MULTI-SCALE RESIDUAL GRAPH LEARNING WITH CONTEXTUAL MEMORY FOR SLEEP STAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sleep is a vital physiological process. 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. Experimental results on ISRUC-S1, ISRUC-S3, and Sleep-EDF-153 show competitive accuracy under the reported subject-level evaluation protocol (83.1%/80.1%, 84.0%/83.8%, and 85.9%/83.7% Accuracy/F1, respectively). In an additional MIT-BIH evaluation, the model achieves 75.1% accuracy and 74.5% F1-score. The implementation-level runtime measurements provide evidence of deployment potential, while repeated-seed and matched-baseline studies remain necessary before making stronger efficiency or real-time claims.

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. Sleep complaints are common in the general population. A cross-national review estimated that insomnia symptoms are reported by approximately 30% of adults, although the prevalence of clinically defined insomnia is substantially lower Ohayon (2002).Among various diagnostic methods, sleep staging is a central aspect of clinical sleep medicine. By identifying and categorizing sleep stages according to standardized polysomnographic scoring rules, sleep staging provides an essential basis for clinical sleep assessment and the evaluation of sleep disorders Iber et al. (2007b). 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 & 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 deep learning, neural networks have increasingly been used for end-to-end sleep staging from physiological signals. Representative approaches include deep recurrent architectures, attention-based models, and systematic reviews of automatic sleep staging Supratak et al. (2017); Eldele et al. (2021a); Phan & Mikkelsen (2022). 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. (2021b), 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 predictionQue et al. (2023). The fusion method can also affect

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

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 sequences. 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, including multi-domain EEG feature extraction and multi-class EEG-based classification systems Liu et al. (2016); Memar & Faradji (2017). 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 & 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

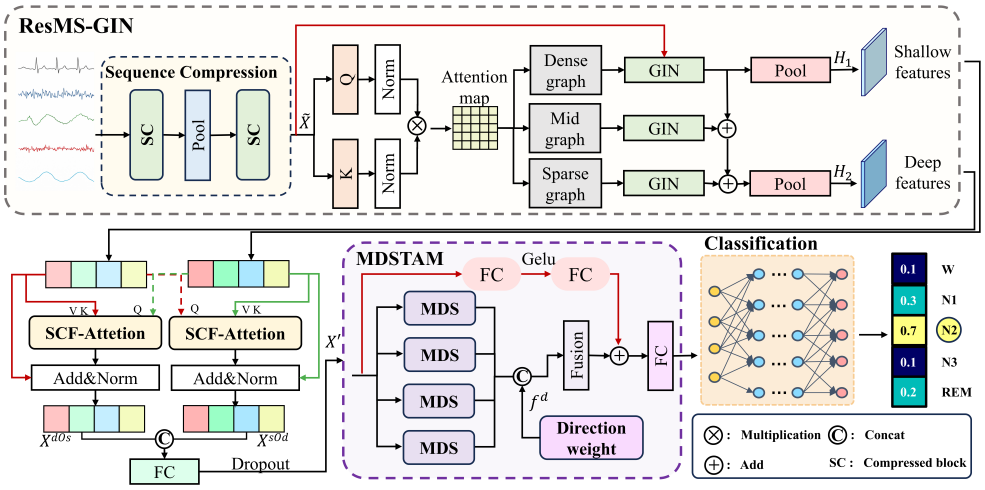

Figure 1: The MSRG-CM model diagram. The original multi-channel data is processed by the model to ultimately produce labels that distinguish between different sleep stages. The ResMS-GIN captures multi-level graph network structures, and SCFA is applied to fuse shallow and high-level networks to obtain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

FC-STGNN model Wang et al. (2024c) 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. First, a CNN-based embedding module reduces the dimensionality of the multi-channel PSG signals to extract local temporal features. An adaptive graph learning mechanism then constructs a channel-wise adjacency matrix A, from which we derive multi-scale sparse matrices $A^{(s)}$ via scale-specific quantile thresholds τ_s:

$$A = \frac{1}{h} \sum_{m=1}^{h} \text{Softmax}\left(\frac{Q_m K_m^\top}{\sqrt{d}}\right), \quad A_{ij}^{(s)} = \begin{cases} A_{ij}, & A_{ij} > \tau_s \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

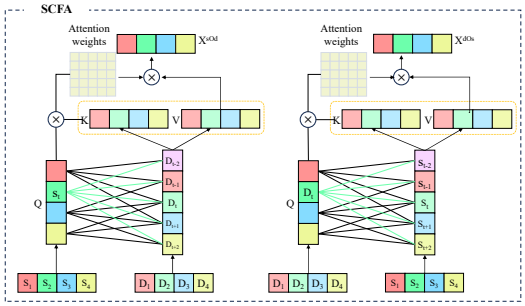

Figure 2: Sparse cross-feature attention module. Local window partitioning bridges shallow morphological features and deep semantic representations. Global dependency modeling is deferred to subsequent modules.

where h is the number of heads, and $s \in \{1, \ldots, S\}$ indexes the scale. We set $S = 3$ and $\tau_s = [0.3, 0.6, 0.9]$ to capture global synchrony, regional networks, and event-specific coupling, respectively. This multi-level topology enables nodes to aggregate information across varying receptive fields.

To enhance nonlinear expressivity over conventional GCNs and alleviate gradient degradation, we incorporate MLPs and residual connections. The node representation $h_i^{(k)}$ at the k-th layer is updated as:

$$h_i^{(k+1)} = \sum_{s=1}^{S} \mathrm{MLP}_s\Big((1+\epsilon)h_i^{(k)} + \sum_{j \in \mathbf{N_s(i)}} A_{ij}^{(s)} h_j^{(k)}\Big) + \mathrm{Skip}_{(k-1)}(h_i^{(k)}) \tag{2}$$

where $\mathbf{N_s(i)}$ denotes the neighbors of node i at scale s, and ϵ is a learnable parameter balancing node self-updating and neighborhood propagation. Architecturally, shallow layers use denser matrices for global collaboration, while deeper layers adopt sparser matrices to focus on strong local connections. This coarse-to-fine strategy, followed by layer-wise channel pooling, yields robust multi-scale spatiotemporal representations for subsequent fusion.

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 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 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 Fig. 2.

