LA-Mamba Net, An Automatic Sleep Staging Model with Temporal Feature Extraction Ability

Jiayu Li
College of Electronic and Information Engineering
Tongji University
Shanghai, China
2230817@tongji.edu.cn

Yangguang Xu
College of Electronic and Information Engineering
Tongji University
Shanghai, China
2432083@tongji.edu.cn

Wei Zhang
College of Electronic and Information Engineering
Tongji University
Shanghai, China
Zhang_wi@tongji.edu.cn
*Corresponding author

Yuan Zhou
College of Electronic and Information Engineering
Tongji University
Shanghai, China
2331979@tongji.edu.cn

2025 10th International Conference on Biomedical Signal and Image Processing (ICBIP) | 979-8-3503-5722-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICBIP66535.2025.11284426

***Abstract*—Most automatic sleep staging models using single-channel electroencephalogram (EEG) signals classify data within 30-second time windows.This approach restricts feature extraction to short-term information, neglecting valuable long-term temporal characteristics from extended historical data. The loss of these temporal features may compromise staging accuracy. The lost historical information contains a large amount of temporal characteristics, which may lead to poor sleep staging results. To address this limitation, the input data is extended beyond the 30-second window to capture richer temporal dependencies. To efficiently process longer sequences without excessive computational overhead, the LA-Mamba network is proposed. Centered on the Mamba architecture, this model integrates multi-core convolutional layers, a local self-attention mechanism, and focal loss optimization. This design enables effective utilization of extended EEG histories for improved sleep staging.Evaluated on the Sleep-EDFx dataset, the proposed model achieves an accuracy of 0.846, a kappa coefficient of 0.839, and an F1-score of 0.780. Experiments further demonstrate enhanced performance in classifying the challenging N1 stage, indicating significant improvements for this critical category.**

Keywords—sleep stage, sequential signal, mamba model

I. Introduction

Modern sleep quality faces significant challenges due to accelerating lifestyles and worsening urban light and sound pollution, driving increased demand for effective sleep assessment. Sleep staging — the temporal segmentation of physiological data collected during sleep — serves as the fundament for both sleep quality evaluation and clinical diagnosis of sleep disorders.

Clinically, sleep staging relies on polysomnography (PSG), known as the "gold standard of sleep". PSG is a collection of data with a large time span and large variety of data, making manual staging labor-intensive and requiring specialized physician expertise — a resource often unavailable in many regions. Consequently, significant interest exists in developing automated, artificial intelligence-based solutions for long-term sleep staging. This project addresses that need by developing an AI model for automatic sleep staging using single-channel electroencephalogram (EEG).

A critical feature of PSG for automated staging is its time correlation, which is different from non-temporal signals likes images. In the sleep staging, the temporal correlation is reflected in the transition relationship between sleep stages. Clinically, when a doctor makes a judgment on the current sleep frame, it is generally assumed that sleep phases change with some regularity, e.g., it is often assumed that wakefulness will be followed by a high probability of a non-rapid eye movement phase I. Therefore, effective AI models must analyze not only the current time window's signal but also relevant historical data to capture the influence of preceding stages. While most current automated models analyze only isolated 30-second windows, this project prioritizes incorporating extended temporal context to better align with clinical reasoning and enhance staging accuracy.

II. Literature Review

In current research on automated sleep staging models, the extraction of temporal correlation features is often performed using network structures such as recurrent neural networks, Transformer, and graph neural networks.

Recurrent neural network (RNN) is specifically designed to explore the temporal correlations in time series signals. Yang et al. employed a combination of CNN networks and bidirectional LSTM networks, with the former for feature learning and the latter for sequence learning, achieving an accuracy rate of 85.87% [4]. Chen et al. utilized LSTM networks and Hidden Markov Models (HMM) for sequence learning, attaining an accuracy rate of 82.71% on the Sleep-EDFx dataset [5]. He et al, extracted features using CNNs and used bi-directional LSTM networks and multi-head attention blocks. They achieved an accuracy rate of 87.2% on the Sleep-EDF-20 dataset [6].

The attention mechanism can capture relevant information and extract key details. For instance, Aozora et al. used the self-attention mechanism to learn the features of single-channel signals and their time-frequency diagrams. They achieved accuracies of 85.73% and 84.83% on Sleep-EDF-20 and Sleep-EDF-78, respectively [7]. Zhu et al. employed the self-attention mechanism to filter the features output by the convolutional network and the correlations between windows, achieving an accuracy of 82.8% on Sleep-EDFx [8]. Pan et al. used the attention mechanism and added causal dilated convolution to enhance temporal features. Their model achieved accuracies of

84.7% and 81.6% on Sleep-EDF-20 and Sleep-EDF-78, respectively [9].

