MoESleepNet: A Multi-view Mixture-of-Experts Model for Single-Channel EEG Sleep Stage Classification

Xi Wu[1], Xinzhong Cui[1], Tiantian Xiao[1], Yaokun Wang[1], Yuhao He[1], Zhiying Long[1,*], Xiangcun Wang[1,*], Yiwei Xu[2,*]

[1]School of Artificial Intelligence, Beijing Normal University, Beijing, China

[2]State Key Laboratory for Fine Exploration and Intelligent Development of Coal Resources, China University of Mining and Technology-Beijing, Beijing, China

Abstract

Accurate sleep staging plays a crucial role in diagnosing patients' sleep health. Numerous studies have confirmed that data from different views can highlight distinct characteristics. Based on the time-domain (TD) EEG, we constructed two additional views: the Power Spectral Density (PSD) and the time-frequency representation (TFR). Notably, there is no universally applicable model suitable for all types of data. That is to say, the characteristics of data should align with the model structure. Therefore, we designed specialized expert models for learning different views. However, directly combining features extracted from different views often results in excessive redundancy, especially for different views of the same data which the underlying data remains essentially identical. To address these issues, we proposed a multi-view mixture-of-experts (MoESleepNet) model for sleep staging, which achieves the best performance on SleepEDF20, SleepEDF78 and SHHS single-channel datasets. This study provides valuable ideas for signal classification.

Keywords: Sleep Staging; Multi-view; Mixture-of-Experts; Single-channel; Signal Classification

Introduction

Sleep health has long been a concern for many, largely due to stress from life and work (Irwin, 2015). The key to assessing sleep quality is the ability to automatically classify the various stages of sleep cycle (Gaiduk et al., 2023). Boostani et al. (2017) reported that numerous efforts have been made to perform sleep staging, but performance is suboptimal due to limitations in data representation and models.

Raw sleep signals are typically recorded by polysomnography tool in 30-second time-domain segments (Mullins et al., 2025). Lots of CNN-based models (Eldele et al., 2021; Sun et al., 2018; Supratak & Guo, 2020; Supratak et al., 2017) have been proposed to capture local temporal dependencies within TD signals. PSD, as another representation of TD, has also been widely employed, with related studies mainly focusing on capturing local frequency dependencies (Li et al., 2020). Meanwhile, TFR can simultaneously characterize both temporal and spectral information, thus holding significant value in sleep signal analysis. Transformer-based models have been extensively applied to capture global temporal dependencies within TFR (Phan et al., 2022). Although the above approaches have achieved effective results to some extent, their model designs often rely on a single data modality, thereby limiting the further enhancement of feature learning. Since different representations can emphasize distinct aspects of information (Tian et al., 2019), there is an urgent need to design new models that jointly use multiple representations.

Deep learning models based on mixture-of-experts (MoE) architecture has achieved significant success in multimodal time-series learning (Alkilane et al., 2024; Hu et al., 2025; Yang et al., 2025). MoE are naturally well-suited for integrating heterogeneous data sources, with each expert learning specific features from a particular modality, thereby fully leveraging the complementary information across modalities and further enhancing the overall feature learning capability (Pan et al., 2025; Yang et al., 2025). The core of MoE lies in designing expert models tailored to data, along with a gating mechanism that effectively integrates learned features from different experts (Cai et al., 2025; Shazeer et al., 2017).

This work focuses on the design of a MoE model, which aims to optimize feature learning by leveraging captured multi-level dependencies within sleep signals. The works by Eldele et al. (2021), Sun et al. (2018), Supratak and Guo (2020), and Supratak et al. (2017) focus on capturing local temporal dependencies within TD data using CNNs. The work by Phan et al. (2022) employs Transformers to jointly consider both temporal and frequency information within TFR. However, they essentially focus on capturing different dependencies within a single modality. While the present study is related to recent approaches based on MoE (Alkilane et al., 2024; Hu et al., 2025; Yang et al., 2025), it capitalizes on integrating the local temporal dependencies within TD, local frequency dependencies within PSD and time-frequency dependencies within TFR from different designed experts.

In this paper, our main contributions are summarized as follows: 1) First, we constructed three well-designed expert models: a CNN-based expert1 model to capture local temporal dependencies within TD, a FC-based expert2 model to capture local frequency dependencies within PSD, and a Mamba with multi-frequency inverse discrete cosine transform (MambaMultiIDCT)-based expert3 model to capture combined temporal and time–frequency dependencies within TFR. 2) Second, we developed a dynamic gating mechanism to reduce the over-redundancy of homogeneous merged features. 3) Last, our MoESleepNet model achieved the best staging accuracy on SleepEDF20 (Fpz-Cz 85.6%, EOG 81.5%), SleepEDF78 (Fpz-Cz 83.4%) and SHHS (C4-A1 84.1%).

