MHFNet: A Multimodal Hybrid-Embedding Fusion Network for Automatic Sleep Staging

Ruhan Liu, Jiajia Li, Yang Wen, *Member, IEEE*, Xian Huang, Bin Sheng, *Member, IEEE*, David Dagan Feng, *Life Fellow, IEEE*, and Ping Zhang, *Senior Member, IEEE*

***Abstract*—Scoring sleep stages is essential for evaluating the status of sleep continuity and comprehending its structure. Despite previous attempts, automating sleep scoring remains challenging. First, most existing works did not fuse local and global temporal information. Second, the correlation for special waves in different signals is rarely used in sleep staging modeling. Third, the logic of scoring rules based on adjacent epochs is not considered in developing sleep staging models. This paper introduces a multimodal hybrid-embedding fusion network (MHFNet), which aims to tackle these challenges in automating sleep stage scoring. MHFNet comprises multi-stream Xception blocks to extract wave characteristics, a hybrid time-embedding module to combine local and global temporal information, a dual-path gate transformer to fuse and enhance attention features, and a refined output header to reconstruct sleep scoring. We perform experiments using three publicly available datasets (SleepEDF-ST, SleepEDF-SC, and SHHS). Experimental results indicate the superiority of MHFNet over baseline approaches in cross-validation. Moreover, at the individual level, MHFNet yielded an average R^2 score improvement of 9% in the testing dataset compared to state-of-the-art models, paving the way for its applications in real-world sleep medicine.**

***Index Terms*—Sleep staging, multimodal fusion, position embedding, transformer-based attention.**

Received 3 January 2024; revised 17 July 2024 and 26 September 2024; accepted 4 January 2025. Date of publication 13 January 2025; date of current version 7 May 2025. This work was supported in part by the Natural Science Fundation of Hunan Province China under Grant 2022JJ80111, in part by the National Natural Science Foundation of China under Grant 6240073549, in part by the Interdisciplinary Program of Shanghai Jiao Tong University under Grant YG2023LC11, in part by the Project of Intelligent Management Software for Multimodal Medical Big Data for New Generation Information Technology, Ministry of Industry and Information Technology of People's Republic of China under Grant TC210804 V, in part by the Postdoctoral Fellowship Program of CPSF under Grant GZC20231604, in part by Shang-hai Science and Technology Commission Research Project under Grant 24YF2731300, and in part by Xiangfu Lab Youth Program under Grant XF052024B01. *(Corresponding author: Xian Huang.)*

Ruhan Liu is with the Furong Laboratory, Central South University, Changsha 410013, China, and also with the Department of Dermatology, Xiangya Hospital, Central South University, Changsha 410013, China (e-mail: 223101@csu.edu.cn).

Jiajia Li is with the School of Chemistry and Chemical Engineering and National Center for Translational Medicine, Shanghai Jiao Tong University, Shanghai 200240, China, also with the Shanghai Artificial intelligence Research institute, Shanghai 200240, China, and also with the Xiangfu Laboratory, Jiashan, China (e-mail: lijiajia@sjtu.edu.cn).

Yang Wen was with the College of Electronics and Information Engineering, Shenzhen 518060, China (e-mail: wen_yang@szu.edu.cn).

Xian Huang is with the Department of Neurology, The Third Xiangya Hospital of Central South University, Changsha 410013, China (e-mail: hxmmeer@163.com).

Bin Sheng is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shengbin@sjtu.edu.cn).

David Dagan Feng is with the School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: dagan.feng@sydney.edu.au).

Ping Zhang is with the Department of Computer Science and Engineering, Ohio State University, Columbus, OH 43210 USA, and also with the Department of Biomedical Informatics, Ohio State University, Columbus, OH 43210 USA (e-mail: zhang.10631@osu.edu).

The code of MHFNet is publicly available at GitHub: https://github.com/Liuruhan/MHET.

Digital Object Identifier 10.1109/JBHI.2025.3528444

I. Introduction

SLEEP stage scoring is essential for analyzing sleep architecture and diagnosing disorders. Experts identify stages using electrical signals from polysomnography (PSG), including EEG and EOG. According to the American Academy of Sleep Medicine (AASM) [1], signals over 4 hours are divided into 30-second epochs, assigned to five sleep stages (W, N1, N2, N3, R). Manual annotation is labor-intensive and time-consuming, taking 2 to 4 hours per subject [2]. Therefore, automatic sleep staging is clinically significant. However, due to complex scoring rules, learning representations and patterns from multimodal signals for data-driven models remains challenging [3].

Recent research shows deep learning algorithms perform well in automated sleep staging [4], [5], [6], [7], [8], [9], [10], [11]. Techniques like CNNs and RNNs are widely used for this purpose. Some methods, called **single-epoch models**, target 30-second epochs as input, constructing CNN models to study time series classification problems [5], [12], [13]. For instance, deep CNN structures using 1D convolutional layers capture wave characteristics across sequential epochs [7], [14]. DeepSleepNet [9] replaces the CNN's linear layer with a Bidirectional Long Short-Term Memory (BiLSTM) layer to enhance sequence representation. Other studies use **multi-epoch models** to capture the interplay across multiple time epochs. SeqSleepNet [8] introduces a hierarchical RNN to understand temporal dependencies across sleep epochs, while U-Time [15] employs a fully connected CNN based on the U-Net [16] architecture to extract temporal-scale features. Additionally, transformer-based models [17], [18], [19] are also used in multi-epoch sleep staging to learn epoch and sequence level information. However, existing methods that implicitly learn temporal dependencies with end-to-end CNN

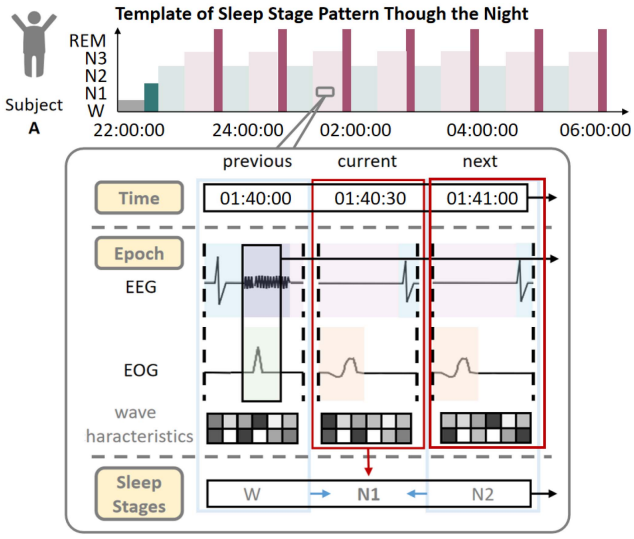

Fig. 1. Example for identifying current sleep stage. The red boxes in the current and subsequent sleep epochs represent the input signal. In the signal features, the two red boxes have similar patterns; however, they belong to different sleep stages due to sleep stage shifting rules [1].

or RNN models may limit the performance in learning temporal patterns, making accurate stage identification challenging.

There are three major challenges in constructing an effective sleep staging model:

Time Information is not Fully Utilized to Learn Temporal Patterns Locally and Globally: As shown in Fig. 1, sleep stages are influenced by both the current wave characteristics and their relative position in each epoch (local patterns), as well as the absolute time (e.g., 22:00:00) across different patients (global patterns). Existing models, such as those based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [9], [20], implicitly learn temporal dependencies. However, they do not consider the time embeddings of global and local temporal patterns in sleep staging modeling.

Sleep Staging Modeling Does not Learn Important Features Correlation for Special Waves in Different Types of Signals.: Different special waves occurring in different signals have logical correlations. For example, in stage N2, K-complex waves, which are well-delineated, negative, sharp waves followed by a positive component in EEG, often co-occur with slow eye movements seen in EOG. Although existing methods [6] have developed multimodal models to extract different wave characteristics, they do not model the correlation between signals.

Complex Logic-Related Sleep Staging Rules are Hard to Learn From Long-Term Sleep Epoch Sequences: Current models do not fully utilize logic-related staging rules based on adjacent epochs. According to AASM rules [1], the sleep stage of the current period is influenced by both the current and the preceding and succeeding periods. Some automatic sleep staging models [8], [15] use sequence models to capture long-term sleep transition rules between multiple 30-second epochs. However, these models, which use more than 30 epoch signals as input, have a large scale of parameters, making them hard to train and learn the complex transition rules.

