DOI: https://doi.org/10.24297/ijct.v25i.9815

CaT-SleepNet: A Cross-Attention and Transformer-Based Hybrid Framework for EEG–EOG Sleep Stage Classification

Jin Peng[1], *Haodong Fang*[1], *Yuanyuan Sheng*[1], *Wensheng Liu*[1], *Yuyue Wu*[1], *Ruiheng Xie*[1], *Li Zhu*[1]

[1]School of Computer Science and Technology, Hangzhou Dianzi University, China

Corresponding author is Li Zhu and email address: zhuli@hdu.edu.cn

Abstract

sleep disorders and understanding sleep mechanisms. However, traditional deep learning models often fail to effectively capture both temporal dependencies within Electroencephalogram(EEG) signals and the semantic correlations between multimodal inputs. In this study, we propose a dual-stream Transformer-based framework that integrates raw EEG and electro-oculogram (EOG) signals and their corresponding time-frequency (TF) representations through a cross-attention fusion mechanism. Each modality is first processed by independent feature extractors, followed by pre-trained channel-level Transformers to capture intra-channel temporal dependencies. Finally, a global Transformer module is used for feature extraction, and the learned representations are classified using an optimized XGBoost classifier to enhance discrimination ability. Experimental results on the Sleep-EDF-20, Sleep-EDF-78, and ISRUC-S3 datasets show that the proposed model achieves accuracies of 88.5%, 86.8%, and 84.0%, respectively, outperforming several state-of-the-art baselines and confirming the effectiveness of the proposed multimodal fusion and hybrid Transformer-XGBoost design for sleep stage classification.

Keywords: Sleep staging; EEG; Transformer; Cross-attention; Multimodal fusion; XGBoost

1 Introduction

Sleep is an indispensable physiological process for humans. Good sleep helps consolidate cognitive function and memory, and facilitates physiological repair and endocrine regulation. Prolonged poor sleep quality can lead to sleep disorders, which are becoming increasingly prevalent worldwide.

Sleep staging, as an effective and core objective indicator for assessing sleep quality, can assist doctors in identifying specific sleep disorders in patients, thereby enabling effective treatment. Sleep staging typically classifies sleep stages using various physiological signals such as EEG, electromyography (EMG), and EOG. In clinical practice, polysomnography (PSG) is an important method for sleep staging. According to the (RK) rules, PSG signals are divided into 30-second segments and classified into 6 sleep stages: Wake, 4 non-rapid eye movement (NREM) stages, and rapid eye movement (REM) stage.

Traditionally, sleep stage experts need to visually interpret PSG signal recordings, but this method relies on the subjective judgment of experts. To address this issue, many studies extract features from the original signals and then select some features for automatic sleep staging using machine learning. Common machine learning classification methods include Random Forest, Hidden Markov Models, and Support Vector Machines. Huang et al. introduced a single-channel EEG sleep staging method that uses a Transition-Optimized Hidden Markov Model (TO-HMM), combining feature extraction with Power Spectral Density (PSD) feature selection and a Gaussian Mixture Model-Hidden Markov Model (GMM-HMM), achieving an accuracy of 86.4% (Huang, Ren, Ji, et al., 2022). Shen et al.

proposed an algorithm using Improved Model-Based Essence Features (IMBEF) and Bagged Trees for classification, demonstrating high accuracy on the Sleep-EDF and DREAMS datasets, and achieving a staging accuracy of 81.65% on the ISRUC dataset (Shen et al., 2020).

Although traditional machine learning algorithms perform well in sleep staging tasks, their feature selection heavily relies on the prior sleep medicine knowledge of researchers. With the development of deep learning technologies, an increasing number of studies have begun to adopt such methods to achieve more end-to-end automatic sleep stage classification. Researchers have applied Convolutional Neural Networks (CNNs) to sleep staging work, also achieving success. Sors et al. proposed a 14-layer CNN that takes the sleep epoch to be classified, along with the two preceding and following epochs, as input, achieving an accuracy of 87% (Sors, Bonnet, Mirek, Vercueil, & Payen, 2018). Meanwhile, Recurrent Neural Networks (RNNs) have also been applied to the field of sleep staging to capture temporal dependencies in PSG data. SeqSleepNet uses bidirectional RNNs to extract epoch-level and sequence-level features. An attention-based recurrent layer captures short-term sequential features, followed by another recurrent layer that learns epoch-level features for long-term modeling, achieving an accuracy of 87% (Phan, Andreotti, Cooray, Chén, & De Vos, 2019). Inspired by Transformer networks, some works have adopted attention mechanisms to replace RNNs due to their high model complexity. Eldele et al. proposed a temporal context encoder using multi-head self-attention layers based on causal convolutions, which was proven to capture temporal dependencies. This method achieved 84.4% on the Sleep-EDF-78 dataset and 81.3% on the Sleep-EDF-20 dataset (Eldele et al., 2021). Recent works such as FlexSleepTransformer (Guo, Nowakowski, & Dai, 2024) adapt to varying input channel configurations, and Vision Transformer-based models like (Lee, Choi, Lee, et al., 2025) offer interpretability for PSG signals.Also,Transformer-based models have gained popularity in recent sleep staging tasks (?, ?). Meanwhile, hybrid frameworks combining machine learning and deep features, e.g. CNN + XGBoost, have also been explored (Y. Wang et al., 2024)