Corresponding authors: Zhiying Long (friskying@163.com), Xiangcun Wang (wangxiangcun@mail.bnu.edu.cn), Yiwei Xu (xuyw137@163.com).

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

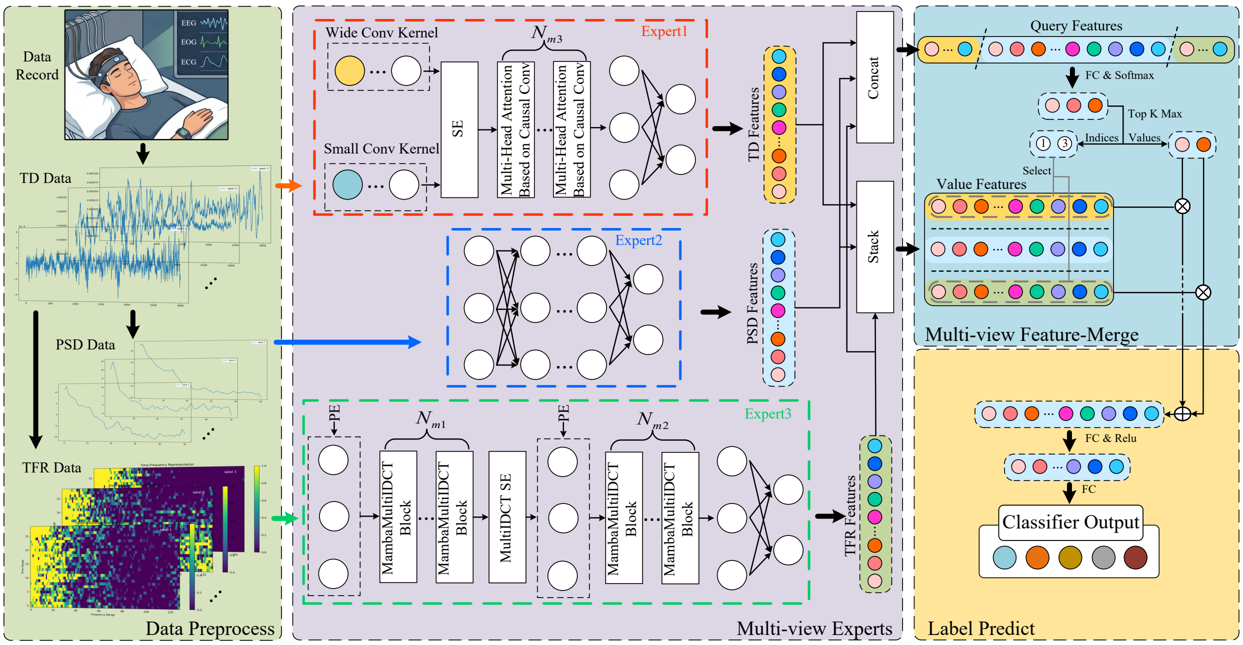


Figure 1: Global architecture and modules details of MoESleepNet. The architecture consists of four parts in reading order (from left to right): Data Preprocess (DP), Multi-view Experts (ME), Multi-view Feature-Merge (MF) and Label Predict (LP). In Part ME, PE denotes Position Encoding and SE denotes Squeeze-and-Excitation.

Methodology

In this section, we introduce proposed MoESleepNet model for single-channel sleep staging. First, we describe the global architecture of model. Then, we provide detailed description of multi-view experts and multi-view feature-merge.

Global Architecture

In Figure. 1, the architecture of the proposed MoESleepNet has four main parts: we convert raw TD view into PSD view and TFR view in part data preprocess (DP); we design distinct expert networks to generate the corresponding feature vectors in part multi-view experts (ME); we design a novel feature selection strategy for multi-view features merging in part multi-view feature-merge (MF); we use merged multi-view features for label prediction in part label predict (LP).

Multi-view Experts

MoE requires different expert models to focus on different data patterns or task subspaces. In part ME, our collection of experts comprises three experts: Expert1, which is designed for EEG TD view; Expert2, which focuses on PSD view; Expert3, which is dedicated to TFR view. The expert1 model is primarily based on the architecture of AttnSleep (Eldele et al., 2021), which is currently the leading model for raw TD EEG signals. Only the final FC layers was adjusted to ensure that the output feature dimensions are consistent with those of other experts. Given the simplicity of PSD data, the architecture of the expert2 model is composed entirely of FC layers. TFR incorporates both time-domain and frequency-domain information. To capitalize on this advantage, we designed a novel expert3 capable of leveraging temporal and time-frequency dependencies captured by MambaMultiIDCT for TFR data.