To address these challenges, we developed a Multimodal Hybrid-Embedding Fusion Network (MHFNet). Unlike previous methods, MHFNet includes four components designed to capture wave characteristics within epochs and infer diagnostic scoring rules from sleep transition patterns. The main contributions of MHFNet for sleep staging are summarized as follows:

- We introduce the MHFNet framework, comprising multi-stream Xception blocks (MXB), a hybrid time embedding (HTE), a dual-path gate transformer (DGT), and a refined output header (ROH) for automatic sleep staging.
- We design MXBs to extract special wave characteristics from multichannel inputs. The HTE includes local and global embeddings (LE and GE) to learn absolute and relative time information. The LE captures ordinal information within a 30-second epoch, whereas GE encodes global time patterns.
- We develop a DGT to fuse and enhance multichannel time-related features, with the DGT generating sleep staging output. The ROH leverages forecasted probabilities across three epochs to refine single-epoch outputs.
- The MHFNet outperforms in cross-validation on three public datasets, including SleepEDF-ST, SleepEDF-SC and SHHS datasets. Additionally, the MHFNet exceeds other methods in individual tests, with an average 9% increase in R^2 scores.

The remaining sections of the paper are organized as follows. Firstly, we review related works on automatic sleep staging methods in Section II. In Section III, we introduce our MHFNet method. Section IV presents the experimental results for MHFNet. Finally, we conclude in Section V.

II. Related Work

Automatic sleep scoring is crucial for analyzing sleep structures and diagnosing sleep disorders. This section introduces two types of automatic sleep staging methods: single-epoch and multi-epoch models. We also conclude and analyze novel transformer models designed for time series classification, suggesting their potential use in automatic sleep scoring for improved performance.

A. Single-Epoch Models for Sleep Staging

Sleep experts assign each 30-second epoch into five sleep stages (W, N1, N2, N3, R). Single-epoch methods are tailored to models focusing on 30-second epochs as their input, commonly employing a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). For instance, DeepSleepNet [9] utilizes CNNs for representation learning and integrates Bidirectional Long Short-Term Memory (BiLSTM) for sequence learning, aiming to discern waveform characteristics and grasp temporal sequencing within individual epochs. Another example is EEGNet [20], which employs diverse two-dimensional convolutional layers to capture prolonged short-term contextual dependencies. Additionally, attention networks based on transformer architectures find application in single-epoch methods. In work [11], an automatic sleep staging model with an enhanced attention module and Hidden Markov Model

(HMM) is proposed for single EEG sleep staging. MultiChannelSleepNet [10] introduces a transformer encoder-based model for automatic sleep staging, utilizing transformer architecture for single-channel feature extraction and multichannel feature fusion. Despite these efforts to extract features within a single 30-second epoch, there remain challenges in effectively learning long-term sleep patterns.

B. Multi-Epoch Models for Sleep Staging

The contextual relationships between consecutive sleep epochs can significantly contribute to the accurate staging of the current epoch. Consequently, several investigations employ multi-epoch models to capture the interplay among multiple sequential time epochs. An example is SeqSleepNet [8], which introduces a hierarchical Recurrent Neural Network (RNN) to comprehend temporal dependencies extending across 20 or more sleep epochs. Another approach, U-Time [15], advocates for a fully connected Convolutional Neural Network (CNN) with a temporal focus, structured upon the U-Net [16] architecture, for the extraction of features at different temporal scales. Nevertheless, prevailing methods in this domain often implicitly acquire temporal dependencies through end-to-end CNN-based or RNN-based models, potentially limiting their capacity to learn intricate temporal patterns effectively. Furthermore, the challenge arises with extended time series data, which may need help discerning accurate sleep stage rules.

C. Transformers in Time Series Classification

Moreover, transformers [17], [18], [19], [21], [22], [23] have demonstrated notable efficacy across diverse time series classification tasks, indicating the prospect of their application in automatic sleep staging to enhance overall performance. For instance, DETRtime [22] introduces a transformer-centric framework specifically designed for time series segmentation, leveraging EEG data. Additionally, gate transformer networks (GTN) [23] employ a dual-tower Transformer architecture, incorporating both time-step-wise and channel-wise attention mechanisms. A trainable gated concatenation mechanism is introduced to integrate these distinct attention-based attributes within GTN. The success achieved by these transformer-based models underscores the potential advantages of integrating transformer-based modules into automated sleep staging. Diverging from these prior works, our transformer-based model employs dual-path attention to enhance feature correlations across multimodal channels and integrates temporal information through hybrid time embedding. Furthermore, we propose a refined output header to refine the single-epoch output using multi-epoch predictive possibilities.

III. Preliminaries

A time series $\boldsymbol{X}_i \in \mathbb{R}^{N\times C}$ records signals in i-th single sleep epoch, where N indicates the number of samples within a single epoch, and C denotes the channel count (i.e. one EEG channels and one EOG channel, $C = 2$). We use $\boldsymbol{\mathcal{X}} = (\boldsymbol{X}_{i-S/2}, \ldots, \boldsymbol{X}_i, \ldots, \boldsymbol{X}_{i+S/2}) \in \mathbb{R}^{S\times N\times C}$ to denote the multi-epoch input of LoH module, where S representing the extent of multi-epoch inputs.

Problem Statement: The sleep staging problem aims to identify the current sleep stage given the multi-epoch inputs. Formally, given a multi-epoch input $\boldsymbol{\mathcal{X}}$ which is centered on $\boldsymbol{X}_i$, our first goal is to learn a mapping function f from the current S step's sleep epochs to predict the possibilities $\boldsymbol{\mathcal{P}} = (\boldsymbol{P}_{i-S/2}, \ldots, \boldsymbol{P}_i, \ldots, \boldsymbol{P}_{i+S/2}) \in \mathbb{R}^{S\times K}$ of sleep stages for S step's, where K is the number of sleep scoring stages.

$$[\boldsymbol{X}_{i-S/2}, \ldots, \boldsymbol{X}_i, \ldots, \boldsymbol{X}_{i+S/2}] \xrightarrow{f} [\boldsymbol{P}_{i-S/2}, \ldots, \boldsymbol{P}_i, \ldots, \boldsymbol{P}_{i+S/2}] \tag{1}$$

Furthermore, given predictive possibilities $\boldsymbol{\mathcal{P}}$ of sleep stages for S steps, our final goal is to learn another mapping function f_r from the current predictive possibilities $\boldsymbol{\mathcal{P}}$ of sleep stages to predict the current stage y_i,

$$[\boldsymbol{P}_{i-S/2}, \ldots, \boldsymbol{P}_i, \ldots, \boldsymbol{P}_{i+S/2}] \xrightarrow{f_r} y_i \tag{2}$$

IV. Methodology

The schematic representation of the MHFNet framework is depicted in Fig. 2. We utilize Multi-stream Xception Blocks (MXBs) to extract multimodal wave features from EEG and EOG signals. Given the periodic nature of nighttime sleep and its close relationship with sleep stage durations, we introduce a Hybrid Time Embedding (HTE) module to encode temporal patterns and integrate both local and global time information. Recognizing that sleep staging depends not only on significant waveform features but also on their temporal order, we propose a Dual-path Gate Transformer (DGT). This module incorporates channel-wise and step-wise attention mechanisms to focus on salient features. To enhance single-epoch classification, we developed a Refined Output Header (ROH). This component utilizes predictive insights from multiple epochs to improve classification accuracy and facilitate the learning of sleep stage scoring rules across segments, thereby reducing training complexity.

A. Multi-stream Xception Blocks (MXBs)

Adhering to the AASM sleep scoring standard [1], it is observed that diverse wave characteristics manifest in EEG and EOG signals even within the same sleep stage. Consequently, employing a singular feature extractor to capture wave characteristics in EEG and EOG signals concurrently proves inadequate in fully exploiting the rich array of features inherent in these signals. For instance, during the wakefulness stage (W), alpha rhythm may be discerned in EEG signals, while slow eye movements may be present in EOG signals. To address the need for capturing distinct wave characteristics in EEG and EOG channels, we introduce MXBs, drawing inspiration from [6], [14]. These MXBs are adept at effectively extracting diverse features from the two types of channels. The intricate architecture of the TXB is illustrated in Fig. 3.

