GPNet: Granularity-Aware Pyramid Network with Graph Aggregation for Sleep Staging and Face-Emotional Recognition Speed Prediction

Congming Tan[1], Ting Pan[2], Yin Tian[1,2,*]
1. School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
2. School of Life and Health Information Science and Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
* Corresponding author: tianyin@cqupt.edu.cn

Abstract

Accurate classification of sleep stages is crucial for sleep quality assessment, health monitoring, and disease prevention. To effectively extract significant waveform features and capture the interactive coupling of features at different layers from single-channel electroencephalogram (EEG) signals, this study proposes the Granularity-Aware Pyramid Network with Graph Aggregation for Sleep Staging (GPNet) model. Specifically, the model first extracts fine-grained time-frequency features from multi-resolution input signals using the feature pyramid. Subsequently, an adaptive deep attention mechanism is incorporated into the layer with the highest depth-wise information to explore the correlations between local and global features. Finally, graph convolution is employed to learn the coupling interactions among high-level features across multiple layers. Comparative experiments conducted on the Sleep-EDF-X datasets demonstrate that GPNet exhibits highly competitive performance compared to other models. Additionally, GPNet predicts post-sleep recognition speed of negative emotions, revealing a negative correlation with REM(%) sleep and suggesting that sleep mitigates negative effects.

Keywords: sleep staging; feature pyramid; adaptive deep attention mechanism; negative emotional expressions

I. Introduction

Sleep is a vital physiological process essential for cognitive function, memory consolidation, emotional regulation, and overall health, occupying about one third of our lives. Good sleep enhances cognitive abilities (Luyster et al., 2012), while disorders like apnea and insomnia can disrupt daily activities and lead to various health issues. Accurately measuring sleep quality and detecting sleep-related disorders are therefore urgent needs. Currently, clinicians manually classify polysomnography (PSG) data using the Rechtschaffen and Kales (R&K) (Wolpert, 1969) or American Academy of Sleep Medicine (AASM) (Berry et al., 2012) standards. However, this manual sleep staging is time-consuming, labor-intensive, and highly dependent on the clinician's expertise (Bagautdinova et al., 2023). Consequently, there is a pressing need for automated and accurate sleep stage classification to reduce human error and lower the risk of sleep-related diseases.

Recent advancements in automatic sleep stage classification technologies have enabled more efficient assessments of sleep quality and disease diagnosis. These approaches are generally categorized into traditional machine learning methods, which rely on manual feature extraction, and deep learning methods that employ end-to-end automatic feature extraction. Traditional methods extract features from physiological signals and classify them using algorithms such as adaptive clustering (Cusinato et al., 2024), multilayer perceptrons (MLP) (Wu et al., 2022), and support vector machines (SVM) (Rahman, Bhuiyan & Hassan, 2018). In contrast, deep learning techniques automatically learn relevant features from raw data, reducing the need for expert knowledge and minimizing time-consuming, labor-intensive processes.

For example, MVF-SleepNet (Li et al., 2022) constructs time-frequency maps and uses graph learning on multimodal physiological signals, leveraging VGG16 and Chebyshev graph convolution to capture spectral-temporal and spatiotemporal features. Zhou et al. (2024) developed a pseudo-siamese network that fuses EEG and EOG features through a shared classifier, enhancing sleep stage classification from single signals. Li et al. (2024) employed multi-scale context learning to capture key EEG wave features and transition rules, addressing class imbalance with a class-adaptive fine-tuning loss function. Bao et al. (2024) combined 1D-CNN and 2D-CNN to extract time-domain and time-frequency features from raw signals and wavelet-transformed images, respectively, integrating them into a temporal convolutional network (TCN) for classification. Additionally, Pan et al. (2024) used causal dilated convolutions and transformers to extract time-correlated sleep features from a single channel, preventing future information leakage.

Most automatic sleep staging studies use multiple physiological signals collected via PSG (Zhang et al., 2024) to classify sleep stages. However, EEG offers advantages over other signals like ECG, EMG, and EOG, including lower cost, portability, and safety, making it the primary clinical standard for sleep staging (Zhang & Wu, 2021). Additionally, collecting multimodal data requires placing various electrodes or sensors on different body parts, which can cause discomfort during sleep (Suetsugi et al., 2007).

