BiTS-SleepNet: An Attention-Based Two Stage Temporal-Spectral Fusion Model for Sleep Staging With Single-Channel EEG

Zhaoyang Cong, *Graduate Student Member, IEEE*, Minghui Zhao, Hongxiang Gao, Meng Lou, Guowei Zheng, Ziyang Wang, Xingyao Wang, Chang Yan, Li Ling, Jianqing Li, and Chengyu Liu, *Senior Member, IEEE*

***Abstract*—Automated sleep staging is crucial for assessing sleep quality and diagnosing sleep-related diseases. Single-channel EEG has attracted significant attention due to its portability and accessibility. Most existing automated sleep staging methods often emphasize temporal information and neglect spectral information, the relationship between sleep stage contextual features, and transition rules between sleep stages. To overcome these obstacles, this paper proposes an attention-based two stage temporal-spectral fusion model (BiTS-SleepNet). The BiTS-SleepNet stage 1 network consists of a dual-stream temporal-spectral feature extractor branch and a temporal-spectral feature fusion module based on the cross-attention mechanism. These blocks are designed to autonomously extract and integrate the temporal and spectral features of EEG signals, leveraging temporal-spectral fusion information to discriminate between different sleep stages. The BiTS-SleepNet stage 2 network includes a feature context learning module (FCLM) based on Bi-GRU and a transition rules learning module (TRLM) based on the Conditional Random Field (CRF). The FCLM optimizes preliminary sleep stage results from the stage 1 network by learning dependencies between features of multiple adjacent stages. The TRLM additionally employs transition rules to optimize overall outcomes. We evaluated the BiTS-SleepNet on three public datasets: Sleep-EDF-20, Sleep-EDF-78, and SHHS, achieving accuracies of 88.50%, 85.09%, and 87.01%, respectively. The experimental results demonstrate that BiTS-SleepNet achieves competitive performance in comparison to recently published methods. This highlights its promise for practical applications.**

***Index Terms*—Sleep staging, single-channel EEG, temporal-spectral feature fusion, two-stage model, feature context learning, conditional random field.**

I. Introduction

SLEEP constitutes approximately one-third of the human lifespan. Adequate quality sleep is essential for sustaining health and optimal functioning in humans. However, modern lifestyles and stresses have significantly increased the prevalence of sleep disorders. Common conditions include insomnia, periodic limb movement disorder, and sleep apnea syndrome [1]. These sleep disorders not only seriously affect the quality of life, but are also closely associated with the onset and progression of a variety of chronic diseases, such as cardiovascular diseases, metabolic disorders and mental health problems [2]. Accurate sleep staging is important for analyzing sleep structure, identifying sleep disorders and evaluating treatment effects. However, it's challenging to achieve accurate sleep staging due to the intricate nature of sleep architecture and individual variability [3].

Sleep staging criteria serve as the foundation for dividing the sleep cycle into distinct stages, primarily based on the Rechtschaffen & Kales (R&K) criteria [4] and the American Academy of Sleep Medicine (AASM) criteria [5]. The R&K criteria classify sleep into six stages: wakefulness (W), rapid eye movement (REM), and non-REM stages S1, S2, S3, and S4 [6]. In contrast, the AASM criteria combine S3 and S4 into a single stage, N3 [7], resulting in five stages: W, N1, N2, N3, and REM. Manual sleep staging, typically performed by trained experts, involves analyzing 30-second sleep segments to determine the corresponding stage. While this method is generally accurate and reliable, it remains cumbersome, labor-intensive, and highly subjective [8], [9].

Received 2 July 2024; revised 28 November 2024; accepted 25 December 2024. Date of publication 30 December 2024; date of current version 7 May 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62171123, Grant 62211530112, Grant 62071241, and Grant 62101129, in part by the National Key Research and Development Program of China under Grant 2023YFC3603600, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20192004. *(Corresponding authors: Jianqing Li; Chengyu Liu.)*

Zhaoyang Cong, Minghui Zhao, Hongxiang Gao, Chang Yan, Li Ling, and Chengyu Liu are with the State Key Laboratory of Digital Medical Engineering, School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: zhaoyang_cong@seu.edu.cn; chengyu@seu.edu.cn).

Meng Lou is with the Department of Computer Science, The University of Hong Kong, Hong Kong, SAR, China.

Guowei Zheng is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150006, China.

Ziyang Wang is with the School of Management, University of Science and Technology of China, Hefei 230026, China.

Xingyao Wang is with the Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore 138632.

Jianqing Li is with the State Key Laboratory of Digital Medical Engineering, School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China, and also with the School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 211166, China (e-mail: ljq@seu.edu.cn).

Digital Object Identifier 10.1109/JBHI.2024.3523908

Polysomnography (PSG) remains the gold standard for sleep staging in medical environments. It integrates multiple physiological signals, including electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), electrooculography (EOG), and respiratory signals [10]. By jointly analyzing these signals, PSG enables accurate classification of sleep stages [9]. Despite its high accuracy, PSG is limited by the complexity and cost of its equipment, as well as its reliance on a controlled laboratory environment. Furthermore, the lack of portability and the time-intensive nature of signal acquisition and analysis impose significant burdens on both patients and healthcare professionals. These limitations hinder the widespread application of PSG in large-scale sleep studies and routine sleep monitoring [11], [12].

Single-channel EEG signals have garnered significant interest due to their portability and accessibility. However, the challenge of limited information necessitates the effective utilization of available data. Sleep staging methods based on single-channel EEG signals are generally divided into two categories: the first is feature engineering-based methods and the second is deep learning-based methods. Specifically, Feature engineering-based methods involve the extraction of handcrafted features and the application of traditional machine learning algorithms for classification. For instance, Liu et al. [13] proposed a method based on multi-domain feature extraction and genetic algorithm optimization, using least squares support vector machines (LS-SVM) for classification. Hassan et al. [14] proposed a method that utilizes tunable Q-factor wavelet transform (TQWT) decomposition, statistical feature extraction, and bootstrap aggregating classifier for efficient and accurate automatic sleep staging. Seifpour et al. [15] developed a sleep staging approach utilizing the statistical behavior features of local extrema, a supervised multi-cluster feature selection algorithm, and multi-class support vector machine (SVM) in single-channel EEG signals. Jiang et al. [16] introduced a method that involves decomposing single-channel EEG signals into multiple modalities to extract frequency and temporal features, initially classifying them using a random forest classifier, and then applying a hidden Markov model (HMM) for rule-free refinement to enhance the overall performance and stability of sleep staging. In contrast, deep learning methods can autonomously learn complex sleep patterns without the need for manual feature extraction. CNNs are predominantly utilized for extracting sleep-related waveform features. Sors et al. [17] developed a 14-layer CNN for automatic sleep staging from raw EEG data. Perslev et al. [18] introduced U-Time, a fully CNN-based model for time series segmentation applied to sleep staging. Supratak et al. [19] designed DeepSleepNet, which combines CNN and Bi-GRU for automatic sleep staging. Eldele et al. [11] integrated CNNs with channel attention to capture feature interdependencies, though their model faced challenges with longer sequences. Qu et al. [20] introduced an self-attention model leveraging residual connections and multi-scale decomposition for enhanced efficiency. Lastly, Mousavi et al. [21] introduced an automated method employing CNNs and a sequence-to-sequence model to capture context relationships between sleep epochs.