Given an input single-epoch $\boldsymbol{X}_i$, we split the input into EEG input $\boldsymbol{X}_i^e \in \mathbb{R}^{N\times 2}$ and EOG input $\boldsymbol{X}_i^o \in \mathbb{R}^{N\times 1}$. We extract

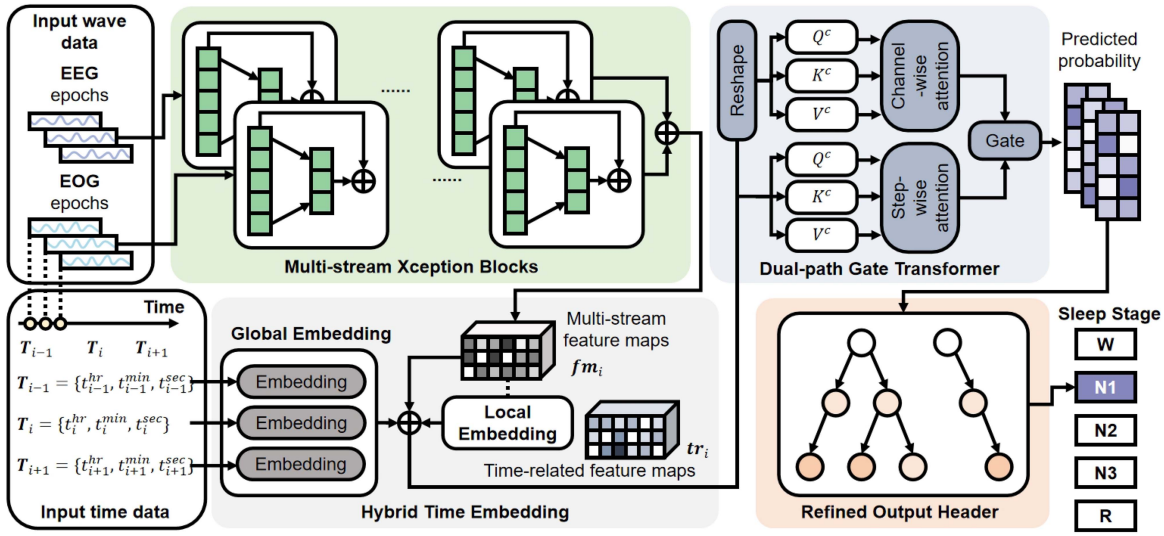

Fig. 2. An overview architecture of MHFNet. MHFNet contains four components: multi-stream Xception blocks (MXBs), a dual-path gate transformer (DGT) module, a hybrid time embedding (HTE) module, and a refined output header (ROH) module. The input data are directed into MXB modules to acquire multi-stream feature maps. Subsequently, these multi-stream feature maps are fed into the HTE module to extract time-related feature maps. Furthermore, the DGT generates the predicted probabilities for sleep stages based on the time-related feature maps. Ultimately, the ROH inputs the multi-epoch predicted probabilities and produces the refined predicted sleep stages as output.

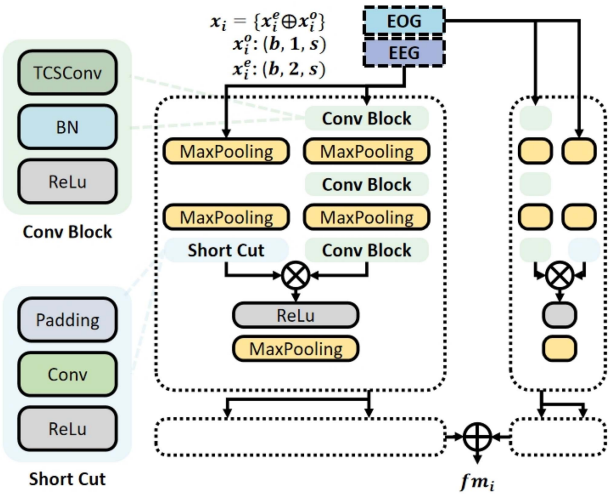

Fig. 3. The structure figure displays multi-stream Xception blocks (MXBs). MXBs are designed to discern wave characteristics in EEG and EOG signals. Distinct channels are directed into separate Xception blocks for the extraction of features. Subsequently, the features from various channels are amalgamated through splicing to generate the final output.

multi-stream (EEGs and EOG) feature maps $\boldsymbol{M}_i \in \mathbb{R}^{h \times w}$ using multiple TXB modules:

$$\boldsymbol{M}_i = T_e(\boldsymbol{X}_i^e) + T_o(\boldsymbol{X}_i^o) \tag{3}$$

where T_e and T_o denote the two MXB modules for EEG and EOG signals, respectively. h and w are the feature length and filter number of $\boldsymbol{M}_i$, respectively.

B. Hybrid Time Embedding (HTE)

To enhance the utilization of time-related information, after the extraction of multi-stream feature maps $\boldsymbol{M}_i$, we devised an HTE module to integrate both absolute and relative time information. Understanding the regulations governing sleep transitions is crucial for automating sleep staging. These rules depend not only on wave characteristics but also on the temporal sequence of these characteristics. Prior studies have utilized RNN or multi-scale CNN models to implicitly capture sleep transition rules, often without incorporating comprehensive temporal information [6], [8]. To effectively leverage both local and global time information, our HTE module is designed to integrate time features within individual epochs and across multiple epochs. We generate time-related feature maps $\boldsymbol{D}_i \in \mathbb{R}^{h \times w}$ by combining local and global time embedding modules:

$$\begin{aligned} \boldsymbol{L}_i &= PE(pos, 2d) + PE(pos, 2d+1) \\ \boldsymbol{G}_i &= FC_h(t_i^{hr}) + FC_m(t_i^{min}) + FC_s(t_i^{sec}) \\ \boldsymbol{D}_i &= \boldsymbol{M}_i + \boldsymbol{L}_i + \boldsymbol{G}_i \end{aligned} \tag{4}$$

where pos is the position and d is the dimension. $PE(pos, 2d) = sin(pos/10000^{2d/l})$ and $PE(pos, 2d+1) = cos(pos/10000^{2d/l})$, details can be seen in [24]. $\boldsymbol{T}_i = (t_i^{hr}, t_i^{min}, t_i^{sec})$ represent the one-hot embeddings corresponding to the current time in the input single epoch, respectively. FC_h, FC_m, and FC_s are three linear layers to encode the t_i^{hr}, t_i^{min}, and t_i^{sec}, respectively. The output time-related feature maps $\boldsymbol{D}_i$ are the sum of two-stream feature maps $\boldsymbol{M}_i$, local time stamp $\boldsymbol{L}_i$, and global time stamp $\boldsymbol{G}_i$.

C. Dual-path Gate Transformer (DGT)

To encode temporal and channel features within multimodal characteristics, we introduce a DGT. DGT accentuates significant features through both the step-wise attention (SA) and channel-wise attention (CA) domains. The SA module enhances feature maps by emphasizing the importance and correlations

in the time series. Meanwhile, the CA module concentrates on extracting feature importance associated with correlations between channels.

The input of DGT is the time-related feature maps $\boldsymbol{D}_i$, the output of HTE module. Given the input $\boldsymbol{D}_i$, the SA and CA modules first obtain the query, key and value matrices of self-attention oprations as:

$$\begin{aligned} \boldsymbol{Q}_i^s &= \boldsymbol{D}_i^T \boldsymbol{W}_q^s, \boldsymbol{K}_i^s = \boldsymbol{D}_i^T \boldsymbol{W}_k^s, \boldsymbol{V}_i^s = \boldsymbol{D}_i^T \boldsymbol{W}_v^s \\ \boldsymbol{Q}_i^c &= \boldsymbol{D}_i \boldsymbol{W}_q^c, \boldsymbol{K}_i^c = \boldsymbol{D}_i \boldsymbol{W}_k^c, \boldsymbol{V}_i^c = \boldsymbol{D}_i \boldsymbol{W}_v^c \end{aligned} \tag{5}$$

where $\boldsymbol{W}_q^s, \boldsymbol{W}_k^s, \boldsymbol{W}_v^s \in \mathbb{R}^{h \times h}$ are learnable parameters of the query, key and value matrices in SA module. $\boldsymbol{W}_q^c, \boldsymbol{W}_k^c, \boldsymbol{W}_v^c \in \mathbb{R}^{w \times w}$ are learnable parameters of the query, key, and value matrices in CA module.

Then, we apply self-attention operations channel-wise and step-wise to model the correlation between and within single epochs and obtain the attention matrices as:

$$\boldsymbol{A}_i^s = \frac{(\boldsymbol{Q}_i^s)(\boldsymbol{K}_i^s)^T}{\sqrt{h}}, \boldsymbol{A}_i^c = \frac{(\boldsymbol{Q}_i^c)(\boldsymbol{K}_i^c)^T}{\sqrt{w}} \tag{6}$$

The $\boldsymbol{A}_i^s$ and $\boldsymbol{A}_i^c$ are step-wise and channel-wise attention matrics.