While deep learning applied to sleep staging achieves high accuracy, the models generally lack generalization mechanisms, and the staging accuracy can be further improved. Given the limited generalization ability of current methods, we propose CaT-SleepNet, a deep learning model based on a Transformer encoder that uses two channels of EEG and one channel of EOG data for sleep staging, combining an attention mechanism with an XGBoost classifier to improve the accuracy of sleep stage classification. The model first preprocesses the data, performs Fourier transform to obtain spectrograms, and then uses CNN to independently extract features from the raw data channels of EEG and EOG. Subsequently, a cross-attention mechanism is used for bidirectional modality fusion between the tf-data of each channel and the processed raw data, allowing time-domain and frequency-domain features to enhance each other. Then, a pre-trained Transformer is used for channel and global modeling to output feature values. Finally, the feature values are input into XGBoost for training to output classification results. We summarize our main contributions as follows:

- We propose a multi-modal sleep staging framework that effectively integrates information from EEG and EOG, thereby improving the accuracy of sleep staging.
- We employ a pre-trained Transformer strategy to enhance the model's data generalization capability, thereby improving training accuracy.
- We conducted extensive experiments on three datasets to validate the effectiveness of the method.

2 Materials and Methods

2.1 Proposed Framework

Our sleep staging framework integrates raw EEG and EOG signals with their corresponding spectrogram representations to enhance the accuracy of sleep stage classification through effective multimodal feature fusion (shown in the Figure 1). The framework consists of four main modules. First, raw EEG and EOG signals are processed by convolutional neural networks (CNNs) to extract temporal domain features, while the TF data are projected into a suitable feature space for fusion. Then, a bidirectional cross-modal attention mechanism is employed to fuse the time–frequency representations and raw features, dynamically learning the complementary information between the two modalities. The fused features are subsequently modeled by a pre-trained Transformer encoder at the channel level to capture intra-channel dependencies, followed by a global Transformer encoder to learn inter-channel relationships and global contextual information. Finally, the extracted and contextually modeled features are fed into an XGBoost classifier for the final sleep stage prediction. Through dual-stream feature extraction, cross-modal attention fusion, and Transformer-based context modeling, the proposed framework effectively integrates raw and time–frequency representations, thereby improving the accuracy of sleep stage classification.

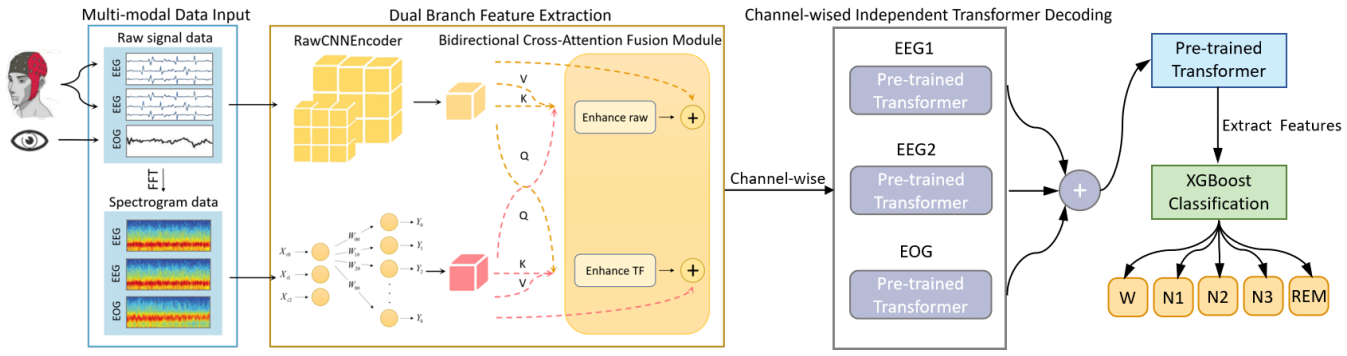


Figure 1: The overall framework of Cat-SleepNet

2.2 Datasets and Preprocessing

This study adopts multiple publicly available polysomnographic (PSG) datasets, including Sleep-EDF Expanded and ISRUC-Sleep Group 3, to comprehensively evaluate the proposed model's stability and generalization ability across different subjects and acquisition conditions. And MASS Dataset is used for preprocessing the transformer. All data were annotated by professional sleep technicians according to the American Academy of Sleep Medicine (AASM) standards, where each 30-second segment (epoch) was categorized into five sleep stages: Wake (W), light sleep (N1, N2), deep sleep (N3), and rapid eye movement (R).

Sleep-EDF Expanded Dataset: The Sleep-EDF Expanded dataset, jointly released by the European Sleep Research Society (ESRS) and the Donders Institute, is one of the most representative datasets in sleep research. Two subsets were used in this study: Sleep-EDF-20 and Sleep-EDF-78. The Sleep-EDF-20 subset includes overnight EEG recordings from 20 healthy subjects, with approximately 8 hours per subject and a sampling rate of 100 Hz. The Sleep-EDF-78 subset contains 153 nights from 78 healthy adults, also sampled at 100 Hz, providing a broader population coverage. Both subsets include two EEG channels (Fpz–Cz and Pz–Oz) and one EOG channel, with sleep stages annotated for every 30-second epoch.

ISRUC-Sleep Dataset: The ISRUC-Sleep dataset, published by the University of Coimbra in Portugal, contains multi-channel PSG recordings from subjects with varying sleep health conditions. Group 3 (ISRUC-S3), is used in this study. ISRUC-S3 includes 10 healthy adults, also recorded at 200 Hz. Both subsets provide multi-channel EEG, EOG, and EMG signals. In this study, only EEG (Fpz-Cz, Pz-Oz) and EOG channels were utilized, and all annotations followed the AASM standards.

