# EEGMAMBA: BIDIRECTIONAL STATE SPACE MODELS WITH MIXTURE OF EXPERTS FOR EEG CLASSIFICATION

## A PREPRINT

Yiyu Gui MingZhi Chen Yuqi Su Guibo Luo Yuchao Yang School of Electronic and Computer Engineerng, Peking University luogb@pku.edu.cn, yuchaoyang@pku.edu.cn

## **ABSTRACT**

In recent years, with the development of deep learning, electroencephalogram (EEG) classification networks have achieved certain progress. Transformer-based models can perform well in capturing long-term dependencies in EEG signals. However, their quadratic computational complexity leads to significant computational overhead. Moreover, most EEG classification models are only suitable for single tasks, showing poor generalization capabilities across different tasks and further unable to handle EEG data from various tasks simultaneously due to variations in signal length and the number of channels. In this paper, we introduce a universal EEG classification network named EEGMamba, which seamlessly integrates the Spatio-Temporal-Adaptive (ST-Adaptive) module, Bidirectional Mamba, and Mixture of Experts (MoE) into a unified framework for multiple tasks. The proposed ST-Adaptive module performs unified feature extraction on EEG signals of different lengths and channel counts through spatio-adaptive convolution and incorporates a class token to achieve temporal-adaptability. Moreover, we design a bidirectional Mamba particularly suitable for EEG signals for further feature extraction, balancing high accuracy and fast inference speed in processing long EEG signals. In order to better process EEG data for different tasks, we introduce Task-aware MoE with a universal expert, achieving the capture of both differences and commonalities between EEG data from different tasks. We test our model on eight publicly available EEG datasets, and experimental results demonstrate its superior performance in four types of tasks: seizure detection, emotion recognition, sleep stage classification, and motor imagery. The code is set to be released soon.

## 1 Introduction

Electroencephalogram (EEG) is a technique of recording brain activity using electrophysiological indicators, which captures the electrical wave changes during brain activity. EEG can be utilized to detect various human physiological activities such as seizure detection, emotion recognition, motor imagery, sleep stage classification, and other physiological related task Shoeibi et al. [2021], Jafari et al. [2023], Altaheri et al. [2023], Sri et al. [2022].

In recent years, with the development of deep learning, EEG classification models based on deep learning have been widely used Chen et al. [2022]. Among them, models based on Convolutional Neural Networks (CNNs) and Transformers are the most representative, each with their own strengths and weaknesses. CNN-based EEG classification networks have the advantage of faster training and inference speeds, and they perform well on short EEG signals. However, due to the lack of global sequence modeling ability, their performance on long EEG signals cannot be guaranteed Sakhavi et al. [2018], Thuwajit et al. [2021], Schirrmeister et al. [2017]. In contrast, Transformer-based EEG classification networks have good capability of global sequence modeling, achieving excellent performance on both short and long EEG signals. However, as the length of the EEG signal increases, the computational complexity of the model increases quadratically, significantly raising the training and inference costs Dai et al. [2023], Xie et al. [2022], Wang et al. [2022].

Recently, State Space Models (SSM) with selection mechanism and efficient hardware-aware design, such as Mamba Gu and Dao [2023], have shown great potential in long sequence modeling. By utilizing selective state space model, it effectively captures the relationships between tokens in a sequence, addressing the limitation of CNNs in modeling long sequences. Moreover, it exhibits linear computational complexity, which outperforms the quadratic complexity of Transformers and provides a strong backbone network for training EEG classification models on long EEG signals.

Existing EEG classification models always focus on solving specific tasks O'Shea et al. [2020], Phan et al. [2022], Algarni et al. [2022], Autthasan et al. [2021]. However, these networks tend to be less universal across different tasks. While some models consider the generality between EEG tasks, such as EEGNet Lawhern et al. [2018], which has been validated on four tasks including P300 visual-evoked potentials, error-related negativity responses (ERN), movement-related cortical potentials (MRCP), and sensory motor rhythms (SMR), they can only address one type of task in a single training session. Therefore, it is essential to design a classification network capable of handling multi-task EEG data simultaneously.