Moreover, we obtain the outputs of SA and CA modules by multiplying corresponding attention matrics with value metrics as:

$$\begin{aligned} CA(\boldsymbol{Q}^c, \boldsymbol{K}^c, \boldsymbol{V}^c) &= softmax(\boldsymbol{A}^c)\boldsymbol{V}^c \\ SA(\boldsymbol{Q}^s, \boldsymbol{K}^s, \boldsymbol{V}^s) &= softmax(\boldsymbol{A}^s)\boldsymbol{V}^s \end{aligned} \tag{7}$$

Based on the outputs of SA and CA modules, we proposed the gate mechanism to enhance the fusion feature. First, we calculate the fusion output $\boldsymbol{H}_i$, and then obtain gate output $\boldsymbol{O}_i$. Finally, we get the predicted probabilities $\boldsymbol{P}_i$ through linear projection as:

$$\begin{aligned} \boldsymbol{H}_i &= CA(\boldsymbol{Q}^c, \boldsymbol{K}^c, \boldsymbol{T}^c) + SA(\boldsymbol{Q}^s, \boldsymbol{K}^s, \boldsymbol{T}^s) \\ (w_1, w_2) &= Softmax(FC(\boldsymbol{H}_i) \\ \boldsymbol{O}_i &= w_1 \cdot CA(\boldsymbol{Q}^c, \boldsymbol{K}^c, \boldsymbol{T}^c) + w_2 \cdot SA(\boldsymbol{Q}^s, \boldsymbol{K}^s, \boldsymbol{T}^s) \\ \boldsymbol{P}_i &= FC(\boldsymbol{O}_i) \end{aligned} \tag{8}$$

where the fusion output $\boldsymbol{H}_i$ is feed into $Softmax(FC(\cdot)$ to output $w_1 = \alpha$ and $w_2 = 1 - \alpha$ as the weights of channel-wise features and step-wise features, respectively. FC represents the linear layer to produce predicted probabilities $\boldsymbol{P}_i$ for each single epoch.

D. Refined Output Header (ROH)

Sleep transition rules are crucial for accurate sleep staging. Assigning a sleep stage to the current epoch often requires considering signals from both preceding and succeeding epochs. Some studies using single-epoch models overlook this temporal relationship [20], [25]. Additionally, certain multi-epoch models employ RNN or CNN architectures to capture time-related rules indirectly, but these approaches may not comprehensively learn diagnostic rules [6], [8].

Thus, we propose a ROH using an XGBoost classifier [27] to refine the multi-epoch output based on classification results from the single-epoch classifier (DGT). The ROH module in MHFNet is refined based on the trained model's outputs. After MHFNet predicts sleep stages for individual epochs, the ROH module further refines these predictions using logical reasoning. It utilizes XGBoost to analyze the probability results across multiple epochs and enhance the accuracy of sleep staging outcomes. This approach aims to improve the model's ability to infer and apply complex sleep stage rules derived from the aggregated predictions of multiple epochs. Given predictive possibilities $\boldsymbol{\mathcal{P}} = (\boldsymbol{P}_{i-S/2}, \ldots, \boldsymbol{P}_i, \ldots, \boldsymbol{P}_{i+S/2})$ of sleep stages for S steps, We merge the predicted possibilities $\boldsymbol{\mathcal{P}}$ to a multi-epoch possibility matric $\boldsymbol{J}_i$. Next, the multi-epoch possibility matric $\boldsymbol{J}_i$ is used to generate predicted sleep stage label $\hat{y}_i$. Formally, the process is derived as:

$$\begin{aligned} \boldsymbol{J}_i &= \boldsymbol{P}_{i-S/2} \oplus \ldots \oplus \boldsymbol{P}_i \oplus \ldots \oplus \boldsymbol{P}_{i+S/2} \\ \hat{y}_i &= XGB(\boldsymbol{J}_i) \end{aligned} \tag{9}$$

where S is the sequence length of considering multiple epochs, XGB is the XGBoost classifier, and y_i is the refined output stage.

E. Loss Function

For sleep scoring, we used a weigted cross-entropy (WCE) loss to balance different size of stages. The WCE loss is as below:

$$L_{WCE} = \sum_{k=0}^{4} -\alpha_k \cdot \log(p_k) \tag{10}$$

where $k \in \{0, 1, 2, 3, 4\}$ represents five sleep stages (W, N1, N2, N3, R), α_k is the class balanced weight and p_k is the predicted output. In our task parameter α_k is the derivative of the data percentage of each sleep stage.

V. Experiments

In this section, we outline the materials, setup, and design of the experiment. We cover the datasets used for training and testing, the data preprocessing methods, the experimental configurations, the evaluation metrics for testing, and the results, including cross-validation performance, individual validation outcomes, and an ablation study.

A. Datasets

1) SleepEDF-SC [28]: This dataset, part of the Sleep-EDF dataset, includes 153 polysomnography (PSG) records from 78 healthy people aged 25 to 101. The PSG records measure brain and eye activity using electroencephalogram (EEG) and electrooculogram (EOG). Each record has 2 EEG signals (Fpz-Cz and Pz-Oz) and 1 horizontal EOG signal, all sampled at 100 Hz. AASM guidelines [1], recordings longer than 4 hours are split into 30-second segments. Note that due to a device error, one record for subjects 13, 36, and 52 is missing.

2) SleepEDF-ST [29]: This dataset is part of the 2013 version of the Sleep-EDF dataset and includes 39 PSG records from 20

TABLE I
BASELINE METHODS, INCLUDING SINGLE-EPOCH MODELS (BILSTM, EEGNET, DEEPSLEEPNET, TINYSLEEPNET, RESCONV, INCEPTIONTIME, XCEPTION, AND CONVLSTM) AND MULTI-EPOCH MODELS (SEQSLEEPNET, UTIME, SALIENTSLEEPNET, SLEEPTRANSFORMER, CORE-SLEEP, AND SLEEPYCO), ARE COMPARED WITH THE MHFNET FRAMEWORK

Method	Input epoch length	Descriptions
BiLSTM [12]	single epoch	a BiLSTM as the feature extractor
EEG-Net [20]	single epoch	depthwise and separable convolutions to construct an EEG model
DeepSleepNet [9]	single epoch	CNNs to extract time-invariant features, and BiLSTM to learn transition rules
TinySleepNet [7]	single epoch	a CNN model based on raw single-channel EEG
ResConv [13]	single epoch	multiple residual CNN blocks as the feature extractor
InceptionTime [25]	single epoch	a deep CNN model, inspired by the Inception-v4 [26]
Xception [14]	single epoch	uses 1D Xception blocks as the feature extractor
ConvLSTM [5]	single epoch	several residual CNN blocks to extract wave features and an LSTM to learn time dependencies
SeqSleepNet [8]	128 epochs	a hierarchical RNN model
UTime [15]	35 epochs	a temporal fully-CNN based on the UNet [16]
SalientSleepNet [6]	20 epochs	a temporal fully-CNN based on the U2-Net architecture
SleepTransformer [17]	21 epochs	a sequence-to-sequence sleep-staging model
CoRe-Sleep [18]	21 epochs	a coordinated representation multimodal fusion network
SleePyCo [19]	10 epochs	a feature pyramid and supervised contrastive learning network for sleep scoring

healthy people aged 25 to 34. Like the SleepEDF-SC dataset, the SleepEDF-ST dataset has recordings from two consecutive nights for each person. The format of the recordings and the sleep stage labels are the same as those in the SleepEDF-SC dataset.

3) SHHS [30]: This large-scale, multi-center database was collected to study the impact of sleep-disordered breathing on cardiovascular diseases [31]. It includes two rounds of PSG records: Visit 1 and Visit 2. For this work, we used data from SHHS-1, which includes 5,791 subjects aged 39-90. The records were manually scored using the R&K guidelines [32]. We selected 1,200 unique subjects from the SHHS database to train and validate all automatic methods.

B. Preprocessing Method

Adhering to the guidelines established by the AASM [1], we combined the N3 and N4 stages into a unified N3 stage and excluded the MOVEMENT and UNKNOWN stages from consideration. The evaluation of our model utilized two EEG channels (Fpz-Cz and Pz-Cz for SleepEDF, EEG(sec) and EEG for SHHS) and one EOG channel (horizontal EOG for SleepEDF and EOG(L) for SHHS). In this study, PSG records from both SleepEDF and SHHS underwent pre-processing with bandpass filters set at 0.3 Hz-100 Hz for EEG and 0.1 Hz-100 Hz for EOG, with all signals resampled to 100 Hz. We applied these pre-processing steps consistently across all methods to ensure a fair comparison with other methodologies. In SleepEDF and SHHS datasets used for training and testing, the awake state signals are dropped.

