MultiChannelSleepNet: A Transformer-Based Model for Automatic Sleep Stage Classification With PSG

Yang Dai, Xiuli Li, Shanshan Liang, Lukang Wang, Qingtian Duan, Hui Yang, Chunqing Zhang, Xiaowei Chen, Longhui Li, Xingyi Li, and Xiang Liao

***Abstract*—Automatic sleep stage classification plays an essential role in sleep quality measurement and sleep disorder diagnosis. Although many approaches have been developed, most use only single-channel electroencephalogram signals for classification. Polysomnography (PSG) provides multiple channels of signal recording, enabling the use of the appropriate method to extract and integrate the information from different channels to achieve higher sleep staging performance. We present a transformer encoder-based model, MultiChannelSleepNet, for automatic sleep stage classification with multichannel PSG data, whose architecture is implemented based on the transformer encoder for single-channel feature extraction and multichannel feature fusion. In a single-channel feature extraction block, transformer encoders extract features from time-frequency images of each channel independently. Based on our integration strategy, the feature maps extracted from each channel are fused in the multichannel feature fusion block. Another set of transformer encoders further capture joint features, and a residual connection preserves the original information from each channel in this block. Experimental results on three publicly available datasets demonstrate that our method achieves higher classification performance than state-of-the-art techniques. MultiChannelSleepNet is an efficient method to extract and integrate the information from multichannel PSG data, which facilitates precision sleep staging in clinical applications.**

***Index Terms*—Automatic sleep stage classification, transformer encoder, deep learning, PSG data, feature extraction and fusion.**

I. Introduction

SLEEP is a significant process for human health, which is related to emotions, behaviors, and physiological functions such as memory and cognition. Studies have suggested that sleep disorders may lead to a range of physical and mental illnesses [1]. Lack of sleep is associated with increased mortality and conditions such as Alzheimer's disease [2].

Sleep staging is a fundamental process in sleep assessment and study. Accurate sleep stage classification can effectively support the analysis of sleep situations. Clinically, polysomnography (PSG) is an important technique to analyze sleep stage and sleep disorder diagnosis. The PSG system generally consists of electroencephalogram (EEG), electrooculogram (EOG), electrocardiogram (ECG), and electromyogram (EMG). Based on the Rechtschaffen and Kales (R&K) rules [3], PSG signals are typically split into segments of 30 s and classified into six sleep stages, which are wakefulness (Wake), four non-rapid eye movement stages (i.e., S1, S2, S3 and S4), and rapid eye movement (REM). Due to limited evidence of a difference between the S3 and S4, the American Academy of Sleep Medicine (AASM) made modifications and additions to the division of sleep stages based on the R&K rules in 2007 [3], [4]. AASM defined only three non-rapid eye movement (Non-REM: N1, N2 and N3) stages, and N3 is merged from S3 and S4.

Traditionally, the classification of sleep stages requires expert visual discrimination of PSG signal recordings, but this is complex and time-consuming, and it relies on the subjective judgment of experts [5]. To solve these problems, many studies have adopted machine learning methods for automatic sleep staging, which usually require some signal processing methods to extract features from raw signals. Some features are then chosen and used by classic machine learning algorithms for classification, including naive Bayes [6], k-nearest neighbor [7],

Manuscript received 8 January 2023; revised 12 May 2023; accepted 31 May 2023. Date of publication 8 June 2023; date of current version 6 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 32127801 and 31925018, and in part by the Chongqing Technology Innovation and Application Development Special Key Project under Grant cstc2019jscx-dxwtBX0010. *(Yang Dai and Xiuli Li contributed equally to this work.) (Corresponding authors: Longhui Li; Xingyi Li; Xiang Liao.)*

Yang Dai is with the Center for Neurointelligence, School of Medicine, Chongqing University, Chongqing 400030, China, and also with the College of Bioengineering, Chongqing University, Chongqing 400044, China (e-mail: valar_d@163.com).

Xiuli Li, Shanshan Liang, and Xiaowei Chen are with the Brain Research Center and State Key Laboratory of Trauma, Burns, and Combined Injury, Third Military Medical University, Chongqing 400038, China (e-mail: lilybrc@aliyun.com; 15340520947@163.com; xiaowei_chen@tmmu.edu.cn).

Lukang Wang, Qingtian Duan, Hui Yang, and Chunqing Zhang are with the Department of Neurosurgery, Xinqiao Hospital, Third Military Medical University, Chongqing 400037, China (e-mail: lkwang2009@aliyun.com; dqttmmu2014@163.com; 13808390069@163.com; cqzhang@tmmu.edu.cn).

Longhui Li, Xingyi Li, and Xiang Liao are with the Center for Neurointelligence, School of Medicine, Chongqing University, Chongqing 400030, China (e-mail: lilonghui@cqu.edu.cn; xingyi_li@cqu.edu.cn; xiang.liao@cqu.edu.cn).

Source code is available at https://github.com/yangdai97/MultiChannelSleepNet.

Digital Object Identifier 10.1109/JBHI.2023.3284160

[8], support vector machine [9], [10], [11], and random forest [12], [13].

Although traditional machine learning algorithms have performed relatively well at the sleep staging task, they are highly dependent on the researchers' knowledge of sleep medicine when selecting appropriate features to distinguish among stages. With the popularity of deep learning, researchers have applied such approaches to perform sleep stage classification; then, given the success of convolutional neural networks (CNN), researchers have adopted architectures based on CNN for sleep staging [13], [14], [15], [16]. Sors et al. [14] proposed a 14-layer CNN, which took as input the sleep epoch to be classified, the two preceding epochs, and the following epoch. With new layers including convolution, competitive, and pooling operators, an unsupervised competition CNN was developed for sleep staging [15]. A CNN framework was designed to process EEG signals after applying a short-time Fourier transform (STFT) or stationary wavelet transform (SWT), for a simpler and smaller deep learning model [16]. Sokolovsky et al. [17] found that a CNN could spontaneously learn specific signal features, which were also used in visual discrimination by experts. In general, a CNN can extract features from raw data and classify them by learning, instead of manually designing and selecting features. However, the principle of CNN is to use convolution kernels of different sizes to capture spatial features and does not take into account the temporal dependence of PSG signals.

The recurrent neural network (RNN) is often used to process sequence data and has been shown to perform well in fields such as natural language processing and voice recognition. Given that sleep is a continuous process, many studies have exploited RNNs to capture temporal dependency within time-series PSG data [17], [18], [19], [20], [21]. SeqSleepNet [18] used bidirectional RNNs to extract both epoch- and sequence-level features. A recurrent layer based on an attention mechanism captured short-term sequential features and was followed by another recurrent layer that learned epoch-wise features for long-term modeling. After fusing manually extracted features and features learned by the network, Sun et al. [19] designed a multi-flow RNN to learn temporal information on feature maps. Models combining CNNs and RNNs have also become common. The DeepSleepNet [20] architecture uses two CNNs with different-sized convolution kernels to extract features from a single sleep epoch, and a long short-term memory (LSTM) network to learn conversion laws among different sleep stages from these epochs. Similarly, SleepEEGNet [21] is composed of CNNs to extract time-invariant features and bidirectional RNNs to capture contextual dependencies between sleep epochs. Korkalainen et al. [22] combined convolutional and LSTM networks for automatic sleep staging to study the effect of obstructive sleep apnea severity on classification accuracy.

While the RNN has shown its advantages in sequence modeling with its recurrent structure, it also has high model complexity and is difficult to train in parallel. Inspired by the transformer network [23], some work has employed an attention mechanism to avoid the use of RNNs. Eldele et al. [24] proposed a temporal context encoder using a causal convolution-based multi-head self-attention layer, which was proven capable of capturing temporal dependencies. Qu et al. [25] proposed a network using multi-head self-attention and a CNN to capture global temporal correlations. The above studies still must use other network structures for feature extraction before using the attention mechanism, which makes the network complex. SleepTransformer [26] used the original transformer [23] as the backbone and offered an interpretable prediction process, but it had only mediocre performance on the sleep staging task.