Consequently, there is a growing trend to use user-friendly wearable devices for home sleep assessments, with single-channel EEG being particularly suitable for wearable sleep

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

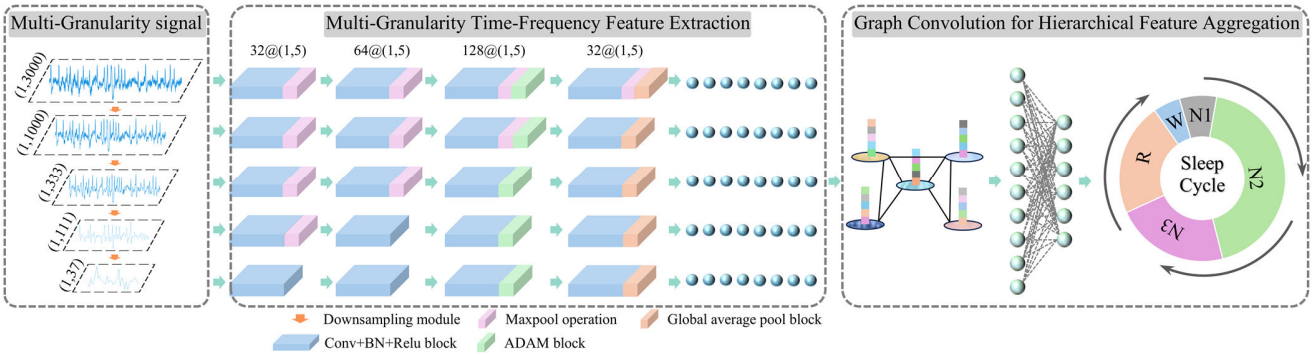


Figure 1: Overall framework of the proposed GPNet network.

systems (Wei et al., 2023). In single-channel sleep staging research, most studies apply Fourier, or wavelet transforms to create frequency-domain or time-frequency views, providing multiple perspectives for sleep stage detection. However, this approach increases model complexity and the number of parameters. Moreover, it does not fully exploit multi-level information from the original signal; for instance, different sleep stages have distinct dominant frequencies, such as spindle waves and K-complexes in the N2 stage and slow-wave signals in the N3 stage. Additionally, effectively aggregating and representing high-level features from different layers remains a challenge. Finally, there is a need to implement existing automatic staging methods in practical settings to ensure their usability.

To address these challenges, we propose GPNet, which leverages time-frequency features from single-channel EEG signals (see Figure 1). Recognizing that different sleep stages exhibit distinct dominant frequencies, we designed a downsampling module to reduce the original EEG signal, generating multi-granularity signals. For each granularity level, multiple convolutional and pooling (ConvPool) layers extract fine-grained features by retaining the maximum values within each stride and discarding weaker ones, thereby reducing the number of model parameters. Additionally, we introduce an Adaptive Deep Attention Mechanism (ADAM) block to integrate global and local information effectively, assigning appropriate weights to enhance key features while suppressing redundant or noisy data. Finally, hierarchical information from different granularities is exchanged and aggregated through graph convolutional layers, capturing both global patterns and local contextual characteristics within the data.

The main contributions of this paper can be summarized as follows:

1. Multi-Granularity Time-Frequency Feature Extraction: We designed a network that leverages the time-frequency features of sleep EEG signals. Using an adaptive downsampling module, we generate hierarchical multi-granularity information. At each level, fine-grained features are extracted and integrated with local and global information through an adaptive deep attention mechanism.
2. Adaptive Deep Attention Mechanism: We introduce an adaptive deep attention mechanism to effectively combine global and local information and fine-tune feature weights for enhanced representation. By leveraging cross-correlation operations, our mechanism explores the relationships between global and local high-level features, enhancing their interactions for more efficient weight allocation.
3. Graph Convolution for Hierarchical Feature Aggregation: We utilize graph convolution to integrate high-level information from different hierarchies by capturing relationships in multi-granularity data. Each granularity level is represented as a node with its extracted high-level features. Features from various hierarchies are dynamically and adaptively fused, enabling effective coupling and interaction of feature information.
4. Validation of Generalization Capability in Practical Scenarios: We demonstrate our algorithm's generalization by identifying a significant negative correlation between the proportion of REM sleep and the recognition speed of negative face-emotional recognition expressions. This finding supports future applications in emotion regulation.

II. Methodology

A. Overview of GPNet Model

The GPNet proposed in this study primarily comprises the following components: 1) a multi-granularity time-frequency feature extraction module, 2) an adaptive deep attention mechanism, and 3) a graph convolution-based multi-level feature fusion module, as illustrated in the overall architecture in Figure 1.

Firstly, we derive five granularity levels of hierarchical information from the original 30-second EEG epochs using a learnable adaptive downsampling module. Next, ConvPool blocks with small convolution kernels extract fine-grained features at each level. Then, the adaptive deep attention mechanism enhances these high-level features by leveraging the correlations between global and local information, reducing redundancy and noise. Finally, dynamic graph convolution integrates and couples high-level features across different granularity levels. The following subsections provide a detailed explanation of each component.

