A Lightweight Segmented Attention Network for Sleep Staging by Fusing Local Characteristics and Adjacent Information

Wei Zhou, Hangyu Zhu, Ning Shen, Hongyu Chen, Cong Fu, Huan Yu, Feng Shu, Chen Chen, and Wei Chen, *Senior Member, IEEE*

***Abstract*— Sleep staging is the essential step in sleep quality assessment and sleep disorders diagnosis. However, most current automatic sleep staging approaches use recurrent neural networks (RNN), resulting in a relatively large training burden. Moreover, these methods only extract information of the whole epoch or adjacent epochs, ignoring the local signal variations within epoch. To address these issues, a novel deep learning architecture named segmented attention network (SAN) is proposed in this paper. The architecture can be divided into feature extraction (FE) and time sequence encoder (TSE). The FE module consists of multiple multiscale CNN (MMCNN) and residual squeeze and excitation block (SE block). The former extracts features from multiple equal-length EEG segments and the latter reinforced the features. The TSE module based on a multi-head attention mechanism could capture the temporal information in the features extracted by FE module. Noteworthy, in SAN, we replaced the RNN module with a TSE module for temporal learning and made the network faster. The evaluation of the model was performed on two widely used public datasets, Montreal Archive of Sleep Studies (MASS) and Sleep-EDFX, and one clinical dataset from Huashan Hospital of Fudan University, Shanghai, China (HSFU). The proposed model achieved the accuracy of 85.5%, 86.4%, 82.5% on Sleep-EDFX, MASS and HSFU, respectively. The experimental results exhibited favorable performance and consistent improvements of SAN on different datasets in comparison with the state-of-the-art studies. It also proved the necessity of sleep staging by integrating the local characteristics within epochs and adjacent informative features among epochs.**

***Index Terms*— Sleep stage, deep learning, EEG, multiple multiscale convolutional neural network, residual squeeze and excitation block, time sequence encoder, multi-head attention.**

Manuscript received 8 May 2022; revised 30 September 2022; accepted 1 November 2022. Date of publication 7 November 2022; date of current version 31 January 2023. This work was supported in part by the Shanghai Municipal Science and Technology Major Project under Grant 2017SHZDZX01; in part by the Greater Bay Area Institute of Precision Medicine, Guangzhou, under Grant IPM2021C002; in part by the Shanghai Municipal Science and Technology International Research and Development Collaboration Project under Grant 20510710500; in part by the National Natural Science Foundation of China under Grant 62001118; in part by the Shanghai Committee of Science and Technology under Grant 20S31903900; and in part by the National Key Research and Development Program under Grant 2021YFC2501404. *(Corresponding authors: Huan Yu; Chen Chen; Wei Chen.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Huashan Hospital under Ethical Permit No. 2021-811.

Wei Zhou, Hangyu Zhu, and Ning Shen are with the Center for Intelligent Medical Electronics (CIME), School of Information Science and Engineering, Fudan University, Shanghai 200433, China (e-mail: wzhou19@fudan.edu.cn; hyzhu20@fudan.edu.cn; nshen20@fudan.edu.cn).

Hongyu Chen is with the Center for Intelligent Medical Electronics, School of Information Science and Technology, Fudan University, Shanghai 200433, China, and also with the Greater Bay Area Institute of Precision Medicine, Guangzhou 510000, China (e-mail: chenhongyudesign@outlook.com).

Cong Fu and Huan Yu are with the Sleep and Wake Disorders' Center, Department of Neurology, Huashan Hospital, Fudan University, Shanghai 200433, China (e-mail: fucong@fudan.edu.cn; dr.yuhuan@163.com).

Feng Shu is with the Shanghai Engineering Research Center of Ultra-Precision Motion Control and Measurement, Academy for Engineering and Technology, Fudan University, Shanghai 200433, China (e-mail: fshu@fudan.edu.cn).

Chen Chen is with the Human Phenome Institute, Fudan University, Shanghai 200433, China (e-mail: chenchen_fd@fudan.edu.cn).

Wei Chen is with the Center for Intelligent Medical Electronics (CIME), School of Information Science and Engineering, Human Phenome Institute, Fudan University, Shanghai 200433, China (e-mail: w_chen@fudan.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2022.3220372

I. Introduction

SLEEP is an important activity for human beings. High-quality night sleep contributes to maintaining physical and mental wellbeing [1]. While lack of sleep, sleep disorders can lead to adverse cardiometabolic risks such as obesity, hypertension, diabetes and cardiovascular disease [2], [3], [4], [5], [6]. Thus, it is necessary to monitor sleep quality and treat sleep disorders expeditiously. In clinical practice, the sleep condition is usually measured using polysomnography (PSG) device, consisting of electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), electrocardiogram (ECG) and so on [7]. Physicians will manually interpret the PSG recording and divide it into the corresponding sleep stage according to the Rechtschaffen and Kales (R&K) [8], which divides sleep into six stages, i.e., wake (W), rapid eye movement (REM) and four non-REM stages (S1, S2, S3 and S4) or American Academy of Sleep Medicine (AASM) [9], which divides sleep into five stages, i.e., wake (W), rapid eye movement (REM) and non-REM stages (N1, N2 and N3). Manual sleep staging is a very tedious and laborious task. It usually takes more than 4 hours to label a full night's sleep

recordings. Therefore, to alleviate the manual interpretation burden on physicians, automatic sleep staging is deemed to be an effective alternative.

The automatic sleep staging methods can be roughly categorized into machine learning-based approach and deep learning-based approach. Whereas, in recent years, deep learning approaches [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23] have gradually replaced traditional machine learning approaches [24], [25] in automated sleep staging. As traditional machine learning methods require extraction of hand-crafted features, which is time-consuming and proven to be unreliable when tested on unseen data. In contrast, deep learning methods can avoid these problems by using neural networks for adaptive and adequate feature extraction. The majority of existing deep learning-based sleep staging approaches are using convolutional neural network (CNN) architecture [26], [27], [28]. To illustrate, Yang et al. extracted features from raw EEG by using CNNs, and applied Hidden Markov Model (HMM) refinement as a post-processing step to correct the unreasonable sleep stage transitions of adjacent EEG epochs [27]. Perslev et al. proposed U-Sleep based on a fully convolutional neural network and evaluated it across several clinical studies [28]. A number of studies are using recurrent neural network (RNN) architectures such as long short-term memory (LSTM) and gated recurrent unit (GRU), where temporal features can be fully learned and explored. For example, H. Phan et al. proposed an architecture named SeqSleepNet to process the sequential signal based on RNN, which exhibited excellent performance, while it also suffered from a considerable amount of time consumption for training [16]. Dong et al. applied multi-layer perception (MLP) and LSTM to address the temporal pattern recognition challenge [14]. A few approaches proposed to combine CNN and RNN in order to extract both temporal and spatial information in the biomedical data [29], [30], [31]. Supratak et al. proposed an architecture named DeepSleepNet which was the combination of the CNN and RNN and the five-class sleep staging results can reach 86.2% [10]. Sun et al. proposed an architecture that considered both automatic and manual features based on CNN and RNN [11].

