Multi-Scale Residual Graph Learning with Contextual Memory for Sleep Staging

Xiaodong Yang[1,2,3]*, **Jieying Hu**[1] and **Dongxin Liang**[1]
[1]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China
[2]Mine Digitization Engineering Research Center of the Ministry of Education
[3]Jiangsu Provincial Industrial Technology Engineering Center for Intelligent Sensing and Emergency IoT in Underground Space
{xyang, ts23170067p31, ts23170073p31}@cumt.edu.cn

Abstract

Sleep is a vital physiological process, and its quality directly impacts human health. Although deep learning-based automated sleep staging methods have achieved significant progress, the studies still face limitations in fully mining spatio-temporal graph dependencies. Existing methods face challenges in capturing implicit spatial topology among multi-channel signals and efficiently fusing multi-level features in long-sequence modeling. To address these challenges, this paper proposes a novel hybrid deep learning framework for sleep staging that progressively and explicitly integrates spatial topology modeling, cross-level feature fusion, and global temporal aggregation. By decoupling shallow and deep spatial features and hierarchically organizing feature interactions, the proposed framework effectively captures implicit spatial dependencies while maintaining computational efficiency in long-sequence modeling. Extensive experimental results on multiple sleep staging datasets (83.1% accuracy, 80.1% F1-score on ISRUC-S1, 84.0% accuracy, 83.8% F1-score on ISRUC-S3, and 85.9% accuracy, 83.7% F1-score on Sleep-EDF-153) demonstrate that the proposed method achieves competitive accuracy with significantly reduced computational latency, highlighting its strong potential for mobile and real-time sleep monitoring applications.

1 Introduction

Sleep is an important physiological process consisting of non-rapid eye movement (NREM) and rapid eye movement (REM) sleep [Umlauf *et al.*, 2015], each cycle lasting approximately 90 to 120 minutes. During sleep, the brain filters and processes information from the day, transferring important memories from temporary storage to long-term memory banks. In addition, adequate sleep enhances immune function and strengthens the body's resistance [Irwin, 2015]. With the rapid pace of modern life, the average sleep duration has decreased by 1.5 hours compared to 50 years ago, while approximately 30% of the global population experiences sleep quality issues [Huang *et al.*, 2023]. Among various diagnostic methods, sleep staging is a central aspect of clinical sleep medicine. By identifying and categorizing the different stages of sleep [Levesque, 1984], it provides foundational support for diagnosing conditions such as insomnia, sleep apnea, and periodic limb movement disorder. Therefore, high-precision, automated sleep staging methods are of significant importance in improving the efficiency and accuracy of sleep disorder diagnoses.

Traditional sleep staging methods [Memar and Faradji, 2017] rely heavily on the interpretation of polysomnographic data by doctors and experts or on the extraction of manual features, which are time-consuming and labor-intensive. With the development of artificial intelligence[Wicaksono *et al.*, 2022], the use of neural network models for automated, end-to-end sleep staging analysis of physiological signals has become a research hotspot. Deep learning is capable of autonomously learning latent temporal and spatial features [Fu *et al.*, 2023]. However, current methods mainly focus on EEG signals [Eldele *et al.*, 2021], with limited research on other peripheral signals. The human body is a whole, and other peripheral signals will also reflect some characteristics of different sleep stages. Therefore, multi-modal physiological signals can offer a more comprehensive prediction[Que *et al.*, 2023]. The fusion method can also affect the quality of the extracted classification features. Additionally, existing methods often focus on isolated time segments, overlooking contextual relationships across the sleep cycle that are essential for understanding sleep stage transitions.

This paper proposes a novel hybrid deep learning sleep staging framework (MSRG-CM) for sleep staging, which mainly consists of three modules: A multi-scale residual graph perception network constructs hierarchical graph representations to extract discriminative spatiotemporal features. The sparse cross-feature attention selects shallow and deep network features as inputs to learn the correlations between local and global features within a time window. In addition, a multi-directional scanning temporal extractor leverages the lightweight and efficient sequence modeling capabilities of the architecture, incorporating a multi-directional scanning strategy to process sequences from multiple directions. It enables comprehensive aggregation of temporal or feature se-

*Corresponding author.

quences. The main contributions are as follows:

- A multi-scale residual graph perception network is designed, aggregating information from different neighboring nodes by constructing multi-scale graphs. The skip connections are established with the previous layer to extract abstract networks layer by layer.
- A sparse cross-feature attention method is proposed, which selects shallow and deep network features for fusion. It is improved by performing attention calculations within a local window range and introducing a neighborhood window mechanism, reducing computational cost.
- A multi-directional feature-scanning aggregator is constructed, which leverages the lightweight nature of the GRU to improve runtime efficiency while capturing information flow paths from different directions.

2 Related work

2.1 Deep Learning Methods

Significant progress has been made in sleep stage classification using machine learning methods [Liu *et al.*, 2016] and [Hached *et al.*, 2021]. However, these methods typically rely on extensive prior knowledge and have limited generalization capabilities. Researchers have begun to shift towards deep learning methods to learn complex sleep patterns. Currently, the most widely used methods are CNN, RNN, and LSTM. For example, some studies [Pan *et al.*, 2024] and [Khalili and Mohammadzadeh Asl, 2021] use CNN to extract local features within each period and employ Bi-LSTM [He *et al.*, 2024] to encode temporal information. Other use complete RNNs to model dependencies for sleep staging [Phan *et al.*, 2023]. These methods either extract features from local regions or model the relationships between sleep stages, without considering both aspects jointly.