Taking shallow features as an example, the shallow feature $\mathbf{X_s}$ and the spatial deep feature $\mathbf{X_d}$ are first linearly projected into query, key, and value spaces ($\in \mathcal{R}^{T \times H}$):

$$\mathbf{Q_s} = \mathbf{W_q X_s} + b_q, \quad \mathbf{K_d} = \mathbf{W_k X_d} + b_k, \quad \mathbf{V_d} = \mathbf{W_v X_d} + b_v \tag{3}$$

where T is the time step and H is the feature dimension. The sequence is decomposed into N locally processed parallel windows with the window parameter set to $W = 8$. To ensure valid neighborhoods at the boundaries, symmetric padding $\widetilde{\mathbf{X}}_\mathbf{s} = \mathrm{Reflect}(\mathbf{X_s}, (K, K))$ is applied. Instead of computing global relationships, SCFA generates a neighborhood index set $\mathcal{I}_i = [i, i+1, \ldots, i+2K]$ for each window i. Thus, K is the neighborhood radius and each query uses at most $2K+1$ neighboring windows (with K set to 16). The sparse cross-attention is then calculated as:

$$X^{sOd} = \mathrm{Softmax}\Big(\frac{\mathbf{Q_s}(\mathbf{K}_d^{neigh})^T}{\sqrt{H}}\Big)\mathbf{V}_d^{neigh} \tag{4}$$

Compared to traditional cross-attention, each window only computes attention across a few adjacent spatial feature windows, which reduces computational overhead and limits noise propagation.

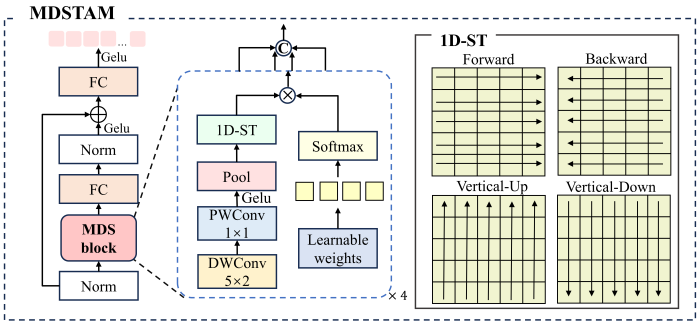

Figure 3: Multi-directional scanning temporal aggregation module. It acquires global temporal features through parallel multi-directional feature extraction, thereby improving the efficiency of the feature acquisition process.

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

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 Fig. 3.

First, the input sequence $\mathbf{X}' = \{\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_L\}$ 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. This structure expands the temporal receptive field, captures short-range dependencies, and is immediately followed by average pooling for temporal downsampling to compress the sequence length:

$$\mathbf{X}_p = \mathrm{AvgPool}\big(\mathrm{GELU}(\mathrm{BN}(\mathrm{PWConv}(\mathrm{DWConv}(\mathbf{X}'))))\big) \tag{5}$$

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 $\mathbf{X}_v$. The computation formulas for each directional scan are as follows:

$$\begin{aligned} \mathbf{H}^{(f)} &= \mathrm{GRU}^{(f)}(\mathbf{X}_p), & \mathbf{H}^{(b)} &= \mathrm{GRU}^{(b)}(\mathrm{Flip}(\mathbf{X}_p)) \\ \mathbf{H}^{(vd)} &= \mathrm{GRU}^{(vd)}(\mathbf{X}_v), & \mathbf{H}^{(vu)} &= \mathrm{GRU}^{(vu)}(\mathrm{Flip}(\mathbf{X}_v)) \end{aligned} \tag{6}$$

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

$$e_t^{(d)} = \mathbf{W}^\top \mathbf{h}_t^{(d)}, \quad \mathbf{f}^{(d)} = \sum_t \alpha_t^{(d)} \mathbf{h}_t^{(d)} \tag{7}$$

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 ($\mathbf{f}'_d$) are concatenated and fed into a fusion network for high-dimensional projection and feature recomposition, enhancing nonlinear representational capacity:

$$\mathbf{f}_{\text{fusion}} = \text{GELU}\big(\text{LayerNorm}(\mathbf{W}_f[\mathbf{f}'_f \| \mathbf{f}'_b \| \mathbf{f}'_{vd} \| \mathbf{f}'_{vu}])\big) \tag{8}$$

To further enhance robustness, a lightweight residual term $\mathbf{r}$ 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 (with the residual scaling coefficient set to $\beta = 0.2$):

$$\mathbf{r} = \mathbf{W}_t\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}'_t\Big), \quad \mathbf{S} = \mathbf{W}_o(\mathbf{f}_{\text{fusion}} + \beta\mathbf{r}) \tag{9}$$

Finally, the fused representation $\mathbf{S}$ 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; an additional MIT-BIH evaluation is provided in Appendix D. These datasets are commonly used as benchmarks for sleep stage classification. All records were resampled to 200 Hz and preprocessed with bandpass filtering (0.3 Hz–35 Hz) and notch filtering (50 Hz). For the pooled evaluation, each continuous record was divided into 6000-point (30-second) windows with a hop length of 3600 points (18 seconds), giving 40% overlap. Windows were generated independently within each record; no window was carried across record boundaries. Each window received the sleep-stage label of its annotated scoring interval at the window start, and each held-out window was counted once in its fold-level confusion matrix.

4.2 Settings

We use PyTorch to implement the model, which is an end-to-end model trained based on back-propagation. The AdamW optimizer is employed with a learning rate of 1×10^{-3}, a weight decay of 0.05, and a batch size of 64 for 100 epochs. 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. Following the common evaluation protocol in sleep staging, k is set to 10. Each fold is trained once. We report the mean $\pm$ standard deviation across the 10 held-out folds; these values summarize split-to-split variation in the present evaluation, with the subject-level fold treated as the reporting unit rather than individual epochs. Because the original experiment log did not retain per-run random-seed metadata, the reported standard deviations do not quantify optimization variability; repeated-seed robustness remains an open experimental question. Model-selection and checkpoint-selection rules should therefore be interpreted as part of the released evaluation protocol rather than inferred from the fold statistics. For comparison with previously published methods, we distinguish our 10-fold evaluation from literature-reported baselines; unless otherwise noted, those baselines are not assumed to share the same codebase, preprocessing implementation, feature normalization, or subject-level fold assignment. Consequently, cross-paper values are used for contextual positioning, not as same-protocol statistical comparisons.

	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.

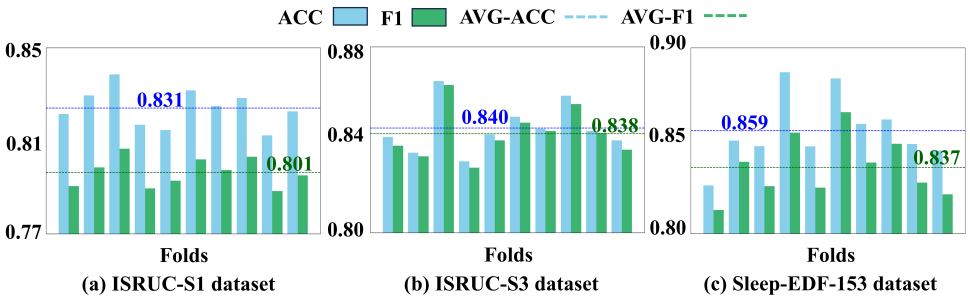

Figure 4: The 10-fold cross-validation result of three datasets.