Although favorable results can be achieved by most of the existing automatic sleep staging approaches, they still face several enormous challenges. Firstly, for those architectures based on one RNN or multiple RNNs, it results in high model complexity mainly caused by computational approach and structural design of RNN [32]. Since the hidden states in RNN can only be calculated in serial, it relies on the information from the previous moment and therefore requires a lot of time to train the model. It is detrimental to transfer the model to new datasets, considering that most existing methods are lack of strong generalization capabilities. On a huge amount of sleep data, the application of RNN undoubtedly increases the computational time and the model complexity significantly. Secondly, in CNN based structures, only the characteristics of whole epoch or adjacent epochs are considered, and the local signal variations within epoch have been ignored [10], [11], [12], [17], [18]. The entire 30s EEG signal is usually fed directly into the network in these works, and features are extracted from the signal by convolutional kernels of different sizes. However, according to the American Academy of Sleep Medicine (AASM) rules [9], sometimes features of different sleep stages appear simultaneously in the same frame of the sleep record. This will then determine which sleep stage the sleep recordings in this frame belong to, based on the length of time that the features of the different sleep stages last. When feeding the 30s EEG signal into the CNN, it may cause some degree of confusion if there is a transitioning in the sleep stage and features are extracted in generalized whole epoch. On the one hand, extracting features from segmented signal can avoid this drawback and yield the contribution of different regions to the decision outcome. On the other hand, the segmentation operation actually divides the model into several submodules and the joint collaboration of multiple submodules facilitates the overall performance.

In this paper, a lightweight segmented attention network (SAN) model for automatic sleep staging is proposed. This model consists of two main constructions: feature extraction (FE) and time sequence encoder (TSE). The FE module is composed of multiple multiscale CNN (MMCNN) and residual squeeze and excitation block (residual SE block). The 30s EEG signal is divided into multiple equal-length segments, and then each segment is processed by a multiscale CNN for feature extraction. Multiscale CNN has both large and small convolutional kernels to fully extract the information in each EEG signal segment. By segmenting the EEG signal before feature extraction, the signal features can be fully extracted, and then features from different regions can be integrated. The residual SE block can adjust the weight of features and enhance them. The time sequence encoder is used to learn the temporal information from the extracted features and its core structure is multi-head attention. The multi-head attention can process data in parallel, greatly improving learning efficiency, which is different from RNNs. We also apply a data augmentation approach to address the imbalance issue in sleep data and improve the generalization ability of the model. The main contributions are summarized as follows:

1) In consideration of exploring extensive characteristics within an epoch, we divided the whole epoch into multiple equal-length segments and fully investigated the local information of each segment and temporal information among segments. By integrating these characteristics, a comprehensive feature that can represent various regions is provided.
2) We propose MMCNN which consists of several multiscale CNN with large and small convolutional kernels to fully extract features from the EEG signal. Features with different temporal frequency resolutions are acquired and then residual SE block is used to focus on the channel-wise informative features.
3) Instead of using RNN, a time sequence encoder that mainly consists of a multi-head attention mechanism is proposed. This will significantly reduce the complexity of the network while ensuring efficiency. It can run in parallel, to learn time sequence information between features. Thus, the model can obtain the contribution of different segments to the classification results.

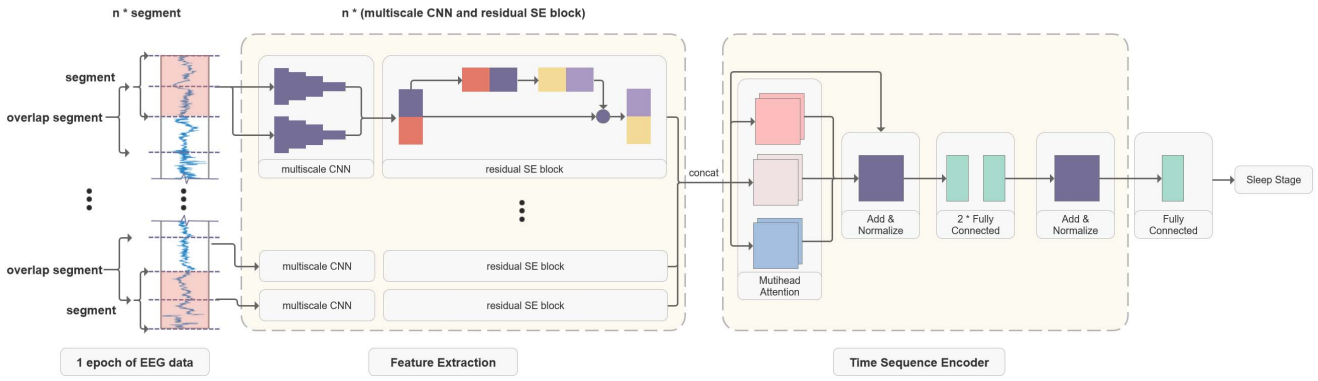

Fig. 1. The overall architecture of the proposed network. First the signal is divided into L segments and L-1 50% overlap segments, which means there is n = 2*L-1 segments in total. Then the segmented signals are fed into the feature extraction module for feature extraction using multiple multiscale CNN and SE block. After that, the feature map is sent to the TSE module to capture the impact of different segments on the weight of subsequent decisions. Finally, the features of multiple segments jointly decide the sleep stage of the signal of this epoch.

This paper is organized as follows: Section II illustrates the details of the proposed model. In Section III, we introduce the datasets, the experimental process and the evaluation indicators. The sleep staging results of the proposed model on different datasets are shown and discussed in Section IV, where we also explore the computation efficiency of the proposed model and compare our approach with that of others. At last, we draw the conclusion in Section V.

II. Method

In this section, we introduce our proposed segmented attention network model for sleep staging using single-channel EEG signal.