MASS Dataset: The Montreal Archive of Sleep Studies is an open-access database containing laboratory-based multi-channel PSG recordings. It aims to provide a standardized and accessible dataset for benchmarking automated sleep analysis methods. The dataset includes full-night recordings from 200 participants, sampled at 256 Hz, with 4 to 20 EEG channels, as well as standard EOG, EMG, ECG, and respiratory signals. All recordings were manually scored by professional sleep technicians according to AASM standards.

For data preprocessing, all datasets, including Sleep-EDF, ISRUC-S3, and MASS, were preprocessed following a unified pipeline. EEG and EOG recordings were segmented into 30-second epochs and standardized using Z-score normalization. For Sleep-EDF, signals were extracted from PSG and hypnogram EDF files and labeled into five sleep stages. For ISRUC-S3, recordings were downsampled from 200 Hz to 100 Hz, and horizontal EOG was derived as the differential signal between ROC-A1 and LOC-A2. For MASS, raw signals sampled at 256 Hz were resampled to 100 Hz to ensure consistency across datasets. All recordings were band-pass filtered between 0.3–35 Hz to remove noise and DC drift.

2.3 Time-Frequency Representation

To extract spectral features from raw EEG and EOG signals, each 30-second epoch was transformed into a TF representation using the Short-Time Fourier Transform (STFT). The continuous signal $x(t)$ was divided into overlapping segments by applying a 2-second Hamming window with 1-second overlap. For each segment, the local frequency spectrum was computed through the discrete Fourier transform:

$$X(f,\tau) = \sum_{t=0}^{T-1} x(t)\, w(t-\tau)\, e^{-j2\pi ft}, \tag{1}$$

where $w(t-\tau)$ denotes the window function centered at time τ, and f represents the frequency index. A 256-point FFT was used to obtain the magnitude spectrum, which was then converted into the logarithmic scale as $20\log_{10}|X(f,\tau)|$ to emphasize relative spectral variations. The resulting spectrograms were resized to a consistent shape of 29×128, corresponding to 29 time frames and 128 frequency bins. Finally, all TF maps were standardized to zero mean and unit variance to ensure comparability across subjects. This process yields a compact two-dimensional representation that preserves both temporal and spectral information, serving as a complementary input to the raw waveform domain.

2.4 Dual-Branch Feature Extraction

2.4.1 Convolutional Neural Network for Domain-Specific Feature Extraction

In the proposed dual-stream framework, raw EEG and its TF representations are processed through two domain-specific feature extractors to capture complementary temporal and spectral information.

For the raw EEG branch, a two-layer one-dimensional Convolutional Neural Network (1D-CNN) is employed to learn hierarchical temporal features. The first convolution layer, with a narrow receptive field, focuses on short-term waveform

dynamics such as sleep spindles and K-complexes, while the second layer integrates higher-level rhythmic patterns across broader temporal contexts. Each convolution block is followed by a ReLU activation function to introduce non-linearity and enhance discriminative representation. To ensure temporal consistency across subjects, an adaptive average pooling layer is applied to map the variable-length outputs into a fixed 29-frame sequence. Such convolutional architectures have been widely validated for EEG temporal modeling due to their local perception and translation-invariant properties .

For the TF branch, each spectrogram patch of size 29×128 is projected into the latent feature space using a linear transformation followed by GELU activation and dropout regularization. This operation aligns the spectral dimension with the temporal embedding dimension of the raw branch, facilitating subsequent multimodal fusion. The linear projection not only compresses redundant frequency components but also enhances the semantic correspondence between temporal and spectral modalities, consistent with practices in recent multimodal EEG models.

These two feature extractors provide domain-optimized embeddings—temporal for raw signals and spectral for TF maps, which serve as the foundation for the subsequent cross-attention fusion and channel-level Transformer encoding.

2.4.2 Bidirectional Cross-Attention Fusion Module

To effectively integrate TF representations and raw-domain signals from EEG and EOG channels, we propose a bidirectional cross-attention fusion module. This module facilitates complementary information exchange between TF and raw domains, enabling frequency-specific patterns to enhance temporal features while allowing temporal dynamics to guide spectral attention. The detailed calculation process is shown in Figure 2.

