Multi-Modal Sleep Stage Classification With Two-Stream Encoder–Decoder

Zhao Zhang, *Graduate Student Member, IEEE*, Bor-Shyh Lin, *Senior Member, IEEE*, Chih-Wei Peng, and Bor-Shing Lin, *Senior Member, IEEE*

***Abstract*— Sleep staging serves as a fundamental assessment for sleep quality measurement and sleep disorder diagnosis. Although current deep learning approaches have successfully integrated multimodal sleep signals, enhancing the accuracy of automatic sleep staging, certain challenges remain, as follows: 1) optimizing the utilization of multi-modal information complementarity, 2) effectively extracting both long- and short-range temporal features of sleep information, and 3) addressing the class imbalance problem in sleep data. To address these challenges, this paper proposes a two-stream encode-decoder network, named TSEDSleepNet, which is inspired by the depth sensitive attention and automatic multi-modal fusion (DSA2F) framework. In TSEDSleepNet, a two-stream encoder is used to extract the multiscale features of electrooculogram (EOG) and electroencephalogram (EEG) signals. And a self-attention mechanism is utilized to fuse the multiscale features, generating multi-modal saliency features. Subsequently, the coarser-scale construction module (CSCM) is adopted to extract and construct multi-resolution features from the multiscale features and the salient features. Thereafter, a Transformer module is applied to capture both long- and short-range temporal features from the multi-resolution features. Finally, the long- and short-range temporal features are restored with low-layer details and mapped to the predicted classification results. Additionally, the Lovász loss function is applied to alleviate the class imbalance problem in sleep datasets. Our proposed method was tested on the Sleep-EDF-39 and Sleep-EDF-153 datasets, and it achieved classification accuracies of 88.9% and 85.2% and Macro-F1 scores of 84.8% and 79.7%, respectively, thus outperforming conventional traditional baseline models. These results highlight the efficacy of the proposed method in fusing multi-modal information. This method has potential for application as an adjunct tool for diagnosing sleep disorders.**

***Index Terms*— Convolutional block attention module, deep learning, multimodal, multiscale extraction, sleep-stage classification.**

Manuscript received 30 December 2023; revised 13 April 2024 and 24 April 2024; accepted 24 April 2024. Date of current version 7 June 2024. This work was supported in part by the Natural Science Foundation Project of Nanping City under Grant N2023J002; in part by the Ministry of Science and Technology in Taiwan under Grant MOST 110-2221-E-A49-096-MY3; in part by the National Science and Technology Council in Taiwan under Grant NSTC 112-2221-E-305-001-MY3; in part by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and Ministry of Education (MOE), Taiwan; in part by the University System of Taipei Joint Research Program under Grant USTP-NTPU-NTOU-112-01; and in part by the Faculty Group Research Funding Sponsorship by National Taipei University under Grant 2024-NTPU-ORD-01. *(Corresponding author: Bor-Shing Lin.)*

Zhao Zhang is with the College of Mechanical and Electrical Engineering, Wuyi University, Wuyishan, Fujian 354300, China, and also with the Department of Computer Science and Information Engineering and the College of Electrical Engineering and Computer Science, National Taipei University, New Taipei City 237303, Taiwan (e-mail: 18259956201@163.com).

Bor-Shyh Lin is with the Institute of Imaging and Biomedical Photonics, National Yang Ming Chiao Tung University, Tainan 71150, Taiwan (e-mail: borshyhlin@nycu.edu.tw).

Chih-Wei Peng is with the School of Biomedical Engineering, College of Biomedical Engineering, and the School of Gerontology and Long-Term Care, College of Nursing, Taipei Medical University, Taipei 11031, Taiwan (e-mail: cwpeng@tmu.edu.tw).

Bor-Shing Lin is with the Department of Computer Science and Information Engineering, National Taipei University, New Taipei City 237303, Taiwan (e-mail: bslin@mail.ntpu.edu.tw).

Digital Object Identifier 10.1109/TNSRE.2024.3394738

I. Introduction

SLEEP plays a crucial role in maintaining health, and the quality of sleep is negatively influenced by the fast pace of life and elevated stress levels prevalent in modern society. These factors can lead to physiological and psychological dysfunction and, in some cases, contribute to the onset of various health conditions [1]. Consequently, research on sleep has become increasingly relevant and has practical significance.

In clinical settings, polysomnography (PSG) is the most commonly used standard physiological technique for assessing sleep-related disorders such as sleep disorders, snoring, epilepsy, and sleep apnea. This procedure is typically conducted in a hospital ward where patients are admitted. Sleep specialists fit a variety of recording devices on patients, including electroencephalogram (EEG), electromyography (EMG), electrooculogram (EOG), and electrocardiogram (ECG) devices. These devices record sleep data in 30-s epochs continuously for 6–8 h. For instance, in a 6-h examination, 720 sleep epochs are recorded. Subsequently, sleep specialists evaluate sleep quality and diagnose potential disorders in patients. Two primary standards are currently used for defining sleep stages. The first, proposed by Rechtschaffen and Kales (R&K) [2] in 1968, categorizes nonrapid eye movement (NREM) into four stages (S1, S2, S3, S3, and S4) based on changes in EEG and EOG signals. The second standard, established by the American Academy of Sleep Medicine (AASM) [3] in 2007, consolidates S3 and S4 of NREM based on the R&K sleep staging. Sleep specialists employ these standards to manually classify the sleep stages, which is both time-consuming and susceptible to subjective interpretation [4]. Therefore, automatic sleep staging is a more effective alternative to manual methods, and it has more clinical value.

Early studies have proposed many machine learning methods based on EEG signals for the classification of sleep stages. Such methods typically involve the extraction of features from sleep stages by using algorithms in the time domain, frequency domain, or time–frequency domain. Examples include sparse superposition coding following time domain analysis [5], the iterative filtering method [6], and the support vector machine method [7]. Additionally, approaches incorporating multiple PSG signals have been explored [8], [9], [10]. In such cases, features from each signal in a sleep stage are combined into a feature vector, which is then used by the classifier. However, these methods have limitations. Manually extracted features exhibit strong correlations with training datasets and are inherently subjective, which hinders their generalizability to a broader population. Moreover, the advancement of conventional machine learning methods for automatic sleep staging has encountered challenges due to the difficulties involved in the manual extraction of optimal features and the inability to adjust the extraction parameters during the classification process.