As a network capable of representing connection relationships, graph neural networks can learn the transition relationships between stages. For instance, Chen used BiRNN and GCN to learn the temporal features of electroencephalogram (EEG) and the transition relationships between stages, and ultimately achieved an accuracy rate of 88.4% and 83.8% was achieved on Sleep-EDF-20 and Sleep-EDF-78 respectively [10]. Wang et al. combined Resnet, LSTM and GCN, and ultimately achieved accuracy rates of 86.96% and 83.79% on SleepEDF-20 and SleepEDF-78 respectively [11].

It can be observed that the design of automatic sleep staging models mainly focuses on two aspects. The first is the extraction and selection of features within a single time frame.At this point, network structures such as CNN and Transformer are often chosen. The second is how to make good use of the temporal features of EEG signals and explore their contextual relationships. In this case, network structures such as RNN, LSTM, and GNN are often selected. At the same time, the computational requirements and training speed during model selection also need to be considered. For instance, Transformer faces the problem of computational explosion when dealing with longer sequences. Therefore, the subsequent model design in this paper needs to focus on three aspects: proximal feature mining, distal temporal feature mining and model training speed optimization.

III. Introduction to Methods

The Mamba network is an improved network model based on the Structured State Space for Sequence (S4) [12]. It boasts the capability to construct state transition models and offers high computational efficiency when processing long input sequences.The capability to construct state transition models stems from its design foundation in the state space model as shown in (1).

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t) \\ y(t) &= Ch(t) + Dx(t) \end{aligned} \tag{1}$$

The high computational efficiency when processing long sequences is attributed to three key design choices in the Mamba network [13]. The first is to change the time-invariant B, C, and discrete Δ matrices in the state transition matrix to time-varying matrices based on the input sequence. This enables content-aware processing of different input sequences, filtering out key information and ignoring irrelevant information, thereby reducing the computational load. The second is to use a scanning operation to achieve parallelized computation in conjunction with the time-varying matrices. The third is to address the computational limitations caused by the transmission speed between hardware components during the computation process by using a hardware-aware algorithm that places the parameters of h and the A, B, and C matrices in SRAM and DRAM, respectively.

The focal loss function is a new loss function designed by Lin et al. to solve the problem of imbalance in the number of samples in each category and the difficulty of classification in each category when dealing with the problem of object recognition [14]. It is based on the classic weighted cross -entropy loss function, and its formula is shown in (2), where p_t represents the probability value of each category.

$$FL - CE(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{2}$$

The design of the focal loss function is mainly aimed at balancing the categories with different levels of classification difficulty. When a category is easy to classify, its probability value will approach 1, and the loss value taken by the above formula will be smaller, thus having a smaller impact on the subsequent model updates. This makes the network pay more attention to those categories with low probability values. Here, γ is a hyperparameter that acts like a scaling factor. When the γ value is larger, the loss values of more categories will approach 0. Meanwhile, the weight α can alleviate the problem of sample imbalance by increasing the loss of data with a smaller sample size.

IV. Method

A. Model Design

LocalAttention-Mamba(LA-Mamba) is a model designed based on the Mamba network and uses the focus loss function to solve the problem of different difficulties of class imbalance. The specific network structure is shown in Figure 1, consisting of three parts: the multi-kernel convolutional layer, the Local Attention-Mamba layer, and the classification layer.

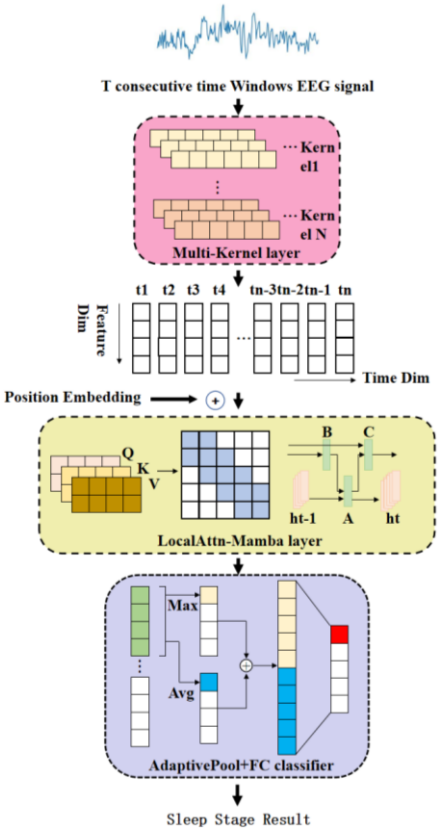

Fig. 1. Schematic diagram of LA-Mamba model network architecture.

