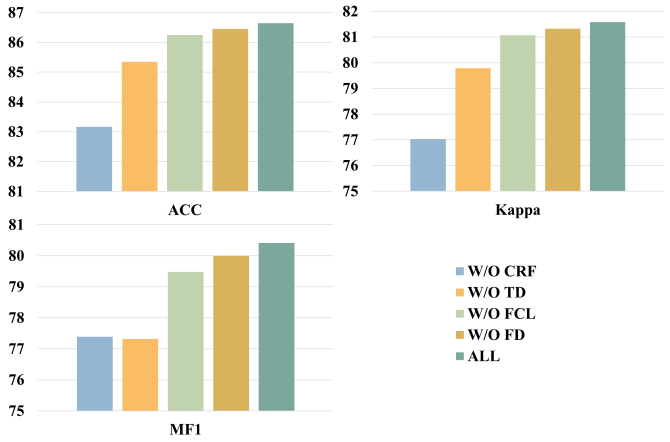


Fig. 8. Ablation study of TSA-Net conducted on Sleep-EDF-20 dataset. The y-axis is the score of the different indices.

1) W/O CRF: TSFE module and FCL module without CRF refinement
2) W/O TD: TSFE module without Time Domain information and the rest of our model.
3) W/O TCL: TSFE module and CRF module without TCL.
4) W/O FD: TSFE module without Frequency Domain information and the rest of our model.
5) ALL: The complete model that we proposed.

Fig. 8 shows that the lack of these critical parts weakens the model from the three perspectives of ACC, Kappa, and MF1. Comparing W/O CRF and all, we can see that the CRF can learn the transition rules during sleep and correct some unreasonable stage changes, showing its remarkable improvement effect. Since the time domain contains abundant original information, we can find from W/O TD and ALL that its lack also has a significant impact. In addition, the feature context learning module can further improve the performance of TSA-Net, and the multi-head self-attention can learn the correlation between features and improve the parallelism ability. The added frequency domain information module can also enhance the overall performance of the model. In general, the comparison between the model without different modules and the complete model verifies that they all make varying degrees of contributions to the TSA-Net.

F. Limitations and Future Work

Nevertheless, there are still some deficiencies in this study that need to be further improved. For example, the classification accuracy in the N1 stage is not enough, and the model cannot deal with the low accuracy caused by insufficient samples. Additionally, there is a lack of consideration of cross-subject differences and the effects of various devices. In the future, we will explore how to solve the data imbalance problem and how to mine the characteristics of N1 to improve its accuracy. In addition, we will further consider the issue of transfer learning [44], [45], as it is crucial to discover common characteristics among different subjects and maintain effective performances across various devices and datasets.

V. Conclusion

In this study, we present a novel dual-input deep temporal-spectral representation and attention-based model called TSA-Net for automatic sleep staging. The TSA-Net includes three major modules: 1) a two-stream feature extractor for extracting salient features from the time domain and frequency domain, 2) a feature context learning module for learning time-dependent dependencies and achieving preliminary staging results, and 3) a conditional random field for optimizing the preliminary results, rationalizing the transition of sleep stages, and obtaining the final classification results. We verify the proposed TSA-Net model on two public sleep staging datasets (i.e., Sleep-EDF-20 and Sleep-EDF-78) and compare them with the existing advanced sleep staging methods. Experimental results demonstrate that the TSA-Net model can not only acquire the transition rules of sleep stages but also achieve better sleep staging performance than state-of-the-art methods.