Deep learning methods can extract the most representative features from large datasets without extensive prior knowledge [11]. These methods have been widely applied in the fields of CV and NLP. From the perspective of data sources, automatic sleep staging can be considered as an intersection of CV and NLP, prompting many researchers to leverage transfer learning from these two fields, yielding notable results. Supratak et al. [12] proposed DeepSleepNet, a classic deep learning sleep-staging method. However, the model has certain limitations. First, it primarily uses a single PSG signal as the original data input, lacking the ability to effectively integrate and utilize multi-modal information. Second, to extract sequence information, DeepSleepNet employs a Long Short-Term Memory (LSTM) network, which may encounter problems such as gradient disappearance or explosion when analyzing lengthy sequences [13]. Therefore, introducing a model that does not use recurrent neural networks is imperative to reducing computational costs. Finally, the model adopts one-to-one input–output mode. Although this approach is straightforward and transparent, it neglects the transition rules present during different sleep periods [14]. Phan et al. sought to overcome this limitation in DeepSleepNet by proposing SeqSleepNet [14], a many-to-many trained model that accepts a sequence of multiple sleep stages as input. However, this model still focuses on a single modality and uses an LSTM network. Although the use of LSTM is intuitive for PSG, which is a sequence of signals, recent advancements in the field of image semantic segmentation have demonstrated the efficacy of fully convolutional networks (FCNs) for sequence learning. Inspired by U-Net [15] in image segmentation, Perslev et al. proposed the U-Time model [16], an FCN network. In this model, all data are resampled, extreme data are eliminated, and multiple sleep stages are combined as one input into the U-shaped codec structure. The authors further optimized the U-sleep model [17] to enhance its performance, but they still did not fully leverage sleep transition rules. To address this problem, ensemble systems have been proposed [18], [19], which combine CNN and RNN to simultaneously extract features in both spatial and temporal domains, generating a more accurate sleep-staging model. Considering that EEG electrodes are distributed in a non-Euclidean space and that the spatial correlation between electrodes is neglected, the limitations of CNN and RNN become apparent. GCNs [20] have demonstrated high capabilities in modeling the topological relationships among EEG electrodes. Li et al. [21] proposed a combination of dynamic and static spatiotemporal graph convolutional networks incorporating multi-temporal attention blocks. This approach effectively captures long-term dependencies between different EEG signals, producing superior performance for sleep staging. Despite their success, methods based on single-channel EEG signals often exhibit limited performance because relying on a single fixed physiological signal may be suboptimal for distinguishing specific sleep stages.

To exploit the complementary potential of PSG signals, researchers have explored the use of multi-modal signals to enhance the performance of sleep-staging models. Multimodal learning is a method of learning using data from a variety of different sensors or interactive modes. The key to multi-modal learning lies in integrating and analyzing data from different sources to gain a more comprehensive and in-depth insight than a single data source. Dong et al. [22] applied a combination of DNN and RNN to extract salient features from EEG and EOG signals. The SeqSleepNet [14] approach solely relied on hierarchical RNN and achieved an overall classification accuracy of 87.1% based on multichannel signals of MASS dataset. These methods primarily focus on extracting features from different PSG signals and combining them through concatenation. However, this approach may not be sufficient to model the complex relationship between multimodal signals. Recent studies have fully integrated multimodal features and demonstrated the different contributions of each modality in identifying specific sleep stages. Examples include the SalientSleepNet [23] and SleepPrintNet [24]. Furthermore, Jia et al. [25] designed a squeeze-and-excitation network to model the heterogeneity of different modalities. MMASleepNet [26] introduced an effective feature fusion module to capture relationships between different modalities. MaskSleepNet [27] combined CNN with the attention mechanism to effectively capture feature information from PSG signals, achieving a sleep-staging accuracy of 85.0% on the Sleep-EDF-153 dataset. However, these models often overlook the complementarity of multi-modal information and multiscale dependence simultaneously and tend to neglect the class imbalance problem in sleep data. These limitations affect the overall performance of sleep models.

To more effectively extract multi-modal features of EEG and EOG signals and to enhance the accuracy of sleep staging, a two-stream encode-decoder network, named TSEDSleepNet, was proposed in this study, which has a two-stream encoder–decoder architecture. The model structure is inspired by the depth sensitive attention and automatic multi-modal fusion (DSA2F) framework [28], which is used on the task of computer vision, as shown in Fig. 1. However, in [28], most

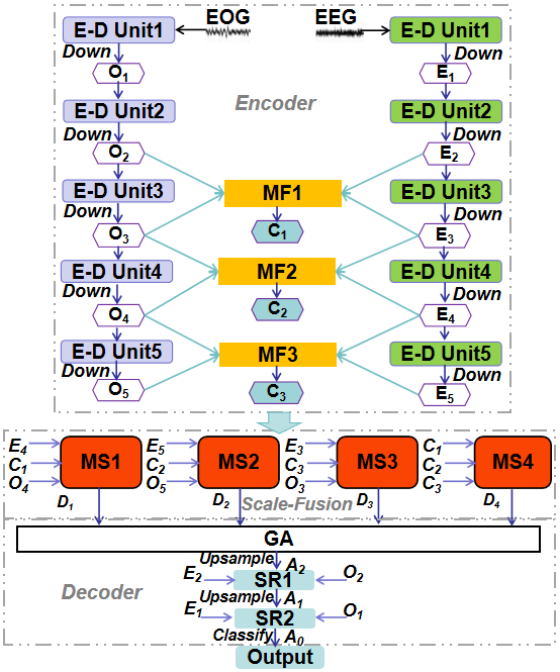

Fig. 1. Structure of the sleep-staging model based on multi-modal information. E-D Unit is the codec module, and MF is the multimodal fusion module. MS is the multiscale fusion module, and GA is the global aggregation module. SR is spatial information restoration module, and the terms in italic letters denote different operations on the feature tensor.

of the sub-modules of the DSA2F model are not detailed, thus we have redesigned all the sub-modules in order to apply to the time-series data of sleep. The architecture of all sub-modules in our proposed model are different from the corresponding sub-modules in DSA2F model. First, in the module of two-stream encoder the depth sensitive attention module in DSA2F model is replaced with U^2 module in our model, which is used to extract multi-scale features from both EEG and EOG channels. A self-attention mechanism is then employed used to fuse these features for extracting more salient features. Second, the coarser-scale construction module (CSCM) [29] is used to combine the multi-scale features and salient features to construct multi-resolution features. In this part, a residual connection is applied to preserve the original information of the multi-scale features and avoid gradient-vanishing. Subsequently, a global aggregation module, utilizing the Transformer architecture, further fuses the multi-resolution features to learn long- and short-range temporal features. Finally, the long- and short-range temporal features are restored with the low-layer detail features and mapped to the predicted classification results. The contributions of this paper are as follows:

1) We propose a novel module for modal fusion at different scales which can more effectively extract the complementary features of EEG and EOG information to improve the classification accuracy of sleep stage. In the module a two-stream encoder based on U^2 structure is used to extract multi-scale features from EEG and EOG channels. A self-attention mechanism is then utilized to select more salient features from the multi-scale features. However, previous models such as MMASleepNet [26] and MaskSleepNet [27] only fuse the multi-modal features of single scale from last layer of the encoder. Thus, our model is able to capture more complementary features of EEG and EOG information from multi-scale features.
2) We propose a combination of CSCM and Transformer modules to extract long and short temporal features from multi-scale features. The CSCM based on multiple convolution layers is used to combine the multi-scale features and salient features from encoder to construct multi-resolution features. Transformer module is then employed to further capture the temporal correlations of different ranges from the multi-resolution features. It is easier to capture long-range dependence by the coarser scales features of multi-resolution features, improving the performance of the model.
3) Experiments were conducted on two public data sets to verify the effectiveness of our model. The results show that our model outperforms all baseline models in automated sleep staging.