2.2 Hybrid Sleep Staging Models

Researchers began to propose hybrid sleep staging models to address the limitations of single-modal models. Ma et al. introduced SleepMG [Ma *et al.*, 2024], which evaluates the classification and domain discrimination performance of each modality, and further defines modality performance indicators. Subsequently, the models capture local temporal features through techniques such as Transformers and attention mechanisms. Zheng et al. proposed MMASleepNet [Yubo *et al.*, 2022], a multi-modal attention network that integrates multimodal features of electrophysiological signals, with a Transformer used to generate attention matrices and extract the interdependencies between multimodal features. TransSleep [Phyo *et al.*, 2023] also uses a novel Transformer to encode different features. CareSleepNet [Wang *et al.*, 2024a] further innovates the multi-modal attention mechanism by constructing a local attention network. With the rise of GNNs, utilizing graphs to interpret the spatiotemporal relationships between different signals has become a research hotspot. GraphSleepNet [Jia *et al.*, 2020] models channel dependencies via a fixed graph with learnable adjacency, while MFST-GCN [Jia *et al.*, 2021a] constructs multi-view graphs based on Euclidean distance priors. Both methods perform convolution on a single shared topology across all layers. Ji et al. proposed the MixSleepNet model [Ji *et al.*, 2024], which designs 3D-CNN modules and GCN modules to explore the correlations of frequency bands and channels along the time and spatial dimensions. The FC-STGNN model [Wang *et al.*, 2024b] constructs a fully connected spatiotemporal graph neural network to capture the spatiotemporal dependencies. Jia et al. proposed the SleepHGNN model [Jia *et al.*, 2023], which constructs a deep graph network composed of heterogeneous graph transformer layers to capture heterogeneity-aware message passing and target-specific aggregation of physiological signal interactions. Li et al. designed the MVF-SleepNet model [Li *et al.*, 2024], which integrates time-frequency images and graph learning representations to improve sleep stage classification performance. Multi-modal time series can extract more valuable information to compensate for the shortcomings of single-modal data. Different modalities are synchronized over time and complementary in terms of semantics. If effective cross-modal information fusion can be achieved, it is expected to enhance the ability to perceive subtle transitions in sleep states.

3 Method

The proposed method is an end-to-end sleep staging model that processes a sequence of sleep epochs and outputs a sequence of predicted sleep stages. We define $x \in \mathcal{R}^{n\times C}$ as a sleep epoch, where n represents the number of sampling points, and C denotes the number of channels. The input multivariate physiological signal sequence is defined as $\mathbf{X} = \{x_1, x_2, \ldots, x_L\}$, where $x_i \in \mathcal{R}^{n\times C}$ represents a sleep epoch, and L denotes the sequence length, with $i \in \{1, 2, \ldots, L\}$. In this experiment, five sleep stages are considered, $Label = 5$ (Wake, N1, N2, N3, and REM), and $\hat{y} \in \{0,1\}^{Label}$ is defined as the ground truth label corresponding to the sleep stage for each sleep epoch. $\hat{\mathbf{Y}} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_L\}$ represents the sleep stage sequence corresponding to $\hat{\mathbf{X}} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_L\}$. Fig. 1 illustrates the overall process of the MSRG-CM Model. It is primarily divided into three parts: the residual multi-scale Graph Isomorphism Network (GIN), the sparse cross-feature attention and the memory-enhanced contextual association module.

3.1 Residual Multi-Scale GIN (ResMS-GIN)

To capture structural dependencies at different granularities, we propose a multi-scale residual-aware graph convolution module, which leverages multiple adjacency matrices to model graph structures at different scales. The network adopts residual connections and integrates features across different scales to enhance representation capability.

Considering the multi-channel nature of PSG signals, a feature embedding module based on CNNs is introduced to reduce the dimensionality of the original long-sequence signals and extract highly robust local correlation features. Subsequently, a dynamic graph learning mechanism adaptively constructs a channel-wise adjacency matrix from the feature space. From this learned matrix, we derive multi-scale adjacency matrices via scale-specific quantile thresholds, yielding

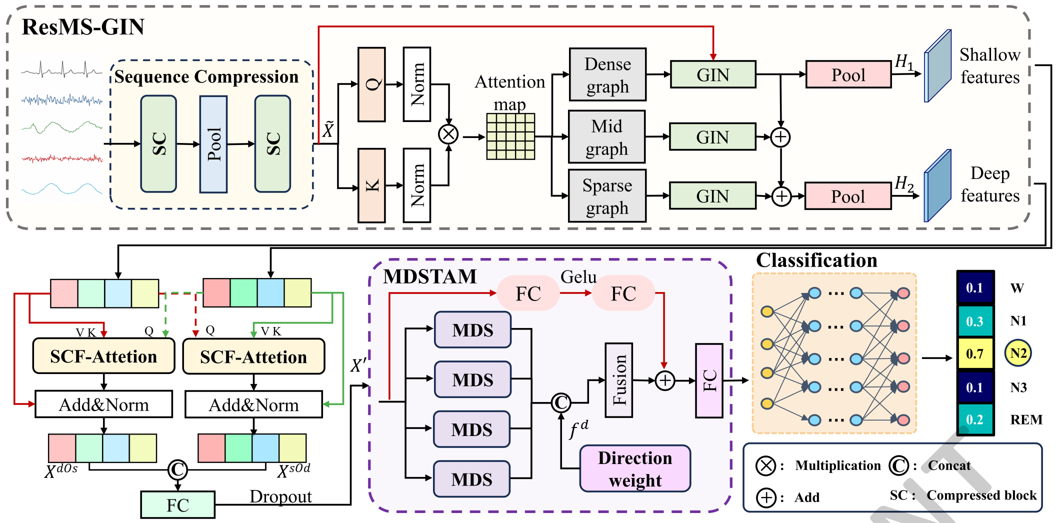

Figure 1: The MSRG-CM model diagram. The original multi-channel data is processed by the model to ultimately produce labels that distinguish between different sleep stages. The ResMS-GIN captures multi-level graph network structures, and SCFA is applied to fuse shallow and high-level networks to obtain spatiotemporal representations. Finally, a multi-directional sequence scanning strategy is introduced to model long-range dependencies across temporal scales and spatial regions.

a set of matrices ranging from dense to sparse.

$$A = \frac{1}{h}\sum_{m=1}^{h} \mathrm{Softmax}\left(\frac{Q_m K_m^\top}{\sqrt{d}}\right) \tag{1}$$

Different adjacency matrices characterize multi-level relationships, ranging from global weak correlations to local strong connections. This enables each node to aggregate information from neighborhoods of varying receptive fields during the graph convolution process.

$$A^{(s)} = \begin{cases} A, & A > \tau_s, \quad s = 1, \ldots, S \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where S is the number of the multi-scale decomposition. In this case, S is set to 3, $\tau_s = [0.3, 0.6, 0.9]$ capture global synchrony, regional networks, and event-specific coupling, respectively—mirroring clinical PSG reading.