A. Overall Structure of the Segmented Attention Network Model

Fig. 1 shows the overall structure of our SAN model. In the process of feature extraction, to preserve as much as possible the local characteristics of the different regions of the signal, we divide the signal into fixed-length segments and maintain a 50% overlap, which helps prevent discontinuities in the signal. We also explore how the variation of segment length impact performance, which is illustrated in Section IV. Then the feature extraction is applied to deal with these segmented signals, which is composed of multiple multiscale CNNs used to extract the feature from the 30-second EEG signal. Multiple multiscale CNNs are designed to better extract comprehensive features at various temporal resolutions. Each multiscale CNN includes small kernel convolutions and large kernel convolutions. It is worth mentioning that in each multiscale CNN there is residual SE block [33], which can make the feature more distinctive. After the feature extraction, the TSE module is employed to learn the time sequence information from the features extracted by multiple multiscale CNNs. The time sequence encoder consists of positional embedding, multi-head attention and feed-forward parts. And the output of the TSE is connected to a fully connected layer with softmax classifier. In this work, to address the imbalance problem in the sleep stages, we adopt various data augmentation strategies to enrich the diversity of the input signals, such as adding Gaussian noise, scaling, etc. In the following subsections, the detail of the blocks is presented.

B. Feature Extraction

An epoch of EEG signal is divided into several segments after data augmentation. Each segmented signal is fed into corresponding multiscale CNN and residual SE block. After the multiple multiscale CNNs and residual SE block, all the features are integrated by a connection layer as the feature information.

1) The Segment of EEG Signal: As shown in Fig. 1, we divide the 30s single-channel EEG signal into segments. With the use of segmentation, which is equivalent to adding windows to the signal, we turn the segment of signal into a quasi stationary. Therefore, the model can learn more stable statistical properties and acquire robust features. Each segment of the EEG signal is fed separately into the multiscale CNN for feature extraction. It is worth mentioning that in order to prevent information loss between segments due to split signals, there is a 50% overlap between two adjacent segments. For the 30s EEG signal, the length of each segment can be calculated as follows.

$$length = \frac{30s}{L} \tag{1}$$

where L represents the number of selected segments. When L is determined, there are $L-1$ overlap segments, and the total number of segments is $n = 2 * L - 1$. We refer to the model with different L segments as SAN-L, and we explored the effect of different number of segments on final results in Section IV.

2) Multiple Multiscale CNN: Fig. 2 shows the specific structure of the multiscale CNNs applied for feature extraction from a segment of 30s single-channel EEG signal. We propose MMCNN to fully extract the features of different sleep stages in 30s single-channel EEG signal. The input of each multiscale CNN is a segmented EEG signal. As shown in the Fig. 2, each multiscale CNN has two branches: one branch with small kernel convolutions is applied to extract the detail features

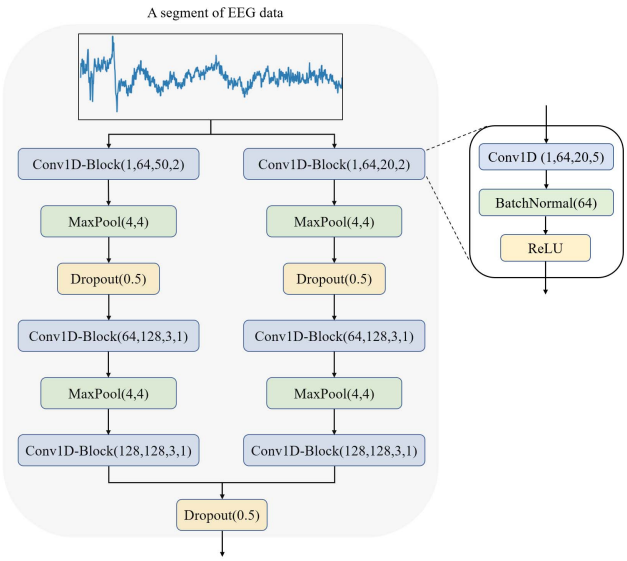

Fig. 2. Structure of the multiscale CNNs. To ensure that the model can capture information at different scales, two CNNs with different kernel and stride are used to extract features of the segmented signal.

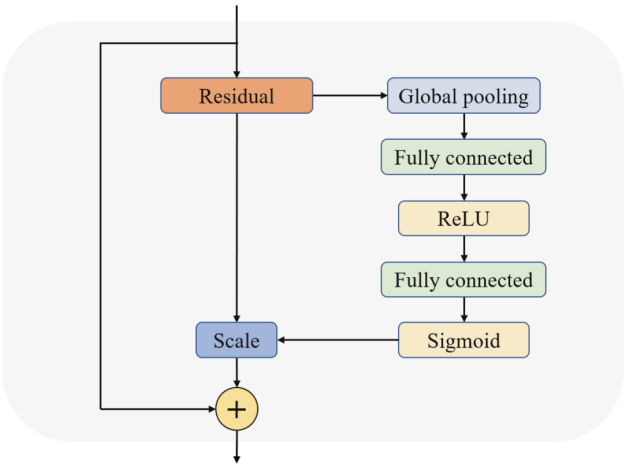

Fig. 3. Structure of the residual SE block. This module enhances the features and prevents gradient disappearance.

and high frequency components of the segmented EEG signal. Another branch with large kernel convolutions is applied to extract the morphological features and low frequency information. In multiscale CNN, three convolution layers and two max-pooling layers are performed for each scale. The first convolutional layer is to reduce the dimensionality of the input signal for subsequent feature extraction. The last two convolutional layers are applied for feature extraction, so the parameters of the last two convolutional layers in both scales are similar. In each convolutional layer, there is a batch normalization layer [34] that aligns the data and a ReLU that acts as an activation function. To prevent overfitting, dropout was performed after the max-pooling layer and the data concatenation of two scales.