We construct a MambaMultiIDCT block by integrating Mamba (Gu & Dao, 2023) and MultiIDCT with squeeze-and-excitation (MultiIDCTSE) through a residual connection. The MultiIDCTSE enables channel-wise scaling by treating the frequency distribution as the importance of each time point, which captures the time-frequency dependencies. First, the IDCT performs the transformation of time-point feature from frequency domain to time domain with selected frequencies.

$$\ddot{x}_i\left(l\right) = \sum_{z \in Coe} \widetilde{x}_i(z) \cdot \phi_z\left(l\right) \tag{1}$$

$$\phi_z\left(l\right) = \cos\left(\frac{\pi}{L}\left(l + \frac{1}{2}\right)z\right) \tag{2}$$

In Eq. 1, $\widetilde{x}_i$ denotes frequency distribution of the i-th time point within TFR, $\ddot{x}_i$ denotes the recovered time-domain feature with selected frequencies of the i-th time point. l denotes the time-domain index. z denotes a frequency index and Coe denotes the set of frequency indices. In Eq. 2, $\phi_z\left(l\right)$ denotes the basis function corresponding to the z-th frequency index.

Let $U = [u_1, \ldots, u_{T-1}] \in \mathbb{R}^T$ denote the amplitude sum vector, where $u_i = \sum_l \ddot{x}_i\left(l\right)$ represents the amplitude sum of $\ddot{x}_i$. Then, a time-wise weight vector $W = [w_1, \ldots, w_{T-1}]$ is generated from the U via two-layer FC layers, formalized by:

$$W = Sigmoid(FC(Relu(FC(U)))) \tag{3}$$

Let $\vec{x}^{(c)} = [\vec{x}_1, \ldots, \vec{x}_{T-1}] \in \mathbb{R}^{T \times L}$ denote time-wise adjusted feature, where $\vec{x}_i$ at the i-th time point is computed by:

$$\vec{x}_i = Conv1D(w_i \cdot \widetilde{x}_i) \quad (4)$$

where $Conv1D$ denotes one-dimensional convolution with padding to preserve the input length L, aiming to capture local dependencies among frequency coefficients.

Notably, we evenly divide the frequency distribution indices Coe into three parts in MultiIDCT: Coe_{high}, Coe_{mid} and Coe_{low}. All features within a specific index range are aggregated by a summation operation, thereby capturing a unified response across the high, middle, low frequency domains.

The MambaMultiIDCT block can be formalized as:

$$\begin{aligned} X_{\text{out}} &= \text{MambaMultiIDCT}(X) \\ &= X + (\text{Mamba}(\text{LayerNorm}(X)))+ \\ &\quad \text{MultiIDCTSE}\left(X + \text{Mamba}(\text{LayerNorm}(X))\right) \end{aligned} \quad (5)$$

The MambaMultiIDCI block not only facilitates selective retention and forgetting of time-point features, but also extracts features from the TFR by leveraging time-frequency dependencies. Notably, we still add a MultiIDCTSE after the first set of N_{m1} MambaMultiIDCT blocks, for adjusting features without discarding irrelevant information. Finally, the expert1 will obtain the TD features, the expert2 will obtain the PSD features and the expert3 will obtain the TFR features.

Multi-view Feature-merge

A key challenge is designing the expert selection mechanism. Ideally, only the most relevant experts are activated for each input, thereby reducing unnecessary computation. Given that both PSD and TFR are derived from the TD data, the three types of features inherently contain a significant amount of redundant information. The direct summation of such redundant features tends to degrade the discriminative capacity. For this, we developed a novel merging mechanism that dynamically identifies the most salient TopK features for utilization.

First, multi-view features from various experts are concatenated to constitute our query features and features from different experts are stacked as our value features. The query features are processed by an FC layer to yield a vector, which has a dimension equivalent to the number of expert models. This vector is then subjected to a softmax function, scaling the values to fall within the range of 0-1, where each value signifies the importance of the respective expert. We shall select the TopK maximum values within the vector to serve as weight coefficients. Then, utilizing the indices of these selected values, we will identify and extract the corresponding features from the value features. The refined value features will be individually multiplied by the corresponding weight factors to yield the merged features, formalized as follows:

$$Z = \sum_{m=1}^{M} \frac{e^{FC(Q)_m}}{\sum_{j=1}^{M} e^{FC(Q)_j}} \cdot V_m \quad (6)$$

where Z denotes the merged features, M denotes the TopK selection indices set, $FC(Q)$ denotes the query features processed by FC and V denotes the value features. In our experiments, K is set to 2.

Table 1: Distribution of sleep stages on three sleep datasets.