B. Multi-Granularity Time-Frequency Feature Extraction

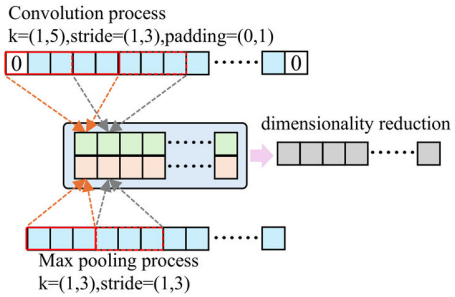


Figure 2: The proposed downsampling module.

Previous studies have shown that multi-granularity feature extraction is effective for both image processing (Li, 2024) and EEG analysis (Thuwajit et al., 2021; Pradeepkumar et al., 2024). In this study, we use four consecutive downsampling modules to partition the original single-channel EEG signal into different resolution scales, with each layer operating at one-third the resolution of the previous layer to capture distinct features. The downsampled data is then utilized to extract high-level features.

We generate multi-resolution signals using the proposed downsampling module (as illustrated in Figure 2), which consists of two branches. The first branch employs convolution operations with learnable parameters to capture and integrate fine-grained information while downsampling through strides. The second branch uses max pooling to reduce signal size and extract the most salient activation values within local regions, which may result in some loss of fine-grained information. By combining the learnable characteristics of convolution with the nonlinear selection of max pooling, the module extracts rich information while maintaining invariance. This provides a solid foundation for learning high-level features at each granularity level. In this study, we set the downsampling factor to 3, implemented as follows:

$$X_{i+1} = Conv([Conv_{down}(X_i), Max(X_i)]), i = 0,1,2,3 \quad (1)$$

Specifically, $Conv_{down}$ denotes a convolution operation with a kernel size of $(1, 5)$, padding of $(0, 1)$, and stride of $(1, 3)$. Max refers to a max pooling operation with a kernel size of $(1, 3)$ and stride of $(1, 3)$. The brackets $[\cdot]$ indicate concatenation operations along the depth dimension, and $Conv$ represents a convolution used for dimensionality reduction.

We define the input data X_0 with the shape (B, D_0, C, T_0), where B denotes the batch size, D_0 represents the initial depth dimension, C denotes the channel dimension (single-channel EEG, so C = 1), and T_0 represents the time points. In this study, the downsampling module performs downsampling solely on the time points dimension; therefore, the initial dimension T_{i+1} of the sample points at layer $i + 1$ is

$$T_{i+1} = \left\lfloor \frac{T_i}{3} \right\rfloor, i = 0,1,2,3 \quad (2)$$

Here, $\lfloor \cdot \rfloor$ denotes the floor function.

To achieve this, we generate five multi-granularity signals at different resolutions and extract time-frequency features for each layer using convolution operations. For each hierarchy level from top to bottom, we sequentially apply $4 - i$ ConvPool blocks and i convolutional blocks to progressively capture features from low to high levels, as illustrated in Figure 1. Each layer consists of four ConvPool or convolutional blocks, with the number of convolution kernels set to 32, 64, 128, and 32 for D_0 to D_3, respectively, and a uniform max pooling size of (1,3). Notably, unlike ConvPool blocks, convolutional blocks exclude pooling operations and do not perform downsampling. Instead, they adjust the depth dimension to learn fine-grained features. Additionally, since inputs at each layer have varying resolutions and receptive fields, applying the same convolution operations allows the extraction of time-frequency features across different ranges. After the j-th ConvPool block at the i-th layer, the scale is ($Y_i^j \in R^{B \times D_j \times 1 \times t_i^j}$):

$$Y_i^j = MP\left(\Phi\left(Conv\left(Y_i^{j-1}\right)\right)\right), i = 0,1,2,3; j = 1,2,3 \quad (3)$$

$$t_i^j = \left\lfloor \frac{T_i}{3^{j-1}} \right\rfloor \quad (4)$$

Here, $Y_i^0 = X_0$, $Conv\left(Y_i^{j-1}\right)$ denotes a convolution operation with a kernel size of k and padding of p (where k=5 and p=2), and $MP(\cdot)$ represents a max pooling operation with a kernel size of (1, 3), $\Phi(\cdot)$ stands for ReLU function .

Finally, we apply global average pooling (GAP) along the depth dimension for each hierarchical output, consolidating them into 32 features per granularity.

C. Adaptive Deep Attention Mechanism

Extensive research in fields like computer vision (Sun et al., 2024) and EEG analysis (Li et al., 2024) has shown that attention mechanisms effectively highlight important features while reducing information redundancy. Building on this, we propose an ADAM block in the most depth-informative layer Y_i^2 (with a depth dimension of 128) of each layer in our model architecture. This mechanism integrates global and local information, redistributes weights appropriately, and emphasizes the most valuable features. The architecture of the adaptive deep attention mechanism is illustrated in Figure 3.