As described above, many recent studies have adopted a seq-to-seq architecture for the EEG-based sleep staging task [18], [25], [27]. These networks receive a sequence of multiple epochs as input, enabling them to capture the temporal relationship between continuous epochs. However, this structure expands the model size and complexity, and the RNN increases the training cost. To overcome these shortcomings, AttnSleep [24] used only one epoch as the input, achieving good classification results while greatly reducing training time. This suggests that there is still potential to develop one-to-one architectures using a single sleep epoch as input data.

PSG is a system with multiple channels, which intuitively should enable the model to obtain more information and achieve better sleep stage classification results. Based on the multichannel scheme, Chambon et al. [28] presented a neural network that used multivariate and multimodal signals (EEG, EOG, and EMG) as input data. Phan et al. [29] transformed multichannel raw signals to time-frequency images for a proposed CNN and formulated sleep staging as a joint classification and prediction problem. However, most studies still used only one of the recording channels, although other channels in the dataset were also available [25], [26], [30]. One reason is that it can reduce both the complexity and training cost of the model, and another main purpose is the studies of the automatic sleep staging using single-channel EEG would have important application for longitudinal research and homed-based sleep monitoring. By contrast, taking multiple physiological signals into consideration is more closed to the manual operation of sleep experts, and easier to be accepted in clinical applications. Previous studies have compared the results of the proposed models using single-channel and multichannel data as input, and the results showed that multichannel data did not significantly improve experimental results [17], [31]. Indeed, these studies just used a simple method like concatenation to learn from multichannel PSG data, and this led to underutilization of the information contained in the multichannel data. Researchers have begun to notice this problem. Jia et al. [32] used a feature fusion method based on squeeze-and-excitation to integrate features from EEG and EOG signals, and designed CNNs with different structures for classification.

The input signals of existing methods usually have two types of representations, one is the one-dimensional raw signals [20], [24], [28], and another is the time-frequency images [18], [26]. The time-frequency images are obtained from the raw signals processed with time-frequency analysis, such as short-time Fourier transform (STFT), Choi-Williams distribution (CWD) and continuous wavelet transform (CWT). Although both time-frequency images and the raw signals have their own specific features, one advantage of time-frequency images is that they contain both time-domain and frequency-domain features of the signals, which are crucial in the process of sleep staging. In general, time-frequency images are considered as a high-level representation of the raw signals [31].

We present *MultiChannelSleepNet*, a deep learning model using multichannel PSG data including EEG and EOG without convolutional and recurrent structures to classify sleep stages, based on the transformer encoder [23] for single-channel feature extraction and multichannel feature fusion, with a self-attention mechanism and one-to-one architecture, which reduces the size and complexity of the model and enables parallel training. MultiChannelSleepNet accepts time-frequency images as inputs, and configures different PSG channels with the same number of transformer encoders that have the same parameters during feature extraction, which enables it to adaptively capture unique information from different channels with different modality contributions to sleep staging, without having to design specific structures for each channel. The multichannel feature fusion block can integrate features extracted from each channel. Compared with variants that just use one channel as input of MultiChannelSleepNet or use simple feature fusion strategies, it improves sleep staging performance. During training, we use the AdamW optimizer [33] instead of the commonly used Adam [34] or stochastic gradient descent (SGD). In experiments on three publicly available datasets, MultiChannelSleepNet surpasses state-of-the-art approaches in terms of different evaluation metrics.

We summarize our major contributions:

1) We adopt transformer encoders to extract features from single-channel PSG recordings including EEG and EOG, which can adaptively extract intra-epoch information from each channel.
2) We propose a block to fuse the features from different channels. Transformer encoders are used to capture multichannel joint features from concatenated single-channel feature maps. A residual connection is applied to preserve the original information of each channel and avoid gradient-vanishing.
3) Instead of Adam or SGD, we apply the AdamW optimizer during training. It modifies the mechanism of weight decay in the original Adam, so as to better suppress the overfitting of the model while making the loss converge better.

II. Proposed Method

Formally, given a training dataset $\{S_n\}_{n=1}^{N}$ of size N, $S_n = (\{X_1^{(n)}, \ldots, X_C^{(n)}\}, y^{(n)})$ is the n-th data group of C channels from the same 30-s period. $X_i^{(n)} \in \mathbb{R}^{T\times F}$, $1 \leqslant i \leqslant C$, represents a time-frequency image obtained from a 30-s EEG or EOG epoch, with label $y^{(n)} \in \{0,1\}^K$, where $K = 5$, as the task is five-stage sleep classification.

Fig. 1 shows the MultiChannelSleepNet model architecture, which consists of 1) single-channel feature extraction; 2) multichannel feature fusion; and 3) classification. For each EEG or EOG channel, transformer encoders [23] extract the feature map from the time-frequency image. These are independent, so they can adaptively capture different features from multimodal physiological signals. A multichannel feature fusion block integrates the features from different channels. After concatenating feature maps from the single-channel feature extraction block, another smaller set of transformer encoders is used to extract joint features. A residual connection [35] is applied to add the original feature maps from each channel to the joint features, so that the multichannel feature fusion block can integrate both of the commonalities and differences of each channel. Two fully connected layers and a softmax activation function at the end of the network complete the classification. The AdamW optimizer [33] is used to minimize the class-aware loss and learn the model parameters.

A. Single-Channel Feature Extraction

1) Time-Frequency Image Representation: According to the AASM scoring manual [4], EEG, EOG, EMG, major body movement and so on are the basis for sleep staging, and specific waves and frequency components are important features. For instance, a low-frequency component in the range of 4–7 Hz is often observed during the N1 stage, and the waves of sleep spindles (SS) or K-complexes (KC) are symbols of the N2 stage. By applying STFT and logarithm scaling, we converted the raw signals from each channel to time-frequency images as the input of the model, which can efficiently represent those specific waves and frequency components.

2) Transformer Encoder and Position Encoding: We exploit transformer encoders to extract features from time-frequency images. Transformer [23], an efficient model based on self-attention, has been shown to perform well in a variety of sequential modeling tasks, especially those of the natural language processing. In our task, we can also consider a time-frequency image $X \in \mathbb{R}^{T\times F}$ along the time dimension as a sequence including T spectral columns, and each spectral column $W \in \mathbb{R}^{1\times F}$ is treated as a word vector in a sentence processed in the natural language processing task. The original transformer model consists of an encoder and a decoder. Since the decoder is generally used in generation or prediction tasks, we just applied the encoder to this classification task. The transformer encoder architecture is presented in Fig. 2.