One of the significant obstacles for multi-task EEG classification is that different EEG data have varying numbers of channels and signal lengths, which makes it difficult for networks to adapt during a single training. For example, MS-HNN Zhu et al. [2023a] is designed for single-channel sleep data and struggles to adapt other multi-channel EEG signals. While MaskSleepNet Zhu et al. [2023b] can classify EEG signals with different numbers of channels by manually setting the channel parameter, it uses a fixed-parameter Multi-scale CNN that can only process EEG signals with limited input lengths. EEG ConvNet Schirrmeister et al. [2017] is designed with a structure capable of adapting to arbitrary signal lengths, it still requires manual setting in different trainings. Therefore, enabling the model to adapt to different signal lengths and channel counts represents a significant challenge.

On the other hand, a network capable of simultaneously handling multi-task EEG data requires a larger network size, more training data, and the ability to address different tasks pertinently. Mixture of Experts (MoE) is a deep learning model with sparse gate-controlled architecture, consisting of a group of expert models and a gating network Jacobs et al. [1991], Shazeer et al. [2016], Xue et al. [2024]. Multiple experts allow for a large increase in the number of model parameters, while the sparse activation mechanism minimizes the impact on the training and inference processes. The gating network can adaptively select experts based on the input, assigning different tasks to different experts, thus achieving task-specificity needed for multitask classification. Therefore, using MoE to achieve EEG multi-task classification might be a feasible solution.

In general, existing EEG classification models mainly face two challenges. On the one hand, these models find it difficult to balance high accuracy and fast inference speed when dealing with long EEG signals. On the other hand, they often struggle to handle different EEG classification tasks and demonstrate poor generality.

To address the aforementioned two issues, we propose EEGMamba, which utilizes bidirectional Mamba suitable for EEG signals, as well as a Spatio-Temporal-Adaptive (ST-Adaptive) module and Task-aware MoE for targeted processing of multi-task EEG classification. Our model enhances Mamba by employing bidirectional modeling to capture the relationships between tokens in a one-dimensional temporal sequence, achieving high accuracy and fast inference speed. Additionally, we propose an ST-Adaptive module that uses spatio-adaptive convolution to process EEG signals of varying channel numbers and a class token to achieve temporal adaptability without any additional processing. To improve generalizability across EEG tasks, we design a task-aware gating network that accurately directs different EEG task tokens to specific experts for processing, while also employing a universal EEG expert to exploit commonalities among different EEG tasks. In summary, our contributions are as follows:

- (1) We design a bidirectional Mamba for EEG signals, which balances fast inference speed with excellent global perception capability.
- (2) We propose a ST-Adaptive module that can automatically adapt to EEG signals of different lengths and channels, thereby processing them simultaneously in single training session.
- (3) We introduce Task-aware MoE with a universal expert, achieving the capture of both differences and commonalities between EEG data from different tasks.

## 2 Method

EEGMamba primarily consists of the ST-Adaptive module, bidirectional Mamba, and Task-aware MoE. The ST-Adaptive module processes EEG signals of arbitrary lengths and channel numbers through Spatial-Adaptive convolution, Tokenize Layer, and Temporal-Adaptation based on the class token. The features extracted by the ST-Adaptive module are then processed by multiple bidirectional Mamba blocks to perform sequence modeling. Finally, the Task-aware

MoE handles task-specific processing of EEG tokens from different tasks, and a task-aware classifier provides the classification results. The overall model architecture is illustrated in Figure 1.