Conventional graph convolution operations have limited capability in modeling complex structural patterns. To enhance expressive power, the proposed multi-scale graph convolution incorporates a multi-layer perceptron following neighborhood aggregation, thereby strengthening the nonlinear modeling of graph structures and enabling the extraction of multi-scale graph information. By stacking multiple graph convolution layers, the network captures graph structures from local to global scales and learns deep feature representations. Furthermore, residual connections are introduced between layers to alleviate gradient degradation in deep networks and to facilitate information flow across layers.

$$\begin{aligned} h_i^{(k+1)} = \sum_{s=1}^{S} \mathrm{MLP_s}((1+\epsilon)\mathrm{h_i^{(k)}} + \sum_{\mathrm{j}\in \mathbf{N_s(i)}} \mathbf{A^s_{ij}} \mathrm{h_j^{(k)}}) \\ + Skip_{(k-1)}(h_i^{(k)}) \end{aligned} \tag{3}$$

where $h_i^{(k)}$ is the feature of node i in the k-th layer, $\mathbf{N(i)}$ is the set of neighbors of node i, s is the scaling factor and ϵ is a learnable parameter that controls the balance between the features of the node that self-update and the propagation of neighboring features.

In terms of network architecture, shallow layers utilize denser adjacency matrices to model global channel-wise collaboration, while deeper layers progressively adopt sparser adjacency matrices to focus on strong and informative channel connections. This design enables a coarse-to-fine hierarchical modeling strategy, capturing both global and local structural patterns. After each layer, channel-wise pooling is applied to obtain the corresponding spatiotemporal feature representation. Finally, the network outputs multi-scale spatiotemporal-aware features from shallow to deep layers for subsequent fusion and classification.

3.2 Sparse Cross-Feature Attention (SCFA)

After the above module, we obtain the multi-layer network structural features of the entire sleep cycle. This section proposes a sparse cross-feature attention fusion method. Deep GIN features, while robust for stable sleep stages, undergo progressive over-smoothing during hierarchical propagation, which can submerge brief transitional patterns (e.g., N1 spindles lasting only 1-2 epochs). Shallow features preserve fine-grained morphological cues critical for detecting these transitional stages.

Unlike standard window attention (e.g., Longformer) that applies self-attention within a single feature space, SCFA introduces a Cross-Feature Neighborhood Index (Eqs. (7)-(8)) for multi-level feature fusion: shallow features serve as queries attending to deep feature neighborhoods, and vice versa. This bidirectional cross-hierarchical exchange bridges the semantic gap between shallow morphological features

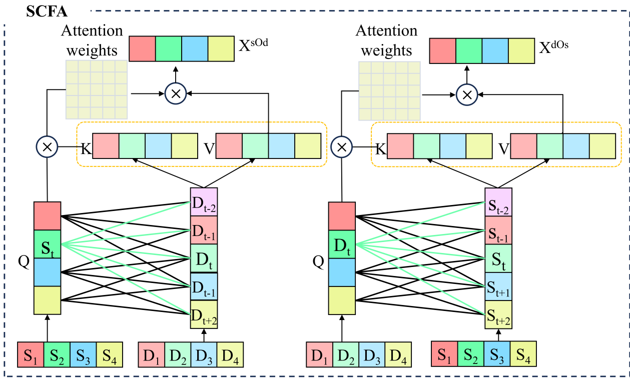

Figure 2: Sparse cross-feature attention module. Local window partitioning bridges shallow morphological features and deep semantic representations. Global dependency modeling is deferred to subsequent modules.

and deep semantic representations within local temporal windows, while delegating the task of global dependency modeling to subsequent modules. The main process is shown in the Fig. 2. Taking shallow features as an example, the temporal feature $\mathbf{X_s}$ is used as the input $\mathbf{Q_s}$ and the spatial feature $\mathbf{X_d}$ is used as the input $\mathbf{K_d}$, $\mathbf{V_d}$.

$$\mathbf{Q_s} = \mathbf{W_q}\mathbf{X_s} + b_q \in \mathcal{R}^{T\times H} \tag{4}$$
$$\mathbf{K_d} = \mathbf{W_k}\mathbf{X_d} + b_k \in \mathcal{R}^{T\times H} \tag{5}$$
$$\mathbf{V_d} = \mathbf{W_v}\mathbf{X_d} + b_v \in \mathcal{R}^{T\times H} \tag{6}$$

where T is the time step and H is the feature dimension. First, the time window is decomposed into locally processed parallel windows, $W = 8$. The corresponding K, V neighbor windows are obtained. The window sizes are symmetrically padded to ensure valid neighborhoods at the boundaries.

$$\widetilde{\mathbf{X}}_\mathbf{s} = \mathrm{Reflect}(\mathbf{X_s}, (\mathrm{K}, \mathrm{K})) \in \mathcal{N}^{\mathrm{N}\times(2\mathrm{k}+1))} \tag{7}$$

where N is the number of local windows and K is the number of neighboring windows, with K set to 16. A neighborhood index is generated for each window, and the neighborhood features $\mathbf{X_s^{neigh}}$ are collected through the index matrix.

$$\mathcal{I} = \begin{bmatrix} 0 & 1 & \dots & 2K \\ 1 & 2 & \dots & 2K+1 \\ 2 & 3 & \dots & 2K+2 \\ N-1 & N & \dots & N+2K-1 \end{bmatrix} \in \mathcal{N}^{N\times(2K+1)} \tag{8}$$

The sparse cross-attention is calculated based on the computed neighborhood shallow features. Compared to traditional cross-attention, each window only computes cross-attention on a few adjacent spatial feature windows, reducing computation and limiting noise propagation.

$$X^{sOd} = \mathrm{Softmax}(\frac{\mathbf{Q_s}\mathbf{K_d^T}}{\sqrt{\mathrm{H}}})\mathbf{V_d} \tag{9}$$

Similarly, the deep feature $\mathbf{X_d}$ is used as $\mathbf{Q_d}$ and the shallow feature $\mathbf{X_s}$ is used as $\mathbf{K_s}$, $\mathbf{V_s}$. Finally, the cross-attention X^{dOs} is computed. Through residual connections, the original features are integrated with the hierarchical features $\mathbf{X}'$ for sleep staging.