II. Materials and Methods

A. Dataset and Preprocessing

This study used two publicly available datasets, namely Sleep-EDF-39 and Sleep-EDF-153 [30]. The two datasets contain a total of 197 overnight PSG sleep recordings, including EEG, EOG, chin EMG, and event markers. Due to some records being corrupted, only 39 and 153 records remained in the two datasets, respectively. For this study, the Sleep Cassette portion of Sleep-EDF-153 and Sleep-EDF-39 were used. The Fpz-Cz EEG and ROC-LOC EOG (horizontal) channels were selected for sleep staging, with a sampling rate of 100 Hz for each sleep epoch. Trained technicians manually classified the corresponding PSG signals (sleep patterns) according to the R&K standard, and each 30-s sleep epoch was categorized into labels {W, N1, N2, N3, N4, REM, MOVEMENT, UNKNOWN}.

Prior to the experiment, data preprocessing was performed. Sampling records labeled "motion" and "unknown" in the original datasets were removed. Additionally, following the convention established by prior studies [31], 30 min of sleep stages (60 in total) in the records were removed when the subject was out of bed, as these sleep stages were considered to be the awake periods. Subsequently, in line with the guidelines provided by the AASM manual [3], N3 and N4 were uniformly labeled as N3, and the classification label $y \in$ {W, N1, N2, N3, REM} was obtained.

The Sleep-EDF datasets exhibit class imbalance, a characteristic inherent to human physiology. The proportional distribution of each sleep stage is detailed in Table I.

B. Model Overview

The structure of the automatic sleep-staging model, which utilizes multi-modal information and is proposed in this paper, is illustrated in Fig. 1. In the training data, each sleep epoch is defined as $x \in R^{n \times c}$, where n is the number of sampling points within a sleep epoch, which is typically set to 3000 (i.e.,

TABLE I
NUMBER AND PROPORTION OF SLEEP STAGES FOR THE DATASETS USED IN THE EXPERIMENT

Datasets	W (%)	N1 (%)	N2 (%)	N3 (%)	REM (%)	Total Samples
Sleep-EDF-39	19.6	6.6	42.1	13.5	18.2	42308
Sleep-EDF-153	33.7	11.0	35.4	6.7	13.2	195479

a sampling rate of 100 Hz), and c is the number of channels of the sleep data (specifically EEG and EOG in this study). The input sleep sequence is localized as $S = \{x_1, x_2, \ldots, x_L\}$, where x_i is a sleep epoch, $i \in \{1, 2, \ldots, L\}$, where L is the length of the sleep sequence (i.e., S contains L sleep epochs).

In this paper, we formulate the sleep-staging problem as follows: learning a mapping F, where $S \xrightarrow{F} \hat{Y}$, and S is the input sleep sequence and $\hat{Y}$ is the output prediction sequence, with $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_L\}$, where $\hat{y}_i$ is the predicted classification result. According to the AASM standard, $\hat{y}_i \in \{0, 1, 2, 3, 4\}$, corresponding to the five sleep stages W, N1, N2, N3, and REM.

This paper draws inspiration from the depth sensitive attention and automatic multimodal fusion (DSA2F) framework used for image saliency detection. The overall framework of the proposed model is depicted in Fig. 1, where EEG and EOG represent the input data from two channels. Blocks of varying colors indicate modules with distinct functions. The TSEDSleepNet is composed of Encoder, Scale-Fusion, Decoder, and Output modules.

C. Model Encoder

1) Two-Stream Encoder: The primary task of the encoder is to convert raw data into a new representation with a greater number of and enhanced features suitable for mapping to the desired output compared with the original data representation. Achieving robust feature capability for the encoder often necessitates training on a substantial amount of data for fine-tuning its parameters. Therefore, researchers usually use pretrained models, such as VGG [32], pretrained Transformer [33], and BERT [34] from Hugging Face to perform this task in the field of CV and NLP. However, given that sleep staging is not as prevalent in deep learning applications, the use of pretrained networks, especially those in the CV field [35], can be challenging. Therefore, in this study, the use of an effective encoder module trained from scratch is required. The U^2-Net network [36] has demonstrated effectiveness in salient object detection even without a pretrained backbone. The SalientSleepNet model, adapted from the U^2-Net network, has also proven effective in the automated sleep-staging task.

The single stream of our encoder adopts the U^2 structure from the U^2-Net network, as illustrated in the upper part of Fig. 2. This structure employs five E-D units connected in series to extract multiscale features from the data. The output E_i of the E-D unit represents the feature of the input data, where E $\in$ {EEG channel, EOG channel} and $i \in \{1, 2, 3, 4, 5\}$. The E-D unit comprises a small and shallow codec, as depicted in Fig. 2. The encoder generates the output e_i,

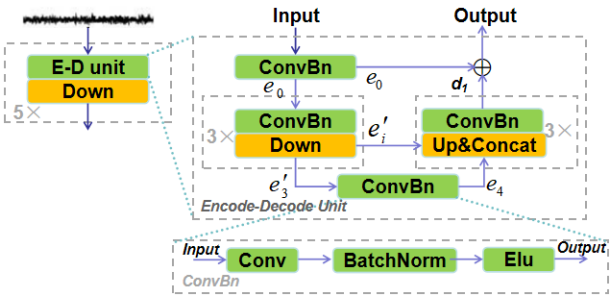

Fig. 2. Basic structure of the encoder (single stream). Down denotes downsampling pooling, Up&Concat denotes the concatenation operation after upsampling, ⊕ refers to tensor element-level addition, Conv is the convolutional layer, BatchNorm is the batch regularization operation, and Elu refers to the Elu activation function.

where $i \in \{0, 1, 2, 3, 4\}$, following the ConvBn operation:

$$e_i' = DownSample\ (e_i), i \in \{1, 2, 3\}. \tag{1}$$

The decoder produces the output d_i by using the operation:

$$d_i = ConvBn\ (UpSample\ (Concat(e, e_i'))), \tag{2}$$

where $i \in \{1, 2, 3\}$, $e \in \{e_5, d_2, d_3\}$, ConvBn module is the combination of convolution calculation and batch normalization, UpSample denotes upsample operation. Concat denotes the concatenation operation.