The transformer encoder has multi-head attention and a feed-forward network. As the transformer encoder discards recurrence, positional encodings are added to the original input to introduce temporal information,

$$\tilde{X} = X + PE, \tag{1}$$

where $PE \in \mathbb{R}^{T\times F}$ represents the positional encoding matrix. We used sine and cosine functions to encode position information [23],

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/F}), \tag{2}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/F}), \tag{3}$$

where pos is the position, $0 \leq pos \leq T$, and $2i$ or $2i+1$ is the dimension of the input, $0 \leq 2i \leq 2i+1 \leq F$. On the one hand, the above formulas can directly calculate the position code in a sleep epoch, each combination of position and dimension

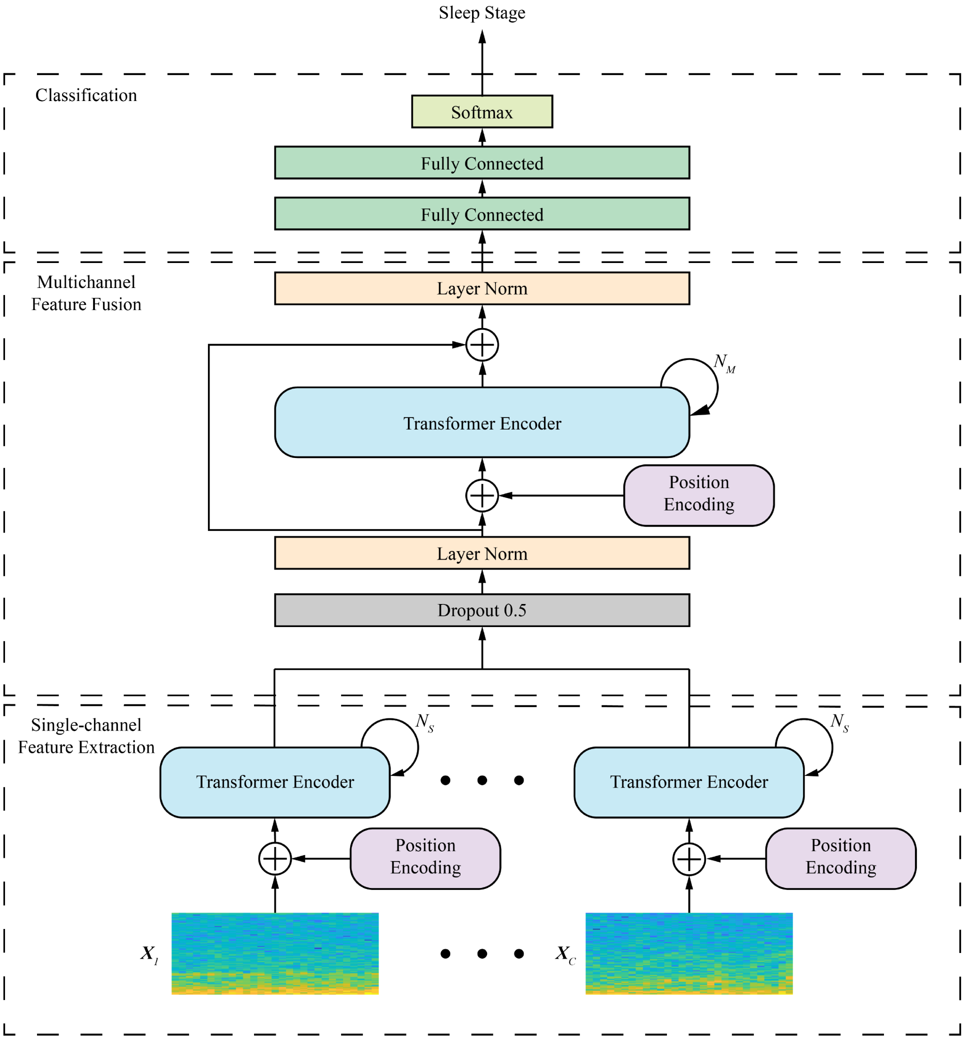

Fig. 1. MultiChannelSleepNet architecture. Model accepts the time-frequency images as input, and the features of each channel are first extracted individually and then integrated with a multichannel feature fusion block. Finally, the classification is completed using two fully connected layers. Here, N_S and N_M represent the number of transformer encoders in a single-channel feature extraction block and multichannel feature fusion block, respectively.

corresponds to a specific value, and it has no parameters that require training. On the other hand, each dimension of the position encoding corresponds to a sinusoid, which allows the model to easily learn the relative positions according to the character of sine and cosine functions. PE_{pos+k} can be represented as a linear function of PE_{pos} for any fixed offset k.

3) Multi-Head Attention: In the transformer encoder, the multi-head attention module captures the dependencies among vectors at different positions. Fig. 3 shows the structure of the multi-head attention module. In this module, $\tilde{X} \in \mathbb{R}^{T\times F}$ in (1) is copied into three copies as the input. Since these three copies have different meanings in the computation, we follow the convention in [23] and call them query, key and value, denoted by Q, K and V, respectively. Here Q, K and V are actually identical ($Q = K = V = \tilde{X}$). After that, each of Q, K and V is divided into H segments in the frequency domain, and the subsequent computation will be performed in each of these segments, which is called that this module contains H heads. This facilitates the model to focus on more detailed information. Specifically, this enables each head to learn features from different frequency domain in our task.

In each head, the segments $Q_i, K_i, V_i \in \mathbb{R}^{T\times F/H}$ from query, key, and value are fed into three single linear layers, respectively. The outputs of the linear layers $\tilde{Q}_i$, $\tilde{K}_i$ and $\tilde{V}_i$ are processed by attention mechanism,

$$\text{Attention}\left(\tilde{Q}_i, \tilde{K}_i, \tilde{V}_i\right) = \text{softmax}\left(\frac{\tilde{Q}_i\tilde{K}_i^{\mathrm{T}}}{\sqrt{F}}\right)\tilde{V}_i. \quad (4)$$

The attention mechanism calculates the relationship between the features at each location and the features at other locations by means of dot product in a sleep epoch. The result of the dot product reflects the strength of the relationship between them, which is so called “attention”.

Then, the results of the attention calculation from H heads are concatenated, and the result is mapped from a higher dimension

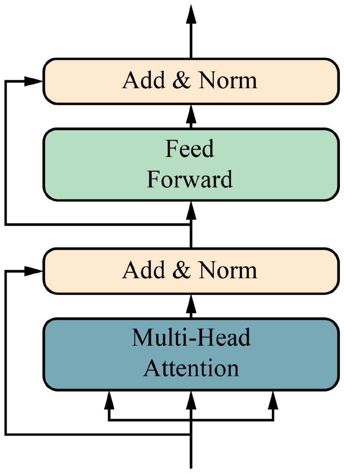

Fig. 2. Transformer encoder architecture. Multi-head attention module extracts dependencies from different positions of input data, and residual connections can prevent gradient vanishing.

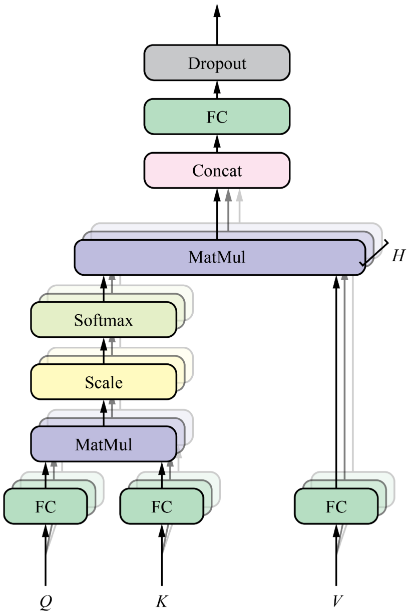

Fig. 3. Structure of multi-head attention module. H is the number of heads.

to the same dimension of the input $\tilde{X}$ by another fully connected layer and processed through a dropout layer.

Finally, in each channel of PSG data, time-frequency images go through N_S transformer encoders:

$$O_i = \text{TransformerEncoder}(O_{i-1}), 1 \leq i-1<i \leq N_S, \quad (5)$$

where $O_1 = \tilde{X}$, and O_{N_S} is the feature map of this channel after single-channel feature extraction.

B. Multichannel Feature Fusion

In this block, the processing aims to integrate the features of each channel and further capture multichannel joint features. Based on the single-channel feature extraction block in Section II.A, the time-frequency images $(X_1, \ldots, X_C)$ are transformed to feature maps $(O_1, \ldots, O_C)$, where $O_i \in \mathbb{R}^{T\times F}$, $1 \leq i \leq C$. We concatenate these feature maps in the column direction to make them the input $O \in \mathbb{R}^{T\times CF}$ of block. Since different PSG channels contain much similar information, a dropout layer [36] is applied to reduce the tendency of the model to overfit, and layer normalization [37] can accelerate the training process. The result of processing the feature map O with these operations is denoted as M.