The multi-core convolutional layer, which is used to extract features from the input EEG signals, can be understood as encoding the data. As shown in Figure. 2, this layer consists of three parallel convolutional sub-networks. Each sub-network is structured with a sequential processing pipeline comprising three convolutional layers followed by a batch normalization layer, an activation layer, and a pooling layer. The difference among the three sub-network layers lies in the first convolutional layer, which uses convolution kernels of different sizes. The lower-level convolutional sub-network uses smaller convolution kernels to extract features with smaller receptive fields; Higher-level convolutional sub-networks use larger convolution kernels to extract features with larger receptive field, ensuring that the features extracted by this network include those with different receptive field.

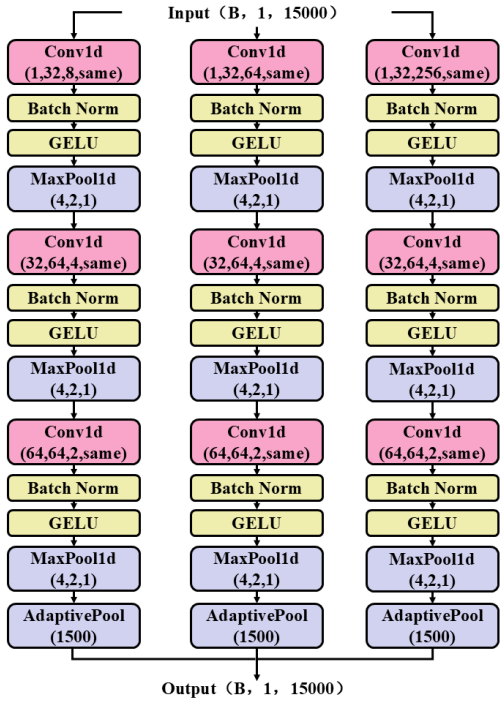

Fig. 2. Schematic diagram of multi-core convolutional layer network architecture.

This design mainly takes into account the fact that in the sleep staging judging criteria, there are both longer duration discriminatory bases, such as lower frequency δ EEG waves; and higher frequency, shorter duration β waves as well as k-complex waves. These features of different duration require different visual fields for extraction.

The GELU activation function is used in the activation layer. The relevant parameters of the above multi-core convolutional layer network structure are shown in Figure. 2.

The LocalAttn-Mamba layer is the core component of the LA-Mamba network, which is responsible for extracting the long-range and short-range correlations of the data and learning the temporal variation patterns of the signals. Its specific structure is shown in Figure. 3.

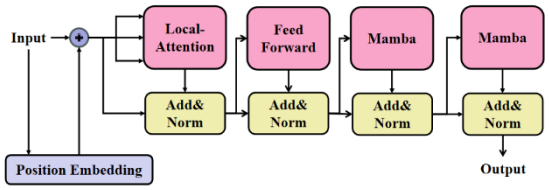

Fig. 3. Schematic diagram of LocalAttn-Mamba layer network structure.

This layer structure consists of one window self-attention layer with position encoding and two Mamba layers. Among them, the Mamba network is used to extract long-range correlations in the time series, while the window self-attention network focuses on short-range correlations in the time series, thereby achieving analysis of the time series signal from both long and short perspectives. At the same time, due to the characteristics of the Mamba model and the reduction of the attention mechanism calculation length in the window attention mechanism, the final model can maintain a certain degree of low computational cost and low computational time.

In addition, on other network structures, to enable the window self-attention layer to access the before-and-after information relationships in time series, the fixed sine positional encoding method is performed on the data before it enters the window self-attention network and is directly added to the input data. After the window attention layer, a Feed Forward network layer is added in the model, which consists of two fully connected layers, and a GELU activation function is set between the two connected layers to extract deeper features. At the same time, to increase stability and prevent gradient vanishing, a residual connection and a Layer Norm layer are added after each of the above network structures.

The purpose of the classification layer is to transform the output features of the LocalAttn-Mamba layer into the final five-classification results. Due to the large output data of the previous layer, it is necessary to reduce the dimensions of each data dimension before inputting it into the classifier to avoid overfitting caused by excessive parameters. In this paper, global pooling and fully connected layers are combined. Global pooling is used for data dimensionality reduction, and the fully connected layer is used for classification.

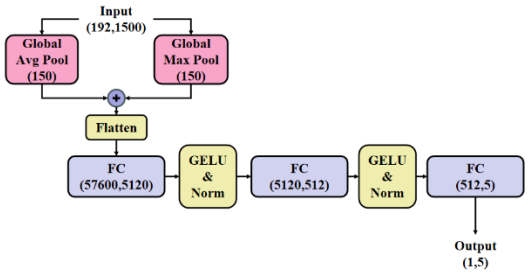

Fig. 4. Schematic diagram of classifier network structure.