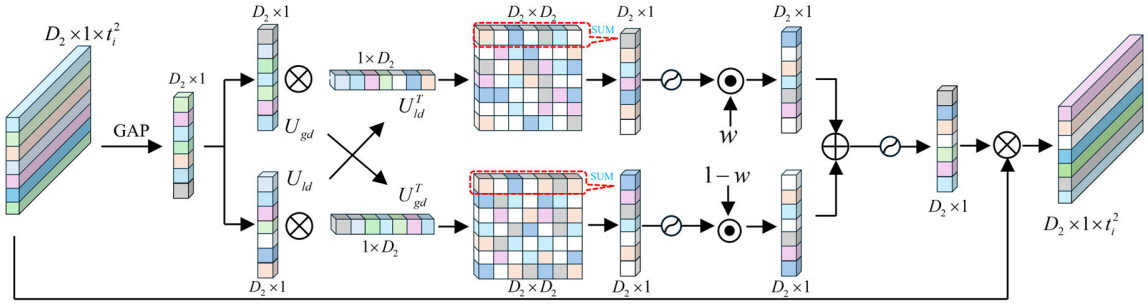
Firstly, we employ a GAP operation to transform feature vectors with a temporal dimension into descriptors that contain only depth information, denoted as $U \in R^{B \times D_2 \times 1}$. This process at the i-th layer can be expressed by the following equation:

$$U = GAP(Y_i^2) = \frac{1}{t_i^j} \sum Y_i^2 \quad (5)$$

Subsequently, we employ a set of learnable vector matrices $A = [a_1, \cdots, a_k]$ for local depth interactions to obtain local depth information U_{ld}:

$$U_{ld} = \sum_{i=1}^{k} U \cdot a_i \quad (6)$$

Where, k is the number of neighboring depths. In this study, this is implemented through 1D convolution, specifically by first performing permutation $U' \in R^{B \times 1 \times D_2}$



⊗ matrix multiplication ⊘ Sigmoid function ⊕ element-wise addition

Figure 3: The proposed adaptive deep attention module.

and then applying a convolutional kernel of size k to achieve local information interaction. Additionally, to enhance the representational capacity of global information, this study employs a fully connected method $B = [b_1, \cdots, b_{D_2}]$ to capture the dependencies among all depths as global information $U_{gd} \in R^{B \times D_2 \times 1}$:

$$U_{gd} = \sum_{i=1}^{D_2} U \cdot d_i \tag{7}$$

In this study, we achieve this using a convolutional layer with D_2 kernels of size 1. We capture the interaction between global and local information through two types of cross-correlation operations, effectively enhancing feature representation and providing a comprehensive understanding of information correlations across different dimensions. The interaction R_{gl} of global information U_{gc} on local information U_{lc} is

$$R_{gl} = U_{gd}^T \cdot U_{ld}, R_{gl} \in R^{B \times D_2 \times D_2} \tag{8}$$

The interaction R_{lg} of local information U_{ld} on global information U_{gd} is

$$R_{lg} = U_{ld}^T \cdot U_{ld}, R_{lg} \in R^{B \times D_2 \times D_2} \tag{9}$$

Then, we obtain the weight vectors U_{ld}^w and U_{gd}^w for the two types of information interactions, and perform adaptive fusion through a learnable factor W:

$$U_{gd}^w = \sum R_{gl} \in R^{B \times D_2 \times 1} \tag{10}$$

$$U_{ld}^w = \sum R_{lg} \in R^{B \times D_2 \times 1} \tag{11}$$

$$W = \sigma\left(\sigma(\theta) \times \sigma\left(U_{gd}^w\right) + \left(1 - \sigma(\theta)\right) \times \sigma(U_{ld}^w)\right) \tag{12}$$

Where θ is a learnable parameter and σ denotes the sigmoid function. The core function of this module is to effectively fuse global and local features by selectively emphasizing useful information while suppressing redundant or irrelevant features, enabling more precise weight allocation. Finally, the adaptively weighted fused feature vector W is applied to the input Y_i^2:

$$Y_i^{2*} = W * Y_i^2 \tag{13}$$

Where Y_i^{2*} is the weighted feature.