The characteristic EEG waveforms associated with sleep typically include δ waves (1–4 Hz) for deep sleep stages, θ waves (4–8 Hz) for light sleep stages, α waves (8-13Hz) for the initial sleep onset, sleep spindles (12–16 Hz), and K-complexes for N2, and β waves (13–30 Hz) for REM [22]. These waveforms have specific characteristics and frequencies in different sleep stages, which help to distinguish and understand each sleep stage. On the one hand, existing single-channel sleep staging deep learning models commonly use a CNN backbone network with large and small convolutional kernels to capture temporal features [11], [12], [19], [23]. And the utilization of multi-scale temporal feature extraction to capture diverse sleep EEG eigenwaveform features may be beneficial in improving the accuracy of sleep staging. In contrast, in addition to temporal information, spectral information is also crucial for the discrimination of sleep staging [12]. However, many current studies have not yet considered the complementary properties of temporal and spectral features. Typically each sleep PSG data collection lasts overnight, about 7–8 hours on average. Sleep staging tasks have a natural contextual relationship of features, and experts often combine features from the upper and lower time periods when determining sleep stages that are difficult to distinguish [24]. Human typical sleep cycles are usually 90 minutes long and repeat multiple times during the entire night. The transition of sleep stages also adheres to a specific pattern, in which the individual undergoes a state transition from light sleep (N1 and N2) to deep sleep (N3 and N4) and then to REM [25]. In summary, extracting multi-scale temporal information, fully combining temporal and spectral features, and utilizing contextual information as well as the patterns of sleep stage transitions are hopeful to further optimize the outcomes of sleep staging.

To address these issues, we propose a novel two-stage temporal-spectral fusion model called BiTS-SleepNet. BiTS-SleepNet is designed to comprehensively leverage temporal and spectral features from single-lead EEG signals. By integrating these features with a temporal context structure-based model, it emulates the decision-making process of clinical sleep staging specialists, enhancing the model's robustness in handling complex transition relationships between sleep stages. The results demonstrate that the proposed BiTS-SleepNet achieves competitive classification performance compared to recently published methods.

II. Method

A. Overview of BiTS-SleepNet Model

This section provides a detailed summary of our proposed BiTS-SleepNet model based on single-channel EEG, with an in-depth description of its overall structure and each block.

As shown in Fig. 1, our BiTS-SleepNet model is divided into two stages. The stage 1 network is mainly to extract discriminative features for sleep staging. First, we designed a dual-stream architecture consisting of two branches: a temporal feature extractor branch (TFEB) and a spectral feature extractor branch (SFEB), to extract temporal and spectral features from a 30-second single-channel EEG, respectively. Then, we developed a cross-attention-based temporal-spectral feature fusion module (TSFFM) to obtain a more discriminative feature representation of sleep staging. The features generated by TSFFM are fed into the fully connected layers to make an preliminary

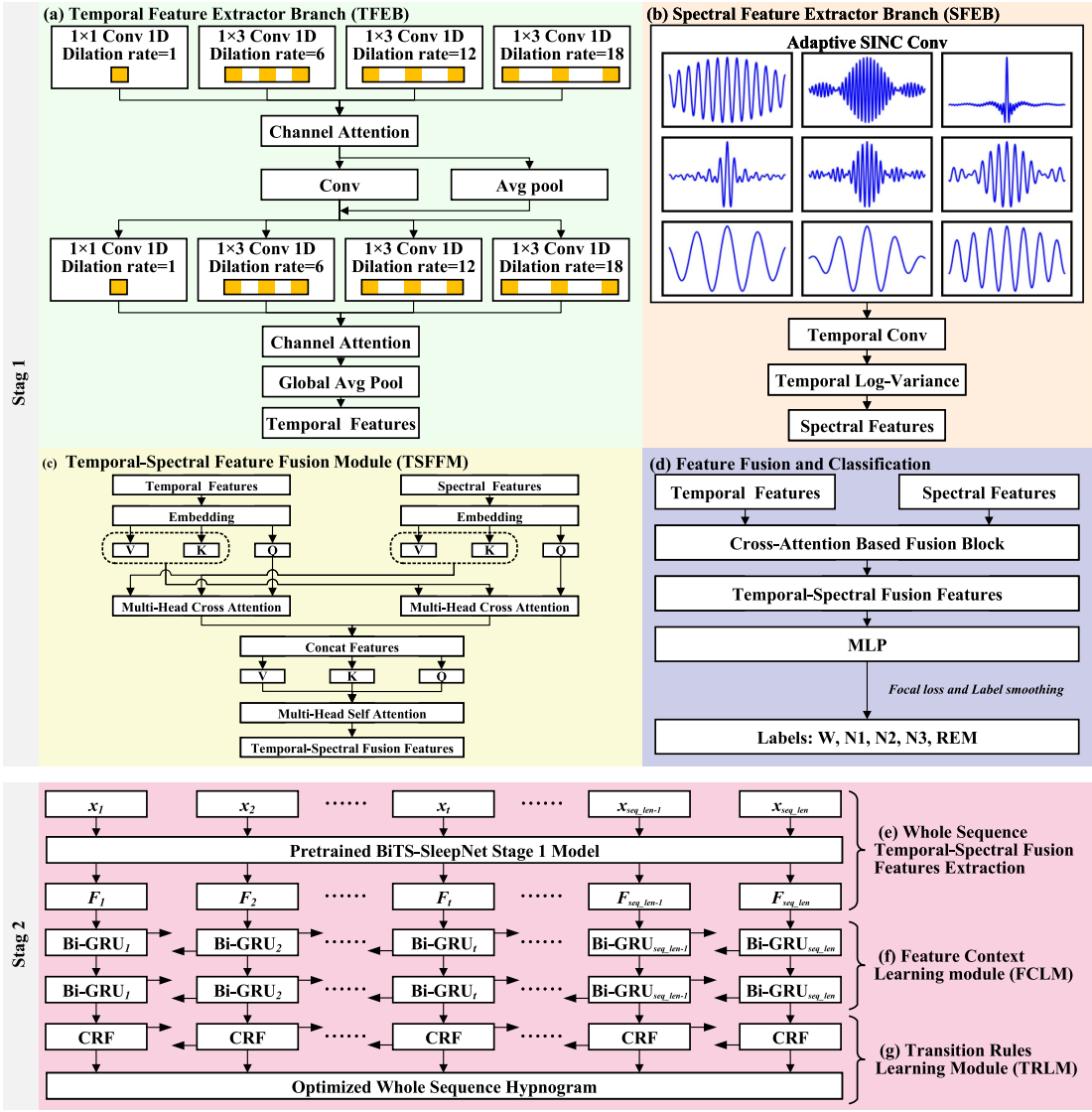

Fig. 1. The complete framework of the proposed BiTS-SleepNet model for automatic sleep staging. (a) TFEB: temporal feature extractor branch; (b) SFEB: spectral feature extractor branch; (c) TSFFM: temporal-spectral feature fusion module; (f) FCLM: feature context learning module; (g) TRLM: transition rules learning module. In the stage 1 network, two branches of raw signals are inputted to the model. One branch is inputted into TFEB to obtain temporal feature, and the other branch is inputted into SFEB to obtain spectral feature. Temporal feature and spectral feature are inputted into cross-attention based TSFFM, and the temporal-spectral fusion feature is obtained. Classification is performed using a focal loss combined with label smoothing. In the stage 2 network, the sequence data of an entire sleep cycle is processed. The temporal-spectral fusion feature is extracted for each sleep epoch by the pre-trained stage 1 model. The extracted features are then fed into Bi-GRU based FCLM for feature context learning. Finally the optimal hypnogram of the whole sequence is generated by the CRF based TRLM.

prediction of sleep stages in the stage 1. We have introduced focal loss with label smoothing to enhance the model's capability to handle unbalanced data and prevent the model from predicting the data with overconfidence, thus to promote the classification robustness. At this point, the model has not yet taken into account the association between ephemeral elements and existing inappropriate transitions. The BiTS-SleepNet stage 2 network is mainly responsible for learning the contextual relationships of sleep staging sequences, simulating the decision-making process of sleep staging experts, and performing sequence optimization on the output of the stage 1 network. The stage 2 network consists of two key modules: a feature context learning module (FCLM) based on Bi-GRU for capturing contextual dependencies, and a transition rules learning module (TRLM) based on Conditional Random Fields (CRF) structures to model stage transitions. After further optimize of the sleep staging results by the BiTS-SleepNet stage 2 network, the overall sleep staging results are ultimately output. The main components of the BiTS-SleepNet model are described as follows.