The global average pooling can extract the average background features in the feature data map, while global max pooling can extract the prominent foreground features in the feature data map.

The design in the fully connected layer is rather classic. It consists of three fully connected networks in total. Between each pair of fully connected networks, the GELU activation function is used for learning nonlinear relationships, and the Layer Norm layer is employed to enhance generalization ability and accelerate training speed.

B. Experimental Procedure and Design

Subsequent experiments used the Sleep-EDFx dataset made publicly available by Physionet [1], [2], which is a dataset of sleep tests performed on healthy Caucasians aged 25-101 years old, including healthy subjects as well as subjects with difficulty falling asleep and taking temazepam medication, for a total of 197 days and nights of sleep polysomnography data including two channels of EEG, one channel of electrooculogram (EOG), and one channel of electromyography. The data were labeled using the R&K standard. Regarding the channel selection, following the study by Radha et al., the electrodes near the frontal area were chosen [12]. For the Sleep-EDFx dataset, the Fpz-Cz channel of the EEG signal was selected (EEG signals for other datasets). The final classification results were categorized into five types based on the AASM sleep discrimination standard, with both stage 3 and stage 4 labeled as stage 3 according to the original K&R standard. After preprocessing the data as described in Section 2.2.4, the training and test sets were set at a ratio of 9:1.

In terms of data preprocessing, to obtain as much historical data as possible, the input signal for the model was chosen to be a single-channel EEG signal lasting 150 seconds, which is the EEG signal of the five-time windows in the clinical sleep staging standard. The final classification result will be based on the last 30 seconds, that is, the signal of the last time window.

Secondly, considering that a complete 24-hour sleep record may contain a large amount of data before falling asleep and after waking up, only the data from the first 30 minutes after entering sleep (i.e., when the N1 stage is marked in the dataset) to the last 30 minutes after waking up (i.e., when the last non-wakefulness stage is marked in the dataset) are taken.

Finally, the extracted data was subjected to denoising. A band-pass filter ranging from 0.5 Hz to 45 Hz was applied to the EEG signals, along with a 50 Hz notch filter to eliminate power line interference. The preprocessed data will form the training and testing sets for the subsequent experiments in this paper.

The output results of the model are compared with the true labels of sleep stages in the test set. Since the automatic sleep staging problem is a multi-classification problem, will be evaluated from multiple metrics including accuracy, recall, precision, F1-score, and Cohen kappa coefficient. The calculation formulas for each of these metrics are as follows, where TP represents the number of correctly predicted positive samples, TN represents the number of correctly predicted negative samples, FP represents the number of falsely predicted positive samples, and FN represents the number of falsely predicted negative samples. In a multi-classification problem, all classes other than the current sample class are regarded as negative samples.

$$\begin{cases} accuracy = \dfrac{TP+TN}{TP+TN+FP+FN} \\ recall = \dfrac{TP}{TP+FN} \\ precision = \dfrac{TP}{TP+FP} \\ F1-score = \dfrac{2PR}{P+R} \end{cases} \tag{3}$$

The Cohen's kappa coefficient is used to evaluate the consistency between the model's classification results and the true labels. It is calculated based on the confusion matrix. The formula is as follows, where f_{ii} represents the value in the row i and column j of the confusion matrix, f_{i+} represents the sum of the row i of the matrix, and f_{+i} represents the sum of the column j of the matrix.

$$\begin{cases} p_0 = \dfrac{1}{n}\sum_{i=1}^{g} f_{ii} \\ p_e = \dfrac{1}{n^2}\sum_{i=1}^{g} f_{i+} f_{+i} \\ kappa = \dfrac{p_0 - p_e}{1 - p_e} \end{cases} \tag{4}$$

The experiments were conducted on a computing platform comprising an Intel® Core™ i9-14900KF CPU and an NVIDIA GeForce RTX 4090 GPU with 24GB of dedicated VRAM, supported by 24GB of system RAM. The software stack utilized Ubuntu 22.04.3 LTS as the operating system, with deep learning implementations accelerated through PyTorch 2.5.1 framework coupled with CUDA 12.4 parallel computing architecture.

The single-channel EEG signal training set was input into LA-Mamba net. During the training process, the ten-fold cross-validation method was adopted. The training set was first divided into ten parts, with one part used as the validation set and the other nine parts used as the training set. This method is used to select the relevant parameters for model training and select the model parameters with the best results in the validation set as the final training results and present the final results in the test set.