C. Experimental Setups

We compared the MHFNet framework against state-of-the-art methods, detailed in Table I. Our evaluation encompassed both single-epoch and multi-epoch models implemented in the PyTorch framework with the Adam optimizer. Performance assessment was conducted using a rigorous 20-fold cross-validation approach. Throughout the training process, we employed the grid search method to optimize hyperparameters. The exploration range for the learning rate was from 0.01 to 1×10^{-7}, while the batch size was explored from 2 to 32.

Additionally, to assess generalization, we employed an alternative data split method on the SleepEDF-SC dataset. This dataset was divided into SC training (123 records from 56 subjects) and SC testing (30 records from 22 subjects) sets, ensuring no overlap between subjects. The SC training set was further split into a developmental subset and a validation subset (9:1 ratio) to optimize model performance. The final model, selected based on performance on the validation set within 60 epochs, was tested on the SC testing dataset to validate its efficacy.

Moreover, we employed the grid search algorithm to ascertain optimal hyperparameters, encompassing batch size, initial learning rate, and the number of layers. Our exploration included batch sizes ranging from 4 to 64 and initial learning rates spanning from 10^{-3} to 10^{-7} . All experiments were conducted on an Intel(R) Xeon(R) CPU E5-2678 v3 @ 2.50G Hz CPU and two NVIDIA GeForce RTX 3090 GPUs operating on the Arch Linux platform.

D. Evaluation Metrics

We used precision (PRE), recall (Recall), and F1-score (F1-score) as metrics for each sleep stage to assess other state-of-the-art methods and our MHFNet. Additional metrics such as accuracy (ACC), macro-averaged F1-score (MF1) [33], and Cohen's kappa (κ) [34] were employed to evaluate the overall model performance in cross-validation and individual tests. MF1 represents the average F1 score of the five sleep stages. In the individual tests, we also use the R-Square coefficient to evaluate sleep stage assessment in each subject.

E. Experiment Results

1) 20-Fold Cross-Validation in Three Public Datasets: We comprehensively compared the MHFNet framework with 14 baseline models, consisting of eight single-step models and six multi-step models. The results of the 20-fold cross-validation

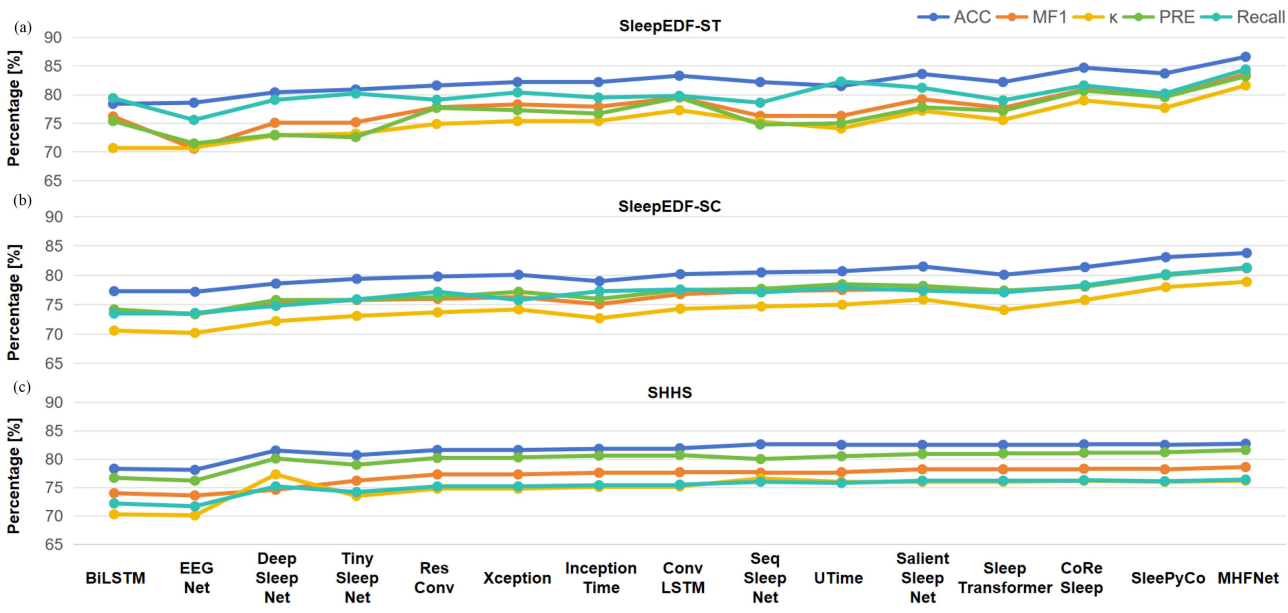

Fig. 4. Comparative analysis shows the state-of-the-art model's performance on three datasets. (a) Comparison performance in the SleepEDF-ST dataset. (b) Comparison performance in the SleepEDF-SC dataset. (c) Comparison performance in the SHHS dataset. The overall validation metrics are evaluated at the 30-second epoch level using a 20-fold cross-validation approach.

for overall metrics are presented in Fig. 4 for the SleepEDF-SC, SleepEDF-ST, and SHHS datasets. The outcomes underscore the superior performance of the MHFNet framework compared to other baseline methods.

Certain single-step CNN models, inspired by architectures utilized in computer vision tasks [13], [14], [26], excel at learning features within individual epochs and exhibit superior performance compared to simpler CNN or RNN structures. The leading single-epoch model (ConvLSTM) achieved an average accuracy of 83.3% in SleepEDF-ST, 80.2% in SleepEDF-SC, and 81.9% in SHHS. Meanwhile, multi-step models leverage multiple epochs as input to generate multi-epoch outputs, enabling them to capture sleep transition rules between sleep epochs. The best-performing multi-epoch model (CoRe-Sleep) attained an average accuracy of 84.7% in SleepEDF-ST, 81.4% in SleepEDF-SC, and 82.6 % in SHHS. The MHFNet framework demonstrated an average accuracy of 86.6% in SleepEDF-ST, 83.8% in SleepEDF-SC, and 82.7 % in SHHS. The observed accuracy improvements for the MHFNet framework were 1.9% in SleepEDF-ST, 0.9% in SleepEDF-SC, and 0.1% compared to the state-of-the-art model. Although the accuracy improvement in SHHS dataset is small, the precision rate in N1 is improved by MHFNet.

Additionally, we present the confusion matrices for the comparison methods (the top four models with the highest accuracy) and MHFNet on the two datasets in Fig. 5. The confusion matrices reveal that MHFNet achieved the highest average precision, reaching 83.2% in the SleepEDF-ST dataset and 81.4% in the SleepEDF-SC dataset. Notably, the results across both datasets indicate that MHFNet improved the accuracy of N1 classification compared to other methods while maintaining high accuracy for W, N2, N3, and R stages.

The F1-scores for each sleep stage class are detailed in Table II. In the SleepEDF-ST dataset, MHFNet performs best in the W, N1, N2, and N3 stages. Moreover, MHFNet secures the best F1 scores in each class of sleep stages in the SleepEDF-SC dataset.

2) Individual Validation in the SleepEDF-SC Dataset: We tested the state-of-the-art models and our MHFNet framework for individual-level validation using the testing SC dataset. The comparison performances were detailed in Table III. In the SC testing set, when considering epoch-level sleep staging, the state-of-the-art model (CoRe-Sleep) achieved an average accuracy of 71.1%. In contrast, the MHFNet attained an average accuracy of 76.3%. Moreover, regarding individual-level sleep staging, the state-of-the-art model (CoRe-Sleep) demonstrated an average accuracy of 72.0%, while the MHFNet showcased an average accuracy of 75.7%. Remarkably, our MHFNet exhibited more substantial enhancements at the individual level (accuracy improvement of 5.2% for epoch level and 3.7% for individual level). Furthermore, to investigate the correlation between predicted sleep stages and actual sleep stages on an individual basis, we computed the average R^2 score for each record and derived the average R^2 score (MHFNet: 0.45vs. state-of-the-art model CoRe-Sleep: 0.36).

To display the real-world diagnosis performance, Fig. 6 further shows the output hypnogram per stage of sleep of a subject of the SleepEDF SC testing dataset (subject ID: SC4201E0-PSG). The output hypnogram figures illustrated that the MHFNet framework outperformed other baseline methods and obtained the best R^2 score corresponding to ground truth.

