MS-HNN: Multi-Scale Hierarchical Neural Network With Squeeze and Excitation Block for Neonatal Sleep Staging Using a Single-Channel EEG

Hangyu Zhu, Laishuan Wang, Ning Shen, Yonglin Wu, Shu Feng,
Yan Xu, Chen Chen, and Wei Chen, *Senior Member, IEEE*

Abstract— **Most existing neonatal sleep staging approaches applied multiple EEG channels to obtain good performance. However, it potentially increased the computational complexity and led to an increased risk of skin disruption to neonates during data acquisition. In this paper, a multi-scale hierarchical neural network (MS-HNN) with a squeeze and excitation (SE) block for neonatal sleep staging is presented in this study on the basis of a single EEG channel. MS-HNN composes of multi-scale convolutional neural network (MSCNN), temporal information learning (TIL) module, and squeeze and excitation (SE) block. MSCNN can extract features from different scales and frequencies, and TIL module can acquire the transition information among adjacent stages. In addition, for these extracted features, SE block can selectively concentrate on informative features and weaken redundant features for achieving better performance. The proposed approach was validated on a clinical dataset involving 64 neonates from the Children's Hospital of Fudan University (CHFU). The proposed network achieves an accuracy of 75.4% and 76.5% for three-class automatic neonatal sleep staging with the single-EEG channel and the eight-EEG channels, respectively. The experimental results show that the proposed method can maintain good performance by making full use of the information in the single channel while reducing the channels to control the computational overhead.**

Index Terms— **Neural network, neonatal sleep staging, squeeze and excitation block, single channel.**

Hangyu Zhu, Ning Shen, and Yonglin Wu are with the Center for Intelligent Medical Electronics, School of Information Science and Technology, Fudan University, Shanghai 200433, China (e-mail: 20110720076@fudan.edu.cn; 20210720085@fudan.edu.cn; 20210720281@fudan.edu.cn).

Laishuan Wang is with the Department of Neonatology, Children's Hospital of Fudan University, Shanghai 201102, China (e-mail: laishuanwang@163.com).

Shu Feng 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).

Yan Xu is with the Department of Neurology, National Children's Medical Center, Children's Hospital of Fudan University, Shanghai 201102, China (e-mail: xuyanfimmu@163.com).

Chen Chen is with the Human Phenome Institute, Fudan University, Shanghai 201203, China (e-mail: chenchen_fd@fudan.edu.cn).

Wei Chen is with the Center for Intelligent Medical Electronics, School of Information Science and Technology, Fudan University, Shanghai 200433, China, and also with the Human Phenome Institute, Fudan University, Shanghai 201203, China (e-mail: w_chen@fudan.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3266876

I. INTRODUCTION

S LEEP, as the predominant activity for neonates, contributes to a crucial role in brain maturation, cognitive and psychical development [1], [2], [3], [4]. Neonates exhibit ultradian rhythms of intermittent repetition in sleeping and waking within 24 hours, and gradually shift to circadian rhythms with a 24-h sleep-wake cycle [5], [6]. Early manifestations of sleep-wake cycling can be observed in electroencephalography (EEG) around 27 weeks postmenstrual age (PMA), and then developed into distinct electrophysiological patterns, which can be observed on the signals recorded by video EEG or polysomnography (PSG) [7]. After 28 weeks PMA, neonatal sleep can generally be divided into active sleep (AS), quiet sleep (QS), and wake stage [8], [9], [10]. Meanwhile, at the onset of term-age, normally after 36 weeks PMA, AS and QS can be further divided into AS I, AS II, and QS I, QS II, respectively. Among these sleep stages, the wake stage in neonatal sleep can be equivalent to the wake stage in adult sleep, QS can be equivalent to the non-rapid eye movement (NREM) sleep stage, and AS can be equivalent to the rapid eye movement (REM) sleep stage [11]. Differentiating the neonatal sleep stages can provide a quantitative evaluation of sleep and assist the assessment of neurological and brain development. However, sleep stage scoring is usually performed by experienced physicians via observing the characteristics of the EEG signals, which is very tedious and laborious. To relieve

the burdens of manual sleep scoring and to provide a more objective assessment of sleep stages, automatic sleep staging methods are proposed.

The existing automatic sleep staging methods can be divided into traditional machine learning-based methods and deep learning-based methods. The traditional machine learning-based methods mainly consist of hand-crafted feature extraction and sleep staging classification using traditional machine learning models such as multi-layer perception machine (MLP) [12], support vector machine (SVM) [13], [14], hidden Markov models (HMM), Gaussian mixture model (GMM) [15], cluster-based adaptive method [16], and etc. Traditional machine learning-based methods are highly dependent on the hand-crafted features, however, the extraction of the features is subjective, which may limit sleep staging performance. In contrast, deep learning-based methods can automatically extract features through networks without medical knowledge, providing a more convenient and accurate method to stage the sleep of neonates. In recent years, Ansari et al. proposed an 18-layer convolutional neural network (CNN) for QS stage detection using multi-channel neonate EEG signals [17]. Afterwards, Ansari et al. optimized the network structure by changing the parameters of the convolutional layers and the depth of the network to better extract the features in the signal [18]. In 2021, Ansari et al. proposed a CNN network combined with Sinc block, which can slightly expand the feature scale to enhance the features to detect QS sleep stage with limited EEG channels [19]. Ghimatgar et al. applied a modified graph clustering ant colony optimization (MGCACO) to extract the feature, and then used a bidirectional-long short term memory (Bi-LSTM) combined with HMM to perform sleep staging using multi-channel EEG signals [20]. Fraiwan et al. used the Bi-LSTM to extract the temporal information in the sleeping signals and stage the sleep of full-term neonates [21].

However, most of existing studies have used multiple channels of EEG signals as the input of network. Although multi-channel EEG signals can improve the performance of the model, additional electrodes may lead to an increased potential risk of skin disruption and discomfort to neonates during acquisition. Moreover, multiple channels of the same modality of signal can result in redundancy of features. The similar features can increase the computational overhead of the model and affect the performance of the model. Additionally, the existing studies focus more on the detection of QS stage and two-class task. For neonates, wake, AS and QS stages are all essential for the growth and development of the neonate. The QS stage should not be separated from the other two stages. Sleep is a continuous process, with a temporal relationship between sleep stages. A whole analysis of sleep can also yield more informative information than only detecting QS stages.