B. Temporal Feature Extractor Branch (TFEB)

This study integrates the principles of Atrous Spatial Pyramid Pooling (ASPP) [26] and residual learning in the TFEB to enhance the analysis of multi-view EEG signals for sleep stage classification [27]. As shown in Fig. 1(a), TFEB leverages ASPP,

utilizing multiple parallel atrous convolutional layers with varying dilation rates to characterize features at multiple scales. This enables the extraction of features across different receptive fields, crucial for decoding the intricate patterns in EEG signals. Incorporating residual blocks within the architecture aids in mitigating the challenges associated with training deep networks. These blocks use shortcuts to facilitate gradient flow, preventing the degradation problem and ensuring stable learning and convergence. The initial layer, a standard convolutional layer, preprocesses the input to prepare it for subsequent extraction and fusion of temporal features and amplify the channel capacity of the input signal to twice its original size. The final component of the architecture is a global average pooling layer (GAP), which transforms the extracted temporal features into a form suitable for feature fusion auxiliary classification. Sequentially arranged, TFEB utilizes ASPP blocks followed by squeeze-and-excitation (SE) layers, further refined through residual connections. These SE layers aim to recalibrate the feature channels adaptively, enhancing the model's focus on relevant signal characteristics.

C. Spectral Feature Extractor Branch (SFEB)

The SFEB consists of three main components, as shown in Fig. 1(b). The initial component is the SINC convolutional layer, which identifies frequency characteristics of single-channel EEG signals using specific SINC band-pass filters. The SINC convolutional block captures information from various EEG signal frequency bands through temporal convolution. In the first layer of SFEB, the input EEG signal undergoes convolutions with a set of band-pass filters, where f_L and f_H denote the learnable low and high cutoff frequencies, k represents the size of convolution kernel, respectively. The filter set, drawing from digital signal processing methods, combines rectangular band-pass filters. The amplitude of the band-pass filter in the frequency domain is expressed as the difference between two low-pass filters:

$$H[f, f_L, f_H] = \text{rect}\left(\frac{f}{2f_H}\right) - \text{rect}\left(\frac{f}{2f_L}\right), \quad (1)$$

where f denotes the signal frequency and rect represents the rectangular window in frequency domain. The band-pass filter's temporal representation, obtained through inverse Fourier transform, is:

$$\text{sinc}(x) = \frac{\sin(x)}{x}, \quad (2)$$

$$h[k, f_L, f_H] = 2f_H \cdot \text{sinc}(2\pi f_H k) - 2f_L \cdot \text{sinc}(2\pi f_L k), \quad (3)$$

To ensure $f_H > f_L > 0$ during training, the actual cutoff frequencies are defined as:

$$f_L^{\text{abs}} = |f_L|, \quad f_H^{\text{abs}} = f_L + |f_H - f_L|. \quad (4)$$

Since an ideal band-pass filter would require an infinite number of convolution samples, it is commonly approximated by multiplying the truncated filter by a window function v, resulting in the final temporal convolution function:

$$h_v[k, f_L, f_H] = h[k, f_L, f_H] \cdot v[k]. \quad (5)$$

In this work, the Hamming window is applied:

$$v[k] = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi k}{W}\right), \quad (6)$$

where W represents the window length, set to $W = 31$. f_L and f_H are initially selected from a uniform distribution spanning 0 to 48 Hz. During optimization, the frequencies are adjusted to stay within the desired range. Since all operations within the SINC convolutional block are fully differentiable, the cutoff frequencies can be optimized alongside other CNN parameters through gradient-based methods.

The second component is the temporal convolutional block, designed to learn multi-scale temporal features from the filtered EEG signals. This component employs hybrid depthwise convolution to capture multi-scale temporal features from EEG data. Each group is processed with convolution using kernels of varying receptive field sizes, starting from the segmented input feature map. The temporal kernel length for each group is determined based on a specific ratio of the EEG sampling rate, using coefficients [0.5, 0.25, 0.125, 0.0625] to capture multi-scale features and enhance computational efficiency [28], [29]. Thus, the input is evenly split into four groups, each learning nine temporal filters. The feature maps produced by these convolutions are then combined to generate the final output.

The final component, the temporal log-variance block, is tasked with computing variance features for each time series. This component focuses on extracting key features from sleep staging data to reduce its temporal dimension. To improve the Gaussianity of the variance features, a logarithmic transformation is applied, $g(x) = \log(x)$. The log-variance of each time series is then calculated in non-overlapping windows of size w.

D. Temporal-Spectral Feature Fusion Module (TSFFM)

There is complementary information that characterizes different sleep stages in the temporal and spectral features of EEG signals. However, these two modalities exhibit significant feature differences and data redundancy. Efficiently learning useful information while ignoring redundant information from temporal and spectral features is crucial for effective fusion of these features. We propose a temporal-spectral feature fusion module (TSFFM) that leverages cross-attention and self-attention mechanisms to effectively integrate temporal and spectral features, thereby enhancing feature representation capabilities.

The attention mechanism is a technique that assigns varying weights to positions in the input sequence based on their respective features. This enables the model to concentrate on the most pertinent parts for the current task, enhancing both efficiency and effectiveness. The attention mechanism can be formulated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (7)$$

where d_k is the dimension of the keys, and Q, K, and V represent the query, key, and value matrices, respectively. The attention mechanism computes the dot product of the query and key matrices, normalizes it, and applies these weights to the value matrix to produce the final output.

The multi-head attention (MHA) mechanism is a key technique for enhancing the model's ability to attend various aspects of the features. By employing multiple parallel attention heads, the model can independently perform attention operations in different subspaces, capturing diverse information from the input features. Each attention head has its own set of query, key, and value transformation matrices. In this work, the MHA mechanism uses 8 heads. The computation of MHA is as follows:

$$\text{MHA}(Q,K,V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O, \tag{8}$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, W_i^Q, W_i^K, and W_i^V are the query, key, and value weight matrices for the ith head, respectively, and W^O is the output transformation matrix.

Since the dimensions of the temporal and spectral features are inconsistent, a linear transformation is first used to adjust the spectral features to match the dimensions of the temporal features. Let the temporal features be F_t (with a dimension of 256) and the spectral features be F_s (with a dimension of 96). The adjustment of the spectral features is given by:

$$F_s' = W_s F_s + b_s, \tag{9}$$

where W_s and b_s denote the weight matrix and bias vector for the linear transformation, respectively, resulting in the adjusted spectral features F_s' with a dimension of 256.

In the cross-attention layer, temporal and spectral features are processed through two cross MHA mechanisms to extract information from each other. Specifically, the steps are as follows: First, the cross-attention from temporal to spectral features is computed as:

$$F_{s\leftarrow t} = \text{MHA}(Q = F_s', K = F_t, V = F_t). \tag{10}$$

Next, the cross-attention from spectral to temporal features is computed as:

$$F_{t\leftarrow s} = \text{MHA}(Q = F_t, K = F_s', V = F_s'), \tag{11}$$

where $F_{s\leftarrow t}$ denotes the spectral features guided by the temporal features, and $F_{t\leftarrow s}$ denotes the temporal features guided by the spectral features. The features obtained from the cross-attention mechanisms, $F_{s\leftarrow t}$ and $F_{t\leftarrow s}$, are concatenated to form a new combined feature representation:

$$F_{\text{concat}} = [F_{s\leftarrow t}; F_{t\leftarrow s}]. \tag{12}$$

Then, the combined features are further enhanced using a self-attention mechanism:

$$F_{\text{enhanced}} = \text{MHA}(Q = F_{\text{concat}}, K = F_{\text{concat}}, V = F_{\text{concat}}). \tag{13}$$

Finally, the enhanced features are mapped to the final output dimensions through a linear layer and layer normalization:

$$F_{\text{fusion}} = \text{LayerNorm}(W_f F_{\text{enhanced}} + b_f). \tag{14}$$

Through these mechanisms, the TSFFM effectively integrates temporal and spectral features, improving feature representation capabilities and thus enhancing the classification performance in sleep staging tasks.