We exploit N_M transformer encoders to further capture joint features from this multichannel feature map. According to Section II.A.2, position encoding is also added to M before transformer encoders. In addition, the input before position embedding and output from transformer encoders are added by a residual connection [35], and another layer normalization is applied. The residual connection retains a copy of the feature map before entering N_M transformer encoders, and adds it to the joint feature output by the transformer encoders, which allows MultiChannelSleepNet to preserve the individual features of each channel to a certain extent, and combine commonalities and differences among features from different channels, and also avoids gradient-vanishing during training. These operations can be formulated as

$$\tilde{M} = M + PE, \quad (6)$$

$$\tilde{O}_i = \text{TransformerEncoder}(\tilde{O}_{i-1}), \quad 1 \leq i-1<i \leq N_M, \tilde{O}_1 = \tilde{M}, \quad (7)$$

$$\tilde{Z} = \text{LayerNorm}(M + \tilde{O}_{N_M}), \quad (8)$$

where PE is the position encoding matrix, as in (2) and (3), and $\tilde{Z} \in \mathbb{R}^{T\times CF}$ is the output of this block.

C. Classification

A block with two fully connected layers accepts the output of the multichannel feature fusion block $\tilde{Z}$. These two fully connected layers can select and combine the features learned in the previous structures to achieve the sleep stage classification. In the first fully connected layer, a ReLU activation function is adopted and is followed by a dropout layer. A softmax function is then employed to generate output probabilities for mutually exclusive classes. We choose AdamW [33] as the optimizer to learn model parameters during training. This improves the weight decay strategy of Adam [34] and replaces the traditional L2 regularization.

III. Experiments

A. Dataset

To evaluate our proposed model, we conducted experiments on three publicly available datasets: SleepEDF-20, SleepEDF-78, and Sleep Heart Health Study (SHHS). These are described in Table I.

SleepEDF-20: This subset of the Sleep-EDF Expanded dataset (2013 version) [38], [39] includes two studies. The Sleep Cassette (SC) study included 20 subjects aged from 25 to 34 years and was designed to study the relationship between age

TABLE I
DETAILS OF EMPLOYED DATASETS (EACH DATA SAMPLE IS A 30-S EPOCH)

Dataset	Subjects	W	N1	N2	N3	REM	Total Samples
SleepEDF-20	20	9118 21.10%	2804 6.50%	17799 41.30%	5703 13.20%	7717 17.90%	43141
SleepEDF-78	79	66822 34.00%	21522 11.00%	69132 35.20%	13039 6.60%	25835 13.20%	196350
SHHS	329	43619 14.30%	10304 3.20%	142125 43.70%	60153 18.50%	65953 20.30%	324854

and sleep in healthy subjects. The Sleep Telemetry (ST) study examined the effect of temazepam on sleep. To avoid the influence of other factors (e.g., drugs), and following previous studies [17], [20], [29], we used the SC subset in this work. SleepEDF-20 contains PSG recordings from each subject for two consecutive days and nights. The recordings of the second night from subject 13 were lost due to a failing cassette or laserdisc. Sleep experts labeled each 30-s epoch in the dataset as one of eight stages {W, S1, S2, S3, S4, REM, MOVEMENT, UNKNOWN} based on R&K rules [3] by visually judging signal characteristics. Similar to previous work [18], [24], [40], S3 and S4 were merged into N3. Moreover, the stages of MOVEMENT and UNKNOWN were removed. In our experiments, the Fpz-Cz EEG, Pz-Oz EEG, and ROC-LOC EOG (horizontal) channels, with a sampling rate of 100 Hz, were adopted for sleep staging. We observed the common setting of keeping only the time window of 60 minutes centered around the in-bed part of a recording [20].

SleepEDF-78: This is a subset of the Sleep-EDF Expanded dataset (2018 version) [38], [39], which has been expanded to contain 78 subjects aged from 25 to 101 years, with 153 whole-night PSG sleep recordings. Similar to SleepEDF-20, the researchers performed PSG recordings on each subject for two nights. Due to a device error, one record was lost for each of subjects 13, 36, and 52. The standard of manual scoring and the labels of 30-s PSG epochs are the same as for SleepEDF-20. We adopted the Fpz-Cz EEG, Pz-Oz EEG, and ROC-LOC EOG (horizontal) channels for sleep staging.

SHHS: This is a multi-center cohort study to research the consequences of sleep-disordered breathing, such as cardiovascular [41], [42]. It contains two parts of PSG records. SHHS visit 1 consists of 6441 subjects aged 40 years and older, and SHHS visit 2 is obtained from 3295 subjects in visit 1. Following previous research [24], [27], 329 subjects who were considered to have normal sleep rhythm were selected for the experiment. In the same setup as the other datasets, N3 was merged from the previous S3 and S4 stages, and the MOVEMENT and UNKNOWN stages were excluded. The C4-A1 EEG, C3-A2 EEG, LOC EOG, and ROC EOG channels were adopted for sleep staging.

B. Parameters

As described in Section II.A, a 30-s epoch from PSG signals was transformed to a time-frequency image by 256-point STFT. A 2-s Hamming window with a 50% overlap rate was chosen as the window function. We used logarithm scaling on the spectra to convert to log-power spectra. This produced an image $X \in \mathbb{R}^{T \times F}$ with $F = 128$ frequency bins and $T = 29$ time points. The time-frequency image calculated from each channel was normalized to zero mean and unit variance across all time-frequency pairs, and input to MultiChannelSleepNet.

In the transformer encoders of our network, we employed $H = 8$ attention heads and 1024 hidden units in a feedforward layer. A dropout rate of 0.1 was used for each sub-layer of transformer encoders. For each channel, $N_S = 16$ transformer encoders were used to extract features. In the multichannel feature fusion block, $N_M = 4$ transformer encoders with the same parameters were used, and a dropout rate of 0.5 was used in the dropout layer. In the last two fully connected layers, 1024 hidden units and a dropout rate of 0.5 were used.

The model was built using PyTorch 1.9.1, and training was executed on a GeForce RTX 3090 GPU. We used subject-wise k-fold cross-validation by dividing the subjects in each dataset into k groups with the three datasets to evaluate the performance of MultiChannelSleepNet, with k values for SleepEDF-20, SleepEDF-78, and SHHS of 20, 10, and 20, respectively, which is the same setup as the previously reported methods. A validation set was left out to evaluate our model during training, and the number of subjects in this validation set is the same as the test set. Cross-entropy loss was adopted as a training objective, and the AdamW optimizer [33] with a learning rate of 5×10^{-6} was used to minimize it. The weight decay of the optimizer was set to 10^{-2}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-2}$. A minibatch size of 64 was used during training. We introduced the early stopping strategy during training, and the training process was stopped if the accuracy for the validation dataset did not increase after 20 epochs.

C. Performance Evaluation Metrics

We adopted precision (Pre), recall (Rec), and F1-score (F1) as per-class metrics to evaluate MultiChannelSleepNet. Other metrics, including the accuracy (ACC), macro-averaged F1-score (MF1) [43], and Cohen's kappa (κ) [44], were used to evaluate the overall model performance. MF1 is the arithmetic mean of the F1-score of the five categories.