To address the aforementioned issues, a multi-scale hierarchical neural network architecture combined with squeeze and excitation (SE) block is proposed. The network consists of multi-scale CNN (MSCNN), SE block, and temporal information learning (TIL) module. MSCNN has two branches with different size convolution kernels to extract features at different scales. The TIL module, which is used to learn the

temporal information from the extracted features, composes of a bidirectional gated recurrent unit (Bi-GRU) [22] with residual structure [23]. The SE block [24] is used to reduce the impact of the redundant features by mapping transformation. It can enhance the informative features and weaken less redundant features. We conducted the experiment on a clinical dataset collected in the Children's Hospital of Fudan University (CHFU). The results indicated that the proposed method maintains good performance in the case of limited channels. The main contribution of this paper can be summarized below:

1. A novel automatic neonatal sleep staging model structure including MSCNN, SE block, and TIL is proposed. A multi-scale feature architecture is presented to adequately extract more informative features at different scales and frequencies from the signal. Meanwhile, temporal information among adjacent stages has been fully considered through a sequence processing model in both forward and backward directions.

2. SE blocks that can selectively concentrate on informative features and weaken less informative features are fused into the hierarchical neural network architecture. It can implicitly reduce feature redundancy and further enhance the performance.

3. To minimize skin disruption and discomfort to neonates during signal acquisition as well as reduce the computational overhead during signal processing, an optimal limited EEG channel is selected by comparing all available EEG channels. In addition, comparisons of performance with one-channel EEG, two-channel EEG and eight-channel EEG are fully investigated to further verify the reliability of the proposed approach.

The remainder of this paper is organized as follows. Section II introduces the method for sleep staging. The specific experiment process and results of the sleep staging are described in Section III. The results are discussed in Section IV. Finally, the conclusion of this paper is given in Section V.

II. METHODS

In this paper, we propose a multi-scale hierarchical neural network (MS-HNN) combined with SE block for neonatal sleep staging. It composes of MSCNN, TIL module, and SE block. The architecture of the network is shown in Fig.1. The following subsections describe the detailed functions of each module in the network architecture.

A. Multi-Scale CNN

The feature extraction capability of CNN has been proved in many studies [25], [26]. Several sleep staging methods for neonates have attempted to extract features using a single-scale CNN network [17], [18]. This demonstrates the feasibility of CNN for neonatal sleep feature extraction. However, a single-scale CNN is only able to extract limited feature information. Inspired by some existing studies [27], [28], [29], a CNN with two different scale branches is used to extract adequate features from different frequencies, as shown in Fig.2. The two branches of CNN extract information on different time and frequency scales in the signal, respectively. Specifically,
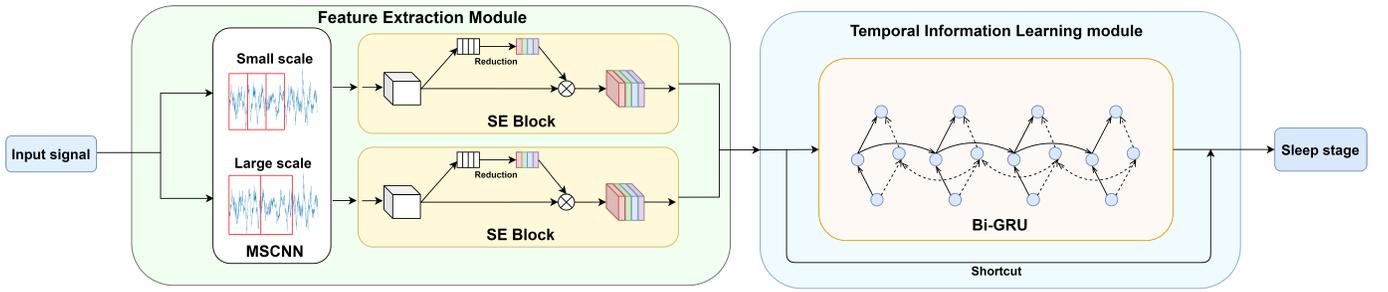
Fig. 1. The overall framework of MS-HNN. The preprocessed EEG signal is first processed by MSCNN at multiple scales for feature extraction. The extracted features are reinforced by SE blocks and then sent to the TIL module for temporal feature learning. The predicted sleep stage is finally output by softmax.
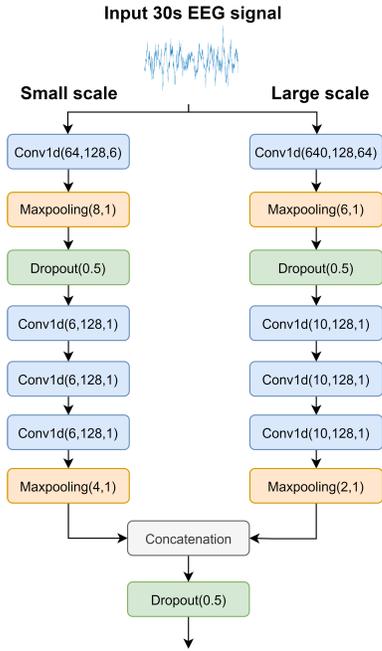
Fig. 2. The structure of MSCNN. MSCNN consists of two scales of CNN. The two scales extract features from different time and frequency spans to make full use of the information in the EEG signals.

Fig. 3. The structure of SE block. The SE block mainly consists of an average pooling layer and two fully connected layers. It reassigns feature weights through these linear mappings.

the coarse-scale CNN branch has a convolutional kernel size of 640 and intercepts the signal with a window length of 5 seconds, extracting low frequency features down to 0.2 Hz. The fine-scale CNN branch has a convolutional kernel size of 64 and intercepts the signal with a window length of 0.5 seconds, extracting features at high frequencies down to 2 Hz. The features of the two branches are then reshaped and concatenated to finally obtain more condensed features. MSCNN expands the limited EEG channels using convolution to obtain multiple subspace feature channels, while effectively utilizing the feature information in the limited EEG channels. Different feature channels contain different scales of sleep information, which provides the basis for feature selection and accurate sleep staging. MSCNN, as the feature extraction module of MS-HNN, sets the tone of MS-HNN performance. Adequate and effective features allow the MS-HNN to perform automatic neonatal sleep staging more accurately. In addition, the other modules of MS-HNN are based on the features extracted by MSCNN for feature optimization and feature

learning. In general, MSCNN is the foundation of MS-HNN for accurate sleep staging.