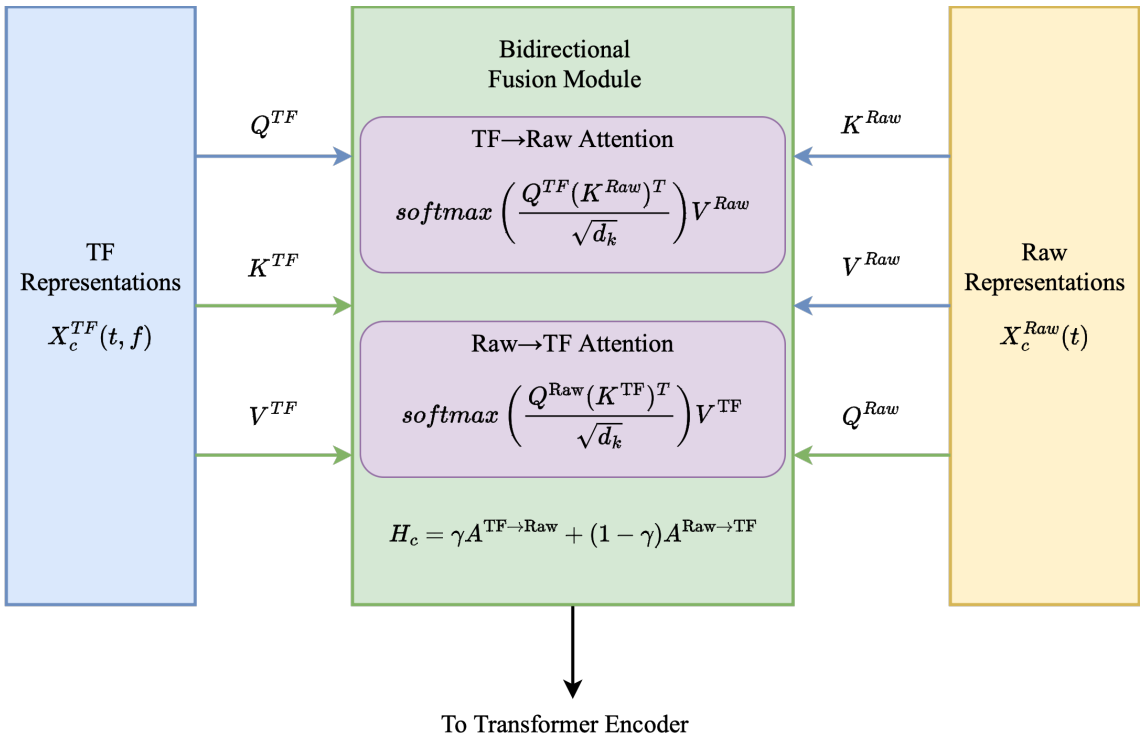


Figure 2: Architecture of bidirectional cross-attention fusion module

The TF representation for channel c is obtained via Short-Time Fourier Transform:

$$X_c^{TF}(t,f) = \sum_{n=0}^{N-1} x_c(t+n) \cdot w(n) \cdot e^{-j2\pi fn/N} \tag{2}$$

where $w(n)$ is the window function of length N, t and f denote time and frequency indices, and $x_c(t)$ is the raw signal. The corresponding spectrogram is $P_c^{TF}(t,f) = |X_c^{TF}(t,f)|^2$.

The raw-domain representation is extracted using a convolutional encoder:

$$X_c^{Raw}(t) = f_{\text{CNN}}(x_c(t); \theta_{\text{CNN}}) \tag{3}$$

where $f_{\text{CNN}}(\cdot;\theta_{\text{CNN}})$ denotes the CNN parameterized by θ_{CNN}.

To address dimensional mismatch between 2D TF and 1D raw representations, we flatten the TF representation along the frequency dimension:

$$\tilde{X}_c^{TF} = \text{Flatten}(X_c^{TF}) \cdot W_{\text{flat}} \tag{4}$$

The bidirectional cross-attention couples these representations. For TF-to-Raw attention:

$$Q^{TF} = \tilde{X}_c^{TF} W_Q^{TF} \tag{5}$$

$$K^{Raw} = X_c^{Raw} W_K^{Raw} \tag{6}$$

$$V^{Raw} = X_c^{Raw} W_V^{Raw} \tag{7}$$

$$A^{TF\to Raw} = \text{softmax}\left(\frac{Q^{TF}(K^{Raw})^T}{\sqrt{d_k}}\right) V^{Raw} \tag{8}$$

For Raw-to-TF attention:

$$Q^{Raw} = X_c^{Raw} W_Q^{Raw} \tag{9}$$

$$K^{TF} = \tilde{X}_c^{TF} W_K^{TF} \tag{10}$$

$$V^{TF} = \tilde{X}_c^{TF} W_V^{TF} \tag{11}$$

$$A^{Raw\to TF} = \text{softmax}\left(\frac{Q^{Raw}(K^{TF})^T}{\sqrt{d_k}}\right) V^{TF} \tag{12}$$

The fused representation combines both attention outputs:

$$H_c = \gamma A^{TF\to Raw} + (1-\gamma) A^{Raw\to TF} \tag{13}$$

where $\gamma = \sigma(w_\gamma)$ is a learnable gating parameter with $\sigma(\cdot)$ denoting sigmoid function, ensuring $0 \le \gamma \le 1$. H_c is forwarded to the Transformer encoder for global modeling.

2.5 Channel and Global Context Modeling with Pre-trained Transformer Module

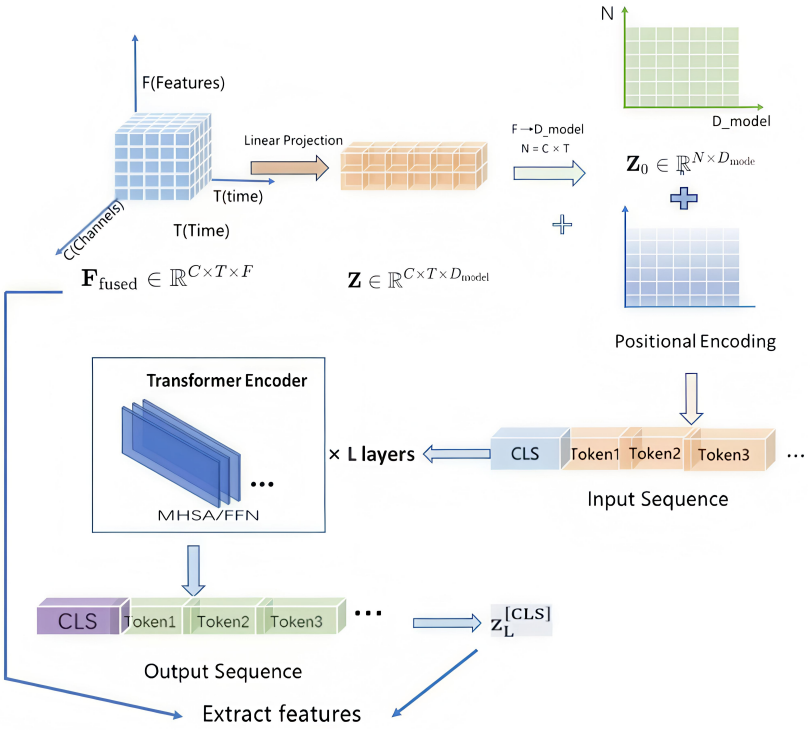


Figure 3: Architecture of transformer decoding model

To comprehensively exploit both intra-channel temporal dependencies and inter-channel correlations in EEG and EOG signals, a hierarchical Transformer-based architecture is designed, consisting of two stages: the Channel-level Transformer and the Global Transformer (shown in Figure 3). This structure ensures that temporal consistency within each channel is preserved while global dependencies across channels are effectively modeled, yielding more discriminative representations of sleep states.