3) Residual Squeeze and Excitation Block: Residual network can prevent gradient disappearance and gradient explosion while the network deepens [33]. Recently it has been improved and enhanced by many researchers. Hu et al. [35] proposed Squeeze and Excitation block (SE block), which can enhance the features that have a significant impact on the results and weakens the features that have a small impact on the results by scaling the extracted features. The structure of the module is shown in Fig. 3. In the residual SE block, it combines residual network and SE block. Given the input $X \in R^{H\times W\times C}$, which is the output of the multiscale CNN. The residual layer is mainly composed of convolutional layers. After the residual layer, we get the $X_1 \in R^{H\times W\times C}$. Next, the SE block compresses the extracted features. The global pooling is used to reduce the dimensionality of features, changing the $X_1 \in R^{H\times W\times C} to X_2 \in R^{1\times 1\times C}$. Afterwards, two fully connected layers and ReLU layer are applied to parameterize the pass selection mechanism, reinforcing the important features of the center and weakening the features of the edge. The following sigmoid activation function is used to give the proportion of weights for each feature. The entire process is shown in the following equation:

$$U = \sigma\left(F_2\left(ReLU\left(F_1\left(X_2\right)\right)\right)\right) \in R^{1\times 1\times C} \tag{2}$$

where the $F_1(\cdot)$ means the first FC layer, the $F_2(\cdot)$ means the second FC layer, the ReLU$(\cdot)$ means the ReLU activation function and the $\sigma(\cdot)$ means the sigmoid activation function. Then, the feature weights are reassigned by matrix multiplication:

$$V = U \times X_2 \in R^{H\times W\times C} \tag{3}$$

Finally, shortcut connection is finally used to superimpose the original special input and the enhanced features. The final input results are as follows:

$$\widetilde{X} = X_1 + V \in R^{H\times W\times C} \tag{4}$$

C. Time Sequence Encoder (TSE)

The function of the TSE module is to perform temporal learning on the extracted features. TSE module consists of a multi-head self-attention layer, an add and normalize layer and a feed forward layer. In the following subsections, the detail of the layers is presented.

1) Multi-Head Attention: Inspired by [36], an attention mechanism to obtain temporal features is proposed. It is more efficient than RNN and consists of several self-attention. Self-attention predicts the final outcome by focusing attention on different features. As shown in the Fig. 4, given the input signal $X \in R^{N\times M}$, the three matrices of Query ($Q \in R^{M\times d_K}$), Key ($K \in R^{N\times d_K}$), and Value ($V \in R^{N\times d_V}$), are obtained by multiplying with the linear transformation matrix $W^Q \in R^{N\times d_K}$, $W^K \in R^{M\times d_K}$, $W^V \in R^{M\times d_V}$. The dimensions of Q and K must be the same, and the dimensions of V and Q can be inconsistent. The lengths of K and V must be the same because K and V essentially correspond to representations of the input signal on different spaces. Finally, the output of self-attention is calculated by the following equation:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

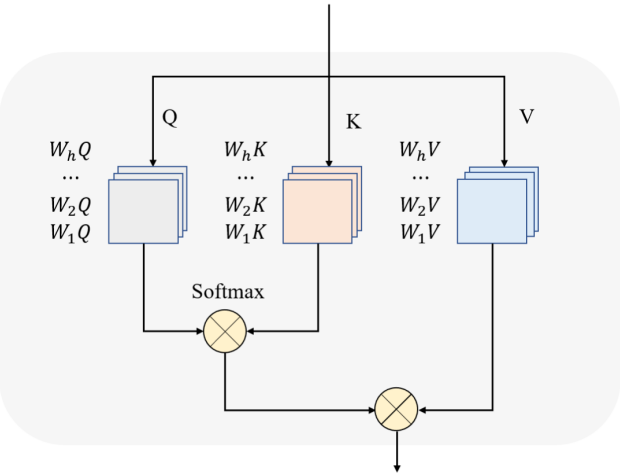

Fig. 4. Structure of the muti-head attention.

where the scaling factor $\frac{1}{\sqrt{d_k}}$ is to make the final distribution result independent of the elements in Q, K and to keep the gradient values stable during the training process.

Multi-head attention splices the results of h self-attention layers:

$$\text{head}_i = Attention\,(Q_i, K_i, V_i) \quad 1 \le i \le h \tag{6}$$

$$Multihead\,(Q, K, V) = Concat(head_1, \ldots, head_h)W^o \tag{7}$$

where W^o is the additional weight matrix, and it will be jointly trained in the model to adjust the weights. Compared with a single self-attention layer, multi-head attention extends the ability of the model to focus on different positions and gives multiple representation subspaces of the self-attention layer, which can find correlations between sequences from different angles, and reduces the dimensionality of each vector when calculating the attention of each head, which can prevent overfitting

2) Add and Normalize Layer: In the TSE module, there are two add and normalize layers. One is after the multi-head attention layer and the other is after the feed forward layer. It adds the input signal to the output via the residual connection, and then normalize the sum. The process can be explained as follows:

$$output = LayerNorm(x + SubLayer(x)) \tag{8}$$

where the x is the input signal of the multi-head attention or the feed forward layer and the SubLayer(x) is the output of the multi-head attention or the feed forward layer. The use of residual connection helps in feature learning, prevents gradient disappearance, and can speed up learning.

3) Feed Forward Layer: Feed forward layer is after the multi-head attention. Feed forward layer contains two linear transformation layers and the activation function between the two linear transformation layers is ReLU. The addition of feed forward layer introduces nonlinearity (ReLU activation function) and transforms the space of multi-head attention output, thus increasing the expressiveness of the model. The operation of the feed forward layer can be defined as follows:

$$\text{FFL}_{output} = F_4(\delta(F_3(x))) \tag{9}$$

4) Mask: In TSE module, for the model to learn only information before the current moment and not to leak information after the current moment, we add the mask function to multi-head attention layer. Specifically, the matrix is made to be a lower triangular matrix after performing the operation. The operation can be defined as follows:

$$Mask\,(X) = \begin{pmatrix} x_1 & 0 & \cdots & 0 \\ x_1 & x_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \tag{10}$$

where $X = \begin{pmatrix} x_1 & \cdots & x_n \\ \vdots & \ddots & \vdots \\ x_1 & \cdots & x_n \end{pmatrix} \in R^{N\times N}$. The operation will be This operation will be performed after the calculation of QK^T. Therefor the equation (5) can be updated to:

$$Attention(Q, K, V) = softmax(\frac{Mask(QK^T)}{\sqrt{d_k}})V \tag{11}$$

In this way, at moment t, which is the t row of the matrix, only information from the first moment to the t moment can be read. Information after the t moment cannot be read.

D. Data Augmentation

In this work, We have made some transformations to the input signal. Specifically, we have designed three ways to perform data augmentation: 1) Adding Gaussian noise. 2) Inverting, that is, multiplying by a factor of -1. 3) Scaling, where the input signal is multiplied by a random factor which is in the range from 0.5 to 2. We use different combinations of the above three methods to produce sufficient signal variation. By applying various transformations to the input signal, we can achieve a more robust model.