B. Squeeze and Excitation Block

The SE block aims to reassign channel feature weights and prevent the vanishing of gradients as the network layers deepen [24]. It calculates the interrelationships between channels and corrects the distribution of weights between channels. SE block is used to optimize the extracted feature weights in this paper. The structure of the SE block is shown in Fig.3. We assume the input of the SE block is $X \in \mathbf{R}^{H \times W \times C}$, which is the output of one branch of the MSCNN. Then the global pooling is used to reduce the dimension of features, changing the input X to $X' \in \mathbf{R}^{1 \times 1 \times C}$. Two fully connected (FC) layers and one ReLU layer are applied to parameterize the pass selection mechanism, enhancing the informative features and weakening the less informative features. The feature weights are then activated and selected by the sigmoid layer. Finally, the optimized feature weights are multiplied by the initial features to obtain the optimized features. The process of SE block can be described as the

following equation:

$$X' = AvgPooling(X) \in \mathbf{R}^{1 \times 1 \times C} \qquad (1)$$

where the $AvgPooling(\cdot)$ is the average global pooling. Then the compressed features will be reassigned weights.

$$S = \sigma(F2(ReLU(F1(X')))) \in \mathbf{R}^{1 \times 1 \times C} \qquad (2)$$

where the $F1(\cdot)$ means the first FC layer, the $F2(\cdot)$ means the second FC layer, the $ReLU(\cdot)$ means the ReLU activation function and the $\sigma(\cdot)$ means the sigmoid activation function. Finally, the feature weights are reassigned by matrix multiplication:

$$\hat{X} = S_{1,1,k} \cdot X_{i,j,k} \qquad (3)$$

where $i = 1, 2 \ldots, H$, $j = 1, 2 \ldots, W$ and $k = 1, 2 \ldots, C$. The SE block obtains the weight information of each layer feature by pooling and nonlinear mapping, and re-optimizes the feature weights based on the weight information. This can be regarded as an attention selection mechanism on the feature dimension. Thus, the SE block can enhance informational characteristics.

C. Temporal Information Learning (TIL) Module

Sleep is a continuous life activity [30], [31]. This means that behaviors during sleep are related rather than independent of each other. Therefore, the temporal information during sleep is useful for sleep scoring, which has been proved in many studies [32], [33]. In this paper, TIL module is proposed for temporal information extraction. Its structure is shown in the Fig.1. The TIL module consists of Bi-GRU with residual structure. Bi-GRU is a type of recurrent neural network (RNN), which has a great ability for temporal information learning. Considering that RNN runs serially, the residual structure is added to accelerate the learning process of the model [23] as well as prevent gradient loss. Assuming that the input of TIL module is F, the output of GRU network and TIL module can be described as follow.

$$H = GRU(F) \qquad (4)$$
$$Y = H + F \qquad (5)$$

where $GRU(\cdot)$ is GRU network, H is the output of the GRU network, and Y is the final probability of the sleep stage. The specific parameters and output sizes for each layer of the network are shown in Table I.

D. Evaluation Metrics

The Leave One Subject Out (LOSO) approach is performed to verify the performance of the proposed method. Accuracy (ACC), F1 score, and Kappa coefficient are adopted to evaluate the experimental results. These metrics are commonly used to evaluate multi-classification problems [25], [34]. Given true positive (TP_i), true negative (TN_i), false positive (FP_i) and false negative (FN_i) for i-th class, along with a total sample size of N, then these parameters can be calculated according to the following equation:

$$ACC = \frac{\sum_i^M TP_i}{N} \qquad (6)$$

TABLE I
THE LAYERS OF THE PROPOSED NETWORK

Layer/Block	Parameters	Input shape	Output shape
Input	-	-	3840*1*1
Conv1	N: 128, k: 64, s: 6	3840*1*1	640*1*128
Maxpool1	k:8, s: 8	640*1*128	80*1*128
Dropout1	α=0.5	80*1*128	80*1*128
Conv1	N: 128, k: 6, s: 1	80*1*128	80*1*128
Conv1	N: 128, k: 6, s: 1	80*1*128	80*1*128
Conv1	N: 128, k: 6, s: 1	80*1*128	80*1*128
Maxpool1	k: 4, s: 4	80*1*128	20*1*128
SE_block1	M: 128, r: 16	20*1*128	20*1*128
Flatten1	-	20*1*128	2560
Conv2	N: 128, k: 640, s: 64	3840*1*1	60*1*128
Maxpool2	k: 6, s: 6	60*1*128	10*1*128
Dropout2	α=0.5	10*1*128	10*1*128
Conv2	N: 128, k: 10, s: 1	10*1*128	10*1*128
Conv2	N: 128, k: 10, s: 1	10*1*128	10*1*128
Conv2	N: 128, k: 10, s: 1	10*1*128	10*1*128
Maxpool2	k: 2, s: 2	10*1*128	5*1*128
SE_block2	M: 128, r: 16	5*1*128	5*1*128
Flatten2	-	5*1*128	640
Concat	-	2560+640	3200
Dense	-	3200	800
Dense	-	800	400
Bi-GRU	l: 2, h: 200, sl:15	400	400
Dropout	α=0.5	400	400
Dense	-	400	200
Dense	Softmax	200	3

[1] Conv1, Maxpool1, SE_block1: represents the convolutional layer, the pooling layer and the SE block in the first CNN branch, Conv2, Maxpool2, SE_block2: represents the convolutional layer, the pooling layer and the SE block in the second CNN branch.

[2] N: number of filters in each convolution, k: kernel size, s: stride, M: number of channels in SE block, r: number of reduction channels in SE block, l: number of layers in Bi-GRU, h: number of hidden layer neurons of Bi-GRU, sl: sequence length in Bi-GRU

$$F1 = \frac{1}{M} \sum_i^M \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \qquad (7)$$

$$Kappa = \frac{ACC - P_e}{1 - P_e} \qquad (8)$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \qquad (9)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \qquad (10)$$

$$P_e = \sum_i^M \frac{(TP_i + TF_i) \times (TP_i + FP_i)}{N^2} \qquad (11)$$

where M is the number of the class of sleep stages. Meanwhile, we calculate the precision for each sleep stage to evaluate the results. We trained the model under the TensorFlow 3.6 environment. The optimizer was Adam. The loss function was a weighted cross-entropy function, which can alleviate imbalance on sleep data and achieve a favorable performance [34].

Indicators	Details
Gender (B/G)	32 / 32
GA (wk+d)	38.3 ± 1.8
PMA (wk+d)	40.5 ± 1.7
Weight (kg)	3.3 ± 0.6
Sleep duration (min)	86.5 ± 34.4
Wake duration (min)	43.0 ± 34.5
Reason for inclusion	Hyperbilirubinemia, sepsis, and etc.