The loss function of the model training is focal loss which is shown as (2), where the α is [1,2,1,1,1] and γ is taken as 1

TABLE I. EXPERIMENTAL TRAINING PARAMETERS

Parameter	Value
BATCHSIZE	16
EPOCHS	50
LEARNING_RATE=0.001	0.001
Optimize	AdamW
AdamW_weight decay	0.01

V. Results and Discussion

A. Results

Figure. 5 shows the variation of the loss function value during the model training experiment. It can be noticed that the value of the loss function gradually decreases until it gradually stabilizes and converges in the the twenty-fifth rounds. This indicates the reliability of the model's training process.

Table 2.5 presents the results of LA-Mamba network on the Sleep-EDFx test set. Overall, the model's overall accuracy reached 84.63%, the Macro-F1 score was 78.01%, and the Cohen kappa value was 0.839. Specifically, the stage Wake had the highest scores for all metrics, with an F1 score of 0.935. The stage N1 had the lowest scores for all metrics, with an F1 score of 0.481. The F1 scores for the remaining stages were all above 0.80. Figure. 6 shows the confusion matrix of the model's results.

TABLE II. Experimental Results of Mamba-Based Automated Sleep Staging Modeling

Sleep Stage	Acc	F1-score	Kappa	Precision	Recall	F1-score
stage W	0.846	0.780	0.839	0.950	0.920	0.935
stage 1				0.494	0.468	0.481
stage 2				0.856	0.883	0.869
stage 3				0.769	0.837	0.801
stage R				0.813	0.816	0.815

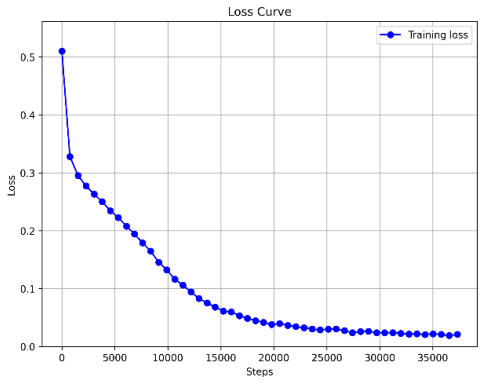

Fig. 5. Model training loss function value change.

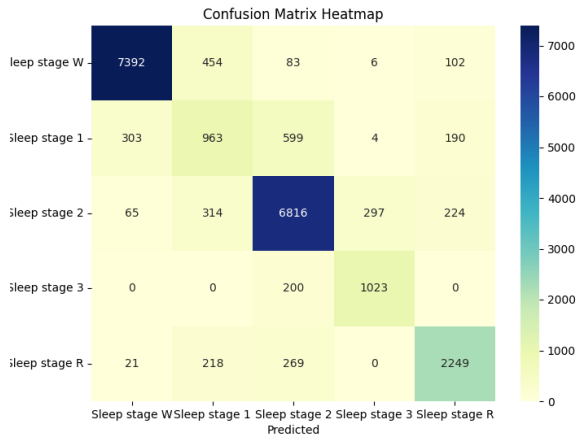

Fig. 6. Confusion matrix for experimental results of Mamba-based automated sleep staging models.

B. Comparison with Other Automated Sleep Staging Models

To demonstrate the superiority of the LA-Mamba net for the problem of automatic sleep staging, the model was compared with other automatic sleep staging models using using single-channel EEG proposed in previous studies on the Sleep-EDFx dataset. The overall classification performance comparison results are shown in Table III. The deepened model name is LA-Mamba net. The underlined items in the classification result evaluation criteria indicate the highest score in the horizontal comparison. Through comparison, it can be found that LA-Mamba net achieves higher accuracy and kappa values on the Sleep-EDFx dataset compared with other models, and the F1 score is only 0.001 lower than the best model, which can be considered basically the same.

TABLE III. Results of the Mamba-Based Automated Sleep Staging Model Compared with Other Models for Overall Staging Effectiveness

Model	F1-score	Acc	Kappa
AttnSleep[15]	0.751	0.813	0.74
TinySleepNet[16]	**0.781**	0.831	0.770
EEGSNet[17]	0.773	0.830	0.770
CausalAttnNet[9]	0.752	0.816	0.748
LSTM-HMM model[5]	0.750	0.827	0.760
BiT-MamSleep[18]	0.7268	0.802	0.7437
LA-Mamba Net	0.7801	**0.8463**	**0.8392**

Table IV shows the F1 scores of different models at each stage, with the highest F1 score at each stage displayed in bold and underlined. It can be observed that LA-Mamba net achieves the highest F1 scores in the wake stage, stage N2 of sleep, and the rapid eye movement stage. Meanwhile, its score in stage N3 of sleep is comparable to those of other models, and it only differs by 0.03 from the highest score in stage N1 of sleep, while it outperforms other models by more than 0.05. Maintaining a certain accuracy rate in the difficult-to-classify stage N1 is attributed to the focal loss function used during model training. This function ensures a better classification effect for stage N1 while minimizing the impact on the classification performance of other stages, thereby enhancing the overall accuracy of classification.