III. Experiment

Our proposed new model is extensively validated on three datasets, including two public datasets and one clinical dataset. In this section, we introduce the database used for the experiment, and the process of our experiment.

A. Database

In this work, we apply Sleep-EDFX and MASS two public datasets and a clinical dataset called HSFU collected in Huashan Hospital, Fudan University, Shanghai, China, during 2019-2020 to validate the effectiveness of the proposed model.

1) Sleep-EDFX: The Sleep-EDFX dataset recorded the sleeping data of healthy subjects (sleep cassette) and people with mild sleep disturbances (sleep telemetry) [37]. In the dataset, the doctor manually divides all the 30-secend sleep periods in to eight stages. The sleeping periods are as follows: W (wake), 1 (S1), 2 (S2), 3 (S3), 4 (S4), R (rapid eye movement), M (body movement) and the 'unscored' (unidentifiable)

are marked with 0, 1, 2, 3, 4, 5, 6 and 9. By discarding the abnormal stages like M and 9, six stages are remained. In our experiments, we used the sleep cassette dataset, which included 78 subjects, and adopted the Fpz-Oz EEG channel.

2) Montreal Archive of Sleep Studies (MASS): Montreal Archive of Sleep Studies (MASS) is a large dataset which was collected from a number of different hospitals [38]. It has the whole-night sleep recording from 200 subjects (97 females and 103 males) aged from 18 to 76 years old. It has five subsets: SS1-SS5. The epoch of the recordings was manually labeled based on the AASM standard [9] and the R&K standard [8]. The length of the epoch in SS2, SS4 and SS5 is 20 seconds and the length of the epoch in SS1 and SS3 is 30 seconds. Each epoch recorded the EEG signals, EOG signals, EMG signals ECG signals and other signals. In our experiments, we used SS3 subset and adopted the C4 EEG channel.

3) Huashan Hospital Fudan University (HSFU): A non-public database collected in Huashan Hospital, Fudan University, Shanghai, China, during 2019-2020. The research was approved by the Ethics Committee of Huashan Hospital (ethical permit no. 2021-811). It consists of 26 clinical PSG recordings, which were acquired on patients diagnosed with obstructive sleep apnea, insomnia, and restless legs syndrome. The PSG recordings were annotated by one qualified sleep expert according to the AASM standard. We adopted the C4 EEG channel in this study.

B. Data Preprocessing

In this experiment, all used EEG signals are filtered by a notch filter and bandpass filter to eliminate industrial frequency interference. Then signals are resampled to 100 Hz to fit the model. EEG signals are normalized to zero mean and standard deviation of one to reduce differences between individuals. All the EEG signals were split into 30s epochs without overlap between each epoch. Each epoch of the EEG signal has a corresponding sleep stage label.

C. Evaluation Indicators

To evaluate the model performance, we adopt a series of commonly used evaluation metrics. Accuracy (Acc) shows the proportion of correctly predicted samples to the total samples. Macro-F1 score (MF1) is an evaluation metric that takes into account both precision and recall, and can evaluate model performance in multi-classification problems on imbalanced datasets. Cohen Kappa (κ) assesses the consistency of classifying the samples. Specificity (Spec) and Sensitivity (Sens) measure the ability of the model to correctly classify in positive and negative cases, respectively. They are calculated as follows.

$$Acc = \frac{1}{N}\sum_{i=1}^{K} TP_i \tag{12}$$

$$MF1 = \frac{1}{K}\sum_{i=1}^{K} \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i} \tag{13}$$

$$Specificity = \frac{1}{K}\sum_{i=1}^{K} \frac{TN_i}{TN_i + FP_i} \tag{14}$$

$$Sensitivity = \frac{1}{K}\sum_{i=1}^{K} \frac{TP_i}{TP_i + FN_i} \tag{15}$$

where True Positives (TP_i), False Positives (FP_i), True Negatives (TN_i) and False Negatives (FN_i) mean the number of correct or incorrect categories identified for the i-th class. $Precision_i = \frac{TP_i}{TP_i+FP_i}$, $Recall_i = \frac{TP_i}{TP_i+FN_i}$. N is the total number of samples and K is the number of sleep stages.

We also evaluated the running time of each network to choose an efficient and expeditious model. The average time for each model to run a fold is recorded as an evaluation reference.

D. Baseline Networks and Setup

In this experiment, we compared the proposed approach with several baseline networks with good performance, namely DeepSleepNet [10], SeqSleepNet [16] and SimpleSleepNet [20]. A brief description of these networks is given below.

- DeepSleepNet [10]: An architecture proposed in 2017 used for sleep staging, which consists of a multiscale CNN and an LSTM with shortcut residual connection. This structure combines the capabilities of two networks for feature extraction and temporal learning.
- SeqSleepNet [16]: A hierarchal bi-directional RNN structure. SeqSleepNet converts the raw EEG signal into power spectrum images by Short-time Fourier transform (STFT), which allows the signal to be characterized in both the time and frequency domains.
- SimpleSleepNet [20]: It consists of two bidirectional Gated Recurrent Unit structure. It also converts the raw EEG signal into power spectrum images and the channels and frequency of the power spectrum images are recombined after STFT. This network has few parameters and small hidden layer size so that it runs very fast.

To avoid serendipity as well as to accurately test the performance of the models, we took a 10-fold cross-validation approach for each model, on each dataset. In each cross-validation, we tested the models using the leave-one-subject-out method. We finally superimposed the results of 10 cross-validation tests as the final test results of the model. In addition, for the comparison of running times, we calculated the time to train one-fold for each model. We adopted the early stop method and terminated training when the validation set loss does not decrease for a consecutive period.

IV. Result and Discussion

A. Effect of Different Number of Segments

In order to investigate the effect of different number of segments on the final result, we conducted experiments on three different numbers of segments, SAN-0 (no segments), SAN-5 ($L = 5$, each segment length is 6s), SAN-10 ($L = 10$, each segment length is 3s) and SAN-15 ($L = 15$, each segment length is 2s) and then performed 10-fold cross-validation to evaluate the impact of segmentation on model performance.