[1] G: girl, B: boy, GA: gestational age, PMA: postmenstrual age during data collection, S: sleep, W: wake, wk+d: the number of weeks plus the number of days.

Fig. 4. SE blocks are placed at different positions to explore the optimal feature optimization position. (a) SE blocks are placed after each CNN branch. (b) The SE block is placed after CNN features of two scales are concatenated.

III. EXPERIMENT AND RESULTS

In this paper, the CHFU dataset is used to evaluate the performance of the proposed model. Firstly, the impact of SE block placement on neonatal sleep staging results is investigated. Secondly, we perform sleep staging using different numbers of channels. The results of sleep staging with limited channels are compared with the results with multiple channels. In addition, we compare the effect of different acquisition positions of channels on sleep staging results in the single-channel case. Finally, we compare the proposed method with some baselines and state-of-art methods. In the following subsections, the experiment details and specific results are depicted.

A. Dataset

The CHFU dataset consists of sleep recordings of 64 neonates from the Children's Hospital of Fudan University during 2017-2018. The research ethics committee of the Children's Hospital of Fudan University approved this study (approval No. (2017) 89). These neonates range in PMA from 36 to 43 weeks. And they suffer from different types of diseases such as bloating, hyperbilirubinemia, jaundice, pneumonia, and etc. The recordings include channels F3, F4, C3, C4, P3, P4, T3, T4, and the reference channel Cz. In addition, electrooculography (EOG) signals, electromyography (EMG) signals, and electrocardiograph (ECG) signals are also recorded. These signals are recorded by a Nicolet device at a sampling rate of 500 Hz. Based on these physiological signals, professional doctors classify the sleep of neonates into three sleep stages: wakefulness, QS, and AS. All details about the dataset are shown in Table II.

B. Data Preparation

In this paper, a 50Hz notch filter and a 0.3-35Hz band-pass filter are applied to remove the interference signal. Then the signals are downsampled to 128Hz, normalized to zero mean and standardized to standard deviation of one to minimize individual variability. Afterwards, we divide the signal into 30-s epochs [9], [11], [35]. To verify the performance of the proposed method in the limited channel case, one-channel signal (P4), two-channel signal (P3, P4), and eight-channel
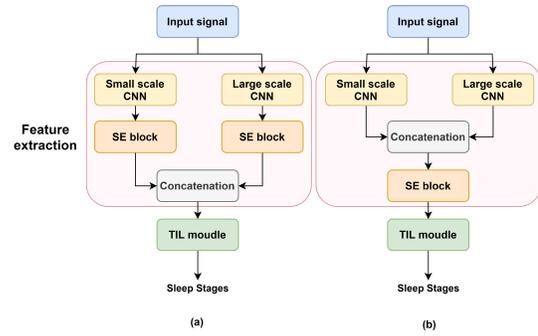
signal (F3, F4, C3, C4, P3, P4, T3, T4) are used as input signals for model training and validation, respectively. The selection of channels is based on previous research as well as experience [9], [19]. Moreover, we also explore the effect of electrode position of the one-channel signal on sleep staging results.

C. The Effect of SE Block Placement

To explore a more optimal location for feature optimization, the SE block is placed after each CNN branch and after the feature concatenation, respectively. The specific location is shown in Fig.4. These two placement methods present the optimization at each scale and the optimization after the concatenation. Table III shows the sleep staging results without SE block (SE_non), with SE block after each CNN branch (SE_each), and with SE block after feature concatenation (SE_con) for the one-channel, two-channel and eight-channel cases.

The addition of SE blocks has definite improvement on the overall sleep staging results in both placement methods, compared with the results without SE block. Additionally, the SE block placed after each branch have the best sleep staging performance. Specifically, in the one-channel case, adding the SE block after each branch improves the sleep staging accuracy from 73.6% to 75.4%, F1 score from 0.737 to 0.758, and Kappa from 0.705 to 0.728. In the two-channel case, adding the SE block after each branch improves the sleep staging accuracy from 74.4% to 75.9%, F1 score from 0.745 to 0.760 and Kappa from 0.715 to 0.731. In the eight-channel case, adding the SE block after each branch improves the sleep staging accuracy from 75.7% to 76.5%, F1 score from 0.757 to 0.763 and Kappa from 0.728 to 0.735. Moreover, adding the SE block after each branch results in significant improvement in the classification accuracy for each sleep stage. However, when the SE block is added after the feature concatenation, the accuracy improvement for each sleep stage is unstable. This phenomenon may be related to the way in which the features are concatenated. The concatenation of features is directly downscaled through the fully concatenated layer. After two different scales of features are concatenated, the interdependence information of features becomes more complex. This increases the difficulty of feature analysis by

TABLE III

SLEEP STAGING RESULTS WITHOUT SE BLOCK (SE_NON), WITH SE BLOCK AFTER EACH CNN BRANCH (SE_EACH), AND WITH SE BLOCK AFTER FEATURE CONCATENATION (SE_CON) FOR THE ONE-CHANNEL, TWO-CHANNEL AND EIGHT-CHANNEL CASES

Channel numbers	SE block	Acc	F1 score	Kappa	Pre (W)	Pre (AS)	Pre (QS)	Avg time
	SE_non	0.736	0.737	0.705	0.738	0.669	0.802	6.43 hrs
One-channel (P4)	SE_each	**0.754**	**0.758**	**0.728**	0.737	**0.670**	**0.878**	6.45 hrs
	SE_con	0.745	0.747	0.715	**0.776**	0.649	0.820	6.46 hrs
	SE_non	0.744	0.745	0.715	0.735	0.677	0.823	8.87 hrs
Two-channel (P3, P4)	SE_each	**0.759**	**0.760**	**0.731**	**0.754**	0.701	0.822	8.89 hrs
	SE_con	0.753	0.754	0.725	**0.731**	**0.702**	**0.830**	8.89 hrs
	SE_non	0.757	0.757	0.728	0.764	0.710	0.792	12.08 hrs
Eight-channel (F3, F4, C3, C4, P3, P4, T3, T4)	SE_each	**0.765**	**0.763**	**0.735**	**0.773**	**0.725**	0.789	12.11 hrs
	SE_con	0.761	0.762	0.734	0.767	0.719	**0.800**	12.10 hrs