D. Graph Convolution for Hierarchical Feature Aggregation

After passing through the multi-granularity time-frequency feature extraction module, the original EEG signals produce five sets of 32-dimensional high-level feature vectors $Y_i^{out} \in R^{B \times D_3}$. In this study, we use residual graph convolution layers to integrate and aggregate information from different granularities and hierarchical levels, effectively capturing the data's information coupling. The specific implementation process is as follows:

Firstly, we treat each granularity level as a node, with its high-level features Y_i^{out} serving as node attributes. We initialize a learnable adjacency matrix $K \in R^{5 \times 5}$ with all values set to 1 to aggregate high-level representations from different levels, forming $H^0 \in R^{B \times 5 \times 32}$. The graph convolution at the l-th layer is defined as:

$$H^l = \sigma(KH^{l-1}HW^l) + H^{l-1} \tag{14}$$

Where W^l is the learnable parameter for the l-th layer, and $\sigma(\cdot)$ is the ReLU activation function. This process aggregates information between nodes and their neighborhoods, integrating multi-level data and capturing feature interactions across different levels to enhance feature representations. To keep the network simple and reduce the number of parameters, we use only one graph convolution layer. Finally, two fully connected layers are applied to generate the probabilities for each sleep stage category.

III. Results

A. Datasets

We evaluated the proposed GPNet on the Sleep-EDF-39 (Kemp et al., 2000) and Sleep-EDF-153 (Goldberger et al., 2000) datasets. Sleep-EDF-39 consists of 39 full-night recordings from 20 healthy subjects, while Sleep-EDF-153 includes 153 recordings from 78 healthy subjects. Each 30-second epoch is annotated with one of five sleep stages (W, N1, N2, N3, REM) according to AASM standards. For sleep stage detection, we used the Fpz-Cz EEG channel with a sampling rate of 100 Hz.

To validate our model's generalization capability, we recruited nine healthy university students (six males) and obtained informed consent for sleep EEG data collection.

EEG recordings were conducted during sleep using a 16-channel Ag/Ag-Cl electrode cap arranged according to the 10-20 system, sampled at 1000 Hz, and referenced to the vertex. A professional physician, Y1, manually annotated all 30-second EEG epochs for sleep staging. Additionally, participants completed face-emotional recognition tasks (positive, neutral, and negative) before and after sleep. These data were exclusively used to investigate the impact of sleep staging on the recognition speed of emotional facial expressions post-sleep. The study was approved by the university's ethics committee and complies with the World Medical Association's Declaration of Helsinki.

Table 1: Overall performance comparison of GPNet and other algorithms on the Sleep-EDF-39 dataset.

Method	Param (M)	Chan.	ACC	MF1	κ
AttnSleepNet	0.5	1	81.25	**73.11**	72.66
MMASleepNet	0.6	2	79.18	71.87	70.02
TransSleep	22.82	1	81.94	73.18	73.45
Cross-modal transformer	0.75	1	80.52	71.00	71.20
		2	81.42	72.50	72.74
CareSleepNet	19.54	2	80.20	72.27	71.14
PSEENet	2.92	1	80.25	71.56	70.83
		2	80.53	71.02	73.08
GPNet (ours)	0.4	1	**83.17**	72.08	**74.35**

Table 2: Overall performance comparison of GPNet on the Sleep-EDF-153 dataset.

Method	Chan.	ACC	MF1	κ
AttnSleepNet (21')	1	77.78	69.72	69.18
MMASleepNet (22')	2	76.76	68.94	68.02
TransSleep (23')	1	78.86	70.29	70.29
Cross-modal transformer (24')	1	76.91	68.90	67.65
	2	79.17	71.84	70.97
CareSleepNet (24')	2	78.09	67.45	69.54
PSEENet (24')	1	78.28	69.87	69.41
	2	80.30	72.71	72.21
GPNet (ours)	1	**80.95**	**73.35**	**73.03**

B. Baselines and Experimental Setup

In this study, GPNet was compared with six baseline models: AttnSleepNet (Eldele et al., 2021), MMASleepNet (Zheng et al., 2022), TransSleep (Phyo et al., 2023), Cross-modal Transformers (Pradeepkumar et al., 2024), CareSleepNet (Wang et al., 2024), and PSEENet (Zhou et al., 2024).

To ensure a fair comparison, we utilized the open code for each baseline model and trained them in the same environment and configuration as our proposed model. We employed single-epoch EEG signals from the Sleep-EDF-X public dataset as input for all models, applying identical preprocessing and data processing steps. A five-fold cross-validation scheme was adopted, randomly partitioning the data into five groups with no overlapping subjects across folds. Some baseline models were evaluated separately on single-modality EEG (1) and multimodal EEG + EOG (2) data, as specified in their respective publications.

C. Sleep Stage Classification

We compared GPNet with six baseline methods on the Sleep-EDF-X dataset, evaluating model parameters, accuracy (ACC), macro F1-score (MF1), and Cohen's kappa (κ). Tables I and II present the performance of Sleep-EDF-39 and Sleep-EDF-153, respectively. GPNet demonstrates highly competitive performance across most metrics. On Sleep-EDF-39, GPNet significantly outperforms the latest Cross-modal Transformer in all metrics while maintaining a similar number of parameters. Although PSEENet achieves higher precision, GPNet uses less than one-seventh of its parameters, highlighting the efficiency and effectiveness of our approach.