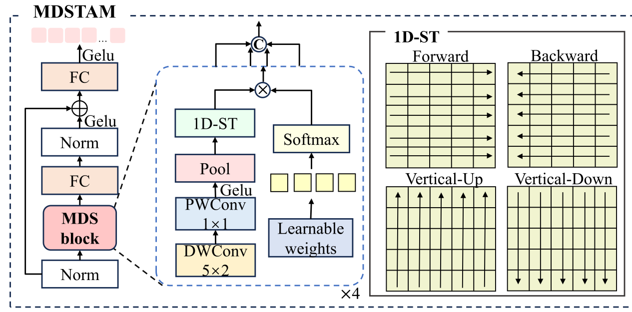

Figure 3: Multi-directional scanning temporal aggregation module. It acquires global temporal features through parallel multi-directional feature extraction, thereby improving the efficiency of the feature acquisition process.

3.3 Multi-Directional Scanning Temporal Aggregation Module (MDSTAM)

To fully exploit the cross-scale dependency structures in complex temporal data, we introduce a multi-directional feature-scanning aggregator. This module integrates convolutional modeling, directional sequence scanning, attention aggregation, and adaptive fusion. It achieves robust long-range dependency perception within a lightweight architecture by optimizing image-based scanning strategies [Zhu *et al.*, 2024] specifically for time-series analysis. Compared with methods such as Bi-LSTM, the proposed module is computationally efficient, parameter-friendly, and capable of preserving the precision of global feature extraction. The overall architecture is shown in the Fig. 3.

First, the input sequence $\mathbf{X}' = \{\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_L\}$, is normalized to enhance training stability. To strengthen the ability to learn local patterns, we adopt a depthwise separable convolutional structure consisting of depthwise and pointwise convolutions, enabling efficient local feature extraction. This convolutional module not only expands the temporal receptive field but also captures short-range dependencies.

$$\mathbf{X}_{dw} = \mathrm{DWConv}(\mathbf{X}') \tag{10}$$
$$\mathbf{X}_c = \mathrm{GELU}\left(\mathrm{BN}\left(\mathrm{PWConv}(\mathbf{X}_{dw})\right)\right) \tag{11}$$

Next, average pooling is applied for temporal downsampling to compress the sequence length and reduce computational complexity, while emphasizing long-range pattern modeling:

$$\mathbf{X}_p = \mathrm{AvgPool}(\mathbf{X}_c) \tag{12}$$

Unlike traditional RNNs or transformers that model sequences only in the forward direction, our proposed module incorporates four complementary information flow paths: forward sequential scanning, backward sequential scanning, vertical scanning, and vertical-backward scanning. This design overcomes the limitations of a single temporal perspective when capturing complex sleep patterns. Specifically, horizontal scanning captures transient features and variations across continuous time segments, while vertical scanning introduces temporal dilation to capture non-contiguous, strided long-range dependencies. Together, these paths establish multi-granularity temporal views. Lightweight GRUs are employed for sequence modeling, and an attention pooling mechanism is introduced to automatically select the most informative hidden states within each direction. This procedure preserves directional structure while improving representational efficiency. To enable vertical scanning, the pooled features are rearranged with an interleaved index set to form a re-ordered sequence X_v The computation formulas for each directional scan are as follows:

$$\begin{aligned} \mathbf{H}^{(f)} &= \mathrm{GRU}^{(f)}(\mathbf{X}_p) \\ \mathbf{H}^{(b)} &= \mathrm{GRU}^{(b)}(\mathrm{Flip}(\mathbf{X}_p)) \\ \mathbf{H}^{(vd)} &= \mathrm{GRU}^{(vd)}(\mathbf{X}_v) \\ \mathbf{H}^{(vu)} &= \mathrm{GRU}^{(vu)}(\mathrm{Flip}(\mathbf{X}_v)) \end{aligned} \tag{13}$$

A learnable attention mechanism is applied to each direction $d \in \{f, b, vd, vu\}$. The attention score and global directional feature are computed as:

$$e_t^{(d)} = \mathbf{W}^\top \mathbf{h}_t^{(d)} \tag{14}$$

$$\mathbf{f}^{(d)} = \sum_t \alpha_t^{(d)} \mathbf{h}_t^{(d)} \tag{15}$$

where $\alpha_t^{(d)}$ is the normalized attention weight. After obtaining the global features from the directional scans, a softmax function is applied to derive adaptive directional weights, allowing the model to emphasize the most discriminative scanning directions based on data characteristics. The weighted directional features are concatenated and fed into a fusion network for high-dimensional projection and feature recomposition, enhancing nonlinear representational capacity:

$$\mathbf{f}_{\text{concat}} = \left[\mathbf{f}'_f \parallel \mathbf{f}'_b \parallel \mathbf{f}'_{vd} \parallel \mathbf{f}'_{vu}\right] \tag{16}$$

$$\mathbf{f}_{\text{fusion}} = \mathrm{GELU}\left(\mathrm{LayerNorm}\left(\mathbf{W}_f \mathbf{f}_{\text{concat}}\right)\right) \tag{17}$$

where f'_d denotes the weighted feature for direction d. To further enhance robustness, a lightweight residual term based on the mean of the input sequence is introduced. This allows the final output to retain global statistical properties of the input, preventing over-reliance on any single scanning direction:

$$\mathbf{r} = \mathbf{W}_t \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{x}'_t \right) \tag{18}$$

$$\mathbf{S} = \mathbf{W}_o \left(\mathbf{f}_{\text{fusion}} + \beta \mathbf{r} \right) \tag{19}$$

where the residual scaling coefficient is set to $\beta = 0.2$. Finally, the fused representation is transformed by an output linear layer to match the original feature dimensionality, serving as the final global representation.

4 Experiment

4.1 Datasets and Preprocessing

To evaluate the performance of the proposed model, experiments were conducted on two subsets of the public ISRUC datasets [Khalighi *et al.*, 2016], and the Sleep-EDF-153 [Goldberger *et al.*, 2000] dataset. These datasets are commonly used as benchmarks for sleep stage classification. Note that all sleep records used in the experiment were preprocessed with bandpass filtering (0.3 Hz–35 Hz) and notch filtering (50 Hz).

4.2 Settings

We use PyTorch to implement the model, which is an end-to-end model trained based on backpropagation. The Adam optimizer is employed with a learning rate set to 1e-3. The number of training epochs is set to 100, with a batch size of 64. To avoid data leakage, a subject-based k-fold cross-validation strategy is adopted, where the subjects in each dataset are divided into k groups, as done in existing methods. To ensure a fair comparison with existing methods, k is set to 10.