Dataset	W	N1	N2	N3	REM	Total
SleepEDF20	9118 21.1%	2804 6.5%	17799 41.3%	5703 13.2%	7717 17.9%	43141 –
SleepEDF78	66822 34.0%	21522 11.0%	69132 35.2%	13039 6.6%	25835 13.2%	196350 –
SHHS	46319 14.3%	10304 3.2%	142125 43.8%	60153 18.5%	65953 20.3%	324854 –

Experiments

Datasets

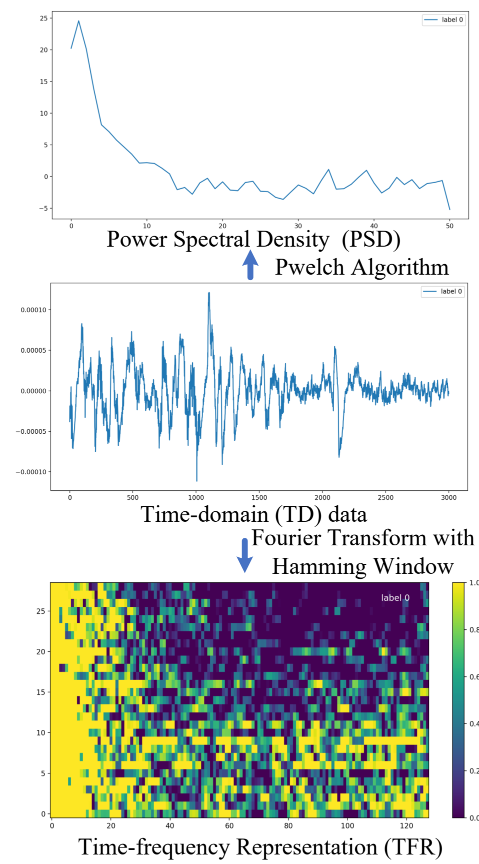


Figure 3: Comparison of preprocessing methods and data views for the Wake (label 0) Stage.

We evaluated the proposed sleep model using the Sleep-EDF (SleepEDF20), Sleep-EDF Expanded (SleepEDF78), Sleep Heart Health Study (SHHS) datasets as shown in Table 1.

SleepEDF20 (Kemp et al., 2000) includes 39 day-night recordings from 20 subjects. To avoid influence of other factors, Fpz-Cz EEG and EOG channel were used. We retained the 30s epochs labeled as W, N1, N2, N3, N4, and REM. We merged N3 and N4 into a N3 with AASM rules. We only included 30-minute wake epochs both before and after sleep.

SleepEDF78 (Kemp et al., 2000) is an extended version of SleepEDF20, including 78 subjects. Similar to SleepEDF20, we adopted the same labeling criteria and channel selection.

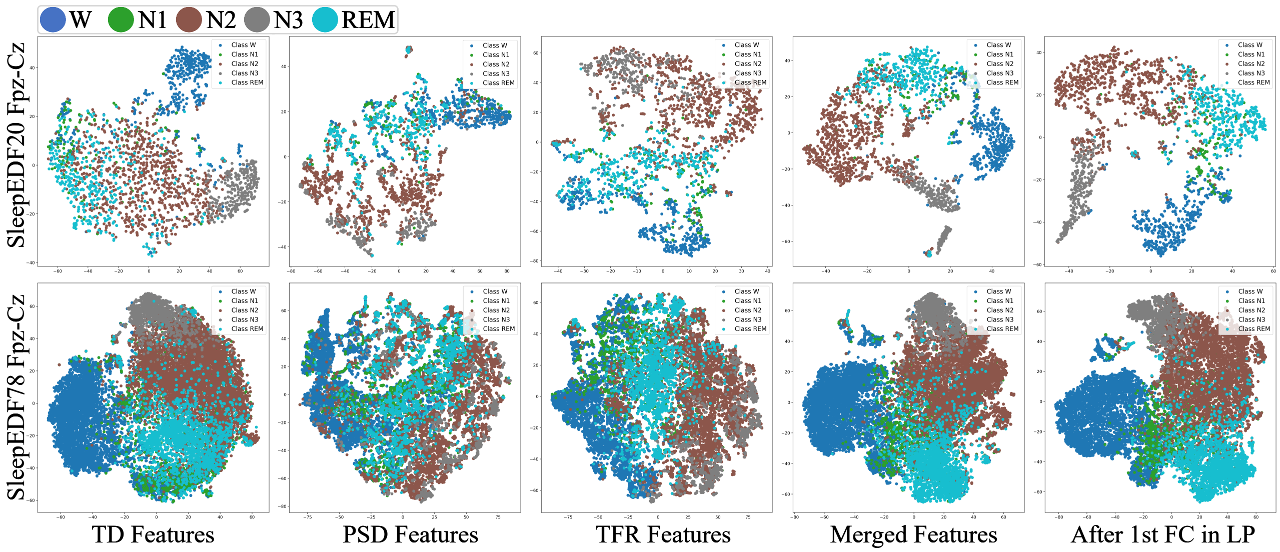


Figure 2: The t-SNE feature maps on SleepEDF20, SleepEDF78 Fpz-Cz channel.