D. Experimental Results

1) Sleep Staging Performance: Tables II–IV use confusion matrices to show the performance of MultiChannelSleepNet. Rows and columns represent ground truth and predicted results, respectively. Bolded numbers indicate the correctly classified

TABLE II
PERFORMANCE METRICS AND CONFUSION MATRIX OF THE SLEEPEDF-20 DATASET

	Predicted					MultiChannelSleepNet			SleepContextNet [27]			AttnSleep [24]		
	W	N1	N2	N3	REM	PR	RE	F1	PR	RE	F1	PR	RE	F1
W	**8399**	357	94	25	243	**93.4**	**92.1**	**92.8**	91.2	88.0	89.6	89.6	89.7	89.7
N1	377	**1244**	625	5	553	**55.0**	44.4	49.1	46.9	**54.7**	**50.5**	47.1	39.1	42.8
N2	85	397	**16074**	591	652	**89.5**	**90.3**	**90.0**	88.0	88.7	88.4	89.1	88.6	88.8
N3	16	3	573	**5107**	4	89.1	89.5	89.3	**90.5**	86.7	88.5	80.7	**89.8**	**90.2**
REM	113	261	584	5	**6754**	82.3	**87.5**	**84.8**	**82.5**	81.6	82.0	76.1	82.2	79.0

TABLE III
PERFORMANCE METRICS AND CONFUSION MATRIX OF THE SLEEPEDF-78 DATASET

	Predicted					MultiChannelSleepNet			SleepContextNet [27]			AttnSleep [24]		
	W	N1	N2	N3	REM	PR	RE	F1	PR	RE	F1	PR	RE	F1
W	**62180**	3284	487	67	804	**95.0**	**93.1**	**94.0**	93.0	92.5	92.8	92.3	91.8	92.0
N1	2437	**10480**	6497	94	2014	**58.1**	**48.7**	**53.0**	50.0	48.0	49.0	45.3	39.2	42.1
N2	350	2730	**62204**	1965	1883	84.0	**90.0**	**86.9**	**84.3**	85.4	84.8	83.5	86.5	85.0
N3	20	15	2484	**10516**	4	**83.1**	80.7	81.8	80.2	81.0	80.6	82.3	**82.0**	**82.1**
REM	436	1533	2380	19	**21467**	**82.0**	**83.1**	**82.6**	78.6	79.1	78.9	73.1	75.3	74.2

TABLE IV
PERFORMANCE METRICS AND CONFUSION MATRIX OF THE SHHS DATASET

	Predicted					MultiChannelSleepNet			SleepContextNet [27]			AttnSleep [24]		
	W	N1	N2	N3	REM	PR	RE	F1	PR	RE	F1	PR	RE	F1
W	**42925**	985	1299	159	951	**91.7**	**92.7**	**92.2**	88.8	89.6	89.2	90.3	83.3	86.7
N1	1469	**3584**	2138	8	3103	48.4	34.8	40.5	**51.4**	**52.7**	**52.0**	30.6	36.7	33.2
N2	1388	2203	**124462**	5261	8811	**89.6**	87.6	**88.6**	87.0	**88.3**	87.6	87.4	86.7	87.1
N3	40	1	8345	**51713**	54	**90.5**	86.0	**88.2**	86.6	82.1	84.3	87.5	**87.6**	87.5
REM	967	628	2712	22	**61624**	82.7	**93.4**	87.7	**89.0**	89.6	**89.3**	80.5	83.8	82.1

epochs with our model. On the right side of these tables, we list the per-class evaluation matrix of MultiChannelSleepNet and baselines, where the optimal value of each evaluation metric is bolded.

From Tables II– IV, we observe that most stages are accurately classified except for N1, which has the fewest epochs. From those confusion matrices, we can see that MultiChannelSleepNet outperforms AttnSleep [24] and SleepContextNet [27] on the three datasets in terms of almost all three metrics (i.e., precision, recall, and F1-score). Notably, on the SHHS dataset, our model performs slightly worse than SleepContextNet [27] for N1 because N1 is the first stage transitioning between wakefulness and sleep, and it has some of the same features as stages W and N2. SleepContextNet [27] refers more to the context, so that it achieves better performance on N1 when a large amount of data is available for training.

Fig. 4 shows the loss and accuracy curves during training, which indicate that MultiChannelSleepNet can converge fast and reach stable performance after a few epochs. Over the training process, accuracy increases and loss decreases continuously on the training set, and both accuracy and loss on the validation set tend to stabilize after several training epochs, which shows that MultiChannelSleepNet is robust to suppress model overfitting.

2) Hypnogram: Fig. 5 shows the hypnogram of subject SC4061E0 from SleepEDF-20. The top denotes the ground truth, and the bottom denotes the classification result from MultiChannelSleepNet. It can be seen that the predicted hypnogram is very similar to the ground truth. For this subject, the accuracy is 89.6% and the F1-score is 0.84 when using MultiChannelSleepNet. Most misclassified sleep epochs are N1, which is consistent with the results in Table V. Intuitively, the hypnogram produced by MultiChannelSleepNet may not be as smooth as the ground truth, and includes some abnormal sleep transitions, due to the fact that each epoch is input to MultiChannelSleepNet after shuffling, and does not consider the timing relationship of epochs for a specific subject. Although such a design have this limitation, it greatly reduces the complexity of the model while improving sleep staging performance.

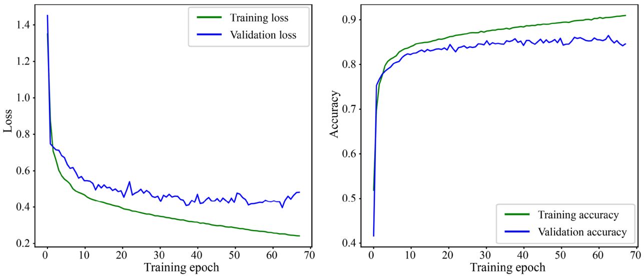

Fig. 4. Accuracy and loss during training of MultiChannelSleepNet for a randomly selected fold (i.e., fold 8) in the SleepEDF-20 dataset.

3) Performance Comparison: We compared the performance of MultiChannelSleepNet and state-of-the-art methods.

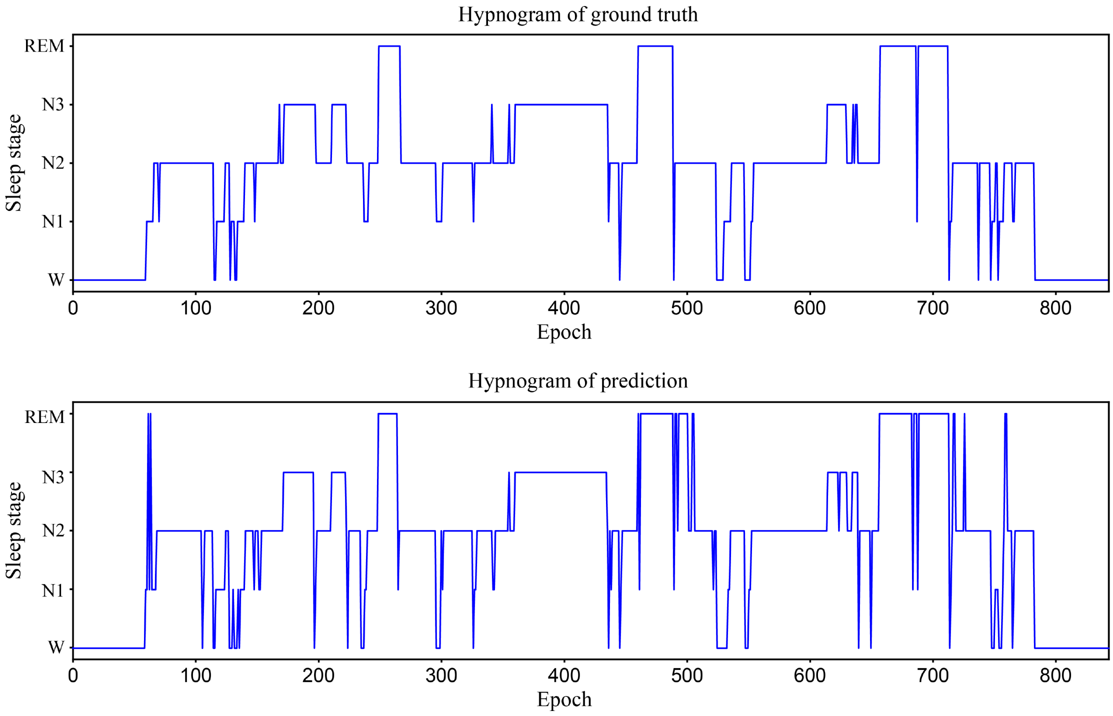

Fig. 5. Hypnograms demonstrate classification accuracy; representative data are from subject SC4061E0 of the SleepEDF-20 dataset. Top row: ground truth of sleep stages; bottom row: stage classification result with MultiChannelSleepNet. Accuracy is 89.6%, and F1-score is 0.84 for this subject.