[1] Pre: precision, Avg time : Average training time / fold.
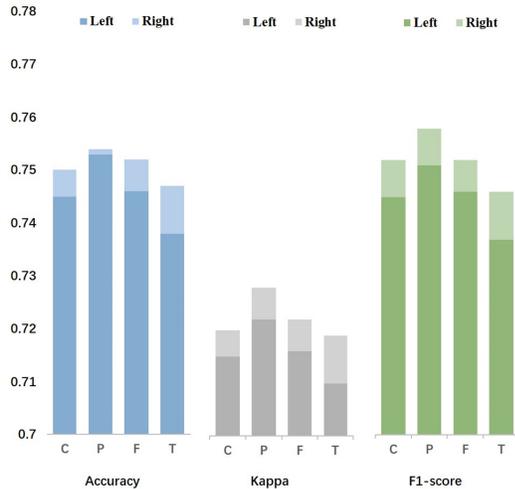
Fig. 5. Results of automatic sleep staging using different channels in the one-channel case (F3, F4, C3, C4, P3, P4, T3, T4), where the left side indicates the electrode located on the left side of the head (F3, C3, P3, T3) and the right side indicates the electrode located on the right side of the head (F4, C4, P4, T4).

the SE block, leading to a decrease in the effectiveness of feature optimization. Therefore, the SE block can well optimize the weight distribution of features. Additionally, a better placement for the SE block is after each branch, where the SE block can better capture the interdependence between features.

D. The Effect of Channel Location in Limited Channel Cases

In this experiment, the effect of different EEG channel locations on the results is also explored in the case of one-channel EEG signal. EEG signals from all eight channels (F3, F4, C3, C4, P3, P4, T3, T4) are fed separately into the network which has SE blocks after each CNN branch (SE_each) for automatic sleep staging. Fig.5 shows the results of different EEG channels on automatic sleep staging. The effect of different channels on the results is not significant, which also proves the robustness of our proposed method. Consistent with previous studies [9], sleep staging using EEG channels closer to the center (such as P, C, and F) obtained more accurate results compared to the farther channel like T.

TABLE IV

METRICS WITH DIFFERENT NUMBER OF CHANNELS

Channel Number	1	2	8
Average time	6.45 hrs	8.89 hrs	12.11 hrs
Mutual Information	0.635	0.643	1.000
Model Complexity	$O(3.77 \times 10^8)$	$O(7.62 \times 10^8)$	$O(3.2 \times 10^9)$

[1] Mutual Information: the mutual information of the features extracted in present channel case and the features extracted in the eight-channel case.

Among the three channels P, F, C near the central location of the brain, the P channel achieved the highest sleep staging performance. As proved in many studies, the maturity of the temporo-parietal junction can be used as one of the criteria to assess the degree of development of the neonatal brain [36]. The temporo-parietal junction has an important growth significance in the brain and contains more information. The location of the P-channel acquisition is close to the temporo-parietal junction, and presumably, the P-channel may contain more information for neonatal sleep staging. Meanwhile, using the right EEG channel (F4, C4, P4, T4) acquired better results than the left EEG channel (F3, C3, P3, T3) when the distance to the center is similar. Although the biological mechanisms underlying early neonatal brain development are currently unclear, the complex microstructural changes, such as myelination, increases in dendritic arborization, axonal elongation and thickening, synaptogenesis, etc., are key processes in brain development [37], [38]. One study using tomographic scans of blood flow in the neonatal brain found a right-hemisphere trend from 1 to 3 years of age, with the trend shifting to the left hemisphere after 3 years of age [39]. This also confirms that electrodes in the right hemisphere have a better sleep staging effect.

E. The Effect of Channel Numbers

The number of channels affects the running speed of the model, the number of features and the sleep staging results, etc. Table IV shows the comparison of metrics with a different number of channels, where the SE block is placed after each CNN branch. The direct effect of the reduced number of channels is the running time of the model. Fewer channels lead

to faster running speed and less model complexity. Moreover, we output the features extracted by MSCNN and SE block for different channel cases. Taking into account that the extracted features in the eight-channel case are the most comprehensive, the features extracted in the single-channel and two-channel cases are calculated separately with the feature extracted in eight-channel case for mutual information to explore the reduction of information. The results demonstrate that the proposed feature extraction method is able to extract sufficient features to maintain good sleep staging performance despite the reduction of channels.

Furthermore, the sleep staging results for the one-channel, two-channel and eight-channel signal cases are compared to evaluate whether the proposed method is able to perform sleep staging well with limited channels. Table III shows the specific results. As the number of signal channels fed into the network increases, the accuracy of sleep staging results improves. Notably, when the number of input signal channels is one, the sleep staging results after using the SE block are only slightly worse than the results for the eight-channel signal without the SE block. When the number of input signal channels is two, the sleep staging results after using the SE block are even better than the results of the eight-channel signal without the SE block. In addition, as the number of signal channels increases, the enhancement effect of the SE block becomes less significant. The SE block after feature concatenation and SE block after each CNN branch have nearly identical enhancements to the results. These are because when the number of signal channels is increased, the extracted features will also increase. However, the parameters and output shape of the network are not changed with the number of signal channels. When the shape of the output features is fixed, the more features are extracted, the more valid features and the less redundant features will be in the final output. There-fore, as the number of the input signal channels increases, fewer and fewer redundant features can be optimized, and the enhancement effect of SE blocks becomes weaker and weaker. The results indicate that the SE block can optimize the features and improve the sleep staging performance. Moreover, the enhancement is more significant in the case of limited channels.

F. Baseline and State-of-Art Methods Comparison

In this experiment, we compare our proposed method with several state-of-the-art and baseline methods. The details of these methods are presented as follows.

• Conv-2d [17]: Conv-2d [17] is an 18-layer CNN that performs downscaling and feature extraction of the signal by deep convolution and pooling.

• Conv-2d [18]: Conv-2d [18] is optimized on the basis of Conv-2d [17]. It optimizes part that assigns weights to the extracted features.

• DeepSleepNet [26]: DeepSleepNet [26] is a network architecture for adult sleep staging that has achieved good results in several public sleep datasets.

• AttnSleep [34]: AttenSleep [34] is proposed for adult sleep staging, which consists of CNN and multi-head attention

TABLE V
RESULTS COMPARISON OF STATE-OF-ART METHODS, BASELINES, AND THE PROPOSED METHODS

Methods	Acc	F1 score	Kappa	Specificity	Sensitivity
Conv-2d [17]	0.535	0.531	0.489	0.768	0.536
Conv-2d [18]	0.523	0.519	0.411	0.761	0.523
DeepSleepNet [26]	0.698	0.689	0.644	0.805	0.703
AttnSleep [34]	0.680	0.646	0.659	0.839	0.650
THNN [28]	0.694	0.686	0.638	0.848	0.698
MSCNN	0.693	0.691	0.650	0.847	0.691
MS-HNN	**0.754**	**0.758**	**0.728**	**0.876**	**0.755**

(MHA). Instead of applying RNN, MHA is presented to learn the temporal information between sleep stages, which greatly improves the efficiency of the model. In this paper, DeepSleepNet and AttnSleep are used to explore the results of direct transfer of adult sleep models to neonatal sleep data.