At the channel-level, an independent Transformer encoder is assigned to each channel to capture temporal dependencies within individual signals. Each channel input includes the raw time-domain signal and its corresponding TF data. These two modalities are first integrated through a Cross-Attention Module, which dynamically fuses complementary temporal and spectral information into a unified embedding. The fused representation is then processed by the channel-level Transformer for sequential modeling, which were mentioned in the reference (?, ?).

Within each channel-level Transformer, the input features are first linearly projected into a fixed-dimensional embedding space. The Multi-head Self-Attention mechanism then models long-range temporal dependencies by assigning adaptive attention weights to key time steps—such as those containing sleep spindles or K-complexes—that are highly indicative of sleep stages. The Feed-Forward Network (FFN) further enhances representation learning through nonlinear transformation. As a result, each channel produces a high-level temporal representation that encodes both short- and long-term EEG dynamics in a physiologically meaningful manner. This design maintains independent feature learning for each channel while preserving long-range temporal consistency, providing robust intra-channel feature embeddings for subsequent global modeling.

Following channel-level encoding, the model obtains high-dimensional temporal representations from all channels.

Although each channel independently captures its local dynamics, strong physiological correlations exist across them. For instance, during the REM stage, the EOG channel exhibits large eye movements, while EEG channels show mixed-frequency, low-amplitude activity; conversely, during N3 sleep, multiple EEG channels exhibit synchronized δ-wave activity. These cross-channel interactions motivate the introduction of the Global Transformer to capture inter-channel dependencies and global contextual relationships.

In the Global Transformer, the outputs of the channel-level Transformers are concatenated and fed into a pre-trained Transformer encoder, where Multi-head Self-Attention learns interactions across channels and their corresponding temporal features. Each channel representation is treated as a "global token", allowing the model to adaptively estimate the relative importance of different channels under varying sleep stages. For example, during REM, the model assigns greater attention to the EOG channel, while during deep sleep (N3), EEG channels dominate. This adaptive attention mechanism enables dynamic, physiologically consistent information fusion.

2.6 Classification with XGBoost Module

XGBoost is a widely used machine learning algorithm for classification tasks. In this study, we input feature vectors into the XGBoost model for training, and output the classification results of sleep stages. The predictive output of this method can be represented by the following formula:

$$\hat{y}_j = \sum_m \lambda_m x_{jm} \tag{14}$$

Hyperparameters refer to the parts of the model that need to be learned from the data. The hyperparameters of XGBoost can be divided into three categories: General Parameters (GP), Booster Parameters (BP), and Learning Task Parameters (LTP). The objective of model training is to find the optimal hyperparameters λ for the training set $\mathcal{JS}^{(tr)}$ and its corresponding labels y_j. To evaluate the model's fit to the training set $\mathcal{JS}^{(tr)}$, we use the following objective function:

$$\text{obj}(\lambda) = L(\lambda) + \Omega(\lambda) \tag{15}$$

where L denotes the loss function during training, and Ω is the regularization term. Since this paper deals with a multi-class classification task, the 'multi-class log loss' ('mlogloss') is chosen as the error evaluation metric, which is defined as follows:

$$L = -\frac{1}{M}\sum_{j=1}^{M}\sum_{k=1}^{N} y_{jk}\log\left(p_{jk}\right) \tag{16}$$

Here, M is the number of samples in the training set $\mathcal{JS}$, p_j is the predicted probability of the XGBoost model for sample j, y_j is the true label, and N is the number of classes. To prevent overfitting, we introduce a regularization term, expressed as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\sum_{i=1}^{T}\omega_i^2 \tag{17}$$

where γ and λ are tunable hyperparameters, T represents the number of leaf nodes in the tree structure, and ω is the score vector on the leaf nodes. XGBoost employs an additive strategy to optimize the objective function step by step, meaning that the training in each round depends on the results of the previous round. Substituting the multi-class log loss function and the regularization term into the equation yields the objective function at the t-th round:

$$\text{obj}^{(t)} = L^{(t)} + \Omega(f_t) + C \tag{18}$$

where $L^{(t)}$ is the multi-class log loss at the t-th round, $\Omega(f_t)$ is the regularization term at the t-th round, and C is a constant.

2.7 Experimental Setup

We adopt 10-fold cross-validation for sleep staging in this experiment. In each fold, the dataset is divided into ten parts, with nine parts used for training and one for testing, and the process is repeated until all data are tested once. Follow previous work (Fu et al., 2023), we use the cross-entropy loss function with class weights of 0.3, 0.4, 0.3, 0.2, and 0.3 for W, N1, N2, N3, and REM, respectively, to handle the imbalance among sleep stages.

During the hyperparameter tuning phase, we selected the tree structure booster: 'dart'. This booster introduces a random drop mechanism to the model, forcing newly generated trees to compensate for the overall shortcomings, thereby making the model more robust and effectively mitigating overfitting. Learning task parameters are used to configure the training environment. The 'objective' parameter is set to 'multisoftmax' suitable for multi-class classification, with its required parameter 'num_class' set to 5, corresponding to the five sleep stages. The evaluation metric 'eval_metric' uses 'mlogloss' during both training and testing phases, while XGBoost's early stopping mechanism 'early_stopping_rounds' is set to 20. To construct an effective fitness function, we use classification accuracy as the core evaluation metric, thereby ensuring that particles with better fitness achieve higher accuracy in the classification task.