To preserve the original information of each channel and mitigate gradient vanishing, residual connections are employed [37]. To address the potential loss of information during up and down sampling, a residual join is performed before the module output is obtained:

$$Output = d_1 \oplus e_0 \tag{3}$$

The ConvBn module serves as the fundamental unit for convolution calculations. Its structure is depicted in Fig. 2, and the operation is expressed as follows:

$$Output = Elu\ (BatchNorm\ (Conv(Input))) \tag{4}$$

where Elu refers to the Elu activation function, BatchNorm is the batch regularization operation, Conv denotes convolution calculation.

2) Multimodal Fusion Module: The SalientSleepNet model introduces a method to fuse multimodal information, expressing using the following formula:

$$Output = Attention\ (e + o + (e \odot o)) \tag{5}$$

where Attention represents the self-attention operation, e and o represent the EEG and EOG encoder outputs of SalientSleepNet, respectively, and $\odot$ represents tensor element-wise multiplication. Although the fusion method is intuitive and easy to implement, the fusion is too simple, which may lead to information loss.

The proposed multimodal fusion (MF) module facilitates the fusion of multimodal features from multilayers, as depicted in Fig. 1. This approach is different from SalientSleepNet, which exclusively performs multimodal fusion on the output tensor of the last layer in the encoder.

The encoder incorporates three MF modules, and the internal structure of the MF module is detailed in Fig. 3. An MF

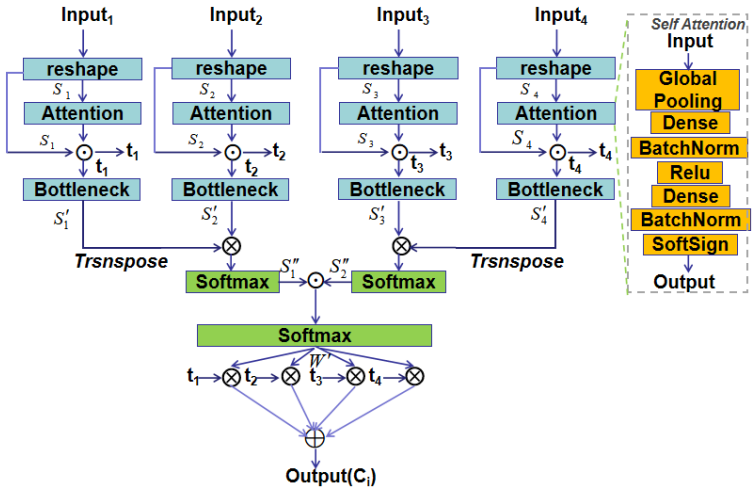

Fig. 3. Structure of a multimodal fusion (MF) module. Global Pooling in Self-Attention compresses tensors to one dimension, and the output weight is multiplied by tensor broadcast. ⊗ denotes the matrix multiplication, ⊙ denotes the element-wise multiplication of tensors, and ⊕ denotes the element-wise addition of tensors.

module processes multimodal information from two scales within four input streams. First, the tensors of input streams are reshaped to the same shape:

$$S_i = reshape\ (Input_i), \tag{6}$$

$i \in \{1, 2, 3, 4\}$.

Subsequently, the self-attention mechanism (depicted in the top right corner of Fig. 3) is employed to assign weights to the input stream itself to obtain more useful features:

$$t_i = S_i \odot Broadcast\ (Attention(S_i)) \tag{7}$$

where ⊙ represents the matrix element-wise multiplication, Broadcast indicates the broadcast operation which is used to expand the dimension of the feature tensor extracted by Attention module, $i \in \{1, 2, 3, 4\}$. Global pooling in self-attention refers to the global average pooling operation, which compresses the input tensor to one dimension for connecting to the fully connected network for weight extraction.

$$S_i' = Bottleneck\ (t_i) \tag{8}$$

where Bottleneck denotes the bottleneck operation, $i \in \{1, 2, 3, 4\}$. Bottleneck is composed of convolution calculation model with convolution kernel of 1×1, which is used to reduce the calculation parameters to conserve computing resources. The final Softsign layer is used to replace the commonly used Sigmoid activation function. Subsequently, tensors from the same scale perform as follows:

$$S_j'' = Soft\max\ (Transpose(S_{2j-1}') \otimes S_{2j}' \tag{9}$$

where Transpose is the transpose operation, Softmax is the activation function, ⊗ is the matrix multiplication, and $j \in \{1, 2\}$. Thereafter, the S tensors are concatenated as follows:

$$W' = Soft\max(S_1'' \odot S_2'') \tag{10}$$

where W' is the final weight tensor, which is multiplied by and added to t_i obtain the final fused output feature:

$$C_j = \sum_{i=1}^{4} (t_i \otimes W'),\ j \in \{1, 2, 3\} \tag{11}$$

3) Multiscale Module: A "macroscale" concept is introduced, which represents a layer encompassing multiple scales. The paper proposes a multiscale (MS) module derived from CSCM, where in the "macroscale" performs a single-scale fusion, which is crucial for performing global multiscale fusion in the decoder. The structure of MS is depicted in Fig. 4, and it receives inputs from three streams: $A \in \{E_4, E_5, E_3, C_1\}$, $B \in \{C_1, C_2, C_3, C_2\}$, and $C \in \{O_4, O_5, O_3, C_3\}$, where E, O, and C denote the output of the EEG encoder, output of the EOG encoder, and output of the MF module, respectively. The MS module stacks these three inputs together and passes them through a dense fully connected layer for reducing the number of channels, conserving computational resources. Subsequently, the tensor is processed through multiple convolutional layers with the same kernel size for iterative operations:

$$L_i = ConvBn\ (L_{i-1}), \quad i \in \{1, 2, 3\} \tag{12}$$

Here, each tensor harbors different scales after undergoing multiple convolution processes. A lower value of i indicates that the tensor encompasses more features from the bottom layer and more detailed information. Conversely, a higher i value implies that the tensor comprises more high-level features, but it also indicates the loss of detail information.

Next, we concatenate S_i to construct multi-resolution features (to ensure that the output tensor has features at each scale) and then use a fully connected layer to boost the number of tensor channels and regularize the output tensor:

$$D_j = LayerNorm(Dense(Concat(S_i, S_{i-1}, \ldots, S_0))) \tag{13}$$

where Concat denotes the concatenation operation, Dense denotes the fully connected layer, and LayerNorm denotes the layer regularization operation, $j \in \{1, 2, 3, 4\}$.