Within clinical practice, the proportions of sleep stages serve as pivotal metrics for evaluating sleep conditions. We compared the actual and predicted sleep stage proportions in MHFNet and

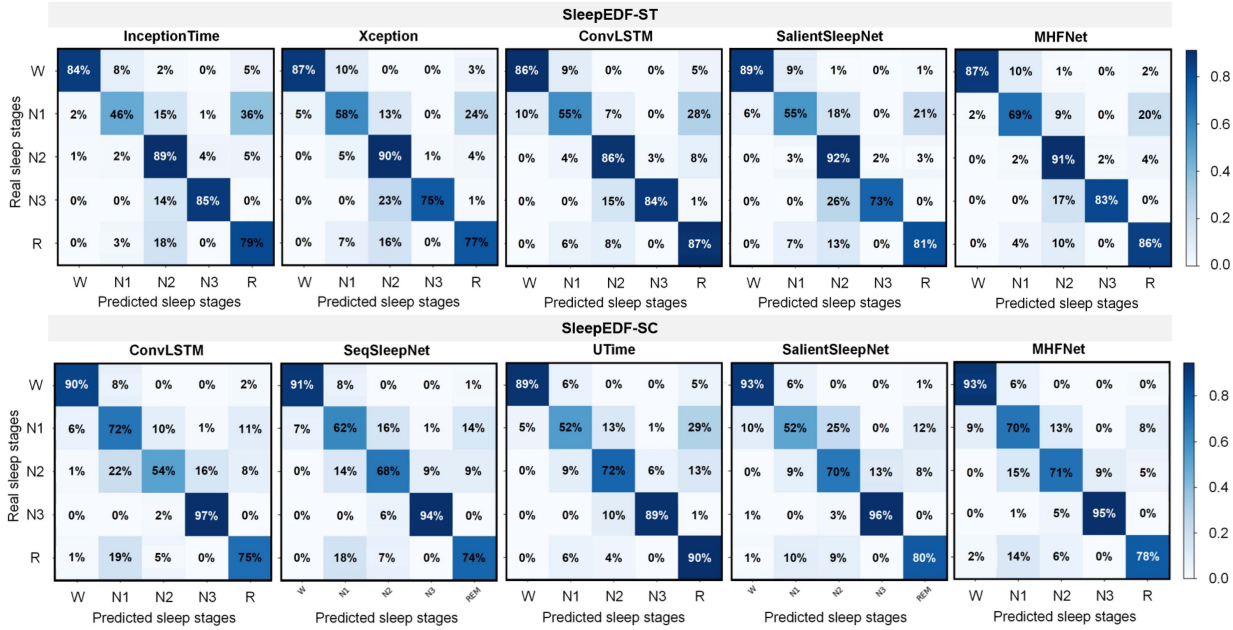

Fig. 5. Confusion matrixes of the comparison methods (top four models with the highest accuracy) and the MHFNet are shown. The first line shows the confusion matrixes of MHFNet and the four methods (InceptionTime [25], Xception [14], ConvLSTM [5], and SalientSleepNet [6]) with the highest accuracy ranking on the SleepEDF-ST dataset. The second line shows the confusion matrix of MHFNet and the four methods (ConvLSTM [5], SeqSleepNet [8], UTime [15], and SalientSleepNet [6]) with the highest accuracy ranking on the SleepEDF-SC dataset.

TABLE II
COMPARATIVE ANALYSIS DISPLAYS THE STATE-OF-THE-ART MODEL'S PERFORMANCE ON SLEEPEDF-SC, SLEEPEDF-ST, AND SHHS DATASETS

Method	SleepEDF-ST					SleepEDF-SC					SHHS				
	F1-score for each class					F1-score for each class					F1-score for each class				
	W	N1	N2	N3	R	W	N1	N2	N3	R	W	N1	N2	N3	R
BiLSTM	90.6	54.8	80.8	80.6	74.2	92.2	53.4	66.2	90.8	66.2	80.6	44.4	82.4	83.0	79.8
EEGNet	90.1	23.2	84.0	84.1	71.9	91.2	51.6	65.7	91.2	67.2	80.9	41.8	82.1	81.8	81.2
DeepSleepNet	90.5	43.5	85.5	81.3	74.9	92.3	54.9	66.8	91.1	71.1	82.8	47.8	84.7	84.8	86.3
TinySleepNet	90.3	47.2	85.6	73.4	79.8	92.7	55.2	68.0	91.3	72.0	81.9	46.5	84.4	83.5	85.0
ResConv	90.6	47.1	85.4	84.4	81.5	94.0	59.7	66.7	91.1	68.3	83.5	47.7	84.9	84.5	85.8
Xception	91.6	50.7	86.4	84.3	78.8	93.6	55.3	67.1	90.7	74.6	83.5	47.4	84.8	84.5	86.4
InceptionTime	89.9	50.2	86.2	86.1	77.1	93.4	53.1	69.8	89.0	70.3	83.6	47.8	85.1	84.7	86.6
ConvLSTM	90.0	52.4	87.5	**86.9**	80.5	93.0	62.3	63.0	90.8	75.2	83.8	47.9	85.1	84.9	86.8
SeqSleepNet	88.3	40.9	87.7	85.3	79.4	93.6	58.7	68.6	91.8	74.0	84.9	44.1	85.8	86.1	86.8
UTime	90.9	59.5	85.4	62.1	83.5	93.0	57.7	71.8	91.1	74.1	85.2	47.3	**86.4**	84.5	85.1
SalientSleepNet	92.3	52.1	87.2	81.7	82.6	93.7	56.2	67.0	91.9	79.5	84.2	48.4	85.6	85.3	87.3
SleepTransformer	84.3	53.7	88.0	80.6	81.7	91.6	60.0	68.5	89.3	76.6	84.3	48.7	85.6	85.3	87.1
CoRe-Sleep	89.2	59.0	89.3	83.9	83.0	93.7	61.8	68.0	88.2	79.0	**84.5**	48.8	85.6	**85.5**	87.3
SleePyCo	87.1	57.1	88.9	82.8	82.2	**94.3**	64.8	71.3	89.4	80.6	84.4	48.8	85.6	**85.5**	87.3
MHFNet	**92.5**	**64.2**	**89.6**	86.8	**84.9**	94.1	**65.5**	**72.4**	**92.7**	**81.3**	84.1	**50.3**	85.6	85.3	**87.6**

The F1 scores in five sleep stages (W, N1, N2, N3, and R) are evaluated at the 30-second epoch level using a 20-fold cross-validation approach.

other baseline methods. The MHFNet framework showcased the lowest errors in sleep stage proportion (MAE: 4.78 ± 3.33).

3) Ablation Study: To evaluate the effectiveness of each module that constructs our MHFNet framework, we designed an ablation study to explore the contribution of each module.

- BA: The baseline model consists of a multi-stream Xception structure with a linear layer serving as the output header. EEG and EOG signals are input into MXBs to extract features specific to each signal type. These extracted features undergo further processing through a fully connected layer, which generates predictions for the five sleep stages based on the learned representations from the MXBs.
- BAH: The BAH model is the baseline model with the hybrid embedding module.
- BAHD: The BAHD model adds a dual-path gate transformer based on the "BAH" model.
- Baseline+HTE+DGT+ROH (MHFNet): The model uses MXBs as the feature extractor, the HTE module to capture time information, the DGT module to focus on channel-wise and step-wise important features, and the ROH module to refine single-epoch output based on multi-step predictive probabilities.