During model training, we employ the Adam optimizer with an initial learning rate of 5×10^{-6}. The batch size is set to 16, and the model is trained for up to 150 epochs with early stopping to prevent overfitting . The dropout rate of channal transformer is set to 0.3 and The dropout rate of global transformer is 0.1. The Transformer encoder has an embedding dimension of 128, 16 attention heads, 16 encoder layers for single-channel feature extraction, and 4 encoder layers for multi-channel feature fusion. These Settings are the same as the previous work.(Dai et al., 2023) All experiments are implemented in PyTorch and conducted on an NVIDIA GPU using CUDA acceleration.

3 Results and Discussion

3.1 Performance of Sleep Staging

Table 1: Sleep Stage Classification Performance Comparison

Dataset	Model	Overall			Per-class F1				
		Acc	**MF1**	**K**	**W**	**N1**	**N2**	**N3**	**R**
S-EDF-20	DeepSleepNet(Supratak, Dong, Wu, & Guo, 2017)	0.840	0.780	0.780	0.871	0.444	0.879	0.882	0.824
	AttnSleepNet(Eldele et al., 2021)	0.813	0.732	0.740	0.920	0.420	0.850	0.821	0.741
	SalientSleepNet(Jia et al., 2021)	0.863	0.805	0.810	0.907	0.498	0.890	0.847	0.884
	XSleepNet(Phan, Chén, et al., 2022)	0.841	0.813	0.798	0.915	0.478	0.879	0.798	0.819
	MMASleepNet(Zheng, Luo, Zou, Zhang, & Li, 2022)	0.873	0.826	0.826	0.922	0.548	0.897	**0.922**	0.864
	TSA-Net(Fu et al., 2023)	0.866	0.804	0.816	0.905	0.469	0.892	0.901	0.853
	DynamicSleepNet(W. Wang et al., 2023)	0.879	0.832	0.835	**0.934**	0.572	0.879	0.901	0.872
	Ours	**0.885**	**0.856**	**0.845**	0.909	**0.659**	**0.910**	0.897	**0.904**
S-EDF-78	DeepSleepNet	0.803	0.752	0.730	0.915	0.460	0.829	0.792	0.764
	TSA-Net	0.822	0.735	0.751	0.917	0.301	0.849	0.800	0.810
	AttnSleepNet	0.813	0.732	0.740	0.920	0.420	0.850	0.821	0.741
	XSleepNet	0.760	0.778	0.779	0.840	**0.778**	0.798	0.726	0.671
	SleepTransformer(Phan, Mikkelsen, et al., 2022)	0.849	0.789	0.788	0.935	0.485	0.865	0.809	0.846
	Sleep-SMC(Ma et al., 2025)	0.816	0.756	0.745	0.924	0.429	0.836	0.800	0.790
	Ours	**0.868**	**0.830**	**0.811**	**0.957**	0.625	**0.879**	**0.866**	**0.870**
ISRUC-S3	AttnSleepNet	0.766	0.748	0.699	0.862	0.517	0.766	0.875	0.720
	Sleep-SMC	0.793	0.782	0.734	0.876	0.572	0.775	0.872	0.812
	XSleepNet	0.671	0.644	0.577	0.789	0.428	0.691	0.805	0.508
	MixSleepNet(Ji, Li, Wen, Barua, & Acharya, 2024)	0.830	**0.821**	**0.782**	0.899	**0.625**	0.819	0.899	0.860
	Ours	**0.840**	0.817	0.781	**0.907**	0.570	**0.830**	**0.915**	**0.861**

As shown in Table 1, the method achieved consistently superior performance compared with several state-of-the-art baselines. On the Sleep-EDF-20 dataset, it reached an accuracy of 88.5% and a macro-F1 score of 85.6%, outperforming DynamicSleepNet and MMASleepNet. The improvement in the N1 stage demonstrates its ability to capture subtle sleep transitions. On the Sleep-EDF-78 dataset, the model achieved 86.8% accuracy and 83.0% macro-F1, surpassing SleepTransformer and other Transformer-based methods, indicating the effectiveness of the proposed multimodal fusion and cross-attention mechanism. On the ISRUC-S3 dataset, the model obtained 84.0% accuracy and 81.7% macro-F1, slightly higher than MixSleepNet, confirming its generalization ability across datasets with different recording conditions. Overall, the results verify that combining raw and time-frequency representations with pre-trained Transformers and XGBoost classification enhances the robustness and discriminative capability of sleep stage recognition.