D. Model Decoder

1) Global Aggregation Module: Following the multiscale fusion performed by the MS module at each "macro scale,"

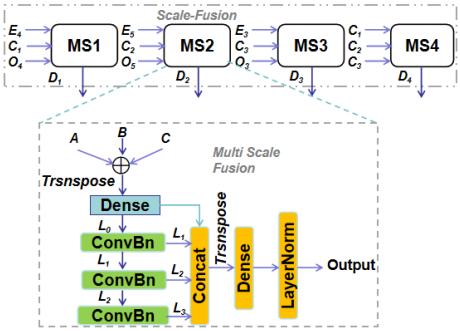

Fig. 4. Multiscale module, where A, B, and C represent the input data, and their specific types are indicated by the arrows on the left side of the figure. E represents the output of the EEG encoder stream, O represents the output of the EOG encoder stream, and C represents the output of the MF module. D denotes the output of the MS module, and the italic letters denote the corresponding operations on the tensor. Dense denotes a fully connected layer, Concat denotes the concatenation operation, and LayerNorm denotes hierarchical regularization.

we introduce a global aggregation (GA) module to further integrate all the multiscale information. The structure of GA is depicted in Fig. 5. Drawing inspiration from the Pyraformer framework [29], we employ a Transformer module to extract long- and short-range temporal features from the multi-resolution features by using its small codecs. The D_i tensor, after being reshaped to the same shape, is connected to a Transformer module, which learns long- and short-range temporal features at each scale through its small internal codec structure.

Drawing inspiration from SalientSleepNet, the output of the Transformer module is subsequently fed into a multiscale extraction module named MSE, whose internal structure is detailed in the relevant part of Fig. 5. The output of the Transformer module is defined as D_i', $i \in \{1, 2, 3, 4\}$, and the calculations in MSE are given by:

$$D_i'' = BN(Bottleneck(Concat(BN(dConv_1(D_i)), \ldots, BN(dConv_N(D_i))))) \tag{14}$$

where BN is the BatchNorm operation, dConvj represents the Atrous Convolution layer with subscript $j \in \{1, 2, \ldots, N\}$, and j represents the dilation rate. Atrous Convolution is another method to improve the model's receptive field. It enables the conservation of computational resources without the need for increasing the number of model parameters, as opposed to increasing the convolution kernel size and weights. However, it poses the disadvantage of potentially losing the information in the "hole" [38].

In this model, Atrous Convolution and Transformer operate complementarily. Transformer compensates for the loss of information in Atrous Convolution, and our model uses multiple expansion rates to compute the same tensor, addressing the potential loss of information. Finally, the GA module splices the output of each MSE submodule and employs a bottleneck layer to reduce the dimensionality of the output to obtain the module's final output results:

$$A_2 = Concat(D_1'', D_2'', D_3'', D_4'') \tag{15}$$

where Concat denotes the concatenation operation.

2) Spatial Information Restoration Module: During encoding, multimodal fusion, and multiscale fusion, operations such as downsampling, pooling, and Atrous Convolution are often performed to obtain a larger receptive field for learning deeper or more abstract features.

However, these operations may result in the loss of detailed information from the sleep sequence. Therefore, we introduce a spatial information restoration (SR) module to compensate for the loss of details from previous operations, the structure of which is presented in Fig. 6. As depicted in Fig. 6, the model consists of two SR modules, which use the output tensor E_i and O_i of the first and second layer of the encoder, respectively, to compensate for the loss of the decoded tensor A_i.

Each input flow is calculated through the convolution layer with different kernel size after they are reshaped to the same shape:

$$E_i^k = Conv^k(E_i) \tag{16}$$
$$O_i^k = Conv^k(O_i) \tag{17}$$
$$A_i^k = Conv^k(A_i) \tag{18}$$

where the superscript k represents the kernel size, and $i \in \{1,2\}$. Subsequently, an element-by-element addition operation is performed on the output tensors of the convolutional layers with the same kernel size:

$$Z_i^1 = E_i^{k_1} \oplus O_i^{k_1} \oplus A_i^{k_1} \tag{19}$$
$$Z_i^2 = E_i^{k_2} \oplus O_i^{k_2} \oplus A_i^{k_2} \tag{20}$$
$$Z_i^3 = E_i^{k_3} \oplus O_i^{k_3} \oplus A_i^{k_3} \tag{21}$$

Finally, the results of the addition operations are concatenated as the output of the SR module:

$$A_{i-1} = Concat(Z_i^1, Z_i^2, Z_i^3) \tag{22}$$

E. Output

The output features of GA traverse the two SR modules, and the output tensor A_0 of SR2 is obtained, which is then classified. In the classification process, the number of channels as well as the length and width of the feature tensors are reduced through a series of convolution operations (with only a single long dimension). This process can be interpreted as a classifier that maps the features to a prediction:

$$\hat{Y} = Soft\max(Classifier(A_0)) \tag{23}$$

III. Experimental Design

A. Experimental Configuration

The proposed network is developed using Python 3.7, with Pytorch 1.11 as the backend. The training and evaluation experiments for our proposed model are conducted in the hardware and software environment outlined in Table II.

Deep learning methods still involve certain metrics that require manual tuning, which are referred to as hyperparameters. The hyperparameter settings for the model are illustrated as follows. Given that machine learning is essentially a process of solving an optimization problem, we define an optimizer to

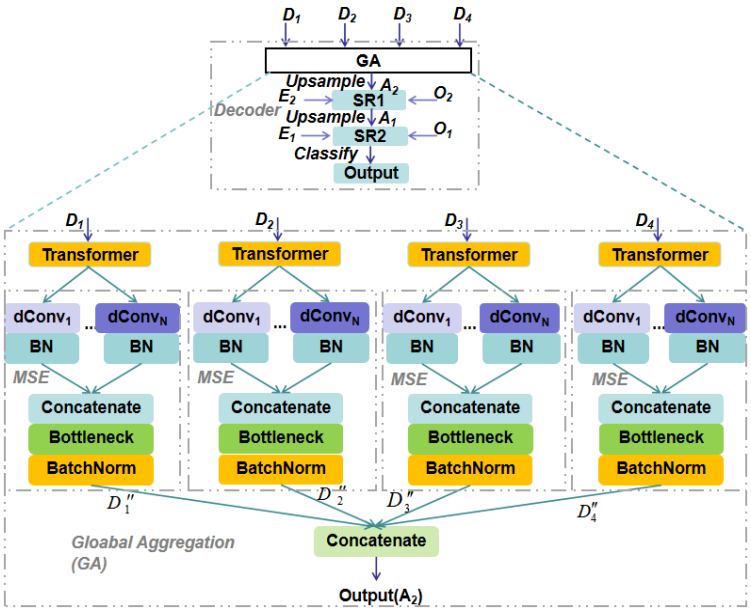

Fig. 5. Global aggregation module. Di is the output of each MS module, and MSE is the multiscale extraction module. dConvj denotes the Atrous Convolution module, with the subscript denoting the expansion rate of the Atrous Convolution. The color intensity of the dConv module reflects the expansion rate, with lighter colors corresponding to lower expansion rates. BN denotes the BatchNorm module.