E. Loss Functions, Focal Loss and Label Smoothing

In machine learning, especially in multi-class classification tasks, a common objective is to minimize the discrepancy between the predicted probabilities and the true label distribution of labels across various classes. Traditionally, cross-entropy loss has been the cornerstone for these tasks, defined for each class c and sample i as:

$$L_{CE} = -\sum_{c=1}^{N_{\text{classes}}} y_{i,c} \log(p_{i,c}). \tag{15}$$

Here, i represents the index of a sample in the dataset, allowing us to individually calculate the contribution of each sample to the overall loss. $y_{i,c}$ denotes the true label distribution, indicating whether class c is the correct classification for sample i, while $p_{i,c}$ is the model's predicted probability for class c.

While cross-entropy loss effectively encourages models to predict accurate probabilities for each class, it does not adequately address the challenges of class imbalance or focus on hard-to-classify examples. This limitation can lead to models that perform well on majority classes but poorly on minority classes, which are often of greater interest.

To tackle the aforementioned problems, a weighted cross-entropy loss applicable to sleep staging task was introduced [11]:

$$L_{WCE} = -\frac{1}{M}\sum_{c=1}^{N_{\text{classes}}}\sum_{i=1}^{M} w_c y_{i,c} \log(p_{i,c}), \tag{16}$$

$$w_c = \mu_c \cdot \max\left(1, \log\left(\frac{\mu_c M}{M_c}\right)\right), \tag{17}$$

where μ_c refers to an adjustable parameter, w_c denotes the weight given to class c, and M_c indicates the sample count within class c.

In object detection task, focal loss is shown to be more effective than weighted cross-entropy loss in handling category imbalance [30]. It improves the performance of the model in detecting small targets and few categories of targets. Focal loss incorporates a difficulty adjustment factor, now denoted as $d_{i,c} = (1 - p_{i,c})^\gamma$, to adjust the loss dynamically according to the confidence level of the prediction. The focal loss is defined as:

$$L_{FL} = -\sum_{c=1}^{N_{\text{classes}}} d_{i,c} y_{i,c} \log(p_{i,c}). \tag{18}$$

Unlike weighted cross-entropy loss, which requires manual setting of category weights, focal loss uses an adjustment factor that automatically modulates the loss contribution of samples. This eliminates the need for manual weight tuning and reduces the complexity of parameter optimization.

However, despite the advantages of focal loss in focusing on hard-to-classify examples, it may still result in overconfidence in the model's predictions. This is where label smoothing [31] comes into play, adjusting the true label distribution to prevent

overconfidence. The adjusted labels y' are introduced as:

$$y'_{i,c} = \begin{cases} 1-\epsilon & \text{if } c = \text{the true class} \\ \frac{\epsilon}{N_{\text{classes}}-1} & \text{otherwise} \end{cases}, \tag{19}$$

where ϵ is the label smoothing parameter. L_{FLS}, integrating the principles of focal loss and label smoothing, is employed to not only focus on hard-to-classify examples but also to prevent overconfidence in predictions. The comprehensive loss function becomes:

$$L_{FLS} = -\sum_{c=1}^{N_{\text{classes}}} d_{i,c} y'_{i,c} \log(p_{i,c}). \tag{20}$$

By combining focal loss with label smoothing, the model enhances both the robustness and the generalization across diverse classification scenarios, focusing on difficult examples while avoiding overconfidence in its predictions.

F. Feature Context Learning Module (FCLM) and Transition Rules Learning Module (TRLM)

To address the contextual feature learning challenge in sleep staging tasks, we introduce a feature context learning module (FCLM) in the BiTS-sleepNet stage 2 network. The FCLM is composed of a two-layer Bi-GRU design, intended to leverage the temporal context dependencies inherent in sleep data and improve the precision of sequence modeling.

The FCLM architecture commences with an input layer configured to process sequences of a predetermined maximum length and corresponding feature size. The Bi-GRU layers, with the first layer comprising 704 units and the second layer comprising 352 units, are employed to extract features from both past and future contexts effectively:

$$\overrightarrow{l_t} = \overrightarrow{GRU}(i_t, \overrightarrow{l_{t-1}}), \quad \overleftarrow{l_t} = \overleftarrow{GRU}(i_t, \overleftarrow{l_{t+1}}), \tag{21}$$

$$L_t = [\overrightarrow{l_t}; \overleftarrow{l_t}], \tag{22}$$

where i_t denotes the input feature at time step t, L_t signifies the concatenated hidden states from both directions at time step t, $\overrightarrow{l_t}$ and $\overleftarrow{l_t}$ represent the hidden states of the forward and backward GRU, respectively.

Following FCLM, we introduce a transition rules learning module (TRLM) composed of a CRF layer. This CRF layer is utilized to model the label sequence by considering the likelihood of the label sequence conditioned on the input sequence, thereby capturing the interdependencies between labels:

$$P(s|x) = \frac{\exp(\sum_{t=1}^{T}\sum_{j=1}^{J}\alpha_j\phi_j(s_{t-1}, s_t, H_t))}{\sum_{s'\in S}\exp(\sum_{t=1}^{T}\sum_{j=1}^{J}\alpha_j\phi_j(s'_{t-1}, s'_t, H_t))}, \tag{23}$$

where s denotes the label sequence, x represents the input sequence, S is the set of all possible label sequences, T indicates the sequence length, J is the number of feature functions, ϕ_j are the feature functions, and α_j are the learned weights.

The integration of FCLM and TRLM models both high-level features and the transition dynamics between output labels. This is beneficial for sleep staging, which require understanding sequential features and contextual label information.

III. Experiment

A. Datasets

Sleep-EDF-20 [32], [33], [34]: The Sleep-EDF-20 dataset, obtained from PhysioBank, comprises 39 PSG records of 20 healthy adult Caucasians aged 25 to 34 years. It includes two parts: Sleep Cassette (SC) and Sleep Telemetry (ST) groups, with this study utilizing the SC segment. The dataset features dual-channel EEG, one horizontal EOG, one EMG, and one oro-nasal respiration signal, with EEG and EOG sampled at 100 Hz. Sleep stages are manually classified by specialists into several categories according to the R&K standard.

Sleep-EDF-78 [32], [33], [34]: An extension of the Sleep-EDF-20, The Sleep-EDF-78 dataset comprises 153 PSG records from 78 subjects aged between 25 and 101. It mirrors the Sleep-EDF-20in terms of data structure and classification, with the primary difference being the larger number of subjects and the inclusion of a broader age range. Similar to Sleep-EDF-20, it offers a rich set of signals for sleep stage classification.

SHHS (Sleep Heart Health Study) [35], [36]: The SHHS dataset originates from a multi-center cohort study focusing on subjects with various medical conditions, such as cardiovascular and lung diseases. Following the criteria outlined in [37], we selected subjects with normal sleep patterns, defined as an apnoea-hypopnoea index (AHI) below 5. A total of 329 subjects were included, with EEG data recorded from the C4-A1 channel at a sampling rate of 125 Hz.

None of the subjects in the three datasets were patients with neurological disorders. The preprocessing of the three datasets involved the following steps [11], [19]. First, all epochs labeled as UNKNOWN, which do not correspond to any specific sleep stage, were excluded. Second, stages N4 were merged into stage N3 in line with AASM guidelines. Third, to enhance focus on sleep stages, 30 minutes of wakefulness were included both before and after sleep periods. The sleep stage information for the three datasets utilized in our study are shown in Table I.

B. Implementation Details

Model training was performed on NVIDIA GeForce RTX 3090 GPU. For stage 1 training, a batch size of 512 and the Adam optimizer were used. The Adam optimizer's weight decay was configured to 1e-3 with beta1 and beta2 of 0.9 and 0.999, respectively. A cosine annealing learning rate decay with warmup was used. The warmup step was set to 10, the maximum learning rate to 5e-4, and the model was trained for 100 epochs with early stopping and a patience of 10 epochs. All convolutional layers are initialized with a Gaussian distribution having a mean of 0 and a variance of 0.02. For stage 2 training, the batch size was set to 64 with the Adam optimizer, and the maximum number of training epochs was 200, with an early stopping patience of 10 epochs. The learning rate was initialized to 1e-3, with an L2 regularization coefficient of 1e-2. The stage 2 model was trained

TABLE I
SUMMARY OF DATASETS, THEIR CHARACTERISTICS, AND SLEEP STAGE DISTRIBUTION

Dataset	Sampling Rate	Subjects	Total Samples	Channel	Number of Folds	W	N1	N2	N3	REM
Sleep-EDF-20	100Hz	20	42308	Fpz-Cz	20-Fold	8285 19.58%	2804 6.63%	17799 42.06%	5703 13.48%	7717 18.24%
Sleep-EDF-78	100Hz	78	195479	Fpz-Cz	20-Fold	65951 33.74%	21522 11.01%	69132 35.37%	13039 6.67%	25835 13.21%
SHHS	125Hz	329	324854	C4-A1	20-Fold	46319 14.26%	10304 3.17%	142125 43.76%	60153 18.52%	65953 20.30%

using the CRF loss function to optimize transition rules learning performance.