SHHS (Quan et al., 1997) dataset consists of 329 subjects with regular sleep. Similar to SleepEDF20, the same label adjustments are applied. The C4-A1 EEG channel was adopted.

Preprocessing and Implementation Details

As shown in the Figure 3, the raw TD view is further transformed into the PSD and TFR views during the data preprocessing. Different views highlight distinct characteristics of the data. For example, TD view focuses on voltage variations over time, the PSD view emphasizes the frequency distribution, and the TFR view captures how the frequency content evolves over time. Specifically, for PSD view, we used the pwelch (window size: 100) in MATLAB to convert raw TD EEG into PSD distribution with shape (51,1). 30-second TD segments are transformed into time-frequency representations with shape (29, 128) by Fourier Transform using a Hamming window (size: 200, overlap: 100). We use epoch-wise K-fold cross-validation with results averaged, (20 for SleepEDF20, SHHS and 10 for SleepEDF78). The model is trained for 200 epochs using AdamW optimizer (learning rate: 5e-6, batchsize: 64) with early stopping. The manually designed architecture parameters (N_{m1}, N_{m2}, N_{m3}) are set to (16, 4, 2).

Baselines

In this experiment, we compared the proposed model with five baseline models. A brief introduction to each baseline is provided below:

1)Mamba (Gu & Dao, 2023): The model uses the vanilla Mamba module for feature extraction from time–frequency representations, and the extracted features are processed through FC layers to achieve automatic sleep staging.

2)Transformer (Phan et al., 2022): The model adapts the original architecture for single-channel tasks, using multiple Transformer blocks to optimize features from time–frequency representations for sleep staging.

3)AttnSleep (Eldele et al., 2021): The model employs a multi-resolution CNN with adaptive feature recalibration to extract low- and high-frequency features, followed by multi-head attention to capture temporal dependencies among the features from time–domain representations.

4)TinySleepNet (Supratak & Guo, 2020): The model integrates CNN and LSTM to achieve a trade-off in sleep staging while maintaining minimal parameters and computational overhead from time–domain representations.

5)ResnetLSTM (Sun et al., 2018): The model leverages a residual network to extract multi-level features from time–domain representations and employs LSTM to learn the state transition dynamics within sleep signals.

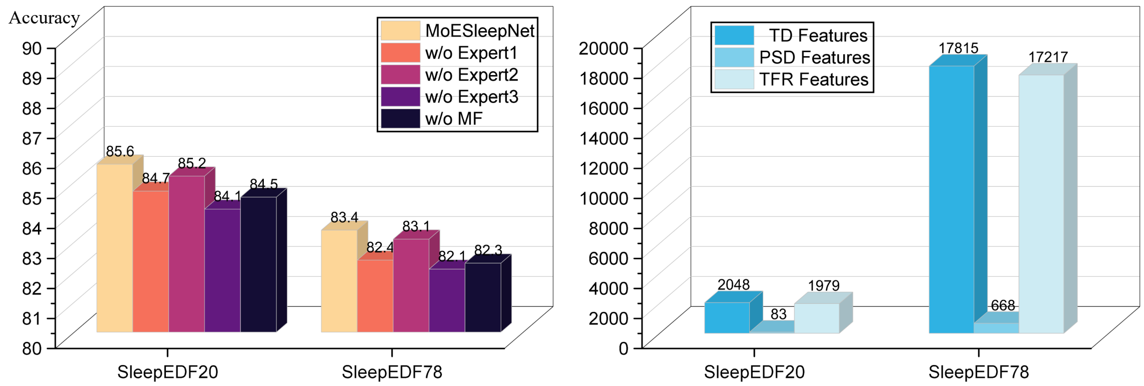
6)DeepSleepNet (Supratak et al., 2017): The model employs a convolutional neural network (CNN) to extract time-invariant features from time–domain representations and a bidirectional long short-term memory network (BiLSTM) to automatically learn the transition rules between sleep stages within EEG.

Evaluation Metrics

We rely on confusion matrix based statistics to assess our model. True Positive (TP) counts samples correctly predicted as a class, False Positive (FP) counts samples incorrectly predicted as that class, False Negative (FN) counts samples that belong to the class but are predicted otherwise, and True Negative (TN) counts samples that do not belong to the class and are correctly not predicted as it.

Per-class precision (PR) measures the proportion of correctly predicted samples among all samples predicted as belonging to class c. It reflects how reliable the model's positive predictions are for each class:

$$\mathrm{PR}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c} \tag{7}$$



(a) Ablation Studies

(b) Multi-view Selection Count

Figure 4: Ablation studies and selection count on SleepEDF20, SleepEDF78 Fpz-Cz channel.