3.2 Ablation Studies

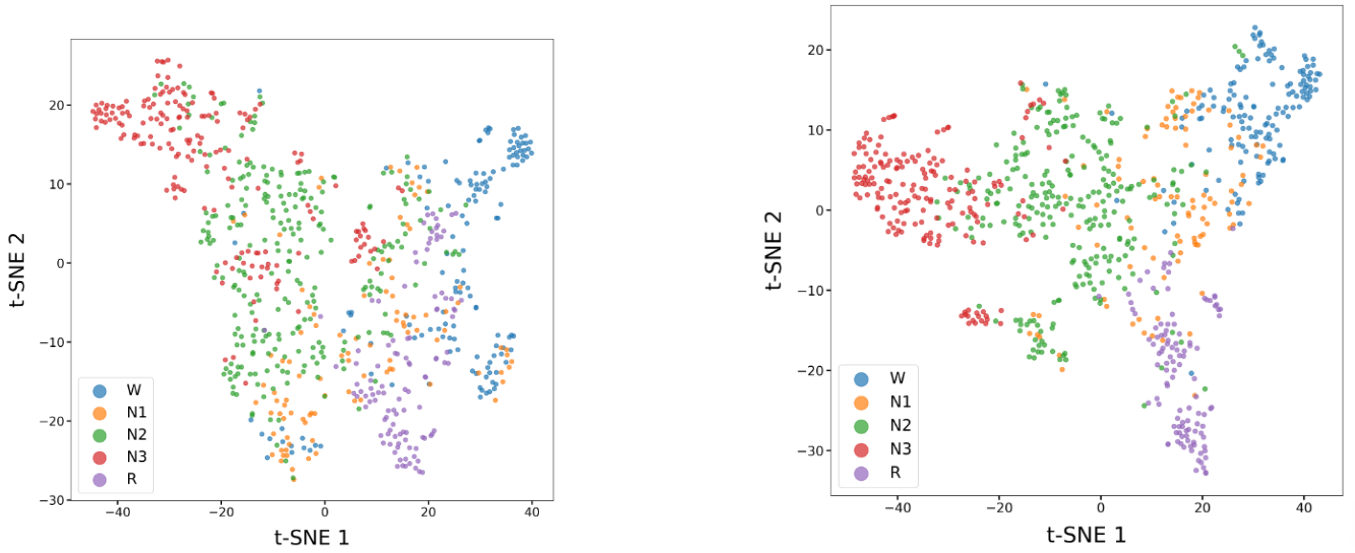
Table 2: Comprehensive ablation study on modality, structural components, pretraining, and classifiers.

Exp. ID	Input	Ch. Trans.	Cross-Att.	Classifier	Accuracy	Macro-F1
A1	Raw	✓	×	Linear	82.6	77.7
A2	TF	✓	×	Linear	85.1	81.3
A3	Raw+TF	✓	×	Linear	85.5	81.4
A4	Raw+TF	×	✓	Linear	86.7	83.2
A5	Raw+TF	✓	✓	Linear	87.5	83.6
A7	Raw+TF	✓	✓	XGBoost	86.9	83.0
A8	Raw+TF (Pret.)	✓	✓	XGBoost	**88.5**	**85.6**

To evaluate the contribution of each component in the proposed framework, a comprehensive ablation study was conducted on the Sleep-EDF-20 dataset, as summarized in Table 2. When using only the raw EEG input (A1), the model achieved an accuracy of 82.6% and a macro-F1 of 77.7%, indicating limited discriminative capacity. Using only the TF representation (A2) improved performance to 85.1% accuracy and 81.3% macro-F1, demonstrating the importance of spectral information. Combining both modalities without cross-attention (A3) further enhanced performance, confirming their complementarity. Introducing the cross-attention module (A4) significantly improved both metrics, highlighting its effectiveness in modeling semantic correlations between modalities. Incorporating channel-level Transformers (A5) provided additional gains, verifying the benefit of temporal context modeling within each channel. Replacing the linear classifier with XGBoost (A7) maintained competitive performance while enhancing stability and interpretability. Finally, applying pretraining to the Transformer backbone (A8) achieved the best results with 88.5% accuracy and 85.6% macro-F1, demonstrating the synergistic effect of multimodal fusion, pretraining, and hybrid classification.

3.3 Visualization and Analysis

To further evaluate the effectiveness and interpretability of the proposed model, we conducted visualization experiments using the ISRUC-S3 dataset, where the improvements were particularly evident. We employed t-distributed stochastic neighbor embedding (t-SNE) to visualize the feature representations learned by the model before and after the Transformer-based preprocessing module.



(a) Without pre-trained Transformer

(b) With pre-trained Transformer

Figure 4: t-SNE visualization of feature distributions on the ISRUC-S3 dataset. (a) Without pre-trained Transformer; (b) With pre-trained Transformer.

As shown in Figure 4a and 4b , the feature distribution without the Transformer preprocessing exhibits significant overlap among different sleep stages, indicating that the features extracted directly from raw signals are not well separated. In contrast, after applying the Transformer-based preprocessing, the clusters corresponding to W, N1, N2, N3, and REM stages become more compact and clearly distinguishable. This demonstrates that the Transformer encoder effectively captures discriminative temporal–spectral dependencies from the EEG signals, leading to more separable representations in the latent space.

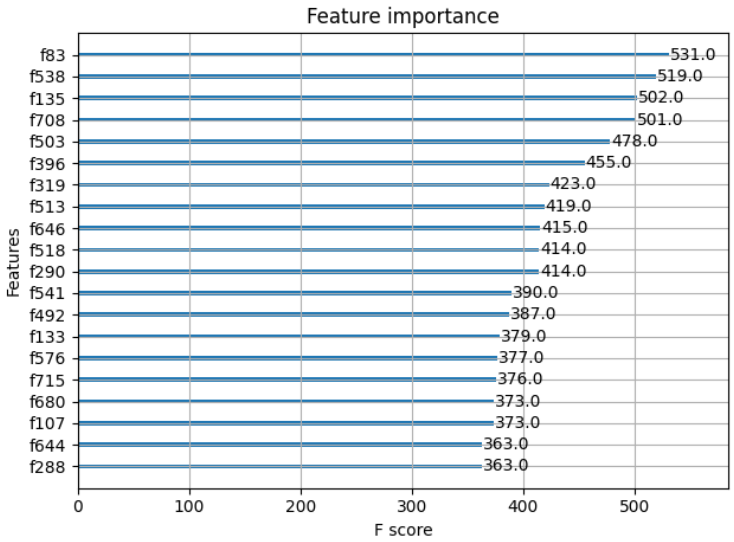


Figure 5: Feature contribution bar chart

In addition, we analyze the feature importance ranking obtained from the XGBoost classifier, as shown in Figure 5. The top20-ranked features primarily correspond to frequency-domain characteristics and high-level embeddings

from the Transformer encoder, confirming that both spectral and learned contextual features contribute substantially to accurate sleep stage classification. The feature importance distribution further validates that the model not only achieves high performance but also aligns with physiological understanding of sleep staging.