As shown in Fig. 5, within a certain range, from SAN-0 to SAN-10, indicators of the model on three datasets, such

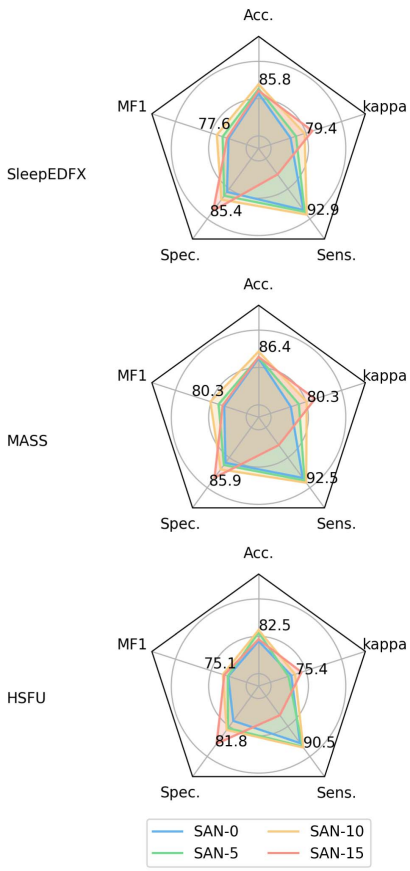

Fig. 5. Comparison with different number of segments on three datasets. The results of each model on different datasets are shown from top to bottom. Within a certain range, from SAN-0 to SAN-10, all the evaluation metrics of the model increase as the number of segments increases. Sequentially when the number of segments increases to a certain point, the performance of the model stabilizes and stops rising. The values of the evaluation indicators show the results of SAN-10.

as the accuracy, MF1 score and kappa coefficient, have increased steadily. As the number of segments increases, more regions will be divided and a relatively comprehensive result originating from these regions is provided. It plays a similar role to ensemble learning, where multiple submodules collaborate together to enhance the overall performance. However, segmentation with shorter duration may destroy the original morphological characteristics, and thus degrade the performance. This is why SAN-15 performs worse than SAN-10. It indicates that the appropriate segment length is also an important parameter. Besides, the running time and complexity of the model gradually increases as the number of segments increases. SAN-5 requires about four times the runtime of SAN-0, and SAN-10 requires about six times the runtime of SAN-0.

B. Effect of the Number of Heads in Multi-Head Attention

We explored the effect of the number of heads on the model performance in our experiments. With other parameters fixed, we will do the validation on the MASS dataset using

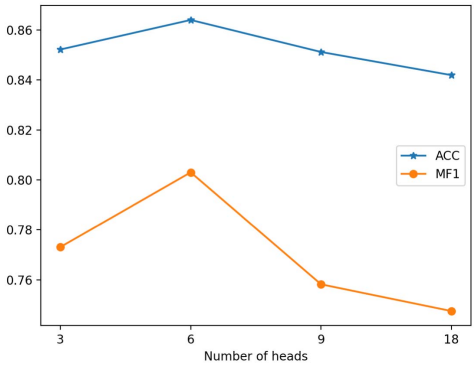

Fig. 6. The performance of SAN-10 on MASS dataset with different number of heads.

models with different number of heads. As shown in Fig. 6, the number of heads does not have a significant impact on the performance of the model, and the values vary only in a small range. However, it can be seen that as the number of head changes, a relatively good setting can be found, which will have some improvement on the model performance. While when the number of heads increases to 18, the model performance decreases a bit. In our experiments, we set the number of heads to 6 in SAN in order to accurately assess the impact of the segmentation we are interested in.

C. Hypnogram

Fig. 7 shows the hypnogram output using our proposed method as well as the real hypnogram and the posterior probability distribution per stage of sleep of a subject of the Sleep-EDFX dataset. It can be seen that the output hypnogram aligns very well with the corresponding ground truth. And the model discriminates the wrong sleep stage mostly in the stage of sleep stage transition. This result suggests that the transitioning sleep stages are much harder to correctly classified compared to the non-transitioning ones. The rationale is that the transitioning epochs often contain information of two or three sleep stages. Even with segmentation of the EEG signal to extract feature information, there are still difficulties in discriminating the sleep stages in the transitioning sleep stages. As a result, these present stages are active as indicated in the probability distribution in Fig 7. However, we need to pick one of them as the final discrete output label for the sleep staging task.

D. Compared With State-of-the-Art Approaches

We compared our proposed method with some state-of-the-art approaches. The accuracy, MF1, kappa coefficient, sensitivity, specificity and runtime of these methods were compared on three datasets.

As shown in the Table. I, compared with state-of-the-art methods, our proposed method obtains the best results on the Sleep-EDFX dataset and HSFU dataset, and only slightly inferior to SeqSleepNet on the MASS dataset. The reason why SAN is inferior to SeqSleepNet in MASS is attributed to the fact that the input to SeqSleepNet is multiple 30s EEG signals that capture the information of adjacent sleep stages.