4.3 Results

The 10-fold cross-validation results are shown in 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 $85.9 \pm 4.84\%$ and $83.7 \pm 4.66\%$, respectively. These standard deviations describe variation across subject-level folds only. Except for the N1 stage, all other stages have higher scores than 78%. N1 is a transitional sleep stage between Wake and N2 and is relatively short within a complete sleep cycle. As seen in 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 higher performance on Sleep-EDF-153 may be related to its more distinct sequential features.

Computational Efficiency Analysis The model runtime performance is presented in Table 1. Compared to traditional LSTM/RNN structures that rely on processing complete sequences, our approach adopts lightweight GRUs combined with sparse attention mechanisms to reduce redundant computations. The theoretical computational cost of our model (FLOPs: 11.6 G) is comparable to that of CareSleepNet Wang et al. (2024a) (11.7 G), while the listed inference time is lower at 2.61 ms versus 16.1 ms. Since reported latency can be affected by hardware, implementation details, batch size, and measurement protocol, we interpret this comparison as implementation-level evidence of efficiency rather than a hardware-normalized benchmark. The small parameter count and memory footprint indicate that the proposed architecture is promising for resource-constrained sleep monitoring.

Variant	Overall Metrics			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 (%).

Comparison with Published Methods The performance of MSRG-CM was compared with representative published methods introduced in the related work. As shown in Table 2, the proposed model gives competitive numerical results under the reported subject-level evaluation protocol. On ISRUC-S1, its overall accuracy is close to CareSleepNet Wang et al. (2024a), while its F1-score is lower. On ISRUC-S3 and Sleep-EDF-153, it obtains the best or near-best reported values on several overall and stage-wise metrics, including N1 performance. These comparisons should not be interpreted as uniform reimplementations: SVM, CareSleepNet, and the other published baselines may use different codebases, preprocessing, feature normalization, or subject-level splits. We therefore use the literature values for contextual positioning only, and do not claim a statistically significant or same-protocol improvement over every baseline. The defensible conclusion is that MSRG-CM is a competitive candidate under the present evaluation setup, subject to confirmation with repeated seeds and matched baselines.

Datasets	Methods	Overall Metrics			Per-class F1-scores				
		Acc	F1	Kappa	Wake	N1	N2	N3	REM
ISRUC-S1	SVM Alickovic & 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 & 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. (2024c)	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 & 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**

† Results obtained from our 10-fold subject-level evaluation.

Table 2: Performance comparison with existing methods on the three datasets (%). Bold indicates the best reported result, while underlined represents the second-best reported result. † denotes results obtained from our 10-fold subject-level evaluation. Other baseline values are taken from the corresponding papers unless explicitly stated otherwise.

Ablation Study Ablation experiments analyze the impact of each module to demonstrate its effectiveness. The five variants are designed as follows: (1) **Ablation 1**: Remove ResMS-GIN; (2) **Ablation 2**: Remove SCFA; (3) **Ablation 3**: Replace SCFA with Cross-Attention; (4) **Ablation 4**: Remove MDSTAM; and (5) **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.

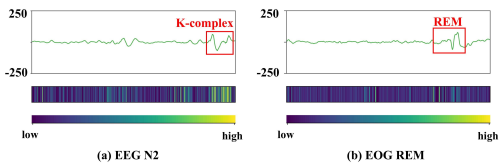

Figure 5: Qualitative attention diagnostic in SCFA. Highlighted regions motivate physiological hypotheses for follow-up testing.

Analysis of Sensitivity We investigated the sensitivity of ResMS-GIN to scale number and MDSTAM to residual ratio (Fig. 9). As shown in Fig. 9(a), using too few scales limits 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. 9(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.

Qualitative Model Diagnostics The SCFA attention maps provide a qualitative diagnostic of the learned representations, as illustrated in Fig. 5. Regions with larger attention values receive greater weight in the corresponding computation, and in several examples they align with patterns commonly associated with K-complexes or rapid EOG fluctuations. We therefore use these maps to motivate physiological hypotheses and guide future event-level testing, rather than to claim a causal mechanism. Fig. 7 presents t-SNE visualizations of learned features on the two ISRUC subsets; the plots offer an intuitive view of representation structure and complement, but do not replace, the quantitative results in Tables 2 and 8.

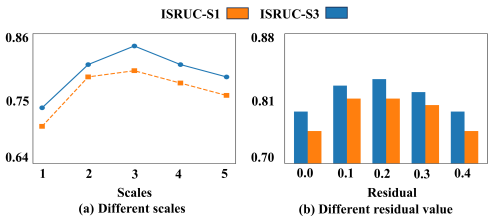

Figure 6: F1-score sensitivity analysis.

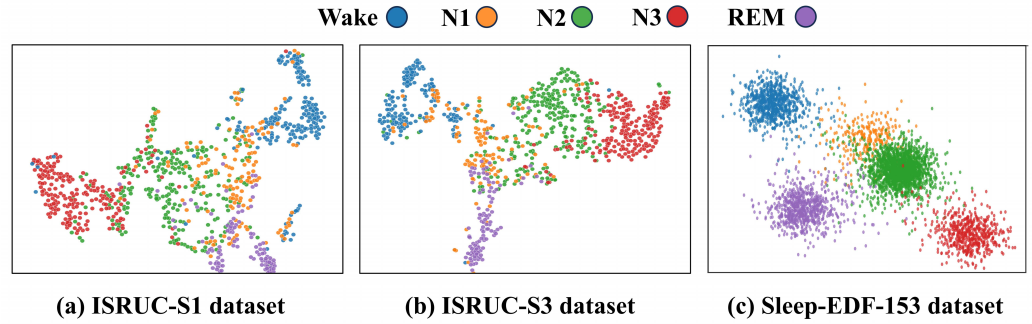

Figure 7: Qualitative t-SNE visualization of learned features, provided as an intuitive complement to the quantitative evaluation.

5 Conclusion

This paper proposes a hybrid deep learning model for automated sleep staging, integrating multimodal information and contextual dependencies through a multi-scale residual graph network, sparse cross-feature attention, and multi-directional scanning temporal aggregation. Experiments on public datasets show competitive performance across the evaluated subject-level folds, together with implementation-level evidence of favorable parameter and inference-time characteristics. These results support MSRG-CM as a promising research candidate for resource-constrained sleep monitoring, but they do not by themselves establish clinical utility, real-time deployment, or superiority under a matched multi-seed benchmark.

Limitations and Future Work The model currently operates on fine-grained unit-level time points, incurring substantial computational cost. We will investigate more efficient multimodal fusion strategies and lighter-weight architectures. The present evaluation uses one training run per subject-level fold, so the reported fold standard deviations do not capture optimization variability; a repeated-seed study is needed before making fine-grained claims about model ranking. In addition, the literature baselines were not uniformly reproduced under the same preprocessing and subject-level split, and the pooled diagnostic tables use an aggregation unit that should be audited against the dataset summary. Sleep stages with limited samples, such as N1, show lower cross-dataset sensitivity, with many samples misclassified into neighboring stages. Our next step is to introduce low-rank learning matrices for bottom-up feature decomposition to strengthen N1-specific representations and improve classification accuracy. Finally, the attention and t-SNE visualizations are qualitative diagnostics rather than causal evidence, and prospective clinical validation remains necessary before any deployment claim.