In summary, it can be seen that both in terms of the overall classification effect and in terms of individual stages, LA-Mamba net exceeds the classification effect of most of the automatic sleep staging models proposed in current studies, especially in the difficult to classify sleep stage 1, which guarantees a high classification effect.

C. Ablation and Comparison Experiments

In order to validate the importance and necessity of each structure in the network model proposed, several ablation and comparison experiments will be designed in this section to

demonstrate the rationality and superiority of each structure and design in the multicore convolutional layer, the LocalAttn-Mamba layer, and the classification layer of the network structure in comparison to other network structures, respectively.

TABLE IV. RESULTS OF THE MAMBA-BASED AUTOMATED SLEEP STAGING MODEL COMPARED WITH OTHER MODELS FOR CLASSIFYING THE EFFECTS OF INDIVIDUAL STAGES

Model	Period	F1-score
TinySleepNet	W	**0.93**
	N1	**0.51**
	N2	0.85
	N3	0.81
	REM	0.80
AttnSleep	W	0.92
	N1	0.42
	N2	0.85
	N3	0.82
	REM	0.74
CausalAttnNet	W	0.92
	N1	0.39
	N2	0.85
	N3	**0.83**
	REM	0.75
BiT-MamSleep	W	0.92
	N1	0.38
	N2	0.84
	N3	0.81
	REM	0.72
LA-MambaNet	W	**0.93**
	N1	0.48
	N2	**0.87**
	N3	0.80
	REM	**0.81**

In terms of the convolutional layer, the experiment tested th e effects of different allocation combinations of sub-layers in th e convolutional network and conducted ablation experiments o n this layer. This section also replaced the multi-core convoluti onal layer with a dilated causal convolutional pyramid structur e for experimentation. The results of the above ablation and co mparison experiments are shown in Table V.

TABLE V. MULTI-CORE CONVOLUTIONAL LAYER ABLATION AND COMPARISON EXPERIMENTAL RESULTS

Ablation setting	Acc	F1-score
Only the first convolutional layer	0.822	0.760
Only the second convolutional layer	0.825	0.741
Only the third convolutional layer	0.827	0.746
Only the second and third convolutional layer	0.832	0.754
Only the first and third convolutional layer	0.834	0.754
Only the first and second convolutional layer	0.834	0.759
No convolutional layer	0.779	0.674
TCN	0.820	0.751

It can be seen that when the number of layers of the multi-core parallel convolutional layer decreases, the model's phase discrimination effect also gradually declines. This also proves that the feature information extracted by different convolutional kernel sizes is different and not redundant. At the same time, without using the convolutional layer for feature extraction and go directly to the subsequent layers, the accuracy drops significantly, which indicates that the convolutional layer plays a crucial role in data information extraction. However, the performance of TCN does not improve, and it is speculated that this is due to conflicts with the subsequent Mamba network and possible information omission when using dilated convolution.

Regarding the LocalAttn-Mamba layer, ablation experimen ts were conducted on the windowed self-attention mechanism a nd the Mamba layer while keeping other structures unchanged. To verify the rationality of the parameter selection for the Mam ba layer and the windowed self-attention layer, comparative ex periments with relevant parameters were also set up in this sect ion. To validate the rationality of the network structure, this sec tion also conducted comparative experiments by replacing the Mamba layer with an LSTM network structure (with 64 hidden units, two layers, and bidirectional) and replacing the windowe d self-attention layer with a global Transformer. These experim ents aim to prove the rationality of the network structure.

The experimental results are shown in Table VI. It can be seen that not using the Mamba layer and the windowed self-attention layer will both lead to a decline in the overall model's classification performance. At the same time, it can be noted that in terms of parameter selection, the effect of using one layer of Mamba is far inferior to that of using two layers. Although the effect of using three layers of Mamba is similar to that of using two layers, considering the runtime and complexity, a simpler network model, that is, a two-layer Mamba network, should be chosen. Other experiments have also proved the rationality of the parameters.