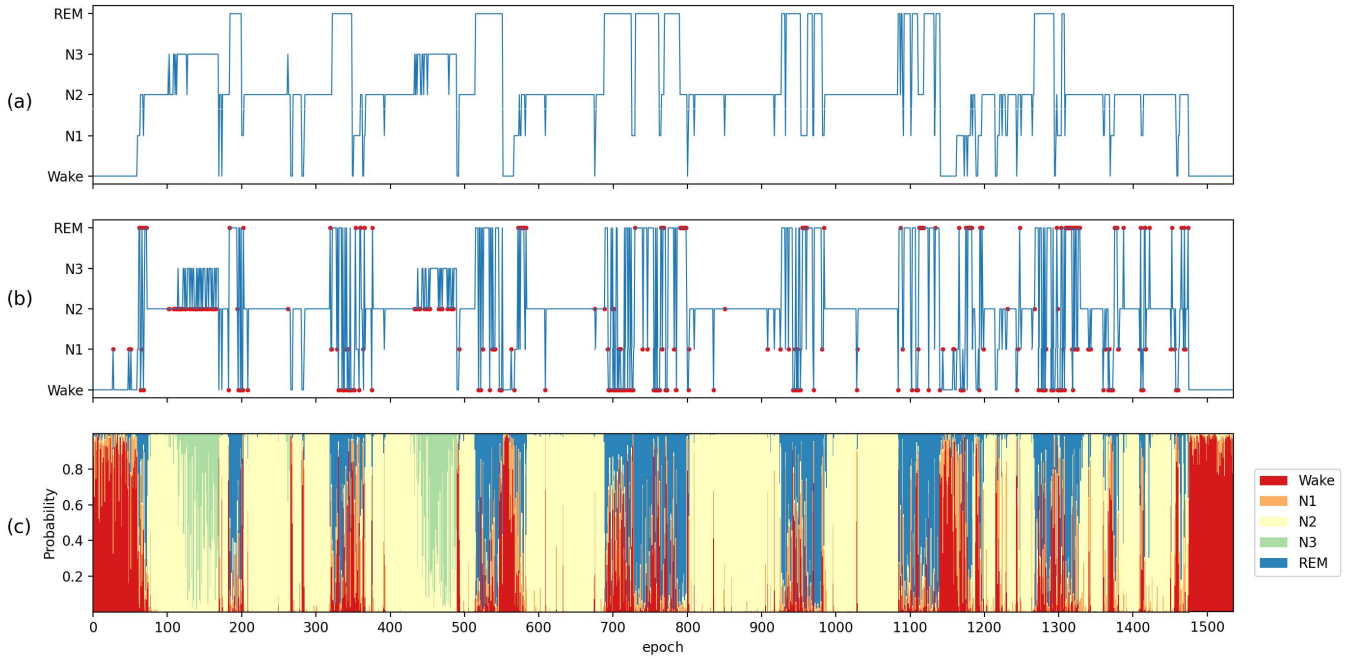

Fig. 7. Visualization of the hypnogram for one subject of Sleep-EDFX. (a) The ground-truth hypnogram; (b) the output hypnogram where · indicates the misclassified epochs; and (c) the probability output.

TABLE I

COMPARED WITH STATE-OF-THE-ART APPROACHES ON THE THREE DATASETS. IN ADDITION TO COMPARING METRICS SUCH AS ACCURACY AND F1 SCORE, THE AVERAGE TRAINING TIME FOR TRAINING AND TESTING DATASETS WERE ALSO COMPARED

	Method	Overall Metrics					Per-Class					Training (hours)	Testing (seconds)
		Acc	MF1	κ	Spec	Sen	W	N1	N2	N3	REM		
Sleep-EDFX	DeepSleepNet [10]	84.5	77.3	83.9	91.2	76.7	70.7	46.2	91.3	89.6	85.8	7.7	2.6
	SeqSleepNet [16]	85.0	79.0	84.7	91.4	78.5	83.3	23.0	**93.2**	90.4	**92.2**	15.1	42.1
	SimpleSleepNet [20]	83.7	**79.3**	83.1	91.3	**79.9**	92.9	**53.3**	84.4	80.6	88.3	10.1	4.8
	SAN (proposed method)	**85.8**	77.6	**85.4**	**92.9**	79.4	**94.5**	37.0	84.5	**96.8**	84.1	**5.7**	**1.4**
MASS	DeepSleepNet [10]	86.1	77.9	85.1	92.1	77.0	81.7	30.2	87.4	**89.0**	**96.5**	5.7	2.8
	SeqSleepNet [16]	**87.9**	**84.8**	**87.5**	**92.7**	**83.6**	93.0	59.4	93.8	80.6	91.0	10.6	44.4
	SimpleSleepNet [20]	84.0	77.6	83.3	90.7	76.9	76.6	46.2	87.4	81.6	92.3	7.7	4.9
	SAN	86.4	80.3	85.9	92.5	80.3	**98.9**	40.3	90.5	81.0	90.7	**4.1**	**1.8**
HSFU	DeepSleepNet [10]	79.1	74.1	78.2	88.3	72.4	77.8	43.5	87.8	63.9	88.9	1.87	2.4
	SeqSleepNet [16]	79.1	72.9	78.6	88.4	72.9	88.0	21.9	88.8	78.2	87.6	5.3	41.6
	SimpleSleepNet [20]	76.8	72.8	75.9	87.2	72.9	80.0	**46.3**	79.4	69.4	**89.4**	4.7	4.8
	SAN	**82.5**	**75.1**	**81.8**	**90.5**	**75.4**	**94.3**	25.9	**89.1**	**85.3**	82.5	**1.5**	**1.4**

Specifically, we observed that SAN is more accurate than other methods in determining wake, N2 and N3 stages. This is made possible by the feature extraction for segmentation of EEG signals and the learning of temporal information by TSE in SAN. However, our proposed method is somewhat less accurate in discriminating the N1 and REM stages. In terms of running time, our proposed method requires the least amount of time, regardless of whether it is SAN-5 or SAN-10. DeepSleepNet, SeqSleepNet and SimpleSleepNet can only run serially during the runtime because of the RNN structure used, which greatly increases the runtime. In contrast, our proposed network only processes the time-domain data and uses CNN combined with attention mechanisms instead of RNN, so it occupies less time. The results show that this segmented attention mechanism is superior to other algorithms in terms of running time.

In out proposed model, the signal of an epoch is divided into different segments, and different segments may have different features of sleep stages. Along with these features, our proposed network model integrates them to output a decision in which all segments contribute, and thus a fairly robust performance can be obtained. If the duration of certain feature is short, the contribution of that segment may be overwritten by other segments. Despite the attention mechanism adopted to try to solve this problem, more satisfactory results are still not obtained for the N1 period. The proposed SAN obtained good results in tests on all three datasets. Although the SAN is slightly less effective than SeqSleepNet on the MASS

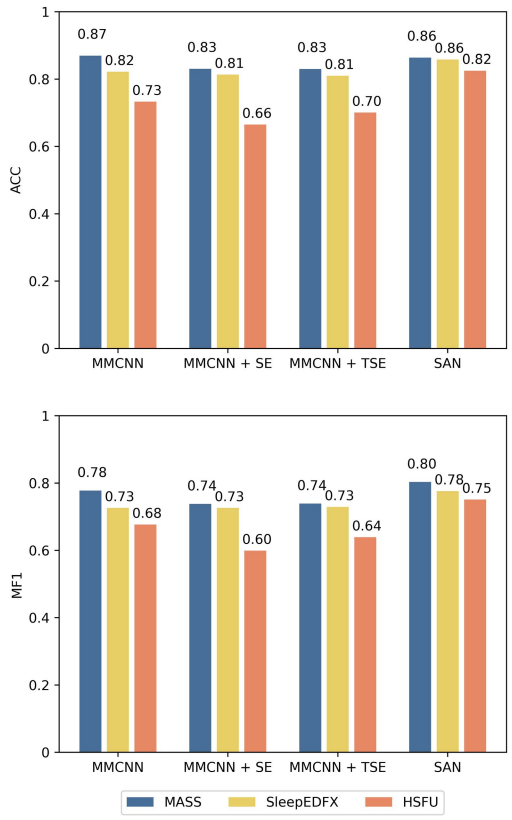

Fig. 8. Ablation study conducted on the three datasets.

dataset, it is noteworthy that SAN can significantly improve the discrimination accuracy of wake, N2 and N3 stages compared to other approaches, especially in HSFU clinical dataset. The excellent performance of the SAN in these stages makes it potentially useful for the diagnosis and prevention of a number of sleep disorders. It is unlikely that SAN trained directly on MASS, Sleep-EDFX, and HSFU would work well for recording sleep disorders because the structure and features of the data samples are different. However, we can use SAN to train on the dataset of sleep disorders or fine-tune transfer learning based on MASS, Sleep-EDFX, HSFU datasets. Whereas the short running time of SAN provides the basis for relatively fast training and transfer learning on new dataset.