TABLE V
Performance Comparison With Previous Methods on Three Datasets

Dataset	System	Overall metrics			Per–class F1–score				
		Acc.	κ	MF1	W	N1	N2	N3	REM
SleepEDF–20	**MultiChannelSleepNet**	**87.2**	**0.82**	**81.2**	**92.8**	49.1	**90.0**	89.3	84.8
	XSleepNet* [31]	86.4	0.81	80.9	–	–	–	–	–
	SeqSleepNet* [18], [31]	86.0	0.81	79.7	–	–	–	–	–
	AttnSleep [24]	84.4	0.79	78.1	89.7	42.6	88.8	**90.2**	79.0
	DeepSleepNet [20]	82.0	0.76	76.9	–	–	–	–	–
	SleepEEGNet [21]	84.3	0.79	79.7	89.2	**52.2**	89.8	85.1	**85.0**
SleepEDF–78	**MultiChannelSleepNet**	**85.0**	**0.79**	**79.6**	**94.0**	**53.0**	**86.9**	81.8	82.6
	XSleepNet* [31]	84.0	0.78	78.7	–	–	–	–	–
	SeqSleepNet* [18], [31]	83.8	0.78	78.2	–	–	–	–	–
	AttnSleep [24]	81.3	0.74	75.3	92.0	42.0	85.0	**82.1**	74.1
	SleepTransformer [26]	81.4	0.74	74.3	91.7	40.4	84.3	77.9	77.2
	CNN + Attention [45]	82.8	–	77.8	90.3	47.1	86	82.1	**83.2**
	U–Time [30]	–	–	76.0	–	–	–	–	–
	SleepEEGNet [21]	80.0	0.73	73.6	–	–	–	–	–
	SleepContextNet [27]	82.7	0.76	77.2	92.8	49.0	84.8	80.6	78.9
SHHS	**MultiChannelSleepNet**	**87.5**	**0.82**	79.3	**92.2**	40.5	**88.6**	**88.2**	87.7
	AttnSleep [24]	84.2	0.78	75.3	86.7	33.2	87.1	87.1	82.1
	SleepContextNet [27]	86.4	0.81	**80.5**	89.2	**52.0**	87.6	84.3	**89.3**

Best metric values are marked in boldface. Superscript * denotes that method used multichannel PSG data.

Table V compares MultiChannelSleepNet and methods in previous studies, evaluating overall performance by accuracy, Cohen's kappa (κ) [44], and MF1 [43], and evaluating the performance of each class with F1-score. The results show that MultiChannelSleepNet achieves higher classification performance than other methods, due to its ability to extract single-channel features and fuse multichannel features.

Some of the methods in Table V use only a single PSG channel as input [20], [21], [24], and some use multiple channels, a similar strategy to ours [31]. The methods using multichannel data did not fuse the information from each channel in a specific way but used a mechanism similar to voting to integrate logits before softmax, or simply concatenated channel data before input. This suggests why MultiChannelSleepNet outperforms other methods that also use multichannel data.

We also observe that MultiChannelSleepNet achieves higher classification performance than other seq-to-seq models [18], [27], even though it only uses a single epoch without context

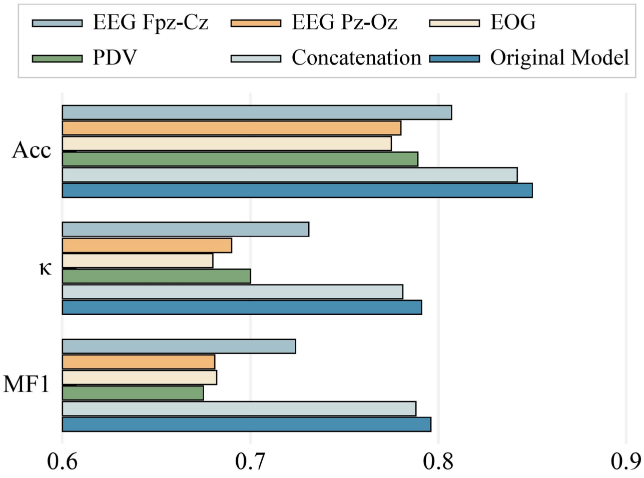

Fig. 6. Performance comparison of the original model and variants with the SleepEDF-78 dataset.

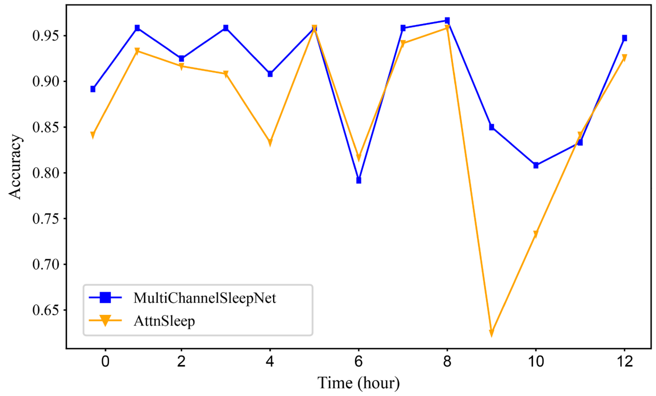

Fig. 7. Accuracies per hour of MultiChannelSleepNet and AttnSleep over a continuous PSG recording (SC4191E0 of the SleepEDF-20 dataset). The average accuracies of MultiChannelSleepNet and AttnSleep are 90.4% and 86.3%, respectively.

as the input. Although the seq-to-seq structure can exploit the sleep transition rules among epochs, due to this limitation, the data cannot be sufficiently shuffled before entering those models, which may affect their generalization ability. Using a single epoch as input, MultiChannelSleepNet performs better by only mining information within a single epoch.

E. Performance Improvement With Multichannel Feature Fusion Block

To analyze the effectiveness of multichannel feature fusion processing, we compared the original model with some variants without the multichannel feature fusion block on SleepEDF-78, where other parameters are the same as the original model. The variants are described as follows:

1) Fpz-Cz EEG: Single-channel feature extraction block for processing Fpz-Cz EEG channel only.
2) Pz-Oz EEG: Single-channel feature extraction block for processing Pz-Oz EEG channel only.
3) EOG: Single-channel feature extraction block for processing EOG channel only.
4) Concatenation: Concatenating three channels before input, without multichannel feature fusion block.
5) Probability distribution voting (PDV): Outputs of 1), 2), and 3) after softmax function are regarded as probabilities, and these are regarded as voting values. They are summed up to obtain the result.
6) Original MultiChannelSleepNet.

Fig. 6 indicates that the original MultiChannelSleepNet is superior to these variants. According to 1), 2), and 3), comparing our original model and those variants using a single channel, the results show that the model with a multichannel feature fusion block is better than any single-channel variant without this module, so we can conclude that fusing multichannel features by an appropriate method is efficient for sleep staging. Another conclusion is that combining multichannel features through simple strategies (i.e., concatenation and PDV) is not significant according to 4) and 5). It is further shown that the multichannel feature fusion block is capable of effectively fusing the features from different channels, which is the key component of MultiChannelSleepNet, to achieve better performance.