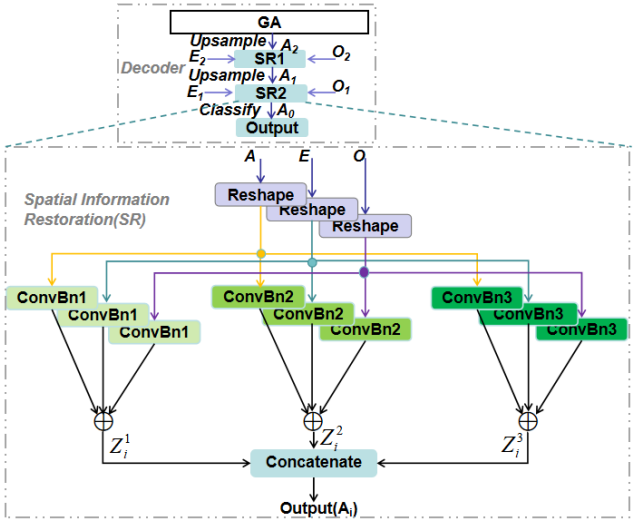

Fig. 6. Structure of the spatial information restoration module. A, E, and O represent the output of SR or), EEG encoder, and EOG encoder, respectively. Different colors for the ConvBn modules indicate varying convolution kernel sizes. Lighter colors correspond to smaller convolution kernel sizes, and ⊕ denotes the element-by-element addition operation.

TABLE II
THE MAIN RUNNING HARDWARE AND SOFTWARE ENVIRONMENT IN THE EXPERIMENTS

Facilities	Configuration
Architecture	x86-64
OS (operating system)	Ubuntu 20.04.1
Processor	Hygon C86 7151 16-core Processor
Graphics Processor	NVIDIA RTX A4000
GPU Driver Version	470.103.01
CUDA version	11.4
Python Version	3.7.11
Deep Learning Framework	Pytorch-1.11.0
Development Environment	Pycharm, Vscode.

guide the model toward a locally optimal solution through gradient descent. In this experiment, the Adam optimizer is used with the following parameter settings: lr = 1e−3, $\beta 1 = 0.9$, $\beta 2 = 0.999$, $\varepsilon = 1e{-}8$. A loss function is mainly used for statistically calculating the deviation between the model's predicted results and the true results:

$$loss = L(Y, \hat{Y}) = L(Y, f(X)) \tag{24}$$

where L is the loss function. The optimizer executes gradient descent on the basis of this loss function. Typically, multi-classification tasks employ the multivariate multi entropy loss function to calculate the model loss. However, for class imbalance tasks such as automatic sleep staging, the multivariate multi entropy function assigns equal weights to each class. This can result in models overfocusing on majority classes and neglecting minority classes, subsequently leading to poor performance in some evaluations.

The SalientSleepNet model, which manually assigns weights to each class in the multivariate multi-entropy function, has demonstrated excellent results on the Sleep-EDF-39 and Sleep-EDF-153 datasets. However, this approach may not be universally applicable to other datasets given its unique characteristics. Therefore, in this experiment, the Lovász loss function [39] is employed, which exhibited superior performance in addressing the class imbalance problem.

In this experiment, a 20-fold multi validation approach is employed. The maximum number of training epochs for each fold is set to 70 in this experiment. The default activation function is Elu, and the default padding operation is "same." The batch size is set to 12 for this experiment.

B. Evaluation Matrices

The model may exhibit a tendency to predict all classifications as the majority class to achieve high accuracy in class imbalance tasks. Although the accuracy approach is intuitive, it may not effectively reflect the model performance under class imbalance problem. To comprehensively assess the classification results for each class, accuracy (*ACC*) and Macro-F1

TABLE III
COMPARISON OF RESULTS FROM OUR MODEL WITH THOSE OF OTHER BASELINE MODELS ON THE SLEEP-EDF-39 DATASET

Model	ACC (%)	MF1 value (%)					
		Macro	W	N1	N2	N3	REM
SVM	76.1	63.7	71.6	13.6	85.1	76.5	71.8
RF	78.1	67.6	74.9	22.5	86.3	80.8	73.3
DeepSleepNet	82.0	76.9	85.0	47.0	86.0	85.0	82.0
SeqSleepNet	86.0	79.7	91.9	47.8	87.2	85.7	85.0
U-Time	85.6	80.5	87.0	52.0	86.0	85.0	82.0
SalientSleepNet	86.3	80.6	92.3	56.2	89.9	87.2	89.2
MMASleepNet	87.3	82.7	92.2	54.8	89.7	**90.2**	86.4
GAC-SleepNet	87.0	82.4	91.1	54.4	89.9	88.7	87.8
TSEDSleepNet	**88.9**	**84.8**	**92.9**	**63.8**	**90.5**	87.8	**91.4**

($MF1$) values are used for evaluation in this experiment. The Macro-Precision (P_i) and Macro-Recall (R_i) for the i-th class, along with the overall ACC and $MF1$ are defined as follows:

$$ACC = \frac{\sum_{i=1}^{K} TP_i}{N} \tag{25}$$

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{26}$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{27}$$

$$MF1 = \frac{1}{K} \sum_{i=1}^{K} \frac{2 \times P_i \times R_i}{P_i + R_i} \tag{28}$$

where TP_i denotes the true positive, FP_i denotes the false positive, FN_i denotes the false negative for the i-th class, K is the number of classes, N is the total number of samples, and $MF1_i$ represents the MF1 value for i-th class.

IV. RESULTS

A. Comparison of Results With Those Obtained Using Baseline Methods

Our TSEDSleepNet was compared with nine other sleep-staging methods based on machine learning, and the classification results on the two datasets are presented in Tables III and IV. A recent paper on GAC-SleepNet [40] only provided test results on the Sleep-EDF-39 dataset, and the paper on MaskSleepNet only provided test results on the Sleep-EDF-153 dataset, including the overall accuracy and the Macro-F1 value. SVM [23] and RF [23] are conventional machine-learning methods. DeepSleepNet and U-Time are single-modal methods. SeqSleepNet, SalientSleepNet, MMASleepNet, GAC-SleepNet, and MaskSleepNet are multimodal methods. The results of MaskSleepNet with EEG and EOG signals as input are adopted in this comparison.

Conventional machine learning methods such as SVM and RF cannot effectively capture the significant features of sleep stages, and they lack processing methods for managing unbalanced data, resulting in low accuracy and low F1 values. By contrast, multimodal models can capture the diversity of different electrophysiological signal features, leading to higher accuracy and F1 values than unimodal models.