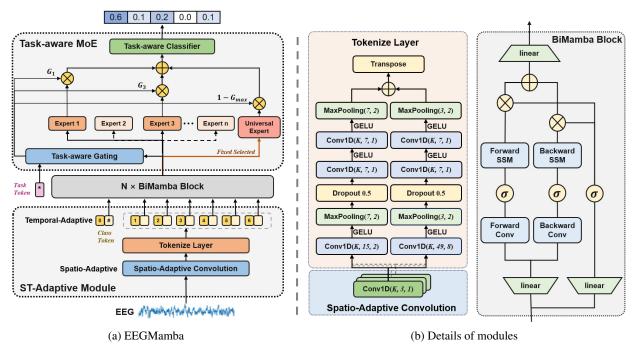


Figure 1: Overall structure of proposed model. The model consists of ST-Adaptive module, Bidirectional Mamba (BiMamba) blocks and Task-aware MoE module.

# 2.1 Preliminary Work

Mamba is inspired by continuous state-space equations. For continuous input  $x(t) \in \mathbb{R}$  in the time domain, the corresponding output  $y(t) \in \mathbb{R}$  is determined by the current hidden state h(t) and input x(t) at time t, as shown in Equation (1). Here,  $A \in \mathbb{R}^{N \times N}$  is the state matrix,  $B \in \mathbb{R}^{N \times 1}$  is related to the system's hidden state, and  $C \in \mathbb{R}^{1 \times N}$  is a parameter associated with the input and output.

$$h'(t) = Ax(t) + Bh(t)$$

$$y(t) = Ch(t)$$
(1)

Mamba discretizes the continuous-time t into discrete time, transforming the continuous state-space equations into discrete state-space equations. Specifically, by introducing a time-scale parameter  $\Delta$ , A and B are transformed into discrete-time parameters  $\bar{A}$  and  $\bar{B}$  respectively. The zero-order hold (ZOH) technique is used as the transformation rule, as shown in Equation (2).

$$\bar{A} = exp(\Delta A)$$

$$\bar{B} = (\Delta A)^{-1} (exp(\Delta A) - I)\Delta B$$
(2)

In practice, following the approach of Gu et al. Gu and Dao [2023], we approximate  $\bar{B}$  using a first-order Taylor expansion, as show in Equation (3):

$$\bar{B} = (\Delta A)^{-1} (exp(\Delta A) - I)\Delta B \approx \Delta B \tag{3}$$

Finally, the discretized form of the continuous state space equation is shown in Equation (4).

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t$$

$$y_t = Ch_t$$
(4)

Based on the mentioned discrete state-space equations, Mamba further introduces data dependency into the model parameters, enabling the model to selectively propagate or forget information based on the sequential input tokens. In addition, it utilizes a parallel scanning algorithm to accelerate the equation-solving process.

## 2.2 ST-Adaptive Module

Signals from different EEG datasets often have different lengths and channel numbers. To address this issue, we design a Spatio-Temporal Adaptive module, which converts input signals of arbitrary lengths and channel numbers into uniform features.

We use Spatial-Adaptive convolutional layers to achieve adaptive channel count. Define  $x \in \mathbb{R}^{B \times C_i \times L_i}$  as the EEG signal,  $C_i$  is the number of EEG channels for the *i*-th task, and  $L_i$  is the length of the EEG signal for the *i*-th task.

$$y_{SA} = CNN_{SA}(x) \in \mathbb{R}^{B \times D \times L_i}$$
(5)

As shown in Equation (5),  $y_{SA}$  is the result obtained through Spatial-Adaptive convolution, where the channel dimension is changed from  $C_i$  determined by the task to a unified D. Then,  $y_{SA}$  is converted into an EEG token sequence through the Tokenize Layer. In order to better extract features from EEG signals, we design a dual-path structure utilizing a small kernel convolution module  $CNN_S$  and a wide convolutional module  $CNN_W$ . Obtain the small kernel feature token sequence  $z_s$  and the wide kernel feature token sequence  $z_w$ , respectively. Finally, we concatenate them in the time dimension to form the EEG token sequence T, as shown in Equation (6).

$$z_{s} = \mathcal{T}(CNN_{s}(y_{SA})) \in \mathbb{R}^{B \times D \times N_{s}}$$

$$z_{w} = \mathcal{T}(CNN_{w}(y_{SA})) \in \mathbb{R}^{B \times D \times N_{w}}$$

$$T = Concat(z_{s}, z_{w}) \in \mathbb{R}^{B \times D \times N}$$
(6)

Among them,  $\mathcal{T}$  represents the transpose operation,  $N_s$ ,  $N_w$ , N are the number of EEG small kernel feature tokens, EEG wide kernel feature tokens, and overall EEG tokens, respectively. Due to the varying lengths of EEG signals, the number of EEG tokens obtained from the Tokenize Layer is inconsistent.