C. Experimental Setup

Model performance was assessed using subject-independent validation and 20-fold cross-validation, following the approaches performed in previous studies [11], [12], [23], [38]. In each round, 19 folds were allocated for training and 1 fold for testing, and the whole process was repeated 20 times. Using the Sleep-EDF-20 dataset as an illustration, each fold involves training with 19 subjects' data and testing with 1 subject's data. Ultimately, we combine the prediction results from all 20 rounds of test samples to compute various performance evaluation metrics.

D. Evaluation Metrics

Sleep stage classification is an imbalanced multi-class task that requires multiple metrics for evaluation. For individual class metrics, Precision (PR) and Recall (RE) are used. When calculating metrics for a single class, data from all other classes are considered as the negative class. Assuming there are N_c classes and N_s samples in total, let TP_i, FP_i, TN_i, and FN_i represent the true positives, false positives, true negatives, and false negatives for the ith class, respectively.

The overall metrics include overall classification accuracy (ACC), macro-averaged F1-score (MF1), Cohen's Kappa Coefficient (κ), and macro-averaged G-mean (MGm). The calculation formulas are as follows:

$$ACC = \frac{1}{N_s}\sum_{i=1}^{N_c} TP_i, \tag{24}$$

$$MF1 = \frac{1}{K}\sum_{i=1}^{K} \frac{2 \times \mathrm{PR}_i \times \mathrm{RE}_i}{\mathrm{PR}_i + \mathrm{RE}_i}, \tag{25}$$

$$\kappa = \frac{ACC - P_e}{1 - P_e}, \tag{26}$$

$$MGm = \frac{1}{K}\sum_{i=1}^{K} \sqrt{\mathrm{SP}_i \times \mathrm{RE}_i}, \tag{27}$$

where $PR_i = \frac{TP_i}{TP_i+FP_i}$, $RE_i = \frac{TP_i}{TP_i+FN_i}$, $SP_i = \frac{TN_i}{TN_i+FP_i}$, $P_e = \frac{1}{N_s^2}\sum_{i=1}^{N_c}(TP_i + FP_i) \times (TP_i + FN_i)$.

Per-class precision, recall, F1-score, and G-mean were also used to evaluate each model. These metrics are calculated similarly to binary classification, treating one class as positive and the remaining four as negative.

IV. RESULTS AND DISCUSSION

A. Model Classification Performance

The evaluation results of BiTS-SleepNet on each of the three datasets, Sleep-EDF-20, Sleep-EDF-78, and SHHS, are shown in Tables III, IV and V. These evaluation results include the confusion matrix and the evaluation metrics for each category. The left side of the table displays the confusion matrix, where rows represent the true labels given by experts and columns represent the labels predicted by the model. The right side of the table shows the performance metrics of the model computed from the confusion matrix for the five sleep stages, including accuracy, recall, F1 score, and Gm. We observe that a large percentage of all true labels are the bold labeled elements of the confusion matrix, proving the effectiveness of our method. Compared to other labels, N1 has lower predictive accuracy and performs poorly in terms of recall, precision, F1 score and MGm. This aligns with the fact that N1 labels are more challenging for clinicians to identify in a clinical setting. N1 is frequently confused with N2, W, and REM.

B. Performance Comparison and Model Analysis

As illustrated in Table II, we present a comparison of the proposed BiTS-SleepNet with various recently published methods on the Sleep-EDF-20, Sleep-EDF-78, and SHHS datasets. BiTS-SleepNet achieves competitive performance on all three datasets, with high per-class F1-scores across five subcategories. In terms of overall metrics, it attains the highest ACC, MF1, κ, and MGm scores. This result demonstrates the model effectiveness in addressing class imbalance, while maintaining high classification accuracy. We attribute the performance of our model to the following key factors: First, in stage 1, our network extracts multi-scale temporal features and multi-band spectral features, and performs feature fusion to enhance the feature representation of the original signal. As shown in the t-SNE visualization in Fig. 2, SFEB and TFEB learn distinct feature representations, while the TSFFM further contributes to the feature representation. The ablation experiments also demonstrate this. Second, in stage 2, the FCLM and TRLM enable the model to simulate the sleep staging patterns of experts. By incorporating contextual information from surrounding epochs and using transformation rules to correct misclassifications from stage 1, the model improves classification accuracy. Specifically,

TABLE II
COMPARISON AMONG BITS-SLEEPNET AND RECENTLY PUBLISHED METHODS

Dataset	Method	Per-Class F1-score					Overall Metrics			
		W	N1	N2	N3	REM	ACC	MF1	κ	MGm
Sleep-EDF-20	DeepSleepNet [19]	86.70	45.50	85.10	83.30	82.60	81.90	76.60	0.76	-
	SleepEEGNet [21]	89.40	44.40	84.70	84.60	79.60	81.50	76.60	0.75	-
	ResnetLSTM [39]	86.50	28.40	87.70	89.80	76.20	82.50	73.70	0.76	81.80
	MultitaskCNN [40]	87.90	33.50	87.50	85.80	80.30	83.10	75.00	0.77	83.10
	AttnSleep [11]	89.70	42.60	88.80	90.20	79.00	84.40	78.10	0.79	85.50
	SeqSleepNet [41]	87.70	43.80	88.20	86.50	84.00	84.60	78.00	0.79	85.30
	MRASleepNet [23]	90.70	46.10	88.70	88.20	80.70	84.50	78.90	0.79	-
	NASNet [42]	86.60	38.80	87.80	88.50	77.60	82.70	75.90	0.76	84.00
	DSNet [43]	90.30	40.10	89.50	**90.60**	81.90	85.80	78.50	0.80	85.30
	CausalAttenNet [44]	89.60	40.10	88.60	**90.60**	80.80	84.70	78.10	0.80	85.50
	LGSleepNet [38]	91.80	49.40	89.60	89.80	82.60	86.00	80.70	0.76	88.20
	TSA-Net [12]	90.54	46.91	89.17	90.10	85.35	86.64	80.41	0.82	86.44
	Zhao et al. [45]	90.40	53.70	88.30	88.30	85.10	85.60	81.10	0.79	-
	TBSTSleepNet [46]	**92.50**	53.30	89.70	88.50	87.70	87.20	82.30	-	-
	CBLSNet [47]	91.00	50.50	89.00	86.90	88.30	86.40	81.10	0.83	-
	BiTS-SleepNet (ours)	92.18	**55.56**	**90.49**	90.48	**88.89**	**88.50**	**83.52**	**0.84**	**88.89**
Sleep-EDF-78	SleepEEGNet [21]	89.80	42.10	75.20	70.40	70.60	74.20	69.60	0.66	-
	ResnetLSTM [39]	90.70	34.70	83.60	80.90	67.00	78.90	71.40	0.71	80.80
	MultitaskCNN [40]	90.90	39.70	83.20	76.60	73.50	79.60	72.80	0.72	82.50
	AttnSleep [11]	92.00	42.00	85.00	82.10	74.20	81.30	75.10	0.74	83.60
	SeqSleepNet [41]	91.80	46.00	85.00	77.50	81.00	82.60	76.30	0.76	84.30
	MRASleepNet [23]	92.00	44.80	84.90	80.10	75.20	81.40	75.40	0.74	-
	NASNet [42]	91.10	39.20	84.00	81.00	68.10	80.00	72.70	0.72	81.20
	DSNet [43]	92.10	33.70	84.60	82.20	74.00	81.00	73.00	0.73	81.23
	CausalAttenNet [44]	92.30	39.40	84.80	82.50	74.80	81.60	75.20	0.75	83.60
	LGSleepNet [38]	92.60	43.70	85.50	**83.00**	74.90	82.30	76.00	0.75	84.90
	TSA-Net [12]	91.71	30.94	84.94	82.21	81.03	82.21	73.57	0.75	82.26
	Zhao et al. [45]	92.90	**52.60**	84.90	78.90	83.70	83.40	78.60	0.77	-
	CBLSNet [47]	92.90	50.90	85.50	80.90	83.50	84.10	78.70	**0.79**	-
	BiTS-SleepNet (ours)	**93.33**	50.23	**87.05**	82.50	**85.33**	**85.09**	**79.69**	**0.79**	**86.27**
SHHS	SleepEEGNet [21]	81.30	34.40	73.40	75.90	77.00	73.90	68.40	0.65	-
	ResnetLSTM [39]	85.10	9.40	86.30	87.00	79.10	83.30	69.40	0.76	76.40
	MultitaskCNN [40]	82.20	25.70	83.90	83.30	81.10	81.40	71.20	0.74	80.40
	NASNet [42]	84.80	36.90	84.40	80.90	80.60	81.90	73.50	0.74	80.70
	AttnSleep [11]	86.70	33.20	87.10	**87.10**	82.10	84.20	75.30	0.78	84.00
	SeqSleepNet [41]	84.20	47.30	87.20	85.40	88.60	85.60	78.50	0.80	85.40
	CausalAttenNet [44]	85.40	27.60	84.80	84.00	82.90	83.10	73.10	0.77	83.00
	Zhao et al. [45]	**88.70**	**51.40**	87.00	83.20	88.60	85.60	79.70	0.79	-
	BiTS-SleepNet (ours)	87.83	49.51	**88.20**	84.77	**90.93**	**87.01**	**80.25**	**0.82**	**86.58**