E. Ablation Study

As shown in Fig. 8, we present an ablation study conducted on three datasets to analyze the effectiveness of each module in our SAN. Our proposed SAN consists of MMCNN, residual SE block and TSE modules. Specifically, we derive four model variants as follows.

1) MMCNN: MMCNN module only.
2) MMCNN+residual SE block: MMCNN and residual SE block without TSE.
3) MMCNN+TSE: MMCNN and TSE without residual SE block.
4) SAN: MMCNN, residual SE block and TSE.

By comparing the results in Fig. 8, we can conclude the following points. First, adding either the SE block or the TSE module alone after the MMCNN leads to some degree of performance degradation. It is difficult for the network to learn the deep features and the connections between these features when using only one of the module. Second, by combining the residual SE block and the TSE module, the model can further improve its performance, and the network can be more efficient by obtaining deep features and internal associations. The results on three datasets illustrate the importance of the combination of these modules.

V. Conclusion

In this paper, we proposed a novel architecture called SAN for sleep stage classification by single EEG channel. We used multiple multiscale CNN for feature extraction of different segments of EEG signal, applied residual squeeze and excitation block to enhance the feature and assigned weights to the features in different regions based on the multi-head attention mechanism. In addition, we added noise to the raw EEG signal for data augementation to solve the class imbalance problem. The method improved the system performance by making decisions based on each segment feature in an integrated manner. The proposed method performed well on two public datasets and one clinical dataset. We compared it with recent state-of-the-art researches and demonstrate the effectiveness of the algorithm. The results showed that our proposed method is competitive and can obtain a better performance on the sleep stage classification. In future work, the idea of object detection could be used to clearly locate the features at different locations in the signal segment, thus achieving higher accuracy identification and facilitating the diagnosis of related sleep disorders.

References

[1] F. S. Luyster, P. J. Strollo, P. C. Zee, and J. K. Walsh, "Sleep: A health imperative," *Sleep*, vol. 35, no. 6, pp. 727–734, Jun. 2012.
[2] M.-P. St-Onge et al., "Sleep duration and quality: Impact on lifestyle behaviors and cardiometabolic health: A scientific statement from the American Heart Association," *Circulation*, vol. 134, no. 18, pp. e367–e386, Nov. 2016.
[3] N. Chanchlani, "Health consequences of shift work and insufficient sleep," *Brit. Med. J.*, vol. 355, Feb. 2017, Art. no. i6599.
[4] O. Itani, M. Jike, N. Watanabe, and Y. Kaneita, "Short sleep duration and health outcomes: A systematic review, meta-analysis, and meta-regression," *Sleep Med.*, vol. 32, pp. 246–256, Apr. 2017.
[5] G. Medic, M. Wille, and M. Hemels, "Short- and long-term health consequences of sleep disruption," *Nature Sci. Sleep*, vol. 9, pp. 151–161, May 2017.
[6] E. Tobaldini, E. M. Fiorelli, M. Solbiati, G. Costantino, L. Nobili, and N. Montano, "Short sleep duration and cardiometabolic risk: From pathophysiology to clinical evidence," *Nature Rev. Cardiol.*, vol. 16, no. 4, pp. 213–224, Apr. 2019.
[7] S. A. Keenan, "An overview of polysomnography," in *Handbook of Clinical Neurophysiology*, vol. 6. Amsterdam, The Netherlands: Elsevier, 2005, pp. 33–50.
[8] J. A. Hobson, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects: A. Rechtschaffen and A. Kales (Editors). (Public Health Service, U.S. Government Printing Office, Washington, D.C., 1968, 58 p.)," *Electroencephalogr. Clin. Neurophysiol.*, vol. 26, no. 6, p. 644, 1969.
[9] C. Iber, S. Ancoli-Israel, A. Chesson, and S. Quan, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Westchester, IL, USA: American Academy of Sleep Medicine, Jan. 2007.

[10] 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.

[11] 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 Informat.*, vol. 24, no. 5, pp. 1351–1366, May 2020.

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

[13] Y. Liao, C. Zhang, M. Zhang, Z. Wang, and X. Xie, "LightSleepNet: Design of a personalized portable sleep staging system based on single-channel EEG," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 1, pp. 224–228, Jan. 2022.

[14] 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.

[15] R. Casal, L. E. Di Persia, and G. Schlotthauer, "Classifying sleep–wake stages through recurrent neural networks using pulse oximetry signals," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102195.

[16] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, 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.

[17] 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.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.

[19] H. Phan, O. Y. Chen, M. C. Tran, P. Koch, A. Mertins, and M. De 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.

[20] A. Guillot, F. Sauvet, E. H. During, and V. Thorey, "Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 9, pp. 1955–1965, Sep. 2020.

[21] 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.

[22] 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, Apr. 2018.

[23] 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.

[24] 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.

[25] G. Zhu, Y. Li, and P. Wen, "Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 6, pp. 1813–1821, Nov. 2014.

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

[27] B. Yang, X. Zhu, Y. Liu, and H. Liu, "A single-channel EEG based automatic sleep stage classification method leveraging deep one-dimensional convolutional neural network and hidden Markov model," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102581.

[28] 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.

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

[30] S. Pathak, C. Lu, S. B. Nagaraj, M. van Putten, and C. Seifert, "STQS: Interpretable multi-modal Spatial–Temporal-seQuential model for automatic Sleep scoring," *Artif. Intell. Med.*, vol. 114, Apr. 2021, Art. no. 102038.

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

[32] G. Tang, M. Müller, A. Rios, and R. Sennrich, "Why self-attention? A targeted evaluation of neural machine translation architectures," 2018, *arXiv:1808.08946*.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.

[35] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[36] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762*.

[37] 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.

[38] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal archive of sleep studies: An open-access resource for instrument benchmarking and exploratory research," *J. Sleep Res.*, vol. 23, no. 6, pp. 628–635, Jun. 2014.