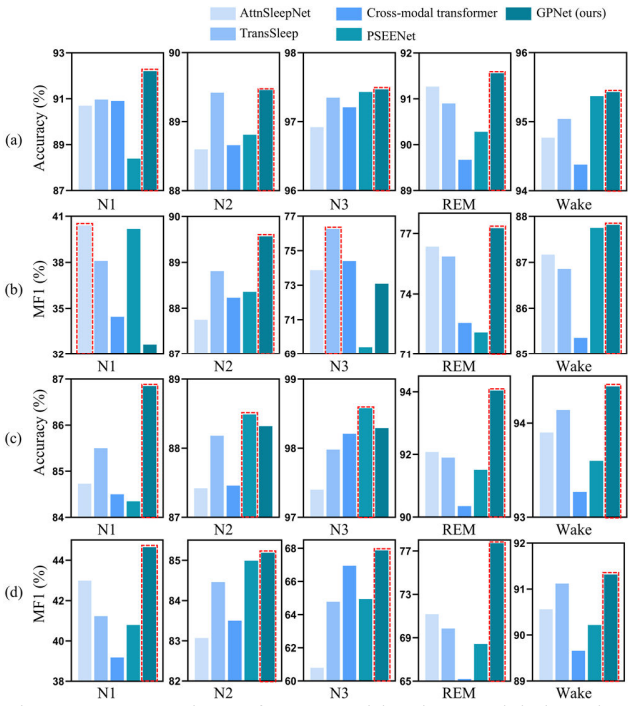


Figure 4: Comparison of GPNet with other models based on accuracy and F1-score for each sleep stage. (a) and (b) present results for Sleep-EDF-39, while (c) and (d) show results for Sleep-EDF-153.

On the Sleep-EDF-153 dataset, GPNet outperforms all EEG-only baseline methods across every metric, achieving state-of-the-art results in ACC, MF1, and κ. Figure 4 illustrates the classification performance for each sleep stage, where GPNet consistently achieves the highest accuracy across all stages. Specifically, on Sleep-EDF-39, GPNet attains a 97.47% accuracy for the N3 stage—the highest among all stages—and over 90% accuracy for the remaining stages, highlighting the effectiveness of our feature pyramid

network's time-frequency representation. On Sleep-EDF-153, GPNet consistently exceeds competing algorithms for every sleep stage. Notably, all baseline methods in this comparison utilize EEG as the input modality.

D. Ablation experiment

We performed a module-based ablation study on the Sleep-EDF-39 dataset to assess the impact of individual modules and verify the effectiveness of each proposed component in GPNet. The study included the following variants:

base: After extracting time-frequency features from single-channel EEG using a feature pyramid, the resulting features are directly passed to fully connected layers for classification.

base + ADAM Block: Building upon the base variant, we introduce the proposed ADAM module in the penultimate layer to enhance the interaction between global and local information, highlight essential features, and discard redundant ones.

GPNet: Extending variant 2, we incorporate residual graph convolution to further integrate high-level features from each layer, followed by fully connected layers for final sleep stage detection.

From these ablation experiments (see Figure 5), we draw three main conclusions: First, the feature pyramid extraction network effectively captures features for each sleep stage. Second, introducing the ADAM block is critical for enriching feature information and eliminating redundancy. Finally, GPNet outperforms the other variants on most metrics, indicating that each module is indispensable and that they work in synergy to achieve superior performance.

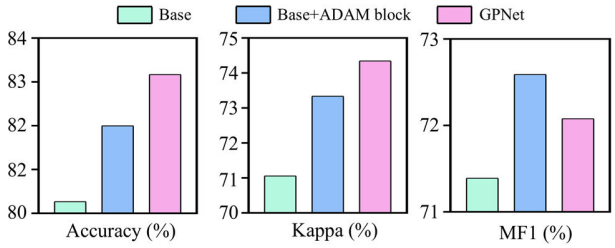


Figure 5: Ablation results of the key modules in the proposed GPNet network.