TABLE III
PERFORMANCE COMPARISONS OF THE STATE-OF-THE-ART MODELS SHOW IN SC TESTING SET

Method	SC testing												
	individual-level metrics					epoch-level metrics			F1-score for each class				
	ACC	MF1	κ	R^2	sleep-stage error	ACC	MF1	κ	W	N1	N2	N3	R
BiLSTM	64.1	45.9	44.6	-0.21	10.02 ± 7.42	62.7	55.0	47.1	60.0	19.6	73.5	72.1	50.0
EEG-Net	65.6	48.2	47.6	-0.08	8.35 ± 5.39	65.1	57.3	51.7	71.4	23.6	75.6	65.7	50.4
DeepSleepNet	64.0	54.0	51.7	-0.30	11.08 ± 8.66	64.5	62.7	55.0	84.9	35.0	61.9	74.0	57.5
TinySleepNet	66.9	58.2	54.7	0.22	8.07 ± 5.91	68.3	67.0	59.0	83.7	41.2	69.7	69.1	71.3
ResConv	65.6	57.1	53.8	0.12	9.86 ± 7.89	66.5	65.8	57.4	83.4	39.6	65.0	73.3	67.9
InceptionTime	65.6	55.0	52.3	0.05	8.64 ± 6.34	66.4	64.0	56.5	81.7	34.8	71.0	71.0	61.7
Xception	66.5	57.1	54.2	0.21	9.17 ± 6.76	67.6	66.1	58.6	85.2	42.0	67.1	66.3	70.0
ConvLSTM	70.2	60.2	57.5	0.34	6.30 ± 4.58	71.3	68.9	62.1	85.4	42.0	72.6	70.1	74.2
SeqSleepNet	65.6	55.0	52.3	0.05	8.64 ± 6.34	66.4	64.0	56.5	81.7	34.6	71.0	71.0	61.7
UTime	66.6	56.3	53.6	0.18	8.54 ± 5.70	67.6	65.7	58.0	84.0	38.7	71.1	70.0	64.6
SalientSleepNet	69.7	59.8	57.8	0.17	7.65 ± 5.13	70.8	68.2	61.8	84.2	39.7	73.1	73.3	70.6
SleepTransformer	68.7	58.8	56.9	-0.04	7.76 ± 6.08	68.9	66.5	59.5	81.7	39.3	74.0	75.9	61.4
CoRe-Sleep	72.0	61.7	59.4	0.36	5.91 ± 4.51	71.1	68.8	62.2	85.3	41.7	74.0	76.2	66.8
SleePyCo	70.1	60.6	58.9	0.07	7.43 ± 5.85	70.5	68.3	61.6	85.0	41.7	73.1	76.1	65.5
MHFNet	**75.7**	**63.8**	**64.0**	**0.45**	**4.78 ± 3.33**	**76.3**	**73.2**	**68.2**	**86.9**	**42.0**	**78.2**	**77.3**	**81.6**

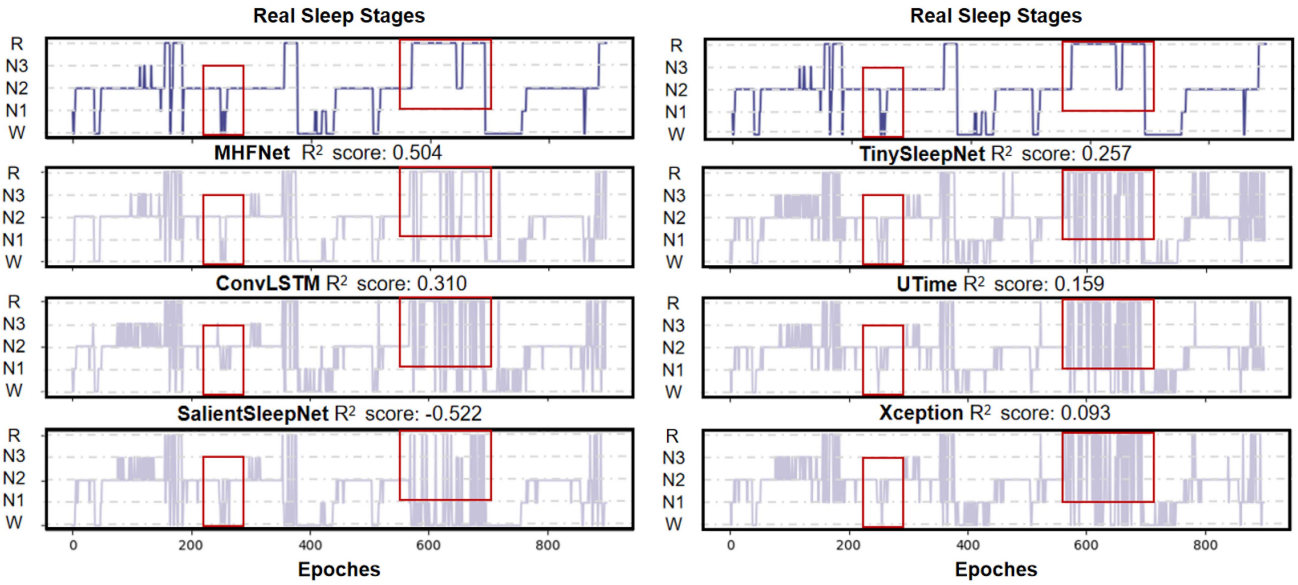

Fig. 6. Compared to the ground truth, the six top-performing methods produce output hypnograms for one example subject of the SleepEDF SC testing set. Red boxes show the outstanding sleep staging performance of the MHFNet compared with other top-performing models.

As depicted in Fig. 7, the HTE module demonstrated an average accuracy improvement of 2.0% at the individual level and 1.5% at the epoch level. Similarly, the DGT module yielded an average accuracy improvement of 2.1% for individual-level and 2.4% for epoch-level. Additionally, the ROH module resulted in an average accuracy increase of 3.7% for individual-level and 3.6% for epoch-level.

VI. DISCUSSION

MHFNet introduces and integrates multi-stream Xception feature extraction, time information embedding, gated multi-attention mechanism, and output refining mechanism to enhance overall performance. The use of multi-channel PSG is widespread in clinical practice, and prior research, such as that presented in [10], has indicated that multimodal automatic sleep staging models like TinySleepNet and EEGNet surpass their single-channel counterparts in classification accuracy. Nevertheless, in the automated classification of sleep stages using multimodal signals, it is essential to account for factors such as the impact of time mode, the extraction and fusion of different feature waveforms across multimodal channels, and the enhancement and fusion of crucial features.

Through 20-fold crossover experiments, we demonstrate the classification performance of the proposed MHFNet over the existing methods on SleepEDF-ST, SleepEDF-SC and SHHS datasets.Furthermore, in individual tests, MHFNet's performance improvements are even more pronounced: a 5.2% improvement in accuracy, a 4.4% improvement in MF1, and a 6.0% improvement in κ, compared with CoRe-Sleep model. We conducted a further comparison to assess the disparities between the predicted stage distributions of the proposed MHFNet model

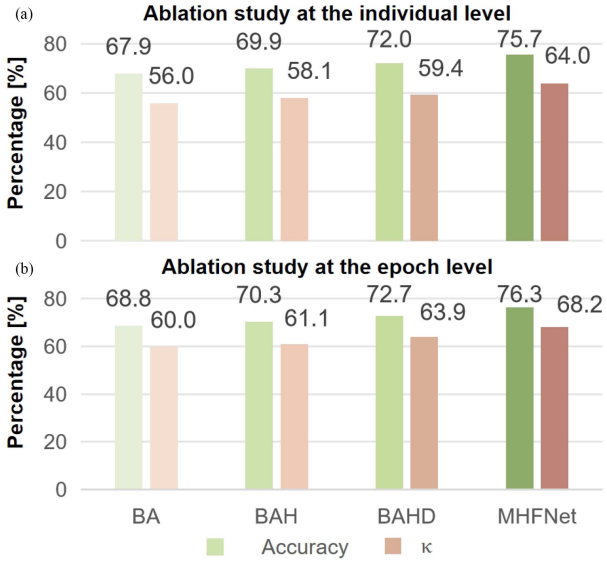

Fig. 7. The accuracy and κ value results are shown in the ablation study. (a) Ablation experiments are shown at the individual level. (b) Ablation experiments are shown at the epoch level.

TABLE IV
ABLATION STUDY ANALYSIS OF MHFNET IN SLEEPEDF ST DATASET

Method	FLOPs (G)	IT (ms)	SleepEDF-ST F1-score for each class				
			W	N1	N2	N3	R
BA	0.695	3.265	86.0	43.9	69.1	71.2	67.1
BAH	0.695	3.838	83.1	40.6	74.2	75.2	65.8
BAHD	0.807	7.342	86.7	43.0	74.4	74.3	75.5
MHFNet	0.984	7.742	86.9	42.0	78.2	77.3	81.6

BA is the baseline model, BAH model is the baseline model with the HTE module, and BAHD is the BAH model with the DGT module. It represents the average inference time for 500 repetitions. The inference batch size is 16.

and those of other state-of-the-art models against the actual distributions individually. Compared with alternative methods, MHFNet exhibits the closest resemblance to the actual distribution at the subject level.

To evaluate the contributions of different MHFNet modules to sleep stage classification, we conducted an ablation study. As shown in Fig. 7, each module improved overall accuracy, MF1, and κ. Table IV details the results on the SleepEDF-SC dataset. Incorporating the HTE module increased accuracy for N2 and N3 stages, likely due to the integration of time information reflecting sleep cycles, though W, N1, and R accuracies slightly decreased. The DGT module further improved feature maps, boosting overall accuracy and classification for most stages, with a minor drop in N1 accuracy due to its similarity to R. Finally, the refining output module, which simulates manual sleep expert review, enhanced performance by integrating information from surrounding epochs.