Our model achieved the accuracies of 88.9% and 85.2%, respectively, on the two datasets, surpassing the accuracies all baseline models. This result demonstrates the effectiveness of our method in fusing multi-modal features. Furthermore, in terms of MF1 values, the proposed model outperforms all baseline models on both datasets, indicating that our model is better able to adapt to unbalanced datasets.

TABLE IV
COMPARISON OF RESULTS FROM OUR MODEL WITH THOSE OF OTHER BASELINE MODELS ON THE SLEEP-EDF-153 DATASET

Model	ACC (%)	MF1 value (%)					
		Macro	W	N1	N2	N3	REM
SVM	71.2	57.8	80.3	13.5	79.5	57.1	58.7
RF	72.7	62.4	81.6	23.2	80.6	65.8	60.8
DeepSleepNet	78.5	75.3	91.0	47.0	81.0	69.0	79.0
SeqSleepNet	83.8	78.2	92.8	48.9	85.4	78.6	85.1
U-Time	82.7	76.0	92.0	51.0	84.0	75.0	80.0
SalientSleepNet	84.1	79.5	93.3	54.2	85.8	78.3	**85.8**
MMASleepNet	82.7	77.6	92.9	49.1	84.9	**81.3**	79.8
MaskSleepNet	85.0	78.3	-	-	-	-	-
TSEDSleepNet	**85.2**	**79.7**	**93.6**	**54.6**	**86.8**	78.4	85.5

TABLE V
VALIDATION OF MULTIMODAL FUSION ON THE SLEEP-EDF-39 DATASET

Model	ACC (%)	MF1 value (%)
Single-Branch Using EEG only	85.1	81.1
Single-Branch Using EOG only	82.0	76.9
TSEDSleepNet Using EEG and EOG	88.9	84.8

TABLE VI
VALIDATION OF MULTIMODAL FUSION ON THE SLEEP-EDF-153 DATASET

Model	ACC (%)	MF1 value (%)
Single-Branch Using EEG only	81.8	76.2
Single-Branch Using EOG only	78.5	75.3
TSEDSleepNet Using EEG and EOG	85.2	79.7

B. Ablation Experiments

We conducted ablation experiments to demonstrate the effectiveness of different modules in our TSEDSleepNet model. Several models were designed to evaluate the effectiveness of different modules in our complete model as follows:

1) **Single-Branch Model**: To validate the effectiveness of multimodal fusion in our model, a single-branch structure was designed. This involved retaining only the EEG single-stream encoder or the EOG single-stream encoder in the complete TSEDSleepNet.

2) **TSEDSleepNet-MF Model**: To verify the effectiveness of the MF module, the MF module was removed from the complete TSEDSleepNet.

3) **TSEDSleepNet-MS Model**: To verify the validity of the MS module in the model, the MS module was removed from the complete TSEDSleepNet.

The results tested on the Sleep-EDF-39 and Sleep-EDF-153 datasets are presented in Tables V and VI, respectively, and they were used to assess the effectiveness of multimodal fusion in our model. Tables V and VI indicate that the ACC and MF1 values of the models using only EEG or EOG signals are lower than the complete model using both EEG and EOG signals by at least 3%. The complete model trained with both EOG and EMG outperformed the models trained with EEG or EOG alone.

The results tested on the Sleep-EDF-39 and Sleep-EDF-153 datasets are presented in Tables VII and VIII, respectively, and they were used to validate the effectiveness of the MF

TABLE VII
VALIDATION OF MF AND MS MODULES ON THE SLEEP-EDF-39 DATASET

Model	ACC (%)	MF1 value (%)
TSEDSleepNet-MF	86.2	82.3
TSEDSleepNet-MS	86.5	83.0
TSEDSleepNet	88.9	84.8

TABLE VIII
VALIDATION OF MF AND MS MODULES ON THE SLEEP-EDF-153 DATASET

Model	ACC (%)	MF1 value (%)
TSEDSleepNet-MF	82.6	78.0
TSEDSleepNet-MS	82.3	77.8
TSEDSleepNet	85.2	79.7

and MS modules. The tables indicate that the models with the MF or MS module removed exhibited decreased sleep-staging accuracy and MF1 values compared with the complete model. This finding indicates that the MF module can effectively perform multimodal fusion, and that the MS module can effectively extract multiscale features from EEG and EOG signals.

V. DISCUSSIONS AND CONCLUSION

In this study, we introduced TSEDSleepNet, an end-to-end convolutional neural network designed for automatic sleep staging by using PSG signals. Our model leverages the DSA2F framework to construct a sleep-staging model specifically tailored for EEG and EOG signals. Notably, our proposed model involves a more efficient feature fusion methodology, distinct from the conventional approach of simply fusing features from different modalities. The experimental findings reveal the presence of complementary features related to sleep stages between the two modalities, indicating the superior performance of our proposed model over conventional baseline models on two publicly available datasets. Conventional machine learning methods exhibit limitations in effectively capturing salient features associated with sleep stages, and their loss functions lack the attention needed for analyzing imbalanced datasets. Consequently, conventional machine learning methods have low MF1 values for certain classes, leading to decreased overall MF1 values. Early deep learning methods such as DeepSleepNet and SeqSleepNet also fail to address the class imbalance problem, resulting in comparatively low MF1 values. U-Time, initiated with a focus on the class imbalance problem, improves the MF1 value in the N1 compared with previous models. However, the overall MF1 value of U-Time is less than that of SeqSleepNet, which is attributed to its utilization of a fully convolutional network, which does not perform as effectively as SeqSleepNet, which is composed of RNN, in multiscale feature extraction. SeqSleepNet is an RNN-based model that has advantages in extracting temporal features from sleep signals, but it is difficult to extract and fuse spatial features of sleep signals. Our model adopts the structure of the combination of CNN and Transformer to simultaneously extract features in the spatial and temporal domains. So our model can better model the complex relationship between multimodal signals in the spatial and temporal domains, and thus achieve a better sleep staging result. To compensate for the disadvantages of a full convolutional network, SalientSleepNet adopts U^2-Net and MSE in multiscale feature extraction and multimodal fusion to learn the saliency features, resulting in considerably enhanced results compared with those obtained using SeqSleepNet. Compared with single-channel methods, our proposed strategy effectively integrates multimodal information, capturing complementary features and thus leading to significant improvements in accuracy and MF1 values for sleep staging.