E. Relationship Between Sleep Staging and Active Emotion Recognition

To validate the generalization capability of our model, we performed a leave-one-subject-out cross-validation (using the Fp1 electrode EEG signals as input) to obtain the proportion of each sleep stage for each participant and then examined the relationship between these proportions and active emotion recognition performance before and after sleep. The results revealed a significant negative correlation (r=-0.73, as shown in Figure 6) between the proportion of REM sleep and the difference in response time before and after sleep, i.e., a higher proportion of REM sleep was associated with faster emotional processing. This finding aligns with previous studies indicating significantly increased activity during REM in subcortical regions such as the amygdala, striatum, and hippocampus, as well as cortical areas including the insula and medial prefrontal cortex (mPFC), all of which are implicated in emotion (Dang-Vu et al., 2010). REM sleep is characterized by mixed-frequency EEG activity (theta, alpha, beta, and gamma) related to cognitive and emotional processing during dreaming (Nir & Tononi, 2010). Additionally, REM sleep has been linked to the consolidation of emotional memories and emotional regulation (Van Der Helm et al., 2011), supporting emotional homeostasis in the brain and preparing the organism for optimal social and emotional functioning the next day (Goldstein & Walker, 2014; Alkalame et al., 2024). Taken together, REM sleep facilitates emotional processing, enhancing the brain's capacity to recognize and process negative emotions, which explains why an increased proportion of REM sleep corresponds to faster recognition of negative emotions.

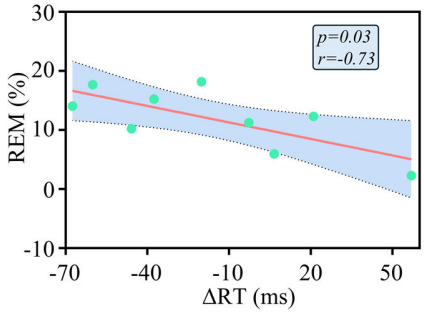


Figure 6: Correlation between REM (%) and the change in facial emotion recognition response post-sleep.

IV. Conclusion

This study introduces GPNet, a novel end-to-end method for sleep stage detection based on single-channel EEG signals. GPNet integrates three key components: a pyramid feature extraction module, an adaptive deep attention mechanism, and a graph convolution-based multi-level feature fusion. The pyramid feature extraction module decomposes the single-channel EEG input into multiple resolutions, allowing for fine-grained time-frequency feature extraction through a series of ConvPool blocks. The adaptive deep attention mechanism enhances salient features by learning interactions between local and global information, effectively suppressing redundant data. Finally, the graph convolutional network integrates high-level features from different resolutions, facilitating cross-level interaction and coupling. Experimental results on the Sleep-EDF-X dataset demonstrate that GPNet achieves superior subject-independent performance compared to state-of-the-art models. Additionally, when applied to a laboratory-collected dataset, GPNet revealed a significant negative correlation between the proportion of REM sleep and the recognition of negative emotions. These findings highlight GPNet's potential for future clinical applications, where AI can collaborate with clinicians to enhance diagnostic accuracy.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62171074, W2411084; the Chongqing Natural Science Fund Innovation and Development Joint Fund (Municipal Education Commission) project under Grant 2024NSCQ-LZX0058 and the funding of Institute for Advanced Sciences of Chongqing University of Posts and Telecommunications under Grant E011A2022327.

References

Alkalame, L., Ogden, J., Clark, J. W., Porcheret, K., Risbrough, V. B., & Drummond, S. P. (2024). The relationship between REM sleep prior to analog trauma and intrusive memories. *Sleep*, *47*, zsae203.

Bagautdinova, J., Mayeli, A., Wilson, J. D., Donati, F. L., Colacot, R. M., Meyer, N., ... & Ferrarelli, F. (2023). Sleep abnormalities in different clinical stages of psychosis: a systematic review and meta-analysis. *JAMA psychiatry*, *80*, 202-210.

Bao, J., Wang, G., Wang, T., Wu, N., Hu, S., Lee, W. H., ... & Wang, G. (2024). 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*, 6641-6652.

Berry, R. B., Budhiraja, R., Gottlieb, D. J., Gozal, D., Iber, C., Kapur, V. K., ... & Tangredi, M. M. (2012). Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. *Journal of clinical sleep medicine*, *8*, 597-619.

Cusinato, R., Gross, S., Bainier, M., Janz, P., Schoenenberger, P., & Redondo, R. L. (2024). Workflow for the unsupervised clustering of sleep stages identifies light and deep sleep in electrophysiological recordings in mice. *Journal of Neuroscience Methods*, *408*, 110155.

Dang-Vu, T. T., Schabus, M., Desseilles, M., Sterpenich, V., Bonjean, M., & Maquet, P. (2010). Functional neuroimaging insights into the physiology of human sleep. *Sleep*, *33*, 1589-1603.

Eldele, E., Chen, Z., Liu, C., Wu, M., Kwoh, C. K., Li, X., & Guan, C. (2021). 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.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, *101*, e215-e220.

Goldstein, A. N., & Walker, M. P. (2014). The role of sleep in emotional brain function. *Annual review of clinical psychology*, *10*, 679-708.

Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A., & Oberye, J. J. (2000). Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, *47*, 1185-1194.

Li, W., Liu, T., Xu, B., & Song, A. (2024). SleepFC: Feature Pyramid and Cross-Scale Context Learning for Sleep Staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *32*, 2198-2208.