We used UMAP and t-SNE methods to visualize feature dimensionality reduction after key modules (Fig. 8). As shown in Fig. 8, both methods reveal a mixed pattern of features for each sleep stage after the MXBs, HTE, and DGT modules, highlighting the inherent challenge of sleep staging. The improvement

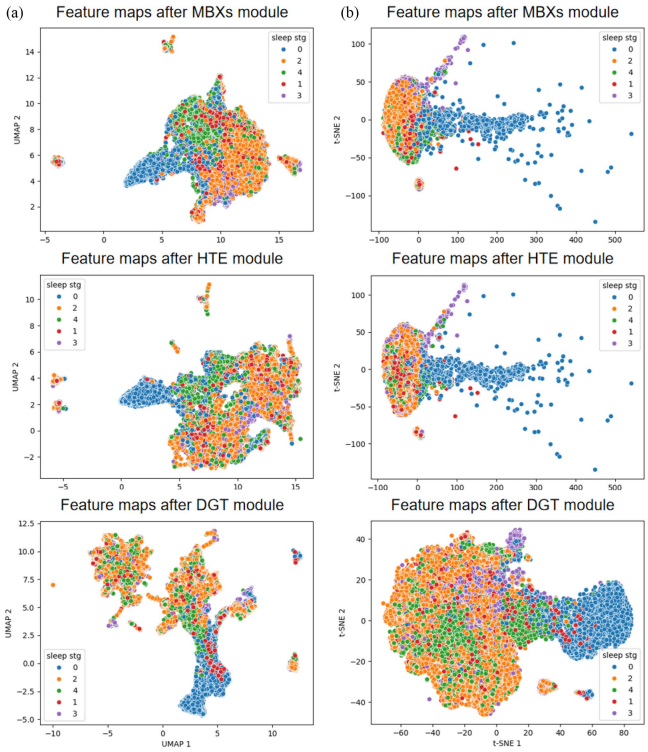

Fig. 8. Feature visualization results on the SleepEDF-SC testing set. (a) UMAP visualization of SleepEDF-SC testing set. (b) t-SNE visualization of SleepEDF-SC testing set.

in feature clustering is minimal, supporting the conclusion that relying solely on single-stage features leads to suboptimal results. Incorporating features from both the previous and next sleep stages, with diagnostic logic, is more likely to improve diagnostic accuracy than enhancing single-stage models.

This study highlights limitations that point to future research opportunities. Adding more modal data could improve sleep stage classification accuracy. PSG, commonly used in clinics, includes multiple EEG leads, mandibular EMG, and ECG data, offering deeper insights into sleep patterns. The multi-stream module in this study can incorporate such data, and future work will focus on validating it with diverse clinical sources. Additionally, exploring multi-epoch models based on diagnostic logic may shed light on sleep stage transitions, enhancing the accuracy and interpretability of classification algorithms.

VII. DECLARATIONS

Conflict of interest The authors declare that they have no conflict of interest.

VIII. CONCLUSION

This study introduces a novel multimodal hybrid-embedding fusion network (MHFNet) designed for automated sleep staging. Our approach represents a pioneering framework integrating local and global time embeddings into the sleep staging. Additionally, we propose a dual-path gate attention mechanism to augment time-related features. Furthermore, introducing a refining output header, adhering to scoring rules, and leveraging

adjacent epochs enhances the precision of the current staging outcome. Experimental results demonstrate that the MHFNet framework achieves state-of-the-art performance, particularly at the individual level. This improvement can potentially mitigate false positive predictions, thereby enhancing the clinical utility of the model.

REFERENCES

[1] R. B. Berry et al., "The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications. version 2.6," American Academy of Sleep Medicine, 2023. [Online]. Available: https://aasm.org/clinical-resources/scoring-manual/

[2] A. R. Hassan and M. I. H. Bhuiyan, "Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating," *Biomed. Signal Process. Control.*, vol. 24, pp. 1–10, 2016.

[3] M. Gaiduk et al., "Current status and prospects of automatic sleep stages scoring: Review," *Biomed. Eng. Lett.*, vol. 13, pp. 247–272, 2023.

[4] Z. Jia et al., "GraphSleepNet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 1324–1330.

[5] E. Efe and S. Özsen, "CoSleepNet: Automated sleep staging using a hybrid CNN-LSTM network on imbalanced EEG-EOG datasets," *Biomed. Signal Process. Control.*, vol. 80, no. Part, 2023, Art. no. 104299.

[6] Z. Jia et al., "SalientSleepNet: Multimodal salient wave detection network for sleep staging," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 2614–2620.

[7] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG," in *Proc. IEEE 42nd Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2020, pp. 641–644.

[8] P. Huy et al., "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.

[9] S. Akara et al., "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.

[10] Y. Dai et al., "Multichannelsleepnet: A transformer-based model for automatic sleep stage classification with PSG," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 9, pp. 4204–4215, Sep. 2023.

[11] J. Huang, L. Ren, X. Zhou, and K. Yan, "An improved neural network based on senet for sleep stage classification," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 10, pp. 4948–4956, Oct. 2022, doi: 10.1109/JBHI.2022.3157262.

[12] J. Xu and H. Sun, "Sleep analysis during light sleep based on k-means clustering and BiLSTM," in *Proc. Web Inf. Syst. Appl.: 18th Int. Conf.*, 2021, vol. 12999, pp. 207–214.

[13] M. Z. Alom et al., "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," *J. Med. Imag.*, vol. 6, no. 1, 2018, Art. no. 014006.

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1800–1807.

[15] M. Perslev et al., "U-Time: A fully convolutional network for time series segmentation applied to sleep staging," in *Proc. NeurIPS*, 2019, pp. 4417–4428.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput-Assist. Interv.-MICCAI 2015: 18th Int. Conf.*, 2015, pp. 234–241.

[17] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, "Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 8, pp. 2456–2467, Aug. 2022.

[18] K. Kontras, C. Chatzichristos, H. Phan, J. Suykens, and M. De Vos, "CoRe-Sleep: A multimodal fusion framework for time series robust to imperfect modalities," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 840–849, 2024.

[19] S. Lee, Y. Yu, S. Back, H. Seo, and K. Lee, "SleePyCo: Automatic sleep scoring with feature pyramid and contrastive learning," *Expert Syst. With Appl.*, vol. 240, 2024, Art. no. 122551.

[20] V. J. Lawhern et al., "Eegnet: A compact convolutional network for EEG-based brain-computer interfaces," 2016, *arXiv:1611.08024*.

[21] J. Wang et al., "RTFormer: Efficient design for real-time semantic segmentation with transformer," in *Proc. NeurIPS*, 2022, pp. 7423–7436.

[22] L. Wolf et al., "A deep learning approach for the segmentation of electroencephalography data in eye tracking applications," in *Proc. Int. Conf. Mach. Learn.*, 2022, vol. 162, pp. 23912–23932.

[23] M. Liu et al., "Gated transformer networks for multivariate time series classification," 2021, *arXiv:2103.14438*.

[24] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:13756489

[25] H. I. Fawaz et al., "Inceptiontime: Finding alexnet for time series classification," *Data Min. Knowl. Discov.*, vol. 34, no. 6, pp. 1936–1962, 2020.

[26] C. Szegedy et al., "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[27] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *SIGKDD*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:4650265

[28] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 10123, pp. e215–20, 2000.

[29] B. Kemp et al., "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.

[30] S. F. Quan et al., "The sleep heart health study: Design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–85, 1997, Online; Accessed: Aug. 7, 2024.

[31] G.-Q. Zhang et al., "The national sleep research resource: Towards a sleep data commons," *J. Amer. Med. Informat. Assoc. : JAMIA*, vol. 25, no. 10, pp. 1351–1358, 2018, Online; Accessed: Aug. 7, 2024.

[32] T. Hori et al., "Proposed supplements and amendments to 'a manual of standardized terminology, techniques and scoring system for sleep stages of human subjects', the Rechtschaffen & Kales (1968) standard," *Psychiatry Clin. Neurosciences*, vol. 55, no. 3, pp. 305–10, 2001, Online; Accessed: Aug. 7, 2024.

[33] D. D. Lewis, R. E. Schapire, J. P. Callan, and R. Papka, "Training algorithms for linear text classifiers," in *Proc. 19th Annu. Int.ACM SIGIR Conf. Res. Develop. Inf. Retrieval, SIGIR'96*, H. Frei, D. Harman, P. Schäuble, and R. Wilkinson, Eds. Zurich, Switzerland: ACM, Aug. 1996, pp. 298–306.

[34] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, pp. 37–46, 1960. [Online]. Available: https://api.semanticscholar.org/CorpusID:15926286