AI USE STATEMENT

The authors used generative AI tools, including ChatGPT/Codex, for language editing, LaTeX formatting, consistency checks across tables and sections, citation and reference sanity checks, and reviewer-style feedback on limitations and reproducibility. The tools were not used to generate experimental results, create or modify datasets, fabricate references, or replace the authors' scientific judgment. All AI-assisted text and edits were reviewed and verified by the authors, who remain responsible for the manuscript, experiments, claims, citations, and released artifacts.

ETHICS STATEMENT

This work uses only publicly available sleep-staging benchmark datasets under their respective licenses and access conditions. No new human-subject data were collected, and all experiments report aggregate model performance without attempting to identify individuals. We use the physiological recordings solely for automated sleep-stage classification research. Because sleep data are sensitive biomedical signals, we avoid any attempt at participant re-identification and do not release derived artifacts that would expose individual-level private information. The proposed model is intended as a research contribution for sleep-stage recognition and computational efficiency analysis, not as a standalone clinical diagnostic system. Any future clinical or home-monitoring deployment should undergo prospective validation and comply with the relevant medical, privacy, and regulatory requirements.

REPRODUCIBILITY STATEMENT

We support reproducibility through transparent dataset descriptions, explicit model configurations, and detailed implementation notes in the Appendix.

1 Datasets. Experiments are conducted on public sleep-staging benchmarks, including ISRUC-S1, ISRUC-S3, Sleep-EDF-153, and MIT-BIH. Dataset statistics, sampling rates, scoring standards, and class distributions are summarized in Appendix C.

2 Preprocessing. All records are resampled to 200 Hz and processed with the filtering and channel-mapping protocol described in the main paper and Appendix A. For evaluation, a 6000-point (30-second) window is moved with stride 3600 points (18 seconds), producing 40% overlap. The number of windows in a record of length L is $\max\{0, \lfloor (L - 6000)/3600 \rfloor + 1\}$. Windows are created separately within each record, and a window label is inherited from the annotated scoring interval containing its start. This window-level definition is distinct from the raw epoch counts reported in the dataset summary.

3 Model and Training. The architecture is specified through the ResMS-GIN, SCFA, and MD-STAM modules. Appendix C reports the key hyperparameters, including layer depth, hidden dimensions, activation and normalization choices, kernel sizes, and dropout settings. Training uses AdamW with the learning rate, weight decay, batch size, and epoch number reported in the experimental settings.

4 Evaluation. We use subject-level 10-fold cross-validation. The reported mean $\pm$ standard deviation values are computed across the 10 held-out folds, with the subject-level fold treated as the reporting unit rather than individual epochs. Accuracy, F1-score, Cohen's kappa, per-class F1, balanced accuracy, pooled 0–1 error, and MCC are defined or reported in the main paper and Appendix D.

5 Repeated Seeds. Each fold is trained once in the present evaluation. Because the original run did not retain complete random-seed metadata, we report fold-level variation directly and treat repeated-seed robustness as a planned follow-up evaluation.

6 Code Release. We will release the training and evaluation code, configuration files, and reproducibility instructions upon acceptance, subject to the redistribution terms of the public datasets used in this study. The release should include the exact random seeds, fold assignments, checkpoint selection rule, feature-normalization configuration, runtime measurement script, and the scripts used to generate the pooled confusion matrices. Until those artifacts are available, the present results should be interpreted as a subject-level empirical evaluation rather than a fully audited benchmark.

Author Contributions

For the anonymous review version, we describe contributions without identifying individual authors. The authors jointly conceived the study, designed the model, implemented the experiments, analyzed the results, and wrote and revised the manuscript. Specific named author contributions will be provided in the camera-ready version if the paper is accepted.

Acknowledgments

For the anonymous review version, funding and institutional acknowledgments are withheld to preserve double-blind review. They will be restored in the camera-ready version if the paper is accepted.

References

Poorya Aghaomidi and Ge Wang. Ecg-sleepnet: Deep learning-based comprehensive sleep stage classification using ecg signals, 2024. URL https://arxiv.org/abs/2412.01929.

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

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. doi: 10.1109/JBHI.2024.3457969.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150, 2020. URL https://api.semanticscholar.org/CorpusID:215737171.

Iksoo Choi and Wonyong Sung. Single-channel ecg-based sleep stage classification with end-to-end trainable deep neural networks. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pp. 1–4, 2023. doi: 10.1109/EMBC40787.2023.10340302.

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, 2021a. doi: 10.1109/TNSRE.2021.3076297.

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, 2021b.

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.

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.

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. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2024.106695. URL https://www.sciencedirect.com/science/article/pii/S1746809424007535.

C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan. The aasm manual for the scoring of sleep and associated events: Rules, terminology and technical specifications. 2007a.

Conrad Iber, Sonia Ancoli-Israel, Andrew Chesson, and Stuart F. Quan. The aasm manual for the scoring of sleep and associated events: Rules, terminology and technical specifications. *American Academy of Sleep Medicine*, 2007b.

Yuji Ichimaru and George B. Moody. Development of the polysomnographic database on cd-rom. *Psychiatry and Clinical Neurosciences*, 53:175–177, April 1999.

Mohammadreza Iravani and Sadaf Moharreri. Sleep stage classification from ecg using machine learning: Evaluating the impact of signal duration. *Neurobiology of Sleep and Circadian Rhythms*, 20:100140, 2026. ISSN 2451-9944. doi: https://doi.org/10.1016/j.nbscr.2025.100140. URL https://www.sciencedirect.com/science/article/pii/S245199442500029X.

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

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. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2023.107992. URL https://www.sciencedirect.com/science/article/pii/S0169260723006582.

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 (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 1324–1330. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/184. URL https://doi.org/10.24963/ijcai.2020/184. Main track.

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, 2021a. doi: 10.1109/TNSRE.2021.3110665.

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 (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2614–2620, 8 2021b. doi: 10.24963/ijcai.2021/360. URL https://doi.org/10.24963/ijcai.2021/360. Main Track.

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)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095397.

Xuheng Jiang, Xinlin Sun, Yushi Hao, Qing Cai, Hongjun Hou, Lili Xia, Haoyu Li, Guang Li, and Zhongke Gao. Wnotnet: A hybrid wavelet neural operator and transformer framework for enhanced eeg denoising. *IEEE Transactions on Instrumentation and Measurement*, 74:1–15, 2025. doi: 10.1109/TIM.2025.3569916.

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. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2015.10.013. URL https://www.sciencedirect.com/science/article/pii/S0169260715002734.

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. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2021.106063. URL https://www.sciencedirect.com/science/article/pii/S0169260721001383.

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. doi: 10.1109/JBHI.2022.3208314.

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. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2016.06.006. URL https://www.sciencedirect.com/science/article/pii/S1746809416300581.