To achieve temporal adaptation, we introduce a special class token Dosovitskiy et al. [2020]. Specifically, we concatenate this class token with the previously extracted feature token sequence  $t_s^1, t_s^2, \dots$  and  $t_w^1, t_w^2, \dots$  to obtain the token sequence  $T^*$ , as shown in Equation (7).

$$T^* = [t_{cls}, t_s^1, t_s^2, ..., t_s^{N_s}, t_w^1, t_w^2, ..., t_w^{N_w}] \in \mathbb{R}^{B \times (N+1) \times D}$$
(7)

Then, the input token sequence  $T^*$  is processed through a network (using bidirectional Mamba blocks in this study) to integrate EEG token sequence information into the class token. This approach prevents the network from developing biases towards certain tokens in the EEG feature token sequence T due to variations in input length, thereby achieving temporal adaptability.

## 2.3 Bidirectional Mamba Block for EEG

Mamba is designed for Natural Language Processing (NLP), with its output at each moment depends only on the current input and hidden state, without consideration for future time steps. Since NLP is primarily a generative autoregressive task that relies on previous information for judgment, Mamba's single-directional modeling approach is sufficient to complete such tasks. However, EEG classification tasks require simultaneous processing of both preceding and following information, which cannot be learned by single-directional modeling. Therefore, for EEG signals, the original Mamba's single-directional modeling is insufficient.

To address this issue, we design a bidirectional Mamba for one-dimensional temporal signals, which can model the input bidirectionally and more effectively learn the dependencies between time series tokens. We use the features extracted by the ST-Adaptive module as the input for the bidirectional Mamba.

We denote the input of the bidirectional Mamba block as a sequence  $T_{k-1}$  and the output as a sequence  $T_k$ . First,  $T_{k-1}$  is normalized to  $T_{k-1}^{norm}$  by layer normalization. Next, it is mapped by  $Linear_X$  and  $Linear_Z$  to  $X_{k-1}$  and  $Z_{k-1}$ , respectively. Then,  $X_{k-1}$  enters parallel forward and backward sequence modeling modules. The forward module includes forward 1D causal convolution  $Conv_f$  and forward SSM module  $SSM_f$ . Similarly, the backward module includes backward 1D causal convolution  $Conv_b$  and backward SSM module  $SSM_b$ . Then, the results of forward

sequence modeling  $y_{k-1}^f$  and backward sequence modeling  $y_{k-1}^b$  are summed with  $Z_{k-1}$  through gating and then projected through a linear layer Linear to obtain  $T'_{k-1}$ . Finally, the output sequence  $T_k$  is obtained through residual connection. The detailed process is shown in Algorithm 1:

# Algorithm 1 Bidirectional Mamba Block Process

```
Input: token sequence T_{k-1} \in \mathbb{R}^{B \times (N+1) \times D}

Output: token sequence T_k \in \mathbb{R}^{B \times (N+1) \times D}

1: T_{k-1}^{norm} \leftarrow LayerNorm(T_{k-1})

2: X_{k-1} \leftarrow Linear_X(T_{k-1}^{norm}), Z_{k-1} \leftarrow Linear_Z(T_{k-1}^{norm})

3: y_{k-1}^f \leftarrow SSM_f(Conv_f(Transpose(X_{k-1})))

4: y_{k-1}^b \leftarrow Reverse(SSM_b(Conv_b(Reverse(Transpose(X_{k-1})))))

5: T_{k-1}' \leftarrow Linear(Transpose(y_{k-1}^f + y_{k-1}^b) \odot SiLU(Z_{k-1}))

6: T_k = T_{k-1}' + T_{k-1}
```