F. Performance on Consecutive Data

Tables II–V are the offline testing results of MutiChannelSleepNet on 30-s PSG segments, which shows the overall performance of the algorithm and the performance at each sleep stage. To further study the clinical application potential of our model, we evaluated the performance of the model on continuous polysomnographic time series. Fig. 7 shows the accuracies of MultiChannelSleepNet and AttnSleep on SC4191E0 per hour, and SC4191E0 is a 12-hour continuous PSG recording for a 28-year-old male. Each point in Fig. 7 represents the accuracy of MultiChannelSleepNet or AttnSleep at each hour. The average accuracies of MultiChannelSleepNet and AttnSleep are 90.4% and 86.3%, respectively. Overall, the accuracy of our model on continuous data is basically stable (no less than 79%), and it is higher than or equal to that of the compared algorithm (i.e., AttnSleep) at each hour. The results suggest although our model is trained on discontinuous data after shuffling, it still has the ability to process continuous PSG recording in clinical practice.

G. Sensitivity Analysis

Since our model uses transformer as the backbone, and the transformer encoders are applied to the process of feature extraction and integration, it is important to perform sensitivity analysis on the structure and parameters related to the transformer encoder. Specifically, we analyzed the impact of the number of heads H in transformer encoders and the number of transformer encoders (including N_S, the number of transformer encoders in single-channel feature extraction block of each channel, and N_M, the number of transformer encoder in multichannel feature fusion block) on the performance of our model. In each experiment, the parameters are fixed except for the parameter that need to be tested, and the values of H, N_S, and N_M we tested are commonly used values in models using transformer. As we

TABLE VI
PERFORMANCE OF MULTICHANNELSLEEPNET ON THE SLEEPEDF-78 DATASET WITH DIFFERENT VALUES OF N_S, THE NUMBER OF TRANSFORMER ENCODER IN SINGLE-CHANNEL FEATURE EXTRACTION BLOCK OF EACH CHANNEL, AND N_M, THE NUMBER OF TRANSFORMER ENCODER IN MULTICHANNEL FEATURE FUSION BLOCK

N_S	N_M	Acc.	κ	MF1
4	2	84.0	0.77	78.2
8	2	84.3	0.78	78.6
16	2	84.7	0.78	79.1
4	4	84.1	0.78	78.7
8	4	84.6	0.78	78.9
16	4	**85.0**	**0.79**	**79.6**
4	8	84.0	0.76	78.6
8	8	84.4	0.78	78.9
16	8	84.5	0.78	78.7

TABLE VII
PERFORMANCE OF MULTICHANNELSLEEPNET ON THE SLEEPEDF-78 DATASET WITH DIFFERENT VALUES OF H, THE NUMBER OF HEADS IN TRANSFORMER ENCODER

H	Acc.	κ	MF1
2	84.4	0.78	78.4
4	84.6	0.78	78.8
8	**85.0**	**0.79**	**79.6**
16	84.6	**0.79**	78.9

mentioned in Section II.A, the number of heads H represents that the feature map is divided into H segments in a transformer encoder, hence H must be divisible by the feature dimension $F = 128$. Tables VI and VII show the model performance on the SleepEDF-78 dataset under different parameter settings in terms of accuracy, Cohen's kappa (κ) and MF1. Overall, the model is insensitive to changes in these parameters H, N_S, and N_M. As N_S increases from 4 to 16, the performance of the model improves slightly, that is because it represents an increase in the number of network layers and an increase in complexity. However, the result of N_M value of 4 is slightly better than its value of 2 and 8. This is because our model combines multichannel joint features and individual single-channel features in the multichannel feature fusion block, excessively increasing the number of transformer encoders in this block may have an adverse effect on this combination. The effect of the number of heads H on the model is similar to that of N_M. As H increases, the effect of the model first increases and then decreases. Although more heads expand the ability of the model to find more effective features, it also means that the number of features contained in each attention head decreases. In extreme cases, if $H = 1$, the model can only find features of a certain aspect; if the number of heads is the same as the dimension of the feature map, i.e., $H = F = 128$, then the model will lose the ability to select features and cannot distinguish which features are more important.

IV. DISCUSSION

MultiChannelSleepNet achieved performance improvement by using a multi-head attention mechanism and fusing multichannel information by a specific architecture. Although the methods using single-channel EEG are common due to its potential for application in longitudinal studies and home-based sleep monitoring [14], [20], [25], [26], [27], [46], clinically, sleep stage classification is a process requiring the joint participation of multiple channels, and multimodal signals enhance sleep stage classification with impacts varying by stage. Fig. 6 reveals the importance of feature fusion from different channels. Compared with the single-channel variant of our model using Fpz-Cz EEG as the input, the accuracy of the original model increased by 4.3%, and MF1 increased by 7.2%.

To further quantify the contribution of different channels to different stages, we tested different combinations of channels as the input of MultiChannelSleepNet. EEG Fpz-Cz was selected as the base channel because it performs significantly better than other channels, according to Fig. 6. The overall performance on SleepEDF-78 is shown in Table VIII. Compared to using Fpz-Cz as the input only, the addition of either Pz-Oz or EOG shows a clear improvement in overall accuracy. Pz-Oz is significant to improve the classification performance on the W and N3 stages, according to the per-class F1-scores. With EOG added to the input, the performance of stage W was not further improved significantly. This is because the α rhythm is the primary feature of stage W and is usually most pronounced in the occipital region, while the Pz-Oz channel is closer to the occipital region. For stage N3, slow wave activity is the main criterion for judgment. For the EOG channel, slow eye movements may be present during stage W, but this is not a necessary discriminant feature. The N3 stage usually has no characteristic eye movements, and improper usage of the EOG may adversely affect the interpretation of these two stages. However, the contribution of EOG to improving performance on stage N1 and REM is outstanding. Such result is predictable because rapid eye movements rely on EOG for monitoring. Schemes using Fpz-Cz EEG + EOG has better performance than using Fpz-Cz EEG + Pz-Oz EEG, which also illustrates the importance of EOG in contributing another dimension of information. Besides, these two schemes achieve the same effect on stage N2, which seems to indicate that Pz-Oz and EOG contribute equally to the classification of stage N2. However, when both were added, the F1-score of stage N2 had another increase, which illustrates the potential complementarity between channels. These results present a tight correlation with the physiological features of the sleep stages, which further shows that MultiChannelSleepNet has not simply superimposed different channels when using them but has fused and integrated their internal information, and this process is similar as the manual operation of sleep experts.

Future work could address the limitations of this work, such as by exploring more possibilities on the channel issue. In this work, to maintain consistency, we used EEG and EOG channels in experiments because only these two types of channels are available in SleepEDF-20 and SleepEDF-78. However, some PSG systems include recordings of more channels, such as EMG and respiratory signals, which are also helpful for sleep stage classification to a certain extent [47]. Appropriately introducing these channels may further enhance the performance of the classification. In addition to the PSG system, multichannel EEG recording is commonly used for sleep analysis [48] and usually

TABLE VIII
PERFORMANCE COMPARISON OF MULTICHANNELSLEEPNET USING DIFFERENT INPUT CHANNELS ON THE SLEEPEDF-78 DATASET

Input channels	Overall metrics			Per-class F1-score				
	Acc.	κ	MF1	W	N1	N2	N3	REM
Fpz-Cz	80.7	0.73	72.4	91.6	37.2	84.7	78.7	70.6
Fpz-Cz + Pz-Oz	83.0	0.77	76.7	93.5	45.1	86.1	81.7	75.7
Fpz-Cz + EOG	83.8	0.78	77.4	92.9	48.5	86.2	79.1	80.2
Fpz-Cz + Pz-Oz + EOG	**85.0**	**0.79**	**79.6**	**94.0**	**53.0**	**86.9**	**81.8**	**82.6**

includes 32 or more EEG channels. It is also worth exploring the selection of appropriate channels from this kind of system and their use for the sleep staging task. In addition, since our model is designed as one-to-one structure, i.e., using only a single epoch as input without considering the context, there are some abnormal transitions of specific samples in the model output, which remain to be improved in future work.