• THNN [28]: THNN is composed of CNN and RNN. Firstly, the feature extraction part of the network is trained. Secondly, the temporal learning of the features is attached and the parameters of the feature extraction part are also fine-tuned to achieve optimal performance.

• MSCNN (baseline): MSCNN is the multi-scale CNN, which is a module applied in this paper. The MSCNN is trained separately to explore the effect of TIL module on the enhancement of the sleep staging results. The fully con-nected layer replaces the TIL module to output the sleep stages.

For all methods, we uniformly use one channel EEG signal (P4) as input, and the output results are three sleep stages. Considering that different methods might have early stop signs, we set the iteration to 150, which is consistent with our proposed method. In addition, each method is equally cross-validated with 10 folds. Table V shows the detailed results of all methods. The proposed method outperforms the state-of-the-art methods. Compared with [17] and [18], the neonatal dataset used in this experiment are several times larger than those used in these studies. When the above two networks are validated using the data in this experiment, there are problems such as underfitting and inability to fully capture the features. For the adult sleep model DeepSleepNet [26] and AttnSleep [34], due to the different sleep characteristics and individual differences between neonates and adults, the adult sleeping model suffers from convergence and underfitting during training. Therefore, the performance of transferring the adult sleep model directly to the neonate sleep data could be greatly reduced, and needs to be adjusted according to the characteristics of the neonate sleep. Additionally, the classi-fication accuracy improvement is obvious with the addition of the TIL module comparing the MSCNN and MS-HNN. This indicates that the temporal information in neonatal sleep data is important. Information on transitions between sleep stages can effectively improve the results of sleep staging. Generally, a good neonatal sleep model requires a deeper network architecture compared with adult sleeping model, and need to take into account different scale features and temporal information.
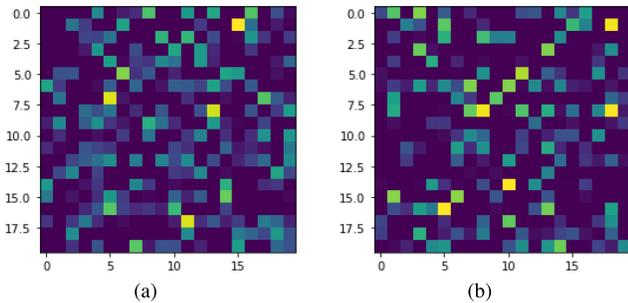
Fig. 6. The feature distribution after the features are concatenated with different SE block placement positions, the brighter the color means the higher the weight. (a) The SE block is placed after each CNN branch. (b) SE block is placed after feature concatenation.

IV. DISCUSSION

In this work, a multi-scale hierarchical neural network for automatic neonatal sleep staging is proposed. This network fully extracts the sleep features in a single-EEG channel from multiple scales and adopts SE block to reduce redundant features and enhance the network training efficiency. In addition, the network takes into account the temporal information and transition information among adjacent stages, which is overlooked in most existing studies [13], [14], [17], [18]. Experimental results exhibit that the proposed method outperforms the baseline and state-of-art methods and can achieve favorable results even with limited EEG channels.

A. The Placement of SE Block

The impact of SE blocks placement positions on the final sleep staging results is investigated in this paper. The sleep staging results of placing the SE block after each CNN branch outperform the results of placing the SE block after the features concatenation. For the SE block placed after the CNN branch, the optimized features are not downscaled. This means that the SE block has sufficient channels to explore the interrelationships between features and to perform weight reassignment. The optimization effect of the SE block will be more obvious. For the SE blocks placed after feature concatenation, the optimized features have been concatenated and downscaled. The redundant features and effective features are mapped and then mixed together, which leads to the ineffective feature optimization of the SE block. Fig.6 shows the distribution of features after the feature concatenation with different SE block placement positions. After using SE block for each CNN branch to optimize the features, the concatenated feature distribution is a bit more concentrated compared to using SE blocks after the features concatenated. Similarly, when the number of signal channels increases, the enhancement of the SE block for sleep staging results becomes less significant. This is because the number of channels and the extracted features increases leading to the redundant features decreasing. Therefore, when the network architecture is fixed, the sleep staging results with a limited number of signal channels combined with SE block can be comparable to the sleep staging results with a multi-signal channel. Alternatively, changing the layer size of network is possible to make a

Channel	AMI	Con	P_value	Chi	Acc
P3	0.653	0.609	0.009	195.38	0.753
P4	**0.663**	**0.654**	0.008	**292.32**	**0.754**
C3	0.639	0.567	0.007	181.75	0.745
C4	0.657	0.576	0.009	262.54	0.750
T3	0.660	0.535	**0.005**	176.73	0.737
T4	0.661	0.585	0.005	165.84	0.747
F3	0.622	0.615	0.013	192.00	0.746
F4	0.624	0.612	0.005	267.97	0.752
P4+F3	-	0.657	0.008	305.38	0.759
P4+F4	-	**0.664**	**0.007**	**316.24**	**0.760**
P4+P3	-	0.660	0.008	308.39	0.759
P4+C4	-	0.656	0.009	306.23	0.755
P4+C3	-	0.655	0.009	302.65	0.756
P4+T4	-	0.653	0.010	304.31	0.754
P4+T3	-	0.650	0.009	303.58	0.757
P4+P3+F3+F4	-	**0.678**	**0.001**	**332.53**	**0.763**

[1] AMI: average mutual information, Con: contribution, Chi: chi-square value.

significant improvement in the sleep staging results with SE block in the case of multiple signal channels. However, the training time may also increase, which needs to be taken into account.