Per-class recall (RE) evaluates the proportion of correctly predicted samples among all ground-truth samples of class c. It reflects the model's ability to detect all instances of class c:

$$\mathrm{RE}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c} \tag{8}$$

Per-class F1 score is the harmonic mean of precision and recall for class c. It provides a balanced measure of both correctness and completeness of predictions:

$$\mathrm{F1}_c = \frac{2 \cdot \mathrm{PR}_c \cdot \mathrm{RE}_c}{\mathrm{PR}_c + \mathrm{RE}_c} \tag{9}$$

Overall accuracy (Acc) measures the proportion of correctly predicted samples among all samples:

$$\mathrm{Acc} = \frac{\sum_{c=1}^{C} \mathrm{TP}_c + \mathrm{TN}_c}{\sum_{c=1}^{C} (\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c + \mathrm{TN}_c)} \tag{10}$$

Overall Cohen's Kappa (κ) evaluates the agreement between predictions and ground truth while considering chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{11}$$

where p_0 is the observed agreement and p_e is the expected agreement by chance.

Macro F1 (MF1) computes the mean of per-class F1 scores:

$$\mathrm{MF1} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{F1}_c \tag{12}$$

Results and Analysis

To verify the effectiveness of the proposed model, we reproduced the representative single-channel models with only 30s input like AttnSleep (Eldele et al., 2021), TinySleepNet (Supratak & Guo, 2020), ResnetLSTM (Sun et al., 2018), DeepSleepNet (Supratak et al., 2017). Similarly, we also built models based on the vanilla Mamba (Gu & Dao, 2023) and Transformer (Phan et al., 2022) blocks for performance comparison. Table 2 provides overall metrics, per-class F1, FLOPs and Parameters for the SleepEDF20, SleepEDF78 and SHHS datasets. All models in the table are trained with the same data settings and hyperparameters. From the table 2, we observed that our proposed model outperforms all other models in the overall metrics. For the relatively high-frequency stages W, N2, and REM, our model achieves the best scores in the per-class F1, indicating that it effectively improves performance for high-frequency stages. And REM indicates a significant improvement, with a rise of 1.6 for SleepEDF20 Fpz-Cz and 2.1 for SleepEDF78 Fpz-Cz. Despite these performance improvements, the model maintains computational efficiency, with moderate FLOPs and a manageable number of parameters, making it an effective solution for sleep stage classification. Notably, on large-scale datasets such as SleepEDF78 and SHHS, our model achieves more pronounced performance improvements, indicating stronger generalization capability. We observed that, on SleepEDF78, the accuracy increased to 83.4% compared with 81.9% in AttnSleep, and on SHHS, it rose to 84.1% compared with 83.2% in the Transformer model. This phenomenon indicates that our model can effectively integrate features from different views on more complex datasets, resulting in improved sleep staging accuracy. We observed that the model contains 7.425M parameters, primarily due to the design of multiple expert models that handle inputs from different views. Although this leads to a slight increase in parameter count, the model achieves a substantial improvement in performance.

To validate the effectiveness of the proposed modules, we visualized the accuracy variations observed in the ablation experiments in Figure. 4(a). The abbreviation "w/o" in the figure stands for "without," indicating that the module is removed from the model. We observe that removing any expert leads to a drop in accuracy, indicating that each expert contributes useful modality-specific information to the overall feature representation. And removing TFR and TD has a more significant impact compared to removing PSD. We observed that the part MF in MoE effectively facilitates the integration of valuable features while minimizing redundancy among them. In Figure. 4(b), we can observe that MoESleep-

Table 2: Performance comparison with other models on three sleep datasets.