## 2.4 Task-aware MoE with Universal Expert

## 2.4.1 Sparsely-activated MoE

Transformer-based MoE commonly use a sparse activation MoE layer to replace the Feed-Forward Neural Network (FFN) inside the Transformer Fedus et al. [2022]. Each MoE layer consists of several experts, and each expert is typically represented as a Multi-Layer Perceptron (MLP) whose activation is controlled by a gating network Shazeer et al. [2016].

We define  $N_e$  as the number of experts,  $E_i$  as the *i*-th expert, and G as the gating network. For each input EEG class token  $t_{cls}^*$ , the output  $y_{cls}$  of MoE can be expressed as Equation (8):

$$y_{cls} = \sum_{i=1}^{N_e} e_i(t_{cls}^*) * E_i(t_{cls}^*)$$

$$e_i(t_{cls}^*) = Top_k(G(t_{cls}^*))_i, t_{cls}^* = T_k [0]$$

$$Top_k(V, k)_i = \begin{cases} v_i, & \text{if } v_i \text{ is top } k \text{ value of } V \\ -\infty & \text{otherwise} \end{cases}$$

$$(8)$$

## 2.4.2 Task-aware Gating Networks

A gating network calculates gating values based on the input tokens and selects K experts for activation, typically implemented using a fully connected layer  $Linear_{Gate}$ . However, this can lead to the problem that only a few experts are trained. To avoid this, we adopted the method from Shazeer et al. [2016], adding noise to the gating value computation process using a fully connected layer  $Linear_{Noise}$ , which increases randomness and helps in balancing the load among the experts.

Furthermore, we propose a task-aware gating network which helps improve the accuracy of experts in processing different types of EEG tokens. Specifically, we encode the EEG task into task tokens  $t_{task}$ , then concatenate  $t_{task}$  with the EEG class tokens  $t_{cls}^*$  to obtain  $t_{cat}$ , which is then sent to the gating network. The gating values calculated in this manner incorporate task information, allowing for better assignment of different tasks to different experts. The working process of the task-aware gating network is shown in Equation (9), where  $\epsilon$  represents standard Gaussian noise.

$$t_{cat} = Concat(t_{cls}^*, t_{task})$$

$$G(t_{cls}^*, t_{task}) = Linear_{Gate}(t_{cat}) + \epsilon * SoftPlus(Linear_{Noise}(t_{cat}))$$
(9)

## 2.4.3 EEG universal expert

EEG signals from different tasks exhibit both differences and commonalities. Only using different experts to process EEG tokens might overlook the connections between tokens from different tasks. Therefore, we design an EEG universal expert that can process EEG tokens from all different tasks and capture their commonalities. To achieve this function, the universal expert is activated for any inputs and not controlled by the gating network's output values.

Overall, our MoE module includes both task experts and universal experts. Task experts can accurately process EEG tokens from different tasks according to gating values, while universal experts can process all EEG tokens. The output of MoE is the weighted sum of these two types of experts. We adopted a weight design scheme similar to Gou et al. [2023], as shown in Equation (10). Here, the output weight  $\omega$  of the universal expert is determined by the maximum gating value:

$$y = \sum_{i=1}^{N_e} e_i(t_{cls}^*) * E_i^t(t_{cls}^*) + \omega * E^u(t_{cls}^*)$$

$$\omega = 1 - Max(e(t_{cls}^*))$$
(10)

# 3 Experimental Setup

## 3.1 Dataset

We evaluate the proposed EEGMamba by using eight datasets from four different tasks, including Bonn Andrzejak et al. [2001], CHB-MIT Shoeb [2009], SleepEDF-20 Kemp et al. [2000], SHHS Goldberger et al. [2000], DEAP Koelstra et al. [2011], SEED Duan et al. [2013], Shu Ma et al. [2022], and BCI-IV-2a Brunner et al. [2008]. Table 1 provides an overview of each dataset. The number of subjects, the number of classes, and the number of channels often varies for different tasks. More details about the datasets can be found in the appendix A.2.