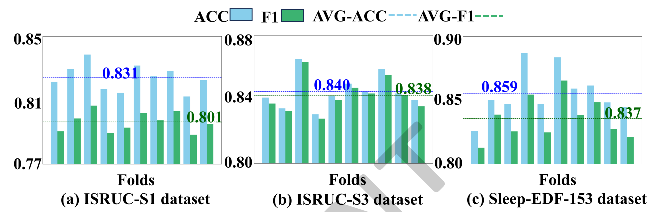

Figure 4: The 10-fold cross-validation result of three datasets.

4.3 Results

The 10-fold cross-validation results are shown in the Fig. 4. On the ISRUC-S1 dataset, the average accuracy and F1 score are $83.1 \pm 3.12\%$ and $80.1 \pm 4.78\%$, respectively. On the ISRUC-S3 dataset, the average accuracy and F1 score are $84.0 \pm 2.67\%$ and $83.8 \pm 2.99\%$. On the Sleep-EDF-153 dataset, the proposed method attains an average accuracy and an F1-score of $85.8 \pm 4.84\%$ and $83.7 \pm 4.66\%$, respectively. Except for the N1 stage, all other stages have higher scores than 78%. Because N1 is a transitional sleep stage from Wake to N2. Within the entire sleep cycle, the N1 stage lasts for a very short duration. As seen in the Fig. 4, each fold is fairly balanced. Performance on ISRUC-S3 surpasses that on ISRUC-S1, as S3 contains only healthy subjects while S1 includes sleep disorder cases with variable sample quality. The highest performance on Sleep-EDF-153 likely results from its more distinct sequential features.

	Params	FLOPs	Inf. Time	Memory
RecSleepNet	0.8 M	86 M	2.07 ms	-
DynamicSleepNet	1.07 M	8.26 G	-	-
MMASleepNet	1.2 M	460 M	8.55 ms	-
CareSleepNet	19.6 M	11.7 G	16.1 ms	-
Ours	1.5 M	11.6 G	2.61 ms	2.4 M

Table 1: Performance comparison on ISRUC-S1 dataset.

Computational Efficiency Analysis The model runtime performance is presented in the table 1. Compared to traditional LSTM/RNN structures that rely on processing complete sequences, our approach adopts optimized lightweight GRUs combined with sparse attention mechanisms, effectively reducing redundant computations. It is worth noting that while the theoretical computational cost (FLOPs: 11.6 G) is comparable to that of the large-scale CareSleepNet [Wang *et al.*, 2024a] (11.7 G), our measured inference time

Datasets	Methods	Overall Metrices			Per-class F1-scores				
		Acc	F1	Kappa	Wake	N1	N2	N3	REM
ISRUC-S1	SVM [Alickovic and Subasi, 2018]	69.3	63.5	0.55	77.0	45.3	65.4	78.9	70.1
	FFTCN [Bao *et al.*, 2024]	77.8	72.3	0.71	88.2	28.89	77.8	82.0	84.6
	MMASleepNet [Yubo *et al.*, 2022]	79.0	77.5	0.73	87.8	54.0	77.0	85.3	83.3
	RecSleepNet[Nie *et al.*, 2021]	79.7	77.9	-	88.8	53.7	79.3	86.8	80.8
	MSTGCN [Jia *et al.*, 2021a]	80.9	78.7	0.75	89.3	53.1	79.9	86.7	84.4
	MixSleepNet [Ji *et al.*, 2024]	81.1	79.0	-	89.5	54.7	79.7	87.6	83.6
	NeuroLingua [Samaee *et al.*, 2025]	81.9	80.2	0.76	86.9	57.6	85.0	86.0	85.7
	MVF-SleepNet [Li *et al.*, 2024]	82.1	80.2	0.77	**90.8**	56.2	81.1	87.1	85.7
	LGBM [Shao *et al.*, 2024]	82.5	80.4	0.77	89.7	51.2	81.4	88.6	87.8
	CareSleepNet [Wang *et al.*, 2024a]	**83.2**	**81.8**	**0.79**	90.2	55.3	**82.7**	**89.6**	**86.0**
	Ours	83.1*	80.1*	0.78	89.8	**56.5**	79.6	88.1	84.1
ISRUC-S3	SVM [Alickovic and Subasi, 2018]	73.3	72.1	0.66	86.8	52.3	69.9	78.6	73.1
	SleepMG [Ma *et al.*, 2024]	78.7	77.5	0.73	88.3	56.9	76.1	86.4	79.5
	SleepHGNN[Jia *et al.*, 2023]	79.0	77.0	-	-	-	-	-	-
	FC-STGNN [Wang *et al.*, 2024b]	80.9	78.8	-	-	-	-	-	-
	MMASleepNet [Yubo *et al.*, 2022]	81.9	80.6	0.77	88.9	59.6	82.0	87.0	86.9
	MSTGCN [Jia *et al.*, 2021a]	82.1	80.8	0.77	89.4	59.6	80.6	89.0	85.6
	DynamicSleepNet [Wenjian *et al.*, 2023]	82.9	81.7	0.79	89.9	62.4	80.7	90.7	76.8
	MixSleepNet [Ji *et al.*, 2024]	83.0	82.1	-	89.9	62.5	81.9	89.9	86.0
	MVF-SleepNet [Li *et al.*, 2024]	**84.1**	79.5	**0.83**	90.0	62.5	83.3	**91.1**	**87.3**
	Ours	84.0*	**83.8***	0.80	**90.8**	**64.1**	**83.4**	91.0	86.3
Sleep-EDF-153	SVM [Alickovic and Subasi, 2018]	71.3	53.4	0.59	80.4	50.3	63.9	82.3	71.1
	LGSleepNet [Shen *et al.*, 2023]	82.3	76.0	0.75	92.6	43.7	85.5	83.0	74.9
	MMASleepNet [Yubo *et al.*, 2022]	82.7	77.6	0.76	92.9	49.0	84.9	81.3	79.8
	DynamicSleepNet [Wenjian *et al.*, 2023]	83.0	77.7	0.76	92.3	48.1	85.8	82.2	80.3
	Multi-task model [Zhao *et al.*, 2024]	83.4	78.6	0.77	92.9	52.6	84.9	78.9	83.7
	SalientSleepNet [Jia *et al.*, 2021b]	84.1	79.5	-	93.3	54.2	85.8	78.3	85.8
	CBLSNet [She *et al.*, 2024]	84.1	78.5	0.79	82.9	50.9	85.5	80.9	83.5
	CareSleepNet [Wang *et al.*, 2024a]	85.1	80.4	0.79	91.2	55.3	85.7	81.6	83.0
	NeuroLingua [Samaee *et al.*, 2025]	85.9	80.0	0.79	**94.1**	53.8	86.6	82.4	85.2
	Ours	**85.9***	**83.7***	**0.80**	88.3	**55.4**	**89.0**	**85.5**	**86.6**