B. Selection of Channels

In this paper, a total of eight monopolar EEG channels located at F3, F4, C3, C4, T3, T4, P3, and P4 sites according to the International 10-20 System were attached. The experimental protocol of the electrode placement incorporates the American Academy of Sleep Medicine (AASM) manual for infant scoring criteria [10] and clinical demands. On the basis of these eight channels, the impact of channel reduction is investigated. Table VI gives the average mutual information between different channels, the contribution of different channel features to the classification results, chi-square test results between different channel features and classification labels, as well as the automatic sleep staging accuracy. The average mutual information is the average of the features of one channel and the features of other channels calculated by mutual correlation. The contribution of features to the classification results is obtained by calculating the correlation coefficient between the features and the results. The chi-square test of different channel features and classification labels is a correlation analysis of the features of different channels with the classification labels. As shown in Table VI, for a single channel, the P4 channel achieves the highest average mutual information, chi-square, contribution value, and accuracy among all the single channels. Thus, P4 can be considered the optimal channel. Based on the selected P4 channel, the forward search is applied to search for the optimal combination of two channels. As shown in Table VI, the optimal combination of two channels is P4+F4. For the

case of three or more channels, the optimal channels can be searched by adding channels in turn to the two-channel result. To illustrate, by using the forward search method, the optimal combination of four channels is P4+P3+F3+F4. With the increase in the number of channels, slight improvements can be observed in contribution, chi-square test, and sleep staging accuracy. In addition, other methods can be used for feature selection and channel selection, such as random forests, Max-Relevance and Min-Redundancy [40], and so on.

For the number of channels, the experiment results in Table IV and Table VI show that with the increase of EEG channel, the sleep staging performance grows slightly. Signals of the same modality may contain similar information, and features extracted from the same modality signal would have redundancy. Thus, comparing the mutual information between channels can contribute to the selection of the number of channels with same modality. If the mutual information values between channels are high, the optimal channel can be used as input. If the mutual information values between the channels are low, all these channels can be used as input. Furthermore, posterior methods such as k-nearest neighbors [41] and random forest can be used for channel selection. Generally, the number of channels requires the calculation of mutual information between channels. The comparison of mutual information can help reduce the redundancy between channel features and enhance the comprehensiveness of channel features.

Abundant channels can provide more spatial information and may gain significant performance improvement. For brain-computer-interfaces applications, normally 256 channels [42] or up to 512 channels can be recorded. However, for the long-time sleep monitoring, with the guideline of AASM manual and without disturbing the natural sleep process, thereby few channels (normally less than 8 monopolar EEG channels) were arranged. Especially for neonatal sleep monitoring, an excessive number of channels would easily lead to skin disruption and discomfort to neonates.

C. The Influence of TIL Module

The TIL module can improve the sleep staging accuracy of the model significantly compared with baseline method. This suggests that information on the temporal between sleep stages is essential in the sleep staging of neonates. This is consistent with the findings in the sleep model for adolescents and adults [28], [43], [44], [45], [46]. In addition, information on the transition between sleep stages has been of interest all the time. Between the AS and QS sleep stages, there is the indeterminate sleep (IS) stage [9]. The IS stage can be used to assist in determining the conversion of some difficult to classify AS and QS sleep stages. Therefore, significant improvement in the results of automatic neonatal sleep staging with the addition of the TIL module is reasonable. However, the TIL module is mainly composed of RNN, which runs serially and needs long training time. Moreover, as the amount of data increases, the training time required for TIL increases dramatically. This is clearly reflected in Table III. When the number of signal channels is eight, the required training time is nearly twice as long as when the number of channel signals

is one. The accuracy rate is only 0.9% higher than that of the one-channel case. This difference in accuracy could be even smaller if only the QS stage is detected. This can demonstrate that SE blocks can greatly improve the accuracy of sleep staging results at the cost of a small amount of complexity and running time. Additionally, the ratio of time spent on training MSCNN alone and training MSCNN and TIL is about 1:15. If the ability to learn temporal information can be added to the MSCNN by optimizing the segment of the dataset, it may be possible to discard the TIL module while retaining the ability of the network architecture to learn temporal information.

D. Limitations and Future Work

Based on the experiments, we need to point out the shortcomings and future directions. Firstly, TIL module may result in a long training time due to the serial operation mechanism of RNN network, which may lead to the inefficiency of the whole model. In the future, the model can be optimized by discarding the TIL module while preserving the learning of temporal information to improve the effectiveness of the model. Alternatively, an efficient temporal learning method like Transformer [47] can be used instead of RNN. Secondly, in this paper, we have only explored the three-class problem. The AS and QS stages can be further divided into ASI, ASII, QSI and QSII. In the future, a five-class task for neonatal sleep staging could be involved. Third, some of the modules used in the paper have been proposed in other fields. In future work, we will aim to explore more innovative modules and methods for neonatal sleep. Fourth, in this paper, we mainly focus on exploring the feasibility and reliability of the proposed method via only using the EEG signals as the input signal for automatic sleep staging. Signals such as EOG, EMG, and ECG were not involved in this paper. However, these signals can be used in future work for the exploration of the impact of multiple modality input signals for automatic neonatal sleep staging. Finally, the deep learning approach is still a black box and does not show well which features in the EEG signal respond better to the sleep stage. Interpretability of deep learning methods remains an important task in future work.

V. CONCLUSION

In this paper, a novel network structure named MS-HNN is proposed for automatic sleep staging of neonates with limited channels. It applies MSCNN to extract signal features from a single EEG channel, optimizes the features using SE blocks, and adopts the TIL module to learn the temporal information among adjacent stages. By incorporating MSCNN, SE blocks, and TIL, the proposed approach can extract more informative features involve the temporal information to enhance the performance. The experimental results show that our proposed method outperforms the baseline and the existing state-of-the-art methods. In addition, the proposed method achieves favorable and comparable results via a single-EEG channel in comparison with that using eight-EEG channels. With these encouraging outcomes, the proposed method is expected to offer a reliable and robust solution for efficient sleep monitoring with limited channels.

References

[1] C. L. A. Bassetti et al., "Narcolepsy b clinical spectrum, aetiopathophysiology, diagnosis and treatment," *Nature Rev. Neurol.*, vol. 15, no. 9, pp. 519–539, 2019.

[2] S. Graven, "Sleep and brain development," *Clinics Perinatology*, vol. 33, no. 3, pp. 693–706, 2006.

[3] M. Mirmiran and R. Ariagno, "Role of rem sleep in brain development and plasticity," *Sleep Brain Plasticity*, vol. 5, pp. 181–187, Jun. 2003.

[4] M. Mirmiran, "The function of fetal/neonatal rapid eye movement sleep," *Behavioural Brain Res.*, vol. 69, nos. 1–2, pp. 13–22, Jul. 1995.