Dataset	Model	Channels	Overall metrics			Per-class F1 score					FLOPs	Parameters
			Acc	κ	MF1	W	N1	N2	N3	REM		
SleepEDF20	**MoESleepNet**	Fpz-Cz	**85.6**	**0.80**	**78.0**	**92.6**	39.7	**89.4**	**86.9**	**81.3**	5.263G	7.425M
	Mamba (Gu & Dao, 2023)	Fpz-Cz	80.0	0.73	72.5	85.2	32.8	86.3	83.9	74.4	0.265G	3.813M
	Transformer (Phan et al., 2022)	Fpz-Cz	84.8	0.79	77.6	91.7	39.7	88.6	88.1	79.7	4.153G	5.918M
	AttnSleep (Eldele et al., 2021)	Fpz-Cz	84.6	0.78	77.7	91.2	41.7	88.8	87.1	79.7	3.896G	0.471M
	TinySleepNet (Supratak & Guo, 2020)	Fpz-Cz	76.0	0.67	63.3	82.3	6.4	82.1	73.2	72.5	2.178G	0.684M
	ResnetLSTM (Sun et al., 2018)	Fpz-Cz	70.6	0.59	59.1	77.7	9.5	77.7	74.5	56.2	-	20.762M
	DeepSleepNet (Supratak et al., 2017)	Fpz-Cz	77.1	0.69	69.5	86.0	37.0	82.3	65.0	77.3	-	9.574M
	MoESleepNet	EOG	**81.5**	**0.75**	74.1	**87.7**	37.1	**84.7**	**82.3**	**78.9**	5.263G	7.425M
	Mamba (Gu & Dao, 2023)	EOG	75.6	0.67	67.1	80.5	32.1	79.7	75.1	67.9	0.265G	3.813M
	Transformer (Phan et al., 2022)	EOG	80.5	0.73	74.3	86.9	42.6	83.9	80.9	77.5	4.153G	5.918M
	AttnSleep (Eldele et al., 2021)	EOG	78.1	0.70	70.4	81.1	35.3	83.5	76.1	76.2	3.896G	0.471M
	TinySleepNet (Supratak & Guo, 2020)	EOG	73.9	0.64	65.9	80.0	31.0	79.0	74.7	64.8	2.178G	0.684M
	ResnetLSTM (Sun et al., 2018)	EOG	70.6	0.58	61.8	76.4	25.6	77.5	75.2	54.2	-	20.762M
	DeepSleepNet (Supratak et al., 2017)	EOG	65.7	0.51	50.8	69.4	0.90	74.2	65.6	44.0	-	9.574M
SleepEDF78	**MoESleepNet**	Fpz-Cz	**83.4**	**0.77**	**76.7**	**93.7**	**43.6**	**86.0**	**82.0**	**78.1**	5.263G	7.425M
	Mamba (Gu & Dao, 2023)	Fpz-Cz	80.3	0.74	72.7	91.0	36.0	84.9	78.2	73.3	0.265G	3.813M
	Transformer (Phan et al., 2022)	Fpz-Cz	81.7	0.74	74.5	92.4	40.0	85.1	80.5	74.5	4.153G	5.918M
	AttnSleep (Eldele et al., 2021)	Fpz-Cz	81.9	0.75	75.2	92.6	41.6	85.2	80.6	76.0	3.896G	0.471M
	TinySleepNet (Supratak & Guo, 2020)	Fpz-Cz	78.2	0.69	69.0	90.7	28.0	81.9	76.6	67.6	2.178G	0.684M
	ResnetLSTM (Sun et al., 2018)	Fpz-Cz	76.0	0.68	67.6	89.7	31.4	80.1	74.2	62.6	-	20.762M
	DeepSleepNet (Supratak et al., 2017)	Fpz-Cz	79.5	0.71	72.3	91.0	38.8	83.0	77.0	71.8	-	9.574M
SHHS	**MoESleepNet**	C4-A1	**84.1**	**0.79**	**72.7**	**88.1**	**22.2**	**85.8**	**86.5**	**80.9**	5.263G	7.425M
	Mamba (Gu & Dao, 2023)	C4-A1	79.9	0.71	66.4	82.3	18.2	81.7	75.3	74.5	0.265G	3.813M
	Transformer (Phan et al., 2022)	C4-A1	83.2	0.77	71.7	86.9	20.5	84.9	85.3	81.0	4.153G	5.918M
	AttnSleep (Eldele et al., 2021)	C4-A1	83.0	0.76	71.6	86.4	20.4	85.0	85.6	80.8	3.896G	0.471M
	TinySleepNet (Supratak & Guo, 2020)	C4-A1	75.4	0.65	63.0	80.7	10.3	79.1	72.5	72.6	2.178G	0.684M
	ResnetLSTM (Sun et al., 2018)	C4-A1	71.9	0.61	61.5	78.2	14.1	78.3	74.1	62.8	-	20.762M
	DeepSleepNet (Supratak et al., 2017)	C4-A1	78.5	0.69	65.5	82.3	17.2	81.4	71.3	75.5	-	9.574M

Net tends to favor selecting TD and TFR features among the three views, indicating that these two representations carry the most informative content. At the same time, it occasionally chooses PSD features, suggesting that for certain sleep epochs, using only PSD features is sufficient for differentiation.