Table 2: Performance comparison with the cutting-edge methods on the three datasets (%). Bold indicates the best effect, while underlined represents the second-best effect. Besides, * indicates $p < 0.05$ in paired t-test against the SVM baseline. Paired t-tests against CareSleepNet further confirm significant F1 improvement on ISRUC-S1 and Sleep-EDF-153 (both $p < 0.05$).

is substantially lower at 2.61 ms versus 16.1 ms. This six-fold latency reduction stems from our architecture's 100% parallelization—unlike CareSleepNet, which suffers from sequential RNN bottlenecks. On edge NPUs and mobile devices, the minimal memory footprint (2.4 MB) and ultra-low inference latency are the primary determinants of real-time viability, not raw FLOPs. These characteristics give our model a distinct competitive advantage over peer models.

Comparison with Cutting-edge Methods The performance of the MSRG-CM was compared with various state-of-the-art methods published in recent years. The comparison model has been introduced in related work. As shown in the table 2, the model demonstrates advantages when compared with the advanced methods of recent years. Its ACC, F1, and κ index surpass most of the models introduced. Although the overall accuracy is slightly lower compared to CareSleepNet [Wang *et al.*, 2024a] and MVF-SleepNet [Li *et al.*, 2024] (with a difference of only about 0.1%), the model achieves the best performance in stage-wise classification. In particular, it outperforms the current models in identifying the N1 stage, indicating that the proposed multi-layer network is capable of capturing deeper discriminative features. Beyond competitive classification accuracy, our model excels in computational efficiency. The model requires only 1.5 M parameters and achieves an ultra-low inference time of 2.61 ms. Compared to large-scale models like CareSleepNet, which suffer from high latency or massive parameter counts, our approach strikes an optimal balance between high accuracy, lightweight design, and real-time capability. This makes it not only accurate but also highly practical for deployment in clinical or portable monitoring devices. Beyond the SVM baseline, paired t-tests against CareSleepNet under identical 10-fold splits confirm significant F1 improvement on ISRUC-S1 ($p < 0.05$) and Sleep-EDF-153 ($p < 0.05$).

Ablation Study Ablation experiments are primarily based on module design to analyze the impact of each module and demonstrate its effectiveness in the model. The module-based ablation experiment design is as follows:

- Ablation 1: Remove ResMS-GIN
- Ablation 2: Remove SCFA
- Ablation 3: Replace SCFA with Cross-Attention
- Ablation 4: Remove MDSTAM

Variant	Overall Metrices			Per-class F1-scores				
	Acc	F1	Kappa	Wake	N1	N2	N3	REM
Ablation 1	82.5	82.3	0.77	92.9	60.1	75.7	88.5	87.0
Ablation 2	82.0	80.3	0.77	91.2	54.7	79.2	88.9	81.3
Ablation 3	82.6	81.5	0.78	91.5	60.2	78.5	90.2	82.5
Ablation 4	81.0	79.5	0.77	91.4	55.5	79.1	89.6	83.0
Ablation 5	82.7	80.9	0.79	89.4	56.0	81.9	89.7	87.5
MSRG-CM	84.0	83.8	0.80	90.8	64.1	83.4	91.0	86.3

Table 3: Ablation studies on the ISRUC-S3 dataset (%).

- Ablation 5: Replace MDSTAM with GRU

The table 3 shows the results of the ablation experiments. Removing MDSTAM resulted in the worst performance, indicating that it could capture potential contextual relationships, playing a crucial role in distinguishing sleep stages. Replacing MDSTAM with the GRU improves the performance, indicating that multi-directional scanning is effective in capturing interactive information generated during different information propagation flows. Replacing SCFA with conventional cross-attention causes a notable drop in performance (Acc from 84.0% to 82.0%), validating our design choice: standard cross-attention attends indiscriminately across the entire time window, introducing global noise and increasing computational overhead, whereas SCFA restricts attention to local neighborhoods and enables shallow-deep cross-level alignment. Removing ResMS-GIN results in a decrease in F1, confirming the necessity of multi-scale hierarchical graph modeling.

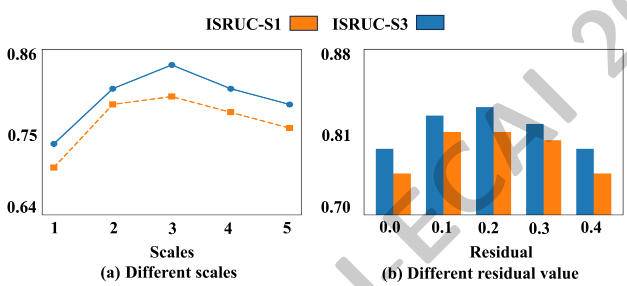

Figure 5: F1-score sensitivity analysis.

Analysis of Sensitivity We investigated the sensitivity of ResMS-GIN to scale number and MDSTAM to residual ratio (Fig. 5). As shown in Fig. 5(a), using too few scales limits its hierarchical feature capture, while too many scales introduce redundancy and overfitting. A three-scale configuration achieves the optimal trade-off across both ISRUC subsets. Fig. 5(b) shows that the optimal residual ratio for MDSTAM is 0.2. Smaller ratios disrupt original spatial feature integrity, whereas larger ratios weaken temporal modeling and long-range dependency capture.

Attention Visualization The SCFA attention scores reveal global features after feature fusion, as illustrated in Fig. 6. In this figure, the red boxes highlight the biomarkers of sleep stages within the physiological signals, and the lower part

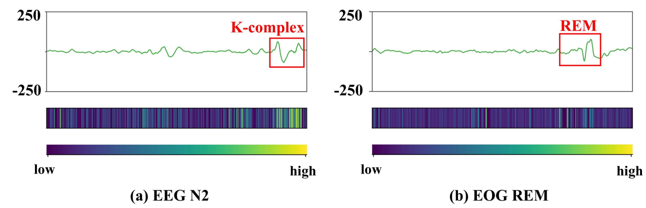

Figure 6: Attention visualization in SCFA.

displays the corresponding attention scores. For instance, the discriminative marker for the N2 stage is the K-complex, and the model assigns higher attention weights to these regions to correctly identify N2. Similarly, for the REM stage, the model captures rapid fluctuations in the EOG signal. Fig. 7 presents the t-SNE visualization results on the two ISRUC subsets. Most classes are roughly separated and form compact clusters, indicating that the proposed method learns discriminative feature representations with good intra-class compactness and inter-class separability.