Note: The Per-Class F1-score, ACC, MF1, and MGm are expressed as percentages (%).
The numbers in bold represent the best performance, and the numbers underlined indicate the second-best performance.

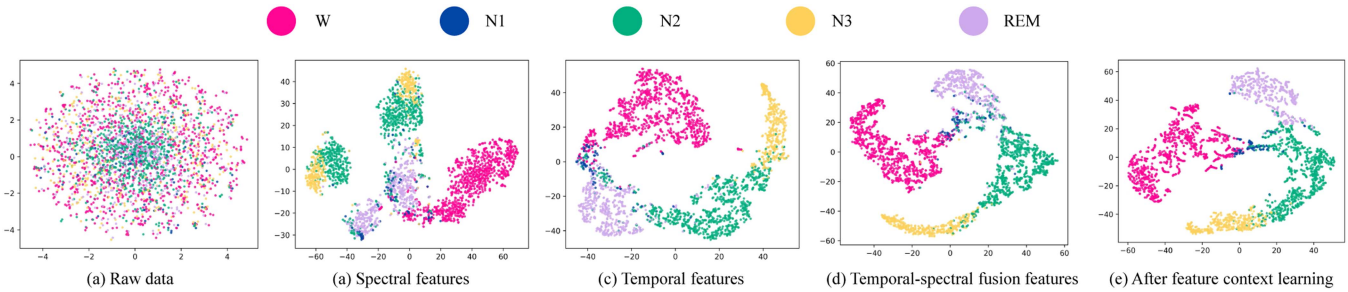

Fig. 2. T-SNE visualization of feature representations from each module in BiTS-SleepNet.

the transition probabilities and matrices in Fig. 4 demonstrate the adaptive learning capabilities of stage 2.

C. Feature Context Learning and Transition Rules Learning

The stage 2 network is dedicated to learning sleep stage context features and transition rules with the aim of correcting the preliminary predictions of stage 1. To observe the effect of stage 2, we visualized the results of stage 1 whole-night sleep staging, the results corrected by the stage 2 transition rules, and the sleep expert staging results. As shown in Fig. 3(a), the stage 1 results have some abnormal sleep stage transitions, especially in the part within the red box where the staging results change frequently. Fig. 3(b) demonstrates the sleep staging results after being enhanced by the stage 2 network

TABLE III
CONFUSION MATRIX AND PER-CLASS METRICS FOR BITS-SLEEPNET ON THE FPZ-CZ CHANNEL OF THE SLEEP-EDF-20 DATASET

	Predicted					Per-class metrics (%)			
	W	N1	N2	N3	REM	PR	RE	F1	GM
W	**7687**	285	220	12	81	92.78	91.58	92.18	95.32
N1	450	**1363**	621	11	359	48.61	64.84	55.56	69.07
N2	155	308	**16256**	568	512	91.33	89.66	90.49	91.84
N3	19	1	482	**5201**	0	91.20	89.77	90.48	94.72
REM	83	145	551	2	**6936**	89.88	87.93	88.89	93.49

Correct predictions in bold.

TABLE IV
CONFUSION MATRIX AND PER-CLASS METRICS FOR BITS-SLEEPNET ON THE EDF-78 DATASET

	Predicted					Per-class metrics (%)			
	W	N1	N2	N3	REM	RE	PR	F1	GM
W	**62087**	2841	564	35	424	94.14	92.53	93.33	95.13
N1	3950	**9533**	6398	75	1566	44.29	57.99	50.23	65.22
N2	566	2836	**62114**	1724	1892	89.85	84.42	87.05	90.39
N3	27	21	2506	**10475**	10	80.34	84.79	82.50	89.17
REM	467	1208	1993	45	**22122**	85.63	85.04	85.33	91.47

Correct predictions in bold.

TABLE V
CONFUSION MATRIX AND PER-CLASS METRICS FOR BITS-SLEEPNET ON THE C4-A1 CHANNEL OF THE SHHS DATASET

	Predicted					Per-class metrics (%)			
	W	N1	N2	N3	REM	RE	PR	F1	GM
W	**42047**	1249	2155	216	652	90.78	85.07	87.83	94.01
N1	1898	**4427**	2856	7	1116	42.96	58.42	49.51	65.22
N2	3252	1486	**126010**	7286	4091	88.66	87.73	88.20	89.51
N3	629	4	9119	**49908**	493	82.97	86.66	84.77	89.76
REM	1599	412	3488	175	**60279**	91.40	90.47	90.93	94.42

Correct predictions in bold.

containing Bi-GRU and CRF. Compared to Fig. 3(a), the results show higher consistency and stability, especially in the unstable region within the red box, and the prediction results are smoother and more coherent. In general, the staging results are more similar between Fig. 3(b) and (c) (expert staging results), which suggests that the stage 2 network is able to effectively characterize the temporal contextual dependencies of the sleep stages, and better approximates the expert's labeling results.

D. Ablation Study

As shown in Fig. 1, our proposed model contains two stages, stage 1 and stage 2, and roughly consists of five parts: TFEB, SFEB, TSFFM, focal loss with label smoothing, FCLM, and TRLM. To assess the impact of each block, we conducted an ablation analysis based on Sleep-EDF-20 dataset, following the procedure outlined in Fig. 1. The specific variant models are designed as follows:

W/O TFEB and TSFFM: Our model removes TFEB and TSFFM in stage 1.

W/O FCLM: Our model removes FCLM in stage 2.

W/O stage 2 network: Our model removes stage 2, leaving only stage 1.

W/O SFEB and TSFFM: Our model removes SFEB and TSFFM in stage 1.

W/O TSFFM: Our model removes TSFFM from stage 1 and uses concat for dual-stream temporal spectral feature fusion.

W/O TRLM: Our model removes TRLM in stage 2.