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, pp. 4004–4013, New York, NY, USA, 2024. ISBN 9798400706868. doi: 10.1145/3664647.3680854. URL https://doi.org/10.1145/3664647.3680854.

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.

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)*, pp. 1458–1461, 2021. doi: 10.1109/BIBM52615.2021.9669687.

Maurice M. Ohayon. Epidemiology of insomnia: What we know and what we still need to learn. *Sleep Medicine Reviews*, 6(2):97–111, 2002. doi: 10.1053/smrv.2002.0186.

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. doi: 10.1109/TCDS.2024.3382109.

Huy Phan and Kaare Mikkelsen. Automatic sleep staging of eeg signals: Recent development, challenges, and future directions. *Journal of Neural Engineering*, 19(4):046002, 2022. doi: 10.1088/1741-2552/ac6049.

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. URL https://doi.org/10.48550/arXiv.2301.03441.

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. URL https://doi.org/10.1109/TCYB.2022.3198997.

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, pp. 127791T. SPIE, 2023. doi: 10.1117/12.2688854. URL https://doi.org/10.1117/12.2688854.

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.

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. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2024.108314. URL https://www.sciencedirect.com/science/article/pii/S0010482524003986.

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. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2024.106010. URL https://www.sciencedirect.com/science/article/pii/S1746809424000685.

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. doi: 10.1109/TIM.2023.3298639.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017. doi: 10.1109/TNSRE.2017.2721116.

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.

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, 2024a. doi: 10.1109/JBHI.2024.3426939.

WeiBo Wang, Dimei Qin, Yu Fang, Chao Zhou, and Yongkang Zheng. Automatic multi-class sleep staging method based on novel hybrid features. *Journal of Electrical Engineering & Technology*, 19(1):709–722, 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, pp. 15715–15724, 2024c.

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. ISSN 1664-042X. doi: 10.3389/fphys.2023.1171467. URL https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2023.1171467.

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.

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. ISSN 1662-453X. doi: 10.3389/fnins.2022.973761. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.973761.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: transformers for longer sequences. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pp. 15, Red Hook, NY, USA, 2020. ISBN 9781713829546.

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. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2023.105615. URL https://www.sciencedirect.com/science/article/pii/S1746809423010480.

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*, pp. 62429–62442, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhu24f.html.

APPENDIX

In this appendix, we present additional implementation details, dataset descriptions, model configurations, supplementary evaluation results, and extended experimental analyses. The detailed contents are organized as follows:

A Reproducibility Guide 15

A.1 Environment Configuration 15

A.2 Preprocessing Process 15

A.3 Model Construction and Training 16

A.4 Reproducing Evaluation Metrics 17

B Explanations of Methods 17

B.1 Algorithm for ResMS-GIN 17

B.2 Further Discussion of SCFA 17

C Details of the Datasets and Model Implementation 18

C.1 Datasets 18

C.2 Model Implementation 18

D Results 18

D.1 Sleep-EDF-153 Dataset 18

D.2 MIT-BIH Dataset 19

D.3 Confusion Matrices of Different Datasets 20

D.4 Cost- and Imbalance-Aware Metrics 21

D.5 Visualisation 22

D.6 Component Analysis 22

D.7 Analysis of Sensitivity 22

A REPRODUCIBILITY GUIDE

This section describes the implementation details needed to reproduce the model. The reported mean $\pm$ standard deviation is computed across the 10 subject-level folds, with one training run per fold; it is not a repeated-seed confidence estimate. The original run did not retain per-run seed metadata, so optimization variability remains unresolved. Results can also vary with software, hardware, feature normalization, and preprocessing implementations. A complete reproduction should therefore record the random seeds, fold assignments, model-selection rule, checkpoint-selection rule, input sequence shape, and runtime measurement protocol in addition to the hyperparameters listed below.

Environment configuration Our proposed model is implemented using Python 3.8.0 and PyTorch 1.8.0. The experiment was conducted on a machine with an Intel Core i9 10900K CPU and an NVIDIA RTX 3080 GPU. It would be ideal if you have GPU support.

Preprocessing process Our input data consist of physiological signals extracted from the dataset, and the input dimensionality can be adjusted according to the number of physiological signal modalities provided by the dataset. For example, ISRUC-S1 contains 6 EEG signals, 2 EOG signals, 1 ECG signal, and 1 EMG signal.

```
Input_data:
    CHANNEL_MAPPING = {
```

```
    'LOC-A2'
    'ROC-A1'
    'F3-A2'
    'C3-A2'
    'O1-A2'
    'F4-A1'
    'C4-A1'
    'O2-A1'
    'X1',
    'X2'
}
```

The preprocessing workflow includes steps like label alignment, so you just need to load the downloaded dataset to run it.

```
if os.path.exists(DATA_ROOT):
    process_dataset(DATA_ROOT, OUTPUT_FILE)
else:
    print(f"error: {DATA_ROOT}")
```

Model construction and training The model mainly consists of three parts:

- ResidualMulti-ScaleGIN(ResMS-GIN)
- SparseCross-FeatureAttention(SCFA)
- Multi-DirectionalScanningTemporal AggregationModule(MDSTAM)

We have integrated them into one model code, and you just need to call this model to train. Here, we have provided the code and input parameters examples for calling the model. The input and output dimensions, among other parameters, can be modified according to your specific experimental setup.

```
    full_dataset = load_data()
    subjects = np.array([item[2] for item in full_dataset])
    unique_subjects = np.unique(subjects)

    kf = GroupKFold(n_splits=10)
    for fold, (train_idx, test_idx) in
        enumerate(kf.split(full_dataset, groups=subjects)):
        optimizer = torch.optim.AdamW(
            model.parameters(),
            lr=1e-3,
            weight_decay=0.05,
            betas=(0.9, 0.999)
        )
        x, y, subj = batch
        x = x.to(device)
        y = y.to(device)
        subj = subj.to(device)

        optimizer.zero_grad(set_to_none=True)

        logits, feat = model(x)
        loss = criterion(logits, y)
        scaler.scale(loss).backward()
        scaler.unscale_(optimizer)
        torch.nn.utils.clip_grad_norm_(
            model.parameters(), max_norm=1.0)

        scaler.step(optimizer)
        scaler.update()


        if scheduler is not None:
            scheduler.step()
```

We have listed the code for the optimizers, learners, and model training that we use. The training epochs, learning rate, and other parameters can be adjusted according to your specific experimental settings.

Reproducing evaluation metrics We provide the code for calling the evaluation model. We hope these results can be of some help to you.

```
model.load_state_dict(
    torch.load(f"fold{fold}_best.pth"))
test_results = evaluate(model, test_loader, device,
     save_attention=False)
```

The above is the complete guide to the model. We hope it will be helpful to you.

B Explanations of Methods

B.1 Algorithm for ResMS-GIN

This section presents the specific code process of ResMS-GIN.