References

[1] C. Hirotsu, S. Tufik, and M. L. Andersen, "Interactions between sleep, stress, and metabolism: From physiological to pathological conditions," *Sleep Sci.*, vol. 8, no. 3, pp. 143–152, Nov. 2015.
[2] L. Besedovsky, T. Lange, and M. Haack, "The sleep-immune crosstalk in health and disease," *Physiological Rev.*, vol. 99, no. 3, pp. 1325–1380, Jul. 2019.
[3] J. G. Klinzing, N. Niethard, and J. Born, "Mechanisms of systems memory consolidation during sleep," *Nature Neurosci.*, vol. 22, no. 10, pp. 1598–1610, Oct. 2019.
[4] K. K. Gulia and V. M. Kumar, "Sleep disorders in the elderly: A growing challenge," *Psychogeriatrics*, vol. 18, no. 3, pp. 155–165, May 2018.
[5] J. V. Rundo and R. Downey III, "Polysomnography," in *Handbook of Clinical Neurolog*, vol. 160. New York, NY, USA: Elsevier, 2019, pp. 381–392.
[6] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, and B. V. Vaughn, "The AASM manual for the scoring of sleep and associated events," *Rules, Terminol. Tech. Specifications, Darien, Illinois, Amer. Acad. Sleep Med.*, vol. 176, p. 2012, Oct. 2012.
[7] J. A. Hobson, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," *Electroencephalogr. Clin. Neurophysiol.*, vol. 26, no. 6, p. 644, Jun. 1969.
[8] S. K. Satapathy, D. Loganathan, and P. Narayanan, "Automated sleep staging analysis using sleep EEG signal: A machine learning based model," in *Proc. Int. Conf. Advance Comput. Innov. Technol. Eng. (ICACITE)*, Greater Noida, India, Mar. 2021, pp. 87–96.
[9] M. Vaezi and M. Nasri, "Application of heuristic feature selection in EEG based sleep stages classification," in *Proc. 6th Iranian Conf. Signal Process. Intell. Syst. (ICSPIS)*, Mashhad, Iran, Dec. 2020, pp. 1–6.
[10] T. F. Zaidi and O. Farooq, "Automatic classification of sleep stages using EEG sub-bands based time-spectral features," in *Proc. Int. Conf. Innov. Intell. Informat., Comput., Technol. (ICT)*, Sep. 2021, pp. 720–725.
[11] C. Guo, F. Lu, S. Liu, and W. Xu, "Sleep EEG staging based on Hilbert–Huang transform and sample entropy," in *Proc. Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Jabalpur, India, Dec. 2015, pp. 442–445.
[12] P. Memar and F. Faradji, "A novel multi-class EEG-based sleep stage classification system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 1, pp. 84–95, Jan. 2018.
[13] L. Siyuan, L. Jingyuan, G. Hangping, and R. Minhua, "Sleep staging prediction model based on XGBoost," in *Proc. Int. Conf. Electron. Inf. Eng. Comput. Sci. (EIECS)*, Changchun, China, Sep. 2021, pp. 350–353.
[14] E. Alickovic and A. Subasi, "Ensemble SVM method for automatic sleep stage classification," *IEEE Instrum. Meas.*, vol. 67, no. 6, pp. 1258–1265, Jun. 2018.
[15] C. Timplalexis, K. Diamantaras, and I. Chouvarda, "Classification of sleep stages for healthy subjects and patients with minor sleep disorders," in *Proc. IEEE 19th Int. Conf. Bioinf. Bioeng. (BIBE)*, Athens, Greece, Oct. 2019, pp. 344–351.

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

[17] P. An, Z. Yuan, J. Zhao, X. Jiang, and B. Du, "An effective multi-model fusion method for EEG-based sleep stage classification," *Knowl.-Based Syst.*, vol. 219, May 2021, Art. no. 106890.

[18] J. Zhang and Y. Wu, "A new method for automatic sleep stage classification," *IEEE Trans. Biomed. Circuits Syst.*, vol. 11, no. 5, pp. 1097–1110, Oct. 2017.

[19] E. Fernandez-Blanco, D. Rivero, and A. Pazos, "Convolutional neural networks for sleep stage scoring on a two-channel EEG signal," *Soft Comput.*, vol. 24, no. 6, pp. 4067–4079, Mar. 2019.

[20] X. Jiang, J. Zhao, D. Bo, A. Panfeng, H. Guo, and Z. Yuan, "MRNet: A multi-scale residual network for EEG-based sleep staging," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[21] 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.