Multimodal methods such as SalientSleepNet, MMASleepNet, and MaskSleepNet focus on fusing the output tensor of the last layer of the encoder, overlooking the class imbalance problem and limiting the overall performance of sleep staging. By contrast, our model integrates multimodal information from deeper multiple layers, enabling the extraction of richer and more valuable features. To compensate for the loss of detailed information during feature extraction, our model adopts the SR module. Furthermore, the Lovász loss function is employed to mitigate the class imbalance problem, resulting in enhanced MF1 values, especially in the N1 stage. GAC-SleepNet utilizes the features in the dual structure of the graph structure and the Euclidean structure for sleep staging. This approach captures the correlation of multiple PSG channels, thereby learning the channel and temporal representations of PSG to enhance classification performance. However, GAC-SleepNet exhibits a slight decrease in performance compared with the latest models based on temporal feature extraction. This decrease is attributed to its emphasis on the channel-wise feature, whereas the most recent models have achieved superior results by temporal feature extraction [41].

Although our proposed TSEDSleepNet exhibits superior performance compared with baseline models, some aspects of our model can be improved. First, our model achieves high accuracy but has increased complexity, resulting in a large number of parameters (11.3 M) and a low training speed. Second, accurately identifying the N1 stage remains challenging, as it can be easily misclassified as other classes [42], [43] due to its limited data and transitional nature between wakefulness and sleep [44]. This indicates the need for more effective techniques to accurately discern the N1 stage. Third, sleep staging often relies on subjective classification and interpretation of physiological signals by professionals, leading to potential variations in conclusions among different experts [44]. Therefore, sleep patterns and characteristics vary between individuals, necessitating a personalized classification method.

Our proposed TSEDSleepNet can enhance the accuracy of sleep staging and provides a new approach to multimodal sleep monitoring. This improved approach can more accurately identify the sleep state; sleep quality can thus be improved through the implementation of relevant interventions. In future research endeavors, we will focus on refining the extraction of significant features related to the N1 stage and enhancing the classification performance in this stage. Moreover, we intend to collaborate with hospitals or research organizations to

validate the scalability and real-world applicability of our proposed method.

REFERENCES

[1] K. Wulff, S. Gatti, J. G. Wettstein, and R. G. Foster, "Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease," *Nature Rev. Neurosci.*, vol. 11, no. 8, pp. 589–599, Jul. 2010.

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

[3] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan. (2020). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specifications*. Accessed: Dec. 30, 2023. [Online]. Available: https://www.sleep.pitt.edu

[4] E. Eldele, M. Ragab, Z. Chen, M. Wu, C.-K. Kwoh, and X. Li, "Self-supervised learning for label- efficient sleep stage classification: A comprehensive evaluation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1333–1342, 2023, doi: 10.1109/TNSRE.2023.3245285.

[5] O. Tsinalis, P. M. Matthews, and Y. Guo, "Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders," *Ann. Biomed. Eng.*, vol. 44, no. 5, pp. 1587–1597, Oct. 2015.

[6] R. Sharma, R. B. Pachori, and A. Upadhyay, "Automatic sleep stages classification based on iterative filtering of electroencephalogram signals," *Neural Comput. Appl.*, vol. 28, no. 10, pp. 2959–2978, Oct. 2017, doi: 10.1007/s00521-017-2919-6.

[7] A. R. Hassan and A. Subasi, "A decision support system for automated identification of sleep stages from single-channel EEG signals," *Knowl.-Based Syst.*, vol. 128, pp. 115–124, Jul. 2017, doi: 10.1016/j.knosys.2017.05.005.

[8] T. Lajnef et al., "Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines," *J. Neurosci. Methods*, vol. 250, pp. 94–105, Jul. 2015.

[9] C.-S. Huang, C.-L. Lin, L.-W. Ko, S.-Y. Liu, T.-P. Su, and C.-T. Lin, "Knowledge-based identification of sleep stages based on two forehead electroencephalogram channels," *Frontiers Neurosci.*, vol. 8, Sep. 2014, Art. no. 263.

[10] S. Güne, K. Polat, and E. 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.

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

[12] 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.

[13] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies," in *A Field Guide Dyn. Recurrent Netw.*, 2001, pp. 237–243.

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

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

[16] M. Perslev, M. Hejselbak Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-time: A fully convolutional network for time series segmentation applied to sleep staging," 2019, *arXiv:1910.11162*.

[17] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, "U-Sleep: Resilient high-frequency sleep staging," *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–12, Apr. 2021.

[18] D. Zhao et al., "A deep learning algorithm based on 1D CNN-LSTM for automatic sleep staging," *Technol. Health Care*, vol. 30, no. 2, pp. 323–336, Mar. 2022.

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

[20] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–14.

[21] M. Li, H. Chen, and Z. Cheng, "An attention-guided spatiotemporal graph convolutional network for sleep stage classification," *Life*, vol. 12, no. 5, p. 622, Apr. 2022.

[22] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, Feb. 2018.

[23] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, "SalientSleepNet: Multimodal salient wave detection network for sleep staging," 2021, *arXiv:2105.13864*.

[24] Z. Jia, X. Cai, G. Zheng, J. Wang, and Y. Lin, "SleepPrintNet: A multivariate multimodal neural network based on physiological time-series for automatic sleep staging," *IEEE Trans. Artif. Intell.*, vol. 1, no. 3, pp. 248–257, Dec. 2020.

[25] 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.

[26] Z. Yubo, L. Yingying, Z. Bing, Z. Lin, and L. Lei, "MMASleepNet: A multimodal attention network based on electrophysiological signals for automatic sleep staging," *Frontiers Neurosci.*, vol. 16, Aug. 2022, Art. no. 1337.

[27] H. Zhu et al., "MaskSleepNet: A cross-modality adaptation neural network for heterogeneous signals processing in sleep staging," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 5, pp. 2353–2364, May 2023.

[28] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1407–1417.

[29] S. Liu et al., "Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–14.

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

[31] S. A. Imtiaz and E. Rodriguez-Villegas, "Recommendations for performance assessment of automatic sleep staging algorithms," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Chicago, IL, USA, Aug. 2014, pp. 5044–5047.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[33] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[35] Z. Jia et al., "Multi-view spatial–temporal graph convolutional networks with domain generalization for sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1977–1986, 2021.

[36] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.

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

[38] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 636–644.

[39] M. Berman, A. R. Triki, and M. B. Blaschko, "The lovasz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, Utah, USA, Jun. 2018, pp. 4413–4421.

[40] T. Li, Y. Gong, Y. Lv, F. Wang, M. Hu, and Y. Wen, "GAC-SleepNet: A dual-structured sleep staging method based on graph structure and Euclidean structure," *Comput. Biol. Med.*, vol. 165, Oct. 2023, Art. no. 107477.

[41] Z. Jin and K. Jia, "SAGSleepNet: A deep learning model for sleep staging based on self-attention graph of polysomnography," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105062.

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

[43] Q. Shen et al., "LGSleepNet: An automatic sleep staging model based on local and global representation learning," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.

[44] R. Li, B. Wang, T. Zhang, and T. Sugi, "A developed LSTM-ladder-network-based model for sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1418–1428, 2023.