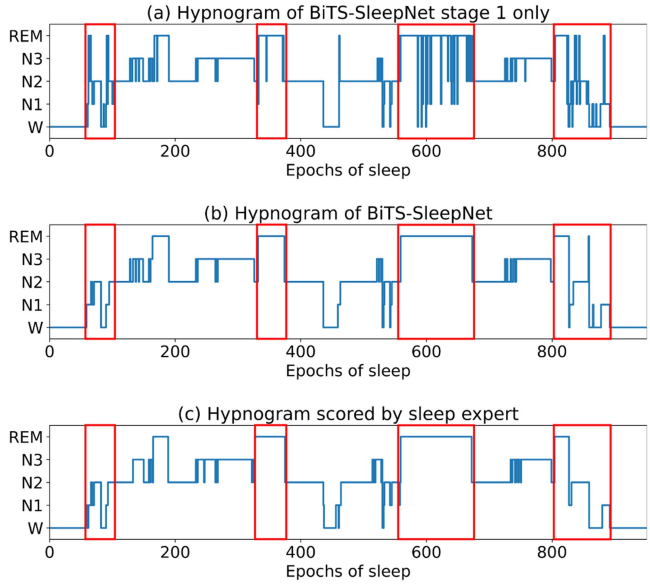

Fig. 3. Hypnograms for a randomized subject (ID: SC4151E0) within Sleep-EDF-20 on the Fpz-Oz channel, including the preliminary outcome hypnogram for BiTS-SleepNet stage 1 (subfigure a), the hypnogram for our model BiTS-SleepNet (subfigure b), the sleep staging expert labeled hypnogram (subfigure c). The red boxes emphasize the obvious incorrect staging.

TABLE VI
OVERALL METRICS FOR ABLATION STUDY OF BITS-SLEEPNET

Experiment	Model	ACC (%)	MF1 (%)	κ	MGm (%)
Exp1	W/O TFEB and TSFFM	79.46	69.40	0.71	77.97
Exp2	W/O FCLM	85.52	77.93	0.80	84.31
Exp3	W/O stage 2 network	85.91	79.02	0.81	85.63
Exp4	W/O SFEB and TSFFM	86.73	80.56	0.82	86.56
Exp5	W/O TSFFM	87.47	81.61	0.83	87.22
Exp6	W/O TRLM	87.87	82.10	0.83	87.45
Exp7	W/O focal loss and label smoothing	88.03	82.41	0.83	87.99
Exp8	BiTS-SleepNet (all)	88.50	83.52	0.84	88.89

W/O focal loss and label smoothing: Our model removes the focal loss and label smoothing used in stage 1 training and replaces it with weighted cross entropy.

BiTS-SleepNet (all): Our proposed completed BiTS-SleepNet model.

From the results of Table VI, we draw the following conclusions: By comparing the results of Exp1, Exp4, and Exp5, we observe that the classification performance in Exp5, which combines both temporal and spectral features, outperforms Exp4 (which uses only TFEB) and Exp1 (which uses only SFEB). This suggests that fusing temporal and spectral features enhances the model's ability to capture the discriminative features relevant for sleep staging from multiple views. Due to the imbalance in sleep staging data and the heterogeneity of some staging results, Exp8, compared to Exp7, enhances the model's ability to handle data distribution imbalance. This is achieved by introducing a loss function based on focal loss with label smoothing, replacing the weighted cross-entropy loss. As a result, the accuracy of sleep staging is improved. Comparing Exp3, Exp2, Exp6, and Exp8 shows that stage 2 of BiTS-SleepNet, which learns feature context relations and transition rules to mimic sleep staging experts, improves overall classification performance. Additionally, it demonstrates that the FCLM aids the TRLM in better

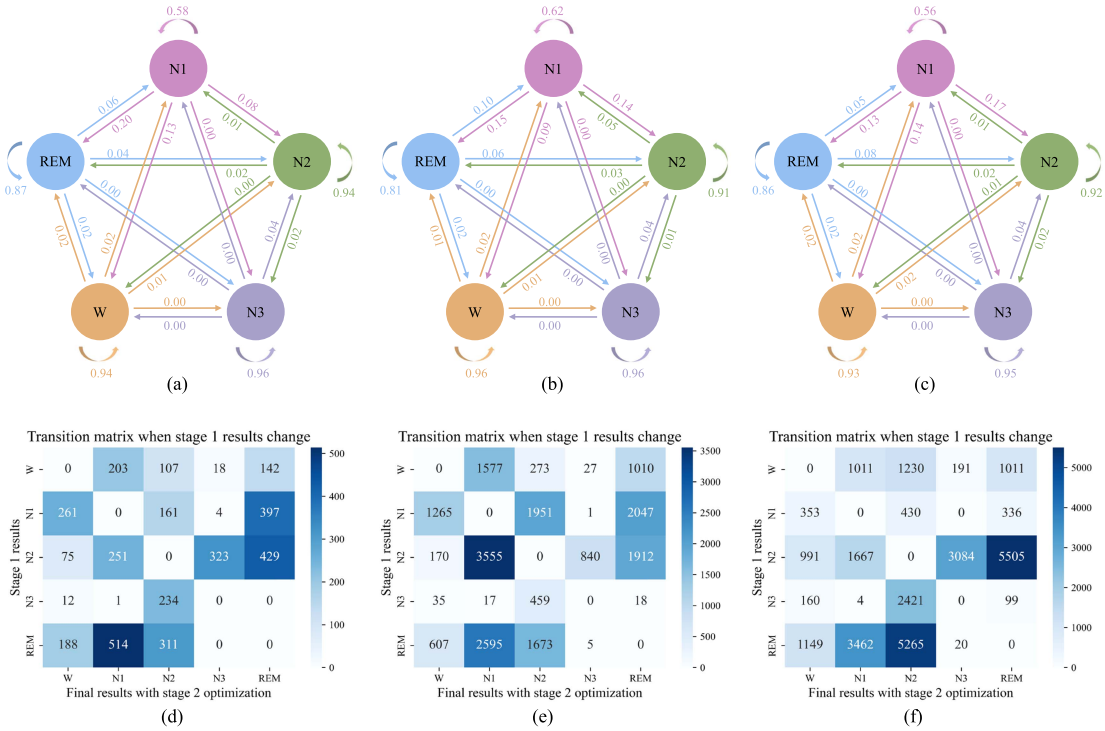

Fig. 4. Transition probabilities and matrices across the three datasets. (a)–(c) represent the transition probabilities of the CRF correction module for Sleep-EDF-20, Sleep-EDF-78, and SHHS, respectively. (d)–(f) represent transition matrices of sleep stages calculated on the Sleep-EDF-20, Sleep-EDF-78, and SHHS datasets when stage 1 results and final results differ.

learning transition rules. Comparing Exp5 and Exp8 shows that TSFFM outperforms concat fusion in fusing dual-stream features from the temporal and spectral domains, leading to better discriminative feature extraction for sleep staging.

E. Limitations and Future Perspectives

Our work still has some limitations. First, despite improving the N1 stage classification accuracy, the overall accuracy for this stage remains low. Our next challenge is to extract distinctive features of the N1 stage without compromising the classification accuracy of other stages. Second, wearable devices require effective daily sleep detection. Although our BiTS-SleepNet, based on single-channel EEG signals, shows potential for wearable applications, it was still trained on data collected from PSG devices. In the future, we will focus on developing lightweight real-time models for wearable devices. Finally, since labeling data by sleep staging specialists is time-consuming and labor-intensive, we plan to explore self-supervised and unsupervised learning strategies to reduce the model's reliance on labeled data in the future.

V. Conclusion

This work presents BiTS-SleepNet, a novel two-stage automated sleep staging model based on single-channel EEG. The first stage of the model comprises two main blocks: (1) a dual-stream feature extractor that captures multi-scale temporal and spectral features, and (2) a cross-attention-based feature fusion block that integrates these temporal and spectral features. To address the data imbalance problem and avoid overconfidence of the model in the classification results, we introduce a focal loss with label smoothing to train the model. The second stage consists of a feature context learning module based on Bi-GRU and a transition rules learning module based on CRF, which were used to learn time-dependent dependencies, rationalize sleep stage transitions, and optimize the staging results of the first stage as the final output. The BiTS-SleepNet model was evaluated on three public sleep staging datasets: Sleep-EDF-20, Sleep-EDF-78, and SHHS. The experimental results demonstrate that the BiTS-SleepNet model outperforms recently published methods in sleep staging performance.