V. CONCLUSION

We proposed MultiChannelSleepNet, which is based on the transformer encoder, for automatic sleep stage classification with multichannel PSG including EEG and EOG. The model consists of single-channel feature extraction, multichannel feature fusion, and classification blocks. In the single-channel feature extraction block, transformer encoders are used to extract features from time-frequency images of each channel independently. During multichannel feature fusion, the feature maps extracted from each channel are fused while further extracting joint features. Transformer encoders and a residual connection are applied in this module. Compared with previous methods using either single-channel or multichannel strategies, experimental results show that MultiChannelSleepNet performs better on various evaluation metrics. To demonstrate the ability of the model to integrate multichannel information, we compared the original MultiChannelSleepNet model with variants using single-channel data and simple fusion strategies. The results showed that the original MultiChannelSleepNet model significantly outperformed these variants. Besides, we evaluated the performance of our model on continuous polysomnographic time series, which demonstrates the potential of MultiChannelSleepNet for clinical applications. Finally, we conducted a sensitivity analysis to analyzed the impact of the number of heads in transformer encoders and the number of transformer encoders in single-channel feature extraction block and multichannel feature fusion block on the performance of our model. The results showed that our model remains stable under different parameter settings.

REFERENCES

[1] A. N. Goldstein and M. P. Walker, “The role of sleep in emotional brain function,” *Annu. Rev. Clin. Psychol.*, vol. 10, pp. 679–708, 2014.

[2] V. K. Chattu, M. D. Manzar, S. Kumary, D. Burman, D. W. Spence, and S. R. Pandi-Perumal, “The global problem of insufficient sleep and its serious public health implications,” *Healthcare*, vol. 7, no. 1, Dec. 2018, Art. no. 1.

[3] E. A. Wolpert, “A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects,” *Electroencephalography Clin. Neurophysiol.*, vol. 26, no. 2, pp. 644–644, 1969.

[4] C. Iber et al., “The AASM manual for the scoring of sleep and associated events: Rules,” *Terminol. Tech. Specification*, 1st. ed., Westchester, IL, USA: Amer. Acad. Sleep Med., 2007.

[5] H. Phan and K. Mikkelsen, “Automatic sleep staging of EEG signals: Recent development, challenges, and future directions,” *Physiol. Meas.*, vol. 43, no. 4, Apr. 2022, Art. no. 04TR01.

[6] S. I. Dimitriadis, C. Salis, and D. Linden, “A novel, fast and efficient single-sensor automatic sleep-stage classification based on complementary cross-frequency coupling estimates,” *Clin. Neuriophysiol.*, vol. 129, no. 4, pp. 815–828, Apr. 2018.

[7] S. Gunes, K. Polat, and S. Yosunkaya, “Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting,” *Expert Syst. Appl.*, vol. 37, no. 12, pp. 7922–7928, Dec. 2010.

[8] S. Abdulla, M. Diykh, R. L. Laft, K. Saleh, and R. C. Deo, “Sleep EEG signal analysis based on correlation graph similarity coupled with an ensemble extreme machine learning algorithm,” *Expert Syst. Appl.*, vol. 138, Dec. 2019, Art. no. 112790.

[9] B. Koley and D. Dey, “An ensemble system for automatic sleep stage classification using single channel EEG signal,” *Comput. Biol. Med.*, vol. 42, no. 12, pp. 1186–1195, Dec. 2012.

[10] E. Alickovic and A. Subasi, “Ensemble SVM method for automatic sleep stage classification,” *IEEE Trans. Instrum. Meas.*, vol. 67, no. 6, pp. 1258–1265, Jun. 2018.

[11] M. Sharma, D. Goyal, A. Pv, and U. R. Acharya, “An accurate sleep stages classification system using a new class of optimally time-frequency localized three-band wavelet filter bank,” *Comput. Bio. Med.*, vol. 98, pp. 58–75, 2018.

[12] X. Li, L. Cui, S. Tao, J. Chen, X. Zhang, and G. Q. Zhang, “HyCLASSS: A hybrid classifier for automatic sleep stage scoring,” *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 375–385, Mar. 2018.

[13] P. Memar and F. Faradji, “A novel multi-class EEG-based sleep stage classification system,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 1, pp. 84–95, Jan. 2018.

[14] 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.

[15] J. M. Zhang and Y. Wu, “Competition convolutional neural network for sleep stage classification,” *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102318.

[16] P. Jadhav and S. Mukhopadhyay, “Automated sleep stage scoring using time-frequency spectra convolution neural network,” *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art no. 2510309.

[17] M. Sokolovsky, F. Guerrero, S. Paisarnsrisomsuk, C. Ruiz, and S. A. Alvarez, “Deep learning for automated feature discovery and classification of sleep stages,” *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 17, no. 6, pp. 1835–1845, Nov./Dec. 2020.

[18] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. D. 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.

[19] C. Sun, C. Chen, W. Li, J. Fan, and W. Chen, “A hierarchical neural network for sleep stage classification based on comprehensive feature learning and multi-flow sequence learning,” *IEEE J. Biomed. Health Inform.*, vol. 24, no. 5, pp. 1351–1366, May 2020.

[20] 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.

[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] 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.

[23] A. Vaswani et al., “Attention is all you need,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, California, USA, 2017, pp. 6000–6010 .

[24] 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.

[25] 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.

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

[27] C. Zhao, J. Li, and Y. Guo, “Sleepcontextnet: A temporal context network for automatic sleep staging based single-channel EEG,” *Comput. Methods Programs Biomed.*, vol. 220, Jun. 2022, Art. no. 106806.

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

[29] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. D. 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.

[30] 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. 4417–4428.

[31] H. Phan, O. Y. Chen, M. C. Tran, P. Koch, A. Mertins, and M. D. Vos, “XSleepNet: Multi-view sequential model for automatic sleep staging,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5903–5915, Sep. 2022.

[32] Z. Jia, X. Cai, and Z. Jiao, “Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging,” *IEEE Sensors J.*, vol. 22, no. 4, pp. 3464–3471, Feb. 2022.

[33] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–10.

[34] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[35] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jun. 2014.

[37] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” 2016. [Online]. Available: https://arxiv.org/abs/1607.06450

[38] 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.

[39] 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. E 215–E220, Jun. 2000.

[40] F. Li et al., “End-to-end sleep staging using convolutional neural network in raw single-channel EEG,” *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102203.

[41] S. F. Quan et al., “The sleep heart health study: Design, rationale, and methods,” *Sleep*, vol. 20, no. 12, pp. 1077–85, Dec. 1997.

[42] G. Q. Zhang et al., “The national sleep research resource: Towards a sleep data commons,” *J. Am Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1351–1358, Oct. 2018.

[43] 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*, 1996, pp. 298–306.

[44] J. Cohen, “A coefficient of agreement for nominal scales,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[45] T. Zhu, W. Luo, and F. Yu, “Convolution-and attention-based neural network for automated sleep stage classification,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 11, pp. 4152, Jun. 2020.

[46] N. Michielli, U. R. Acharya, and F. Molinari, “Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals,” *Comput. Biol. Med.*, vol. 106, pp. 71–81, Mar. 2019.

[47] A.-M. Tautan, A. C. Rossi, R. d. Francisco, and B. Ionescu, “Automatic sleep stage detection: A study on the influence of various PSG input signals,” *IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.*, vol. 2020, 2020, pp. 5330–5334.

[48] A. Piryatinska, W. A. Woyczynski, M. S. Scher, and K. A. Loparo, “Optimal channel selection for analysis of EEG-sleep patterns of neonates,” *Comput. Methods Programs Biomed.*, vol. 106, no. 1, pp. 14–26, Apr. 2012.