To compare the feature learning capabilities of our proposed model at different layers, we visualized the t-SNE feature map changes of MoESleepNet on the SleepEDF20 and SleepEDF78 Fpz-Cz validation datasets in Figure. 2, which we refer to as vertical comparison. We observe, in TD features, W exhibits clear decision boundaries with respect to (N1, N2, N3), whereas REM shows significant overlap with (N1, N2). This indicates that temporal features are effective in capturing stable oscillatory patterns of wakefulness, but are less capable of distinguishing sleep stages characterized by low amplitude, irregularity and weak structural patterns. Specifically, REM, N1, and N2 all exhibit desynchronized, mixed-frequency activity, leading to substantial overlap in the feature space. In contrast, the separability between REM and N2 is significantly improved in PSD and TFR representations. This is mainly because frequency-domain and time–frequency features can better capture the sigma-band energy enhancement associated with sleep spindles in N2. In PSD, this appears as a distinct spectral peak, while in TFR it manifests as localized high-frequency energy bursts. Moreover, TFR further enhances this distinction by capturing the non-stationary and transient nature of spindle events, resulting in clearer inter-class boundaries. In summary, frequency-domain and time–frequency features demonstrate clear advantages in characterizing band-specific distributions and transient structures, thereby improving the discrimination between N2 and REM. These findings are consistent with the neurophysiological properties of sleep and further highlight the necessity of multi-view feature modeling. By integrating temporal, spectral and time–frequency information, the limitations of single-view representations can be effectively mitigated, ultimately enhancing the model's ability to distinguish complex sleep stages. And we can clearly see that the class boundaries for TD features and TFR features are relatively distinct, while the five categories of PSD features are mixed together and difficult to differentiate. Additionally, we observe that merging the features makes the boundaries of the five categories easier to distinguish. These findings suggest that our model is capable of dynamically selecting features and further refining merged features to make more accurate classification decisions.

Conclusion

In this work, we proposed a novel Mixture-of-Experts (MoE) model for multi-view EEG sleep datasets to perform automatic sleep stage classification. Different Expert models were specifically designed for each data view. Notably, for time-frequency images, we introduced the MambaMultiIDCT block, which effectively selects relevant features and adjusts their distribution based on time-time and time-frequency dependencies. Additionally, we developed a dynamic feature selection method to identify the most relevant features for fusion, thereby reducing redundancy across views. Through these innovations, our proposed network, MoESleepNet, achieves the best performance by leveraging multi-level dependencies in single-channel sleep stage classification on the SleepEDF20, SleepEDF78 and SHHS datasets and provides a promising solution for sleep health monitoring.

References

Alkilane, K., He, Y., & Lee, D.-H. (2024). Mixmamba: Time series modeling with adaptive expertise. *Information Fusion*, *112*, 102589.

Boostani, R., Karimzadeh, F., & Nami, M. (2017). A comparative review on sleep stage classification methods in patients and healthy individuals. *Computer methods and programs in biomedicine*, *140*, 77–91.

Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., & Huang, J. (2025). A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*.

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.

Gaiduk, M., Serrano Alarcón, Á., Seepold, R., & Martínez Madrid, N. (2023). Current status and prospects of automatic sleep stages scoring. *Biomedical engineering letters*, *13*(3), 247–272.

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Hu, J., Duan, P., Cao, X., Xue, Q., Zhao, B., Zhao, X., Yuan, X., & Zhang, C. (2025). A multi-energy load forecasting method based on the mixture-of-experts model and dynamic multilevel attention mechanism. *Energy*, *324*, 135947.

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

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*(9), 1185–1194.

Li, X., Wang, Y., Zhang, B., & Ma, J. (2020). Psdrnn: An efficient and effective har scheme based on feature extraction and deep learning. *IEEE Transactions on Industrial Informatics*, *16*(10), 6703–6713.

Mullins, A. E., Pehel, S., Parekh, A., Kam, K., Bubu, O. M., Tolbert, T. M., Rapoport, D. M., Ayappa, I., Varga, A. W., & Osorio, R. S. (2025). The stability of slow-wave sleep and eeg oscillations across two consecutive nights of laboratory polysomnography in cognitively normal older adults. *Journal of Sleep Research*, *34*(1), e14281.

Pan, L., Wang, H., Chen, Z., Huang, Y., Niu, Y., Liu, Z., He, Q., & Liu, X. (2025). Controllable mixture-of-experts for multivariate soft sensors. *IEEE Transactions on Automation Science and Engineering*.

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.

Quan, S. F., Howard, B. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., Rapoport, D. M., Redline, S., Robbins, J., Samet, J. M., et al. (1997). The sleep heart health study: Design, rationale, and methods. *Sleep*, *20*(12), 1077–1085.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Sun, Y., Wang, B., Jin, J., & Wang, X. (2018). Deep convolutional network method for automatic sleep stage classification based on neurophysiological signals. *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1–5.

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.

Supratak, A., & Guo, Y. (2020). Tinysleepnet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg. *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 641–644.

Tian, X., Deng, Z., Ying, W., Choi, K.-S., Wu, D., Qin, B., Wang, J., Shen, H., & Wang, S. (2019). Deep multi-view feature learning for eeg-based epileptic seizure detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *27*(10), 1962–1972.

Yang, X., Li, Y., Zhang, J., Tian, H., Li, S., & Pan, G. (2025). Evomoe: Evolutionary mixture-of-experts for ssvep-eeg classification with user-independent training. *IEEE Journal of Biomedical and Health Informatics*.