[5] M. Mirmiran, Y. G. H. Maas, and R. L. Ariagno, "Development of fetal and neonatal sleep and circadian rhythms," *Sleep Med. Rev.*, vol. 7, no. 4, pp. 321–334, Aug. 2003.

[6] S. Gertner, C. W. Greenbaum, A. Sadeh, Z. Dolfin, L. Sirota, and Y. Ben-Nun, "Sleep–wake patterns in preterm infants and 6 month's home environment: Implications for early cognitive development," *Early Human Develop.*, vol. 68, no. 2, pp. 93–102, Jul. 2002.

[7] M. Liu, H. Zhu, and J. Tang, "B overview of a sleep monitoring protocol for a large natural population," *Phenomics*, vol. 4, no. 5, pp. 100–193, 2023.

[8] V. Bertelle, A. Sevestre, K. Laou-Hap, M. C. Nagahapitiye, and J. Sizun, "Sleep in the neonatal intensive care unit," *J. Perinatal Neonatal Nursing*, vol. 21, no. 2, pp. 149–148, Apr./Jun. 2007.

[9] A. Dereymaeker et al., "Review of sleep-EEG in preterm and term neonates," *Early Hum. Develop.*, vol. 113, pp. 87–103, Oct. 2017.

[10] M. M. Grigg-Damberger, "The visual scoring of sleep in infants 0 to 2 months of age," *J. Clin. Sleep Med.*, vol. 12, no. 3, pp. 429–445, Mar. 2016.

[11] 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," *Amer. Acad. Sleep Med.*, vol. 9, pp. 103–123, Jan. 2007.

[12] S. F. Abbasi et al., "EEG-based neonatal sleep-wake classification using multilayer perceptron neural network," *IEEE Access*, vol. 8, pp. 183025–183034, 2020.

[13] O. De Wel et al., "Complexity analysis of neonatal EEG using multiscale entropy: Applications in brain maturation and sleep stage classification," *Entropy*, vol. 19, p. 516, Sep. 2017.

[14] N. Koolen, "Automated classification of neonatal sleep states using EEG," *Clin. Neurophysiol.*, vol. 128, pp. 1100–1108, Jun. 2017.

[15] K. Pillay, A. Dereymaeker, K. Jansen, G. Naulaers, S. Van Huffel, and M. De Vos, "Automated EEG sleep staging in the term-age baby using a generative modelling approach," *J. Neural Eng.*, vol. 15, no. 3, Jun. 2018, Art. no. 036004.

[16] A. Dereymaeker et al., "An automated quiet sleep detection approach in preterm infants as a gateway to assess brain maturation," *Int. J. Neural Syst.*, vol. 27, Sep. 2017, Art. no. 1750023.

[17] A. H. Ansari et al., "Quiet sleep detection in preterm infants using deep convolutional neural networks," *J. Neural Eng.*, vol. 15, Dec. 2018, Art. no. 066006.

[18] A. H. Ansari et al., "A convolutional neural network outperforming state-of-the-art sleep staging algorithms for both preterm and term infants," *J. Neural Eng.*, vol. 17, Jan. 2020, Art. no. 016028.

[19] A. H. Ansari et al., "A deep shared multi-scale inception network enables accurate neonatal quiet sleep detection with limited EEG channels," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 3, pp. 1023–1033, Mar. 2021.

[20] H. Ghimatgar et al., "Neonatal EEG sleep stage classification based on deep learning and HMM," *J. Neural Eng.*, vol. 17, p. 18, Dec. 2020.

[21] L. Fraiwan and M. Alkhodari, "Neonatal sleep stage identification using long short-term memory learning system," *Med. Biol. Eng. Comput.*, vol. 58, no. 6, pp. 1383–1391, Jun. 2020.

[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[23] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 5987–5995.

[24] 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, Apr. 2020.

[25] 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.

[26] 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.

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

[28] H. Zhu et al., "The masking impact of intra-artifacts in eeg on deep learning-based sleep staging systems: A comparative study," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1452–1463, 2022.

[29] H. Zhu et al., "MasksleepNet: A cross-modality adaptation neural network for heterogeneous signals processing in sleep staging," *IEEE J. Biomed. Health Informat.*, vol. 7, pp. 1–12, Mar. 2023.

[30] R. D. Ogilvie, "The process of falling asleep," *Sleep Med. Rev.*, vol. 5, no. 3, pp. 247–270, Jun. 2001.

[31] H. Matsuura, K. Tateno, and S. Aou, "Dynamical properties of the two-process model for sleep-wake cycles in infantile autism," *Cognit. Neurodynamics*, vol. 2, no. 3, pp. 221–228, Sep. 2008.

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

[33] J. Phyo, W. Ko, E. Jeon, and H.-I. Suk, "TransSleep: Transitioning-aware attention-based deep neural network for sleep staging," *IEEE Trans. Cybern.*, early access, Sep. 5, 2022, doi: 10.1109/TCYB.2022.3198997.

[34] E. Eldele et al., "An attention-based deep learning approach for sleep stage classification with single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 809–818, 2021.

[35] 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., $4.00)," *Electroencephalogr. Clin. Neurophysiol.*, vol. 26, no. 6, p. 644, 1969.

[36] M. Lewis and D. Carmody, "Self-representation and brain development," *Develop. Psychol.*, vol. 44, pp. 1234–1329, Oct. 2008.

[37] G. Li et al., "Computational neuroanatomy of baby brains: A review," *NeuroImage*, vol. 185, pp. 906–925, Jan. 2019.

[38] C. Ortinau and J. Neil, "The neuroanatomy of prematurity: Normal brain development and the impact of preterm birth: The neuroanatomy of prematurity," *Clin. Anatomy*, vol. 28, no. 2, pp. 168–183, Mar. 2015.

[39] C. Chiron, "The right brain hemisphere is dominant in human infants," *Brain*, vol. 120, no. 6, pp. 1057–1065, Jun. 1997.

[40] G. Karakaya, S. Galelli, S. D. Ahipaşaoğlu, and R. Taormina, "Identifying (quasi) equally informative subsets in feature selection problems for classification: A max-relevance min-redundancy approach," *IEEE Trans. Cybern.*, vol. 46, no. 6, pp. 1424–1437, Jun. 2016.

[41] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient KNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018.

[42] S. R. Ashwini and H. C. Nagaraj, "Classification of eeg signal using EACA based approach at SSVEP-BCI," *Int. J. Artif. Intell.*, vol. 10, no. 3, pp. 717–726, 2021.

[43] 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.

[44] 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.

[45] 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.

[46] 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. 2021.

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