Algorithm 1 ResMS-GIN

Input: $\mathbf{X} \in R^{B\times C\times T}$
Parameter: Layers L, scales S, hidden dimension H
Output: Layer-wise features $\{\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(L)}\}$

Reshape for CNN: $\mathbf{X}_{cnn} \leftarrow \text{reshape}(\mathbf{X}, (BC, 1, T))$
Patch embedding:
$\mathbf{X_p} \leftarrow \text{Dropout}\Big(\text{GELU}(\text{BN}(\text{Conv1D}(\mathbf{X}_{cnn})))\Big) \times 2$
Reshape to graph tokens:
$\mathbf{X_p} \leftarrow \text{reshape}(\mathbf{X_p}, (B, C, H, T_{new}))$
Compute multi-scale adjacency $\{A^{(1)}, \ldots, A^{(S)}\}$
Initialize output list $\mathcal{O} \leftarrow [\,]$
for $\ell = 1$ to L **do**
    Select scale $s \leftarrow \min(\ell - 1, S - 1) + 1$
    $\mathbf{F} \leftarrow \text{DenseGINLayer}_\ell(\mathbf{F}, A^{(s)})$
    $\mathbf{F} \leftarrow \text{LayerNorm}(\mathbf{F})$; $\mathbf{F} \leftarrow \text{ReLU}(\mathbf{F})$
    if $\ell > 1$ **then**
        $\mathbf{F} \leftarrow \mathbf{F} + \text{SkipCons}_{\ell-1}(\text{residual})$
    end if
    residual $\leftarrow \mathbf{F}$
    Channel pooling: $\mathbf{Z}^{(\ell)} \leftarrow \text{Pool}_C(\mathbf{F}) \in R^{(BT_{new})\times H}$
    $\mathbf{Z}^{(\ell)} \leftarrow \text{reshape}(\mathbf{Z}^{(\ell)}, (B, T_{new}, H))$
    Append $\mathbf{Z}^{(\ell)}$ to output list: $\mathcal{O} \leftarrow \mathcal{O} \cup \{\mathbf{Z}^{(\ell)}\}$
end for
return $\mathcal{O}$

B.2 Further Discussion of SCFA

The current mainstream feature fusion methods mainly include concatenation and addition, which are simple combination operations between these two types of features. While these methods can integrate spatiotemporal information to a certain extent, they cannot fully explore the potential relationships between different types of features. To better capture the deep interactions between features, cross-attention mechanisms have become an effective technique in recent years. This method, through bidirectional attention mechanisms, adaptively captures the correlations from different feature spaces, helping the model better learn the cross-dependencies between features. However, traditional attention methods are time-consuming and memory-intensive, and the sleep features within specific time windows tend to be concentrated around adjacent sampling points, with no long-range

dependencies. Therefore, directly using cross-attention for sleep staging is not a suitable strategy. Building upon the window-based sparse attention mechanisms introduced in Longformer Beltagy et al. (2020) and BigBird Zaheer et al. (2020). We adapt and refine these concepts to address the limitations of traditional cross-attention, specifically for multi-level feature fusion. The proposed model utilizes this approach to bridge the semantic gap between shallow morphological features and deep semantic representations within local temporal windows, while delegating the task of global dependency modeling to subsequent modules.

Datasets	Subjects	Recordings	Sample Rate	Scoring	All Epochs	Wake	N1	N2	N3	REM
ISRUC-S1	100	100	200	AASM	87,187	19,237	11,110	27,232	18,188	12,420
ISRUC-S3	10	10	200	AASM	8,889	1,674	1,217	2,616	2,016	1,066
Sleep-EDF-153	78	153	100	R&K	195,479	65,951	21,522	69,132	13,039	25,835
MIT-BIH	16	18	250	R&K	6,356	2,746	774	1,960	388	488

Table 4: Summary of the four datasets.

C Details of the Datasets and Model Implementation

C.1 Datasets

The detailed information of the dataset and the distribution of each sleep period are introduced in the Table 4.

ISRUC: The datasets are commonly used as a benchmark for sleep staging, which consists of three subsets. ISRUC-S1 includes 100 overnight PSG recordings from 100 adult sleep disorder patients, some of whom are affected by sleep medications, and most of the participants exhibit sleep apnea events. ISRUC-S3 contains recordings from 10 healthy subjects without sleep disorders or other sleep-related conditions. Each subset in this dataset is scored by sleep experts according to the AASM standards Iber et al. (2007a). Detailed information can be found in Table 4. For this experiment, these two subsets were selected for research. Each PSG recording includes 6 EEG channels, 2 EOG channels, 1 ECG channel, and 1 EMG channel.

Sleep-EDF-153: The Sleep-EDF dataset Goldberger et al. (2000) is a publicly available benchmark dataset for sleep analysis and sleep stage classification. It contains whole-night polysomnography recordings collected from healthy subjects during overnight sleep studies. The signals are sampled at 100 Hz and are segmented into 30-second epochs, following the standard sleep scoring protocol.

MIT-BIH: The MIT-BIH Polysomnographic dataset Ichimaru & Moody (1999) is a multi-channel physiological sleep monitoring dataset released by PhysioNet. It contains 18 PSG recordings collected from 16 subjects in the sleep laboratory at Beth Israel Hospital, with a total duration of over 80 hours. The dataset includes multiple signal modalities such as ECG, EEG, and respiratory signals.

C.2 Model Implementation

This section provides a comprehensive description of the module configurations used in our model to facilitate experiment replication. The experiment was conducted on a machine with an Intel Core i9 10900K CPU and an NVIDIA RTX 3080 GPU. Specifically, we report the key architectural hyperparameters and implementation details for each component in the table 5, such as layer depth, hidden dimension, number of heads/states, kernel sizes, normalization and activation choices, and dropout rates. The final release should additionally provide the numeric hidden dimension, input sequence shape, vertical-scan index mapping, checkpoint-selection rule, and runtime measurement script; symbolic placeholders alone are insufficient for an independent reproduction.

D Results

D.1 Sleep-EDF-153 Dataset

Fig. 8 illustrates the ten-fold cross-validation results and a qualitative representation visualization on the Sleep-EDF-153 dataset. As shown in Fig. 8(a), fold-level accuracy and F1-score vary across