Li, Y. (2024). DPNet: Scene text detection based on dual perspective CNN-transformer. *Plos one*, *19*, e0309286.

Li, Y., Chen, J., Ma, W., Zhao, G., & Fan, X. (2022). MVF-SleepNet: Multi-view fusion network for sleep stage classification. *IEEE Journal of Biomedical and Health Informatics*, *28*, 2485-2495.

Luyster, F. S., Strollo Jr, P. J., Zee, P. C., & Walsh, J. K. (2012). Sleep: a health imperative. *Sleep*, *35*, 727-734.

Nir, Y., & Tononi, G. (2010). Dreaming and the brain: from phenomenology to neurophysiology. *Trends in cognitive sciences*, *14*, 88-100.

Pan, J., Feng, Y., Zhao, P., Zou, X., Hou, A., & Che, X. (2024). Causalattennet: A fast and long-term-temporal network for automatic sleep staging with single-channel eeg. *IEEE Transactions on Instrumentation and Measurement*, *73*, 1-13.

Phyo, J., Ko, W., Jeon, E., & Suk, H. I. (2022). TransSleep: Transitioning-aware attention-based deep neural network for sleep staging. *IEEE Transactions on Cybernetics*, *53*, 4500-4510.

Pradeepkumar, J., Anandakumar, M., Kugathasan, V., Suntharalingham, D., Kappel, S. L., De Silva, A. C., & Edussooriya, C. U. (2024). Towards interpretable sleep stage classification using cross-modal transformers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *32*, 2893-2904.

Rahman, M. M., Bhuiyan, M. I. H., & Hassan, A. R. (2018). Sleep stage classification using single-channel EOG. *Computers in biology and medicine*, *102*, 211-220.

Suetsugi, M., Mizuki, Y., Yamamoto, K., Uchida, S., & Watanabe, Y. (2007). The effect of placebo administration on the first-night effect in healthy young volunteers. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *31*, 839-847.

Sun, H., Wen, Y., Feng, H., Zheng, Y., Mei, Q., Ren, D., & Yu, M. (2024). Unsupervised Bidirectional Contrastive Reconstruction and Adaptive Fine-Grained Channel Attention Networks for image dehazing. *Neural Networks*, *176*, 106314.

Thuwajit, P., Rangpong, P., Sawangjai, P., Autthasan, P., Chaisaen, R., Banluesombatkul, N., ... & Wilaiprasitporn, T. (2021). EEGWaveNet: Multiscale CNN-based spatiotemporal feature extraction for EEG seizure detection. *IEEE Transactions on Industrial Informatics*, *18*, 5547-5557.

Van Der Helm, E., Yao, J., Dutt, S., Rao, V., Saletin, J. M., & Walker, M. P. (2011). REM sleep depotentiates amygdala activity to previous emotional experiences. *Current biology*, *21*, 2029-2032.

Wang, J., Zhao, S., Jiang, H., Zhou, Y., Yu, Z., Li, T., ... & Pan, G. (2024). CareSleepNet: a hybrid deep learning

network for automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics*, *28*, 7392-7405.

Wei, Y., Zhu, Y., Zhou, Y., Yu, X., & Luo, Y. (2023). Automatic sleep staging based on contextual scalograms and attention convolution neural network using single-channel EEG. *IEEE Journal of Biomedical and Health Informatics*, *28*, 801-811.

Wolpert, E. A. (1969). A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. *Archives of General Psychiatry*, *20*, 246-247.

Wu, Y., Jia, Y., Ning, X., Xu, Z., & Rosen, D. (2022). Detection of pediatric obstructive sleep apnea using a multilayer perceptron model based on single-channel oxygen saturation or clinical features. *Methods*, *204*, 361-367.

Zhang, J., & Wu, Y. (2021). Competition convolutional neural network for sleep stage classification. *Biomedical Signal Processing and Control*, *64*, 102318.

Zhang, Z., Lin, B. S., Peng, C. W., & Lin, B. S. (2024). Multi-Modal Sleep Stage Classification with Two-Stream Encoder-Decoder. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *32*, 2096-2105.

Zheng, Y. B., Luo, Y. Y., Zou, B., Zhang, L., & Li, L. (2022). MMASleepNet: A multimodal attention network based on electrophysiological signals for automatic sleep staging. *Frontiers in Neuroscience*, *16*, 973761.

Zhou, W., Shen, N., Zhou, L., Liu, M., Zhang, Y., Fu, C., ... & Chen, C. (2024). PSEENet: A pseudo-siamese neural network incorporating electroencephalography and electrooculography characteristics for heterogeneous sleep staging. *IEEE Journal of Biomedical and Health Informatics*, *28*, 5189-5200.