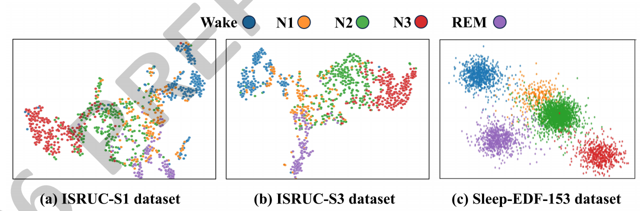

Figure 7: Visualization results using t-SNE.

5 Conclusion

This paper proposes a hybrid deep learning model for automated sleep staging, integrating multimodal information and contextual dependencies through these modules: a multi-scale residual graph network, a sparse cross-feature attention mechanism, and a multi-directional scanning temporal aggregation module. Experiments on public datasets show that the model outperforms most existing methods in accuracy and robustness, while offering significant advantages in inference speed and parameter size. Thus, it provides a new approach for sleep disorder assessment, enabling accurate characterization of complex sleep patterns with computational efficiency suitable for clinical and long-term home applications.

Limitations and Future Work The model currently operates on fine-grained unit-level time points, incurring substantial computational cost. Future work will investigate more efficient multimodal fusion strategies and lighter-weight architectures. Additionally, for sleep stages with limited samples, such as N1, the model shows low cross-dataset sensitivity, with many samples misclassified into other stages. To address this, future work will introduce low-rank learning matrices for bottom-up feature decomposition to enhance N1's distinctive characteristics and improve classification accuracy.

Ethical Statement

This study used only publicly available, de-identified datasets (ISRUC, Sleep-EDF). All data were collected by the original authors with appropriate informed consent and ethical approvals. No new human or animal experiments were conducted. Therefore, this study does not raise any ethical concerns.

Acknowledgments

The project is supported by the Xuzhou Key Research and Development Program (Social Development) (Grant No. KC21304), and the National Natural Science Foundation of China (Grant No. 62176258).

References

[Alickovic and Subasi, 2018] Emina Alickovic and Abdulhamit Subasi. Ensemble svm method for automatic sleep stage classification. *IEEE Transactions on Instrumentation and Measurement*, 67(6):1258–1265, 2018.

[Bao *et al.*, 2024] Jiameng Bao, Guangming Wang, Tianyu Wang, Ning Wu, Shimin Hu, Won Hee Lee, Sio-Long Lo, Xiangguo Yan, Yang Zheng, and Gang Wang. A feature fusion model based on temporal convolutional network for automatic sleep staging using single-channel eeg. *IEEE Journal of Biomedical and Health Informatics*, 28(11):6641–6652, 2024.

[Eldele *et al.*, 2021] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021.

[Fu *et al.*, 2023] Guidan Fu, Yueying Zhou, Peiliang Gong, Pengpai Wang, Wei Shao, and Daoqiang Zhang. A temporal-spectral fused and attention-based deep model for automatic sleep staging. *IEEE transactions on neural systems and rehabilitation engineering*, 31:1008–1018, 2023.

[Goldberger *et al.*, 2000] Ary Goldberger, Luís Amaral, L. Glass, Shlomo Havlin, J. Hausdorg, Plamen Ivanov, R. Mark, J. Mietus, G. Moody, Chung-Kang Peng, H. Stanley, and Physiotoolkit Physiobank. Components of a new research resource for complex physiologic signals. *PhysioNet*, 101, 01 2000.

[Hached *et al.*, 2021] Mustapha Hached, Khalide Jbilou, Christos Koukouvinos, and Marilena Mitrouli. A multidimensional principal component analysis via the c-product golub-kahan-svd for classification and face recognition. *ArXiv*, abs/2105.06924, 2021.

[He *et al.*, 2024] Mingyuan He, Mengmeng Tang, Linyuan Meng, and Zhenhu Liang. Tbstsleepnet: Three-branch spectro-temporal bidirectional lstm based attention model for eeg sleep staging. *Biomedical Signal Processing and Control*, 97:106695, 2024.

[Huang *et al.*, 2023] Bei Huang, Steven WH Chau, Yaping Liu, Joey WY Chan, Jing Wang, Suk Ling Ma, Jihui Zhang, Paul KS Chan, Yun Kit Yeoh, Zigui Chen, et al. Gut microbiome dysbiosis across early parkinson's disease, rem sleep behavior disorder and their first-degree relatives. *Nature communications*, 14(1):2501, 2023.

[Irwin, 2015] Michael R Irwin. Why sleep is important for health: a psychoneuroimmunology perspective. *Annual review of psychology*, 66(1):143–172, 2015.

[Ji *et al.*, 2024] Xiaopeng Ji, Yan Li, Peng Wen, Prabal Barua, and U Rajendra Acharya. Mixsleepnet: A multi-type convolution combined sleep stage classification model. *Computer Methods and Programs in Biomedicine*, 244:107992, 2024.

[Jia *et al.*, 2020] Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1324–1330. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[Jia *et al.*, 2021a] Ziyu Jia, Youfang Lin, Jing Wang, Xiaojun Ning, Yuanlai He, Ronghao Zhou, Yuhan Zhou, and Li-wei H. Lehman. Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1977–1986, 2021.

[Jia *et al.*, 2021b] Ziyu Jia, Youfang Lin, Jing Wang, Xuehui Wang, Peiyi Xie, and Yingbin Zhang. Salientsleepnet: Multimodal salient wave detection network for sleep staging. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2614–2620, 8 2021. Main Track.

[Jia *et al.*, 2023] Ziyu Jia, Youfang Lin, Yuhan Zhou, Xiyang Cai, Peng Zheng, Qiang Li, and Jing Wang. Exploiting interactivity and heterogeneity for sleep stage classification via heterogeneous graph neural network. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[Khalighi *et al.*, 2016] Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 2016.