Module	Layer Type	Configuration Parameters
ResMS-GIN	Patch Embedding	Conv1d(k=64, s=4) → MaxPool(k=4, s=4) → Conv1d(k=3, s=1), Dropout=0.3.
	Dynamic Graph Learner	multi_head = 4, head_dim = hidden_dim/4
	Multi-Scale Thresholding	num_scales=3
	DenseGIN Layers (×3)	hidden_dim , GELU, Dropout=0.1 With Residual Connections
	Channel Attn Pool	-
	Temporal Reshape	reshape(B, T_{new}, hidden_dim)
	Output: Multi-layer features	-
GraphTemporalFusion	SCFA (shallow→deep)	heads=4, window=8, neighbors=16
	SCFA (deep→shallow)	heads=4, window=8, neighbors=16
	Shallow Feature	from ResMS-GIN layer 1
	Deep Feature	from ResMS-GIN final layer
	Output: Feature	fused temporal embedding (B, T, H)
MDSTAM	LayerNorm	-
	Depthwise Conv1D	kernel=5, groups=hidden_dim
	Pointwise Conv1D	kernel=1
	AvgPool1D	kernel=2, stride=2
	GRU (Forward)	input=H, hidden=H/4
	GRU (Backward)	input=H, hidden=H/4
	GRU (Vertical-Down)	input=H, hidden=H/4
	GRU (Vertical-Up)	input=H, hidden=H/4
	Attention Pooling	Linear(hidden/4 → 1), softmax
	Direction Weights	learnable(4), softmax fusion
	Concat	dim = 4 × (H/4) = H
	Fusion	Linear, LayerNorm, GELU
	Residual Projection	GlobalAvgPool(T), Linear(H→H)
Classifier	Linear	64 → 256
	GELU	–
	Dropout	p = 0.3
	Linear	256 → 5

Table 5: Detailed hyper-parameters of the MSRG-CM model.

Datasets	Methods	Overall Metrics			Per-class F1-scores				
		Acc	F1	Kappa	Wake	N1	N2	N3	REM
MIT-BIH	Random Forest Iravani & Moharreri (2026)	48.7	31.6	-	51.0	36.0	51.0	20.0	36.0
	SVM Alickovic & Subasi (2018)	49.3	37.5	0.47	58.7	35.3	65.4	68.9	40.1
	CNN+LSTM Wicaksono et al. (2022)	73.4	69.3	-	-	-	-	-	-
	AMSS Wang et al. (2024b)	74.6	65.1	0.60	-	-	-	-	-
	KNN Aghaomidi & Wang (2024)	74.8	71.0	0.69	78.2	28.9	77.8	76.0	66.5
	ContextNet Choi & Sung (2023)	74.9	-	0.65	73.3	**61.2**	59.1	62.2	55.6
	ECG-SleepNet Que et al. (2023)	76.1	72.0	**0.72**	**85.0**	57.0	**83.0**	**81.0**	**79.0**
	WNOTNet Jiang et al. (2025)	**76.8**	73.1	-	-	-	-	-	-
	Ours†	75.1	**74.5**	0.71	81.1	47.0	79.6	76.1	66.1

† Results obtained from our evaluation.

Table 6: Performance comparison with existing methods on the MIT-BIH dataset (%). † denotes results obtained from our evaluation. Other baseline values are taken from the corresponding papers unless explicitly stated otherwise.

subject-level splits, while the marked averages summarize the overall level. Fig. 8(b) provides a t-SNE view of the learned embeddings; the overlap among neighboring stages is consistent with the ambiguity of transitional sleep boundaries and complements the numerical evaluation.

D.2 MIT-BIH Dataset

We additionally evaluated our model on the MIT-BIH dataset to assess performance on a smaller and more imbalanced sleep dataset. Following the comparison protocol in Table 6, the model obtains an accuracy of 75.1% and an F1-score of 74.5%. These are the authoritative comparison results

		Predicted					Per-class Metrics		
		Wake	N1	N2	N3	REM	Pre	Rec	F1
ISRUC-S1	Wake	**18241**	1260	247	10	112	87.9	91.8	89.8
	N1	1330	**5991**	2673	35	990	58.9	54.4	56.5
	N2	530	2023	**22370**	1772	723	77.7	81.6	79.6
	N3	401	208	2860	**13638**	203	88.2	78.8	88.1
	REM	240	710	643	7	**9617**	82.6	85.7	84.1
ISRUC-S3	Wake	**1439**	55	21	22	15	90.3	91.3	90.8
	N1	73	**718**	179	25	99	62.1	66.4	64.1
	N2	18	174	**1926**	185	35	79.8	85.2	83.4
	N3	18	20	112	**1620**	23	90.2	93.0	91.0
	REM	37	79	38	21	**783**	85.3	89.7	86.3
Sleep-EDF-153	Wake	**42347**	1290	1810	216	651	85.3	91.4	88.3
	N1	1520	**5273**	2456	17	1075	60.7	51.0	55.4
	N2	3276	1470	**126145**	7250	3771	88.1	88.9	89.0
	N3	627	20	7990	**51978**	399	85.9	85.2	85.5
	REM	1869	640	3615	1062	**58654**	88.2	89.1	86.6
MIT-BIH	Wake	**2223**	217	200	35	74	81.3	80.9	81.1
	N1	210	**364**	132	24	44	47.5	46.6	47.0
	N2	200	125	**1560**	29	45	79.5	78.6	79.6
	N3	36	24	24	**298**	5	75.3	77.0	76.1
	REM	64	45	46	10	**325**	65.9	66.3	66.1

Table 7: Pooled confusion matrices and per-class metrics on four datasets (%). The prediction unit is a 6000-point window with a 3600-point stride; each held-out window is counted once and fold-level matrices are pooled without duplication. Overall comparison accuracies are reported in Tables 2 and 6. The matrix-generation audit must reproduce the displayed totals and recompute the derived metrics from these same counts.

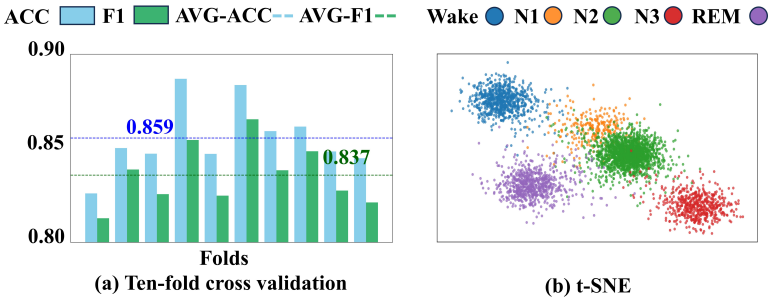

Figure 8: Ten-fold cross-validation and qualitative feature visualization on the Sleep-EDF-153 dataset. The t-SNE panel complements the quantitative metrics.

for MIT-BIH, while the confusion matrix in Table 7 gives a complementary pooled view of class-wise errors. The dataset's smaller sample size and severe class imbalance make this setting more challenging than the other benchmarks, and the present results motivate a dedicated repeated-seed evaluation. Compared with this work Aghaomidi & Wang (2024), although it reports a higher accuracy, the model was trained on only 3,000 samples rather than the full dataset. Given the small size of this dataset, such a setting can easily lead to the model memorizing the patterns of the selected samples instead of learning generalizable features.