TABLE VI. LOCALATTN-MAMBA LAYER ABLATION AND COMPARATIVE EXPERIMENTAL RESULTS

Ablation setting	Acc	F1-score
Not using the Mamba layer	0.827	0.764
Not using the Windowed Self-Attention Layer	0.837	0.774
Use only one Mamba layer	0.830	0.770
Use only three Mamba layer	0.843	0.783
Mamba_state=96	0.834	0.754
Mamba_state=48	0.834	0.768
local_windowsize=300	0.831	0.773
LSTM	0.835	0.674
Transformer	0.841	0.759

In the comparison results with LSTM and Transformer networks, it can be found that the accuracy of the LSTM network is slightly lower than that of the Mamba network, the Transformer network, while differing from the windowed self-attention network by only 0.05 in terms of correctness, takes more than 100 seconds longer to train than the network proposed in this paper in terms of one round of training time. This is consistent with the background and purpose of the Mamba network design. That is, the Mamba network can

ensure a higher accuracy than LSTM and a similar accuracy to Transformer, while using less time for model parameter updates than Transformer.

In terms of the classification layer, The experiments compared the output results of several model designs, including using convolutional networks for dimensionality reduction (mainly along the feature dimension), using only global average pooling or global max pooling for dimensionality reduction, directly outputting five classifications using only global average pooling without fully connected layers.

The results are shown in Table VII. It can be seen from the results that using convolutional neural networks or a single type of global pooling for data dimensionality reduction is inferior to using both global max pooling and global average pooling simultaneously. This also indicates that the final classifier may require both background and foreground feature information, and dimensionality reduction in the time dimension may be more reasonable than in the feature dimension.

TABLE VII. RESULTS OF CLASSIFICATION LAYER ABLATION AND COMPARISON EXPERIMENTS

Ablation setting	Acc	F1-score
Convolutional neural network dimensionality reduction	0.830	0.764
Use only global max pooling for dimensionality reduction	0.818	0.748
Use only global average pooling for dimensionality reduction	0.817	0.747
Classification using global max pooling	0.834	0.768

In LA-Mamba net, the focal loss function is selected as the loss function for the automatic sleep staging model. However, the focal loss function has two parameters, γ and weights, to be chosen. To verify the correctness of the parameter selection and the improvement of the classification effect of the focal loss function on difficult classifications, this section designs relevant comparative experiments.

The loss function is set as follows in the experiment, and all other parameters remain the same as those mentioned earlier:

Experiment 1: Use the classic multi-class cross-entropy loss function;

Experiment 2: Use the weighted cross-entropy loss function and set the weight of period N1 to 2, while the weights of other periods are set to 1.

Experiment 3: Use the focal loss function without setting weights and set the γ value to 1;

Experiment 4: Use the focal loss function without setting weights and set the γ value to 2.

Experiment 5: Use the focal loss function without setting weights, and set the γ value to 3.

Experiment 6: Use the focal loss function, set the weight of stage N1 to 2 and that of other stages to 1, and set the γ value to 1.

Experiment 7: Use the focal loss function, set the weight of stage N1 to 2 and that of other stages to 1, with the γ value set to 2.

All experiments underwent 20 rounds of training, and the results of each round were tested on the test set. The training model with the highest Macro Avg F1-score was selected as the final output result. The F1 score of the N1 stage classification result and the overall classification accuracy were observed. The results are shown in the Table VIII.

TABLE VIII. RESULTS OF CLASSIFICATION LAYER ABLATION AND COMPARISON EXPERIMENTS

	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7
N1stage F1-score	0.4514	0.4972	0.4137	0.4627	0.4677	0.4998	0.4559
Overall ACC	0.8381	0.8126	0.8474	0.8412	0.8375	0.8319	0.8389
Macro Avg F1-score	0.7687	0.7628	0.7665	0.7709	0.7677	0.7782	0.7673

The experimental results show that compared with the most basic cross-entropy loss function, the weighted loss function (Experiment 2) or the addition of the γ term in the focal loss function (Experiments 4, 5, and 6) can both improve the classification effect of stage N1 to a certain extent. Moreover, these modifications maintained the overall F1-score within the range of 0.77 ± 0.08, indicating no statistically significant degradation in overall performance.

An increase in the γ value does not necessarily improve the classification effect of stage N1 (Experiments 6 and 7). This conclusion is also mentioned in the focal loss function paper [14]. When the γ value is too large, it may lead to an excessive number of other stages being classified as stage N1, resulting in a decrease in its F1-score.

The application of focal loss necessitates empirical tuning of both class weights and γ to optimize classification performance on hard-to-classify samples (Experiments 6 and 7). This is also reflected in the focal loss function paper. In this experiment, it can be found that when the weight values are [1, 2, 1, 1, 1] and the γ value is 1, the best classification effect for stage N1 can be achieved.