[Khalili and Mohammadzadeh Asl, 2021] Ebrahim Khalili and Babak Mohammadzadeh Asl. Automatic sleep stage classification using temporal convolutional neural network and new data augmentation technique from raw single-channel eeg. *Computer Methods and Programs in Biomedicine*, 204:106063, 2021.

[Levesque, 1984] Hector J. Levesque. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pages 198–202, Austin, Texas, August 1984. American Association for Artificial Intelligence.

[Li *et al.*, 2024] Yujie Li, Jingrui Chen, Wenjun Ma, Gansen Zhao, and Xiaomao Fan. Mvf-sleepnet: Multi-view fusion network for sleep stage classification. *IEEE Journal of Biomedical and Health Informatics*, 28(5):2485–2495, 2024.

[Liu *et al.*, 2016] Zhiyong Liu, Jinwei Sun, Yan Zhang, and Peter Rolfe. Sleep staging from the eeg signal using multi-domain feature extraction. *Biomedical Signal Processing and Control*, 30:86–97, 2016.

[Ma *et al.*, 2024] Shuo Ma, Yingwei Zhang, Qiqi Zhang, Yiqiang Chen, Haoran Wang, and Ziyu Jia. Sleepmg: Multimodal generalizable sleep staging with inter-modal balance of classification and domain discrimination. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 4004–4013, New York, NY, USA, 2024.

[Memar and Faradji, 2017] Pejman Memar and Farhad Faradji. A novel multi-class eeg-based sleep stage classification system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(1):84–95, 2017.

[Nie *et al.*, 2021] Haodong Nie, Shikui Tu, and Lei Xu. Recsleepnet: An automatic sleep staging model based on feature reconstruction. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1458–1461, 2021.

[Pan *et al.*, 2024] Jiahui Pan, Yangzuyi Yu, Jianhui Wu, Xinjie Zhou, Yanbin He, and Yuanqing Li. Deep neural networks for automatic sleep stage classification and consciousness assessment in patients with disorder of consciousness. *IEEE Transactions on Cognitive and Developmental Systems*, 16(4):1589–1603, 2024.

[Phan *et al.*, 2023] Huy Phan, Kristian P. Lorenzen, Elisabeth R. M. Heremans, Oliver Y. Chén, Minh C. Tran, Philipp Koch, Alfred Mertins, Mathias Baumert, Kaare B. Mikkelsen, and Maarten De Vos. L-seqsleepnet: Whole-cycle long sequence modelling for automatic sleep staging. *CoRR*, abs/2301.03441, 2023.

[Phyo *et al.*, 2023] Jaeun Phyo, Wonjun Ko, Eunjin Jeon, and Heung-Il Suk. Transsleep: Transitioning-aware attention-based deep neural network for sleep staging. *IEEE Trans. Cybern.*, 53(7):4500–4510, 2023.

[Que *et al.*, 2023] Yinqing Que, Pengyi Jiang, Tianyi Zhang, and Yunzhang Cheng. Automated sleep staging based on multi-module neural network using simpler signal: respiratory signal. volume 12779, page 127791T. SPIE, 2023.

[Samaee *et al.*, 2025] Mahdi Samaee, Mehran Yazdi, and Daniel Massicotte. Neurolingua: A language-inspired hierarchical framework for multimodal sleep stage classification using eeg and eog. *arXiv preprint arXiv:2511.09773*, 2025.

[Shao *et al.*, 2024] Yizi Shao, Bokai Huang, Lidong Du, Peng Wang, Zhenfeng Li, Zhe Liu, Lei Zhou, Yuanlin Song, Xianxiang Chen, and Zhen Fang. Reliable automatic sleep stage classification based on hybrid intelligence. *Computers in Biology and Medicine*, 173:108314, 2024.

[She *et al.*, 2024] Yichong She, Di Zhang, Jinbo Sun, Xuejuan Yang, Xiao Zeng, and Wei Qin. Cblsnet: A concise feature context fusion network for sleep staging. *Biomedical Signal Processing and Control*, 91:106010, 2024.

[Shen *et al.*, 2023] Qi Shen, Junchang Xin, Xinyao Liu, Zhongyang Wang, Chuangang Li, Zhihong Huang, and Zhiqiong Wang. Lgsleepnet: An automatic sleep staging model based on local and global representation learning. *IEEE Transactions on Instrumentation and Measurement*, 72:1–14, 2023.

[Umlauf *et al.*, 2015] Mary Grace Umlauf, Anneliese C Bolland, Kathleen A Bolland, Sara Tomek, and John M Bolland. The effects of age, gender, hopelessness, and exposure to violence on sleep disorder symptoms and daytime sleepiness among adolescents in impoverished neighborhoods. *Journal of youth and adolescence*, 44(2):518–542, 2015.

[Wang *et al.*, 2024a] Jiquan Wang, Sha Zhao, Haiteng Jiang, Yangxuan Zhou, Zhenghe Yu, Tao Li, Shijian Li, and Gang Pan. Caresleepnet: A hybrid deep learning network for automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 28(12):7392–7405, 2024.

[Wang *et al.*, 2024b] Yucheng Wang, Yuecong Xu, Jianfei Yang, Min Wu, Xiaoli Li, Lihua Xie, and Zhenghua Chen. Fully-connected spatial-temporal graph for multivariate time-series data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 15715–15724, 2024.

[Wenjian *et al.*, 2023] Wang Wenjian, Xiao Qian, Xue Jun, and Hu Zhikun. Dynamicsleepnet: a multi-exit neural network with adaptive inference time for sleep stage classification. *Frontiers in Physiology*, Volume 14 - 2023, 2023.

[Wicaksono *et al.*, 2022] Pandu Wicaksono, Philip Samuel, Islam Nur Alam, and Sani M Isa. Dealing with imbalanced sleep apnea data using dcgan. *Traitement du Signal*, 39(5):1527, 2022.

[Yubo *et al.*, 2022] Zheng Yubo, Luo Yingying, Zou Bing, Zhang Lin, and Li Lei. Mmasleepnet: A multimodal attention network based on electrophysiological signals for automatic sleep staging. *Frontiers in Neuroscience*, Volume 16 - 2022, 2022.

[Zhao *et al.*, 2024] Caihong Zhao, Jinbao Li, and Yahong Guo. Sequence signal reconstruction based multi-task deep learning for sleep staging on single-channel eeg. *Biomedical Signal Processing and Control*, 88:105615, 2024.

[Zhu *et al.*, 2024] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62429–62442, 21–27 Jul 2024.