4 Conclusions

We proposed a Transformer-based multi-modal framework for automatic sleep stage classification using EEG and EOG signals. By combining raw time-domain and time-frequency features, the model effectively captures complementary information across modalities. The integrated attention and fusion modules enhance feature representation and inter-channel interaction. Experimental results demonstrate that our approach achieves superior performance and strong generalization. Future work will focus on self-supervised pretraining and lightweight model optimization for real-world deployment.

Data Availability (excluding Review articles)

The datasets used in this study are publicly available. The Sleep-EDF Expanded dataset can be accessed from the PhysioNet database at `https://physionet.org/content/sleep-edfx/1.0.0/`. The ISRUC-Sleep dataset is available at `https://sleeptight.isr.uc.pt/?pag`. The Montreal Archive of Sleep Studies (MASS) dataset can be accessed at `https://ceams-carsm.ca/en/MASS/`.

References

Dai, Y., Li, X., Liang, S., Wang, L., Duan, Q., Yang, H., ... Liao, X. (2023). Multichannelsleepnet: A transformer-based model for automatic sleep stage classification with psg. *IEEE Journal of Biomedical and Health Informatics*, *27*(9), 4204-4215.

Eldele, E., Ragab, M., Chen, Z., 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. doi: 10.1109/TNSRE.2021.3076234

Fu, G., Zhou, Y., Gong, P., Wang, P., Shao, W., & Zhang, D. (2023). A temporal-spectral fused and attention-based deep model for automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 1008–1018. doi: 10.1109/TNSRE.2023.3238852

Guo, Y., Nowakowski, M., & Dai, W. (2024). Flexsleeptransformer: a transformer-based sleep staging model with flexible input channel configurations. *Scientific Reports*, *14*, 26312. doi: 10.1038/s41598-024-76197-0

Huang, J., Ren, L., Ji, Z., et al. (2022). Single-channel eeg automatic sleep staging based on transition optimized hmm. *Multimedia Tools and Applications*, *81*, 43063–43081. doi: 10.1007/s11042-022-12551-6

Ji, X., Li, Y., Wen, P., Barua, P., & Acharya, U. R. (2024). Mixsleepnet: A multi-type convolution combined sleep stage classification model. *Computer Methods and Programs in Biomedicine*, *244*, 107992. doi: 10.1016/j.cmpb.2023.107992

Jia, Z., Lin, Y., Wang, J., Wang, X., Xie, P., & Zhang, Y. (2021). Salientsleepnet: Multimodal salient wave detection network for sleep staging. In *Proceedings of the thirtieth international joint conference on artificial intelligence (ijcai-21)* (pp. 2614–2620). International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2021/360

Lee, H., Choi, Y., Lee, H., et al. (2025). Explainable vision transformer for automatic visual sleep staging on multimodal psg signals. *npj Digital Medicine*, *8*, 55. doi: 10.1038/s41746-024-01378-0

Ma, S., et al. (2025). Ubiquitous sleep staging via supervised multimodal coordination (sleepsmc). In *International conference on learning representations (iclr)*.

Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., & De Vos, M. (2019). Seqsleepnet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *27*(3), 400–410. doi: 10.1109/TNSRE.2019.2896659

Phan, H., Chén, O. Y., Tran, M. C., Koch, P., Mertins, A., & De Vos, M. (2022). Xsleepnet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5903-5915. doi: 10.1109/TPAMI.2021.3070057

Phan, H., Mikkelsen, K., Chén, O. Y., Koch, P., Mertins, A., & De Vos, M. (2022). Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, *69*(8), 2456–2467. doi: 10.1109/TBME.2022.3147187

Shen, H., Ran, F., Xu, M., Guez, A., Li, A., & Guo, A. (2020). An automatic sleep stage classification algorithm using improved model based essence features. *Sensors*, *20*(17), 4677. doi: 10.3390/s20174677

Sors, A., Bonnet, S., Mirek, S., Vercueil, L., & Payen, J.-F. (2018). A convolutional neural network for sleep stage scoring from raw single-channel eeg. *Biomedical Signal Processing and Control*, *42*, 107–114. doi: 10.1016/j.bspc.2017.12.001

Supratak, A., Dong, H., Wu, C., & Guo, Y. (2017). 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. doi: 10.1109/TNSRE.2017.2721116

Wang, W., et al. (2023). Dynamicsleepnet: a multi-exit neural network with adaptive inference time for sleep stage classification. *Frontiers in Physiology*, *14*, 1171467. doi: 10.3389/fphys.2023.1171467

Wang, Y., et al. (2024). Research on sleep staging based on support vector machine and extreme gradient boosting algorithm. *Nature and Science of Sleep*, *16*, 1827–1847. doi: 10.2147/NSS.S467111

Zheng, Y., Luo, 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. doi: 10.3389/fnins.2022.973761

Funding Statement

This work was supported by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LQ24F020035. (Corresponding author: Li Zhu.).

Acknowledgments

We express our sincere gratitude to Li Zhu for her invaluable guidance and continuous support.