D.3 Confusion Matrices of Different Datasets

We report the confusion matrices for sleep staging on each dataset, together with per-class evaluation metrics (precision, recall, and F1-score) in Table 7. In each matrix, rows correspond to the ground-truth sleep stages and columns denote the predicted stages. The prediction unit is one 6000-point window generated with a 3600-point stride; therefore, the matrix count is a window count rather than a raw 30-second epoch count. Windows are counted once in the held-out fold in which their subject is tested, and are not duplicated when fold-level matrices are pooled. The diagonal entries represent correctly classified windows, while off-diagonal entries highlight common misclassifica-

tions between stages. The per-class metrics further quantify performance under class imbalance. For cross-dataset comparison, the fold-level accuracies in Tables 2 and 6 remain the primary reported results. The released audit script must report, for every dataset, the raw epoch count, record-level window count, held-out prediction count, discarded tail samples, and pooled matrix total; it must also verify that the pooled error, balanced accuracy, and MCC are recomputed from the same matrix.

Datasets	Pooled Error Rate (0–1 Loss)	Balanced Acc.	MCC
ISRUC-S1	19.6	78.5	0.75
ISRUC-S3	16.2	82.6	0.79
Sleep-EDF-153	12.0	81.7	0.83
MIT-BIH	25.0	70.2	0.64

Table 8: Pooled error and imbalance-aware metrics on four datasets (%). The 0–1 loss is equivalent to the misclassification rate for the pooled confusion matrix. These pooled values are diagnostic and should be recomputed from the released matrix-generation script.

Module	Method	Performance		Efficiency		
		Acc (%)	F1 (%)	Params (M)	Inf. Time (ms)	Memory (M)
(A)	CNN	77.3	73.1	0.7	0.90	1.5
	GCN	80.5	78.2	1.5	0.81	1.2
	GAT	81.3	80.9	2.1	2.61	3.4
	ResMS-GIN	84.0	83.8	1.1	2.09	2.4
(B)	Attention Pooling	80.5	77.1	0.5	0.81	1.3
	Bi-LSTM	83.5	83.1	1.5	2.61	5.9
	MDSTAM	84.0	83.8	1.1	2.09	2.4

Table 9: Comparison between individual components and baseline models on the ISRUC-S3 dataset.

D.4 Cost- and Imbalance-Aware Metrics

Based on the obtained confusion matrices, we computed the following metrics (Table 8), which are commonly used in physiological sequence tasks to provide a more comprehensive evaluation of model performance. The metrics and their corresponding computation methods are described below.

Pooled Error Rate (0/1 Loss): Under the 0/1 loss setting, correct predictions incur zero loss and misclassifications incur unit loss. Let N_{ij} denote the number of samples with ground-truth class i and predicted class j, and let $L_{ij} = \mathbb{1}[i \neq j]$ denote the unit loss. The pooled 0–1 loss is therefore equivalent to the misclassification rate:

$$E_{0/1} = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{K} L_{ij}N_{ij} = \frac{\sum_{i\neq j} N_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{K} N_{ij}} = 1 - \mathrm{Accuracy}_{\mathrm{pooled}}. \tag{10}$$

Balanced Accuracy: This metric computes the average recall across all classes, providing a fairer assessment of model performance under class imbalance by equally weighting each class.

$$\text{Balanced Acc.} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{Recall}_k \tag{11}$$

$$\mathrm{Recall}_k = \frac{C_{kk}}{\sum_{j=1}^{K} C_{kj}}. \tag{12}$$

Matthews Correlation Coefficient (MCC): A correlation-based measure derived from the confusion matrix that jointly considers true and false predictions across all classes, yielding a more robust overall performance evaluation in imbalanced multi-class scenarios.

$$\mathrm{MCC} = \frac{cs - \sum_{k=1}^{K} p_k t_k}{\sqrt{\left(s^2 - \sum_{k=1}^{K} p_k^2\right)\left(s^2 - \sum_{k=1}^{K} t_k^2\right)}} \tag{13}$$

where $c = \sum_{k=1}^{K} C_{kk}$, $t_k = \sum_{j=1}^{K} C_{kj}$, $p_k = \sum_{i=1}^{K} C_{ik}$, $s = \sum_{i=1}^{K}\sum_{j=1}^{K} C_{ij}$.

These pooled metrics provide a complementary view of class imbalance and error patterns. Because pooled counts and fold-level arithmetic means aggregate samples differently, we use them to interpret error structure rather than to replace the primary accuracies in Tables 2 and 6.

D.5 Visualisation

Fig. 5 in the main text shows the qualitative attention diagnostic. Regions with larger values receive greater weight in the attention computation, and several highlighted segments visually resemble patterns commonly associated with K-complexes or rapid EOG fluctuations. We treat these observations as hypotheses for event-level testing rather than as verified evidence of a physiological mechanism.

Fig. 6 in the main text shows t-SNE visualisations for the two ISRUC subsets. The plots offer an intuitive view of the learned feature geometry: overlap between N1 and neighboring stages is compatible with the transitional nature of N1, while the compactness of the other clusters is consistent with the quantitative results.

D.6 Component Analysis

In the table 9, (A) denotes the ResMS-GIN module, and (B) denotes the MDSTAM module. Under this reported protocol, these components obtain competitive values relative to the included reference models. This table is an ablation-style component comparison rather than a controlled proof of superiority, because the reference implementations and measurement conditions are not all identical. ResMS-GIN is designed to capture correlations among physiological signals through shallow and deep graph representations, while MDSTAM uses a four-branch GRU architecture to model temporal dynamics. The observed parameter and runtime values motivate further matched-protocol deployment measurements rather than establishing hardware-independent efficiency.

D.7 Analysis of Sensitivity

Experimental analyses were conducted to investigate the sensitivity of ResMS-GIN to the multi-scale configuration and of MDSTAM to the residual ratio, as illustrated in Fig. 9. Specifically, Fig. 9(a) reports the accuracy variations under different scale numbers for the two ISRUC subsets, indicating that moderate scales provide the best trade-off and the selected setting yields the highest overall performance. Fig. 9(b) further compares different residual ratios in MDSTAM, showing that performance does not increase monotonically with a larger residual contribution. In particular, the optimal result is achieved when the residual ratio is set to 0.2, while overly small or large ratios may weaken either the preservation of original features or the effectiveness of temporal modeling.

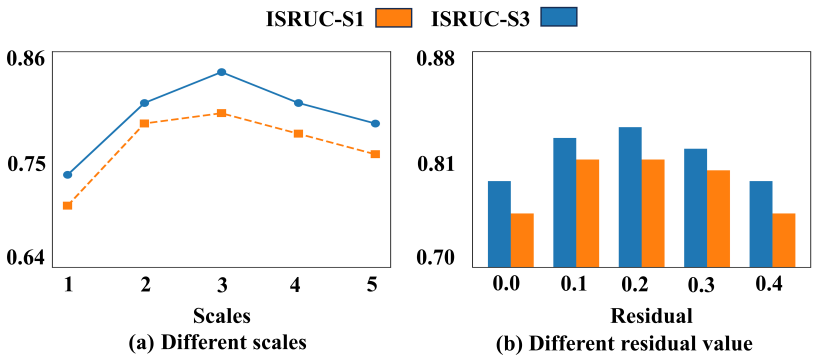

Figure 9: F1-score sensitivity analysis across scales and memory units.