[22] H. Seo, S. Back, S. Lee, D. Park, T. Kim, and K. Lee, "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*, vol. 61, Aug. 2020, Art. no. 102037.

[23] E. Eldele et al., "An attention-based deep learning approach for sleep stage classification with single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 809–818, 2021.

[24] G. Shi, Z. Chen, and R. Zhang, "A transformer-based spatial–temporal sleep staging model through raw EEG," in *Proc. Int. Conf. High Perform. Big Data Intell. Syst. (HPBDIS)*, Macau, China, Dec. 2021, pp. 110–115.

[25] C.-E. Kuo, G.-T. Chen, and P.-Y. Liao, "An EEG spectrogram-based automatic sleep stage scoring method via data augmentation, ensemble convolution neural network, and expert knowledge," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 102981.

[26] V. Gupta and R. B. Pachori, "FBDM based time-frequency representation for sleep stages classification using EEG signals," *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102265.

[27] J. Zhang, "EEG-based sleep staging analysis with functional connectivity," *Int. J. Psychophysiology*, vol. 168, p. S32, Oct. 2021.

[28] R. Zhao, Y. Xia, and Q. Wang, "Dual-modal and multi-scale deep neural networks for sleep staging using EEG and ECG signals," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102455.

[29] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, "SalientSleepNet: Multimodal salient wave detection network for sleep staging," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 2614–2620.

[30] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, Apr. 2018.

[31] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 5999–6009.

[32] M. Kim, "Conditional ordinal random fields for structured ordinal-valued label prediction," *Data Mining Knowl. Discovery*, vol. 28, no. 2, pp. 378–401, Mar. 2014.

[33] D. Aeschbach and A. A. Borbély, "All-night dynamics of the human sleep EEG," *J. Sleep Res.*, vol. 2, no. 2, pp. 70–81, Jun. 1993.

[34] G. D. J. Forney, "The Viterbi algorithm," *Proc. IEEE*, vol. 61, no. 5, pp. 268–278, Mar. 1993.

[35] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[36] L. Fiorillo, P. Favaro, and F. D. Faraci, "DeepSleepNet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2076–2085, 2021.

[37] J. L. Rodríguez-Sotelo, A. Osorio-Forero, A. Jiménez-Rodríguez, D. Cuesta-Frau, E. Cirugeda-Roldán, and D. Peluffo, "Automatic sleep stages classification using EEG entropy features and unsupervised pattern analysis techniques," *Entropy*, vol. 16, no. 12, pp. 6573–6589, 2014.

[38] Y. Zhou, S. Huang, Z. Xu, P. Wang, X. Wu, and D. Zhang, "Cognitive workload recognition using EEG signals and machine learning: A review," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 3, pp. 799–818, Sep. 2022.

[39] B. Yang, W. Wu, Y. Liu, and H. Liu, "A novel sleep stage contextual refinement algorithm leveraging conditional random fields," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[40] D. Zhou et al., "LightSleepNet: A lightweight deep model for rapid sleep stage classification with spectrograms," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 43–46.

[41] B. Yang, X. Zhu, Y. Liu, and H. Liu, "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*, vol. 68, Jul. 2021, Art. no. 102581.

[42] W. Neng, J. Lu, and L. Xu, "CCRRSleepNet: A hybrid relational inductive biases network for automatic sleep stage classification on raw single-channel EEG," *Brain Sci.*, vol. 11, no. 4, p. 456, Apr. 2021.

[43] E. Khalili and B. Mohammadzadeh Asl, "Automatic sleep stage classification using temporal convolutional neural network and new data augmentation technique from raw single-channel EEG," *Comput. Methods Programs Biomed.*, vol. 204, Jun. 2021, Art. no. 106063.

[44] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[45] N. Banluesombatkul et al., "MetaSleepLearner: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 1949–1963, Jun. 2021.

[46] Q. Xiao et al., "Self-supervised learning for sleep stage classification with predictive and discriminative contrastive coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1290–1294.