References

[1] M. G. Umlauf, A. C. Bolland, K. A. Bolland, S. Tomek, and J. M. Bolland, "The effects of age, gender, hopelessness, and exposure to violence on sleep disorder symptoms and daytime sleepiness among adolescents in impoverished neighborhoods," *J. Youth Adolescence*, vol. 44, pp. 518–542, 2015.

[2] D. Zhao, Y. Wang, Q. Wang, and X. Wang, "Comparative analysis of different characteristics of automatic sleep stages," *Comput. Methods Programs Biomed.*, vol. 175, pp. 53–72, 2019.

[3] H. Korkalainen et al., "Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 7, pp. 2073–2081, Jul. 2020.

[4] E. A. Wolpert, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," *Arch. Gen. Psychiatry*, vol. 20, no. 2, pp. 246/247, 1969.

[5] R. B. Berry et al., "The AASM manual for the scoring of sleep and associated events," in *Rules, Terminology and Technical Specifications*, vol. 176. Darien, IL, USA: American Academy of Sleep Medicine, 2012, pp. 1–7.

[6] D. Moser et al., "Sleep classification according to AASM and Rechtschaffen & Kales: Effects on sleep scoring parameters," *Sleep*, vol. 32, no. 2, pp. 139–149, 2009.

[7] L. Novelli, R. Ferri, and O. Bruni, "Sleep classification according to AASM and Rechtschaffen and Kales: Effects on sleep scoring parameters of children and adolescents," *J. Sleep Res.*, vol. 19, no. 1p2, pp. 238–247, Mar. 2010.

[8] T. Mitterling et al., "Sleep and respiration in 100 healthy Caucasian sleepers—A polysomnographic study according to American academy of sleep medicine standards," *Sleep*, vol. 38, no. 6, pp. 867–875, 2015.

[9] S. Chowdhury, T.-A. Song, R. Saxena, S. Purcell, and J. Dutta, "250 AI-supported sleep staging from activity and heart rate," *Sleep*, vol. 44, 2021, Art. no. A101.

[10] S. A. Keenan, "An overview of polysomnography," *Handbook Clin. Neurophysiol.*, vol. 6, pp. 33–50, 2005.

[11] 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.

[12] G. Fu, Y. Zhou, P. Gong, P. Wang, W. Shao, and D. Zhang, "A temporal-spectral fused and attention-based deep model for automatic sleep staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1008–1018, 2023.

[13] Z. Liu, J. Sun, Y. Zhang, and P. Rolfe, "Sleep staging from the eeg signal using multi-domain feature extraction," *Biomed. Signal Process. Control*, vol. 30, pp. 86–97, 2016.

[14] 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, 2017.

[15] S. Seifpour, H. Niknazar, M. Mikaeili, and A. M. Nasrabadi, "A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal," *Expert Syst. Appl.*, vol. 104, pp. 277–293, 2018.

[16] D. Jiang, Y. -N. Lu, M. A. Yu, and W. Yuanyuan, "Robust sleep stage classification with single-channel EEG signals using multimodal decomposition and HMM-based refinement," *Expert Syst. Appl.*, vol. 121, pp. 188–203, 2019.

[17] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J. -F. Payen, "A convolutional neural network for sleep stage scoring from raw single-channel EEG," *Biomed. Signal Process. Control*, vol. 42, pp. 107–114, 2018.

[18] M. Perslev, M. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-time: A fully convolutional network for time series segmentation applied to sleep staging," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 4415–4426.

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

[20] W. Qu et al., "A residual based attention model for EEG based sleep staging," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 10, pp. 2833–2843, Oct. 2020.

[21] S. Mousavi, F. Afghah, and U. R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLoS One*, vol. 14, no. 5, 2019, Art. no. e0216456.

[22] A. Krakovská and K. Mezeiová, "Automatic sleep scoring: A search for an optimal combination of measures," *Artif. Intell. Med.*, vol. 53, no. 1, pp. 25–33, Sep. 2011.

[23] R. Yu, Z. Zhou, S. Wu, X. Gao, and G. Bin, "MRASleepNet: A multi-resolution attention network for sleep stage classification using single-channel EEG," *J. Neural Eng.*, vol. 19, no. 6, 2022, Art. no. 066025.

[24] S. Khalighi, T. Sousa, G. Pires, and U. Nunes, "Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels," *Expert Syst. Appl.*, vol. 40, no. 17, pp. 7046–7059, 2013.

[25] M. A. Carskadon et al., "Normal human sleep: An overview," *Princ. Pract. Sleep Med.*, vol. 4, no. 1, pp. 13–23, 2005.

[26] L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[27] G. Wang, M. Chen, Z. Ding, J. Li, H. Yang, and P. Zhang, "Inter-patient ECG arrhythmia heartbeat classification based on unsupervised domain adaptation," *Neurocomputing*, vol. 454, pp. 339–349, 2021.

[28] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and C. Guan, "TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2238–2250, Jul.–Sep. 2023.

[29] K. Liu, M. Yang, X. Xing, Z. Yu, and W. Wu, "SincMSNet: A Sinc filter convolutional neural network for EEG motor imagery classification," *J. Neural Eng.*, vol. 20, no. 5, 2023, Art. no. 056024.

[30] T. -Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[31] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 4694–4703.

[32] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, Sep. 2000.

[33] M. S. Mourtazaev, B. Kemp, A. H. Zwinderman, and H. A. C. Kamphuisen, "Age and gender affect different characteristics of slow waves in the sleep EEG," *Sleep*, vol. 18, no. 7, pp. 557–564, Sep. 1995.

[34] 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, 2000.

[35] G. -Q. Zhang et al., "The national sleep research resource: Towards a sleep data commons," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1351–1358, 2018.

[36] S. F. Quan et al., "The sleep heart health study: Design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[37] P. Fonseca, N. den Teuling, X. Long, and R. M. Aarts, "Cardiorespiratory sleep stage detection using conditional random fields," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 4, pp. 956–966, Jul. 2017.

[38] Q. Shen et al., "LGSleepNet: An automatic sleep staging model based on local and global representation learning," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 2521814.

[39] Y. Sun, B. Wang, J. Jin, and X. Wang, "Deep convolutional network method for automatic sleep stage classification based on neurophysiological signals," in *Proc. 11th Int. Congr. Image Signal Process., BioMed. Eng. Inform.*, 2018, pp. 1–5.

[40] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction CNN framework for automatic sleep stage classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, May 2019.

[41] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 400–410, Mar. 2019.

[42] G. Kong, C. Li, H. Peng, Z. Han, and H. Qiao, "EEG-based sleep stage classification via neural architecture search," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1075–1085, 2023.

[43] Y. Fang, Y. Xia, P. Chen, J. Zhang, and Y. Zhang, "A dual-stream deep neural network integrated with adaptive boosting for sleep staging," *Biomed. Signal Process. Control*, vol. 79, 2023, Art. no. 104150.

[44] J. Pan, Y. Feng, P. Zhao, X. Zou, A. Hou, and X. Che, "CausalAttenNet: A fast and long-term-temporal network for automatic sleep staging with single-channel EEG," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 2529613.

[45] C. Zhao, J. Li, and Y. Guo, "Sequence signal reconstruction based multi-task deep learning for sleep staging on single-channel EEG," *Biomed. Signal Process. Control*, vol. 88, 2024, Art. no. 105615.

[46] M. He, M. Tang, L. Meng, and Z. Liang, "TBSTSleepNet: Three-branch spectro-temporal bidirectional LSTM based attention model for EEG sleep staging," *Biomed. Signal Process. Control*, vol. 97, 2024, Art. no. 106695.

[47] Y. She, D. Zhang, J. Sun, X. Yang, X. Zeng, and W. Qin, "CBLSNet: A concise feature context fusion network for sleep staging," *Biomed. Signal Process. Control*, vol. 91, 2024, Art. no. 106010.