VI. CONCLUSION

This paper presents a novel single-channel EEG-based automatic sleep staging architecture built upon the Mamba network framework. The proposed model capitalizes on temporal correlations inherent in EEG signals to address sleep stage classification. Meanwhile, for sleep stage one, which is

difficult to classify and has little data, the network is trained by using a focal loss function to improve the classification of this stage. A series of ablation and comparison experiments were conducted on the proposed structure in this chapter, comparing the substructures within the network with causal dilated convolution, global multi-head self-attention, LSTM, and other network structures, thereby demonstrating the rationality of the proposed network structure.

REFERENCES

[1] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, Sep. 2000, doi: 10.1109/10.867928.

[2] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215-220, Jun. 2000, doi: 10.1161/01.cir.101.23.e215.

[3] M. Radha, G. Garcia-Molina, M. Poel, and G. Tononi, "Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2014, pp. 1876–1880. doi: 10.1109/EMBC.2014.6943976.

[4] J. Yang, Y. Meng, Q. Cheng, H. Li, W. Cai, and T. Wang, "Deep Learning Automatic Sleep Staging Method Based on Multidimensional Sleep Data," *IEEE Access*, vol. 12, pp. 168360–168369, 2024, doi: 10.1109/ACCESS.2024.3496721.

[5] W. Chen, Y. Cai, A. Li, Y. Su, and K. Jiang, "Single-Channel Sleep EEG Classification Method Based on LSTM and Hidden Markov Model," *Brain Sci.*, vol. 14, no. 11, p. 1087, 2024, doi: 10.3390/brainsci14111087.

[6] M. He, M. Tang, L. Meng, and Z. Liang, "TBSTSleepNet: Three-branch spectro-temporal bidirectional LSTM based attention model for EEG sleep staging," *Biomed. Signal Process. Control*, vol. 97, p. 106695, Nov. 2024, doi: 10.1016/j.bspc.2024.106695.

[7] A. Ito and T. Tanaka, "SleepSatelightFTC: A Lightweight and Interpretable Deep Learning Model for Single-Channel EEG-Based Sleep Stage Classification," *IEEE Access*, vol. 13, pp. 46263–46272, 2025, doi: 10.1109/ACCESS.2025.3549436.

[8] T. Zhu, W. Luo, and F. Yu, "Convolution- and Attention-Based Neural Network for Automated Sleep Stage Classification," *Int. J. Environ. Res. Public. Health*, vol. 17, no. 11, Art. no. 11, Jan. 2020, doi: 10.3390/ijerph17114152.

[9] J. Pan, Y. Feng, P. Zhao, X. Zou, A. Hou, and X. Che, "CausalAttenNet: A Fast and Long-Term-Temporal Network for Automatic Sleep Staging With Single-Channel EEG," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024, doi: 10.1109/TIM.2024.3453309.

[10] X. Chen, Y. Zhao, S. Shen, and X. Chen, "MMSleepGNet: Mixed Multibranch Sequential Fusion Model Based on Graph Convolutional Network for Automatic Sleep Staging," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–10, 2024, doi: 10.1109/TIM.2023.3298402.

[11] X. Wang and Y. Zhu, "SleepGCN: A transition rule learning model based on Graph Convolutional Network for sleep staging," *Comput. Methods Programs Biomed.*, vol. 257, p. 108405, Dec. 2024, doi: 10.1016/j.cmpb.2024.108405.

[12] A. Gu, K. Goel, and C. Ré, "Efficiently Modeling Long Sequences with Structured State Spaces," Aug. 05, 2022, *arXiv*: arXiv:2111.00396. doi: 10.48550/arXiv.2111.00396.

[13] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," May 31, 2024, *arXiv*: arXiv:2312.00752. doi: 10.48550/arXiv.2312.00752.

[14] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," Jun. 06, 2023, *arXiv*: arXiv:1606.08415. doi: 10.48550/arXiv.1606.08415.

[15] A. Supratak and Y. Guo, "TinySleepNet: An Efficient Deep Learning Model for Sleep Stage Scoring based on Raw Single-Channel EEG," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Jul. 2020, pp. 641–644. doi: 10.1109/EMBC44109.2020.9176741.

[16] "A Deep Learning Method Approach for Sleep Stage Classification with EEG Spectrogram." Accessed: Apr. 25, 2025. [Online]. Available: https://www.mdpi.com/1660-4601/19/10/6322.

[17] X. Zhou *et al.*, "BiT-MamSleep: Bidirectional Temporal Mamba for EEG Sleep Staging," Nov. 21, 2024, *arXiv*: arXiv:2411.01589. doi: 10.48550/arXiv.2411.01589.

[18] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," Apr. 19, 2018, *arXiv*: arXiv:1803.01271. doi: 10.48550/arXiv.1803.01271.