Table 1: Dataset Introduction. '# Subjects' represents the number of subjects, and the same is true for '# Classes' and '# Channels'. For the SHHS dataset, we select data from 392 subjects out of 6441 subjects Fonseca et al. [2016].

| Dataset     | Task                        | # Subjects    | # Classes | # Channels | Sampling Frequency |
|-------------|-----------------------------|---------------|-----------|------------|--------------------|
| Bonn        | Epilepsy detection          | 10            | 5         | 1          | 173.61             |
| CHB-MIT     | Epilepsy detection          | 22            | 2         | 23         | 256                |
| SleepEDF-20 | Sleep stages classification | 20            | 5         | 1          | 100                |
| SHHS        | Sleep stages classification | 329 from 6441 | 5         | 1          | 125                |
| DEAP        | Emotion recognition         | 32            | 2         | 4          | 128                |
| SEED        | Emotion recognition         | 15            | 3         | 62         | 200                |
| Shu         | Motor imagery               | 25            | 2         | 32         | 250                |
| BCI-IV-2a   | Motor imagery               | 9             | 4         | 22         | 250                |
|             |                             |               |           |            |                    |

## 3.2 Implementation Details

In the EEGMamba experiment, we train for 100 epochs. The number of bidirectional Mamba blocks and hidden channels is set to 8 and 256, respectively. We use 8 task experts and one universal expert, 2 experts are activated at a time among regular experts. In addition, to demonstrate the effectiveness of the Mamba-based model, we also conduct EEGMamba experiments for each single dataset. In the following text, Single-task EEGMamba is used to represent this experiment. In this experiment, We train for 200 epochs. The number of bidirectional Mamba blocks and hidden channels is set to 2 and 128, respectively. For all experiments, we set the batch size to 128 and the learning rate to 2e-4. The training and test sets are divided in an 8:2 ratio. All models are trained on Intel(R) Xeon(R) Gold 6342 CPU and Nvidia A100 GPUs 80G.

# 4 Results and Discussion

# 4.1 Single-task EEGMamba Performance Comparison

We compare the performance of EEGMamba with previous classification models EEGNet Lawhern et al. [2018], Attnsleep Eldele et al. [2021], and EEG Conformer Song et al. [2022] on eight datasets and evaluated them using accuracy (ACC), Area Under Curve (AUC), and F1-score. Figure 2 illustrates the performance comparison of various classification models on different datasets. Obviously, EEGMamba outperforms the other three classification networks across all evaluation metrics in most datasets. It is worth noting that on the CHB-MIT dataset, the extremely imbalanced distribution of seizure and non-seizure samples might make accuracy a less appropriate metric, while F1-score performs

better in reflecting the model's performance. Our model, particularly, shows a significant advantage over other models in the F1 evaluation metric.

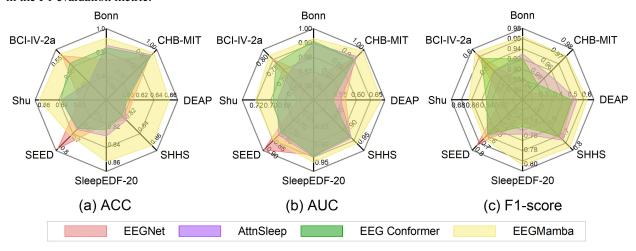


Figure 2: Performance comparison with other EEG classification models on different datasets.

We also discuss the memory-usage and inference speed of Single-task EEGMamba and Transformer-based models, especially when dealing with long sequences, as shown in Figure 3. Obviously, as the signal length increases, the memory usage of AttnSleep and EEG Conformer grows quadratically. When the signal length reaches 10000, the memory usage of Transformer-based models approaches the upper limit. In contrast, the memory usage of Single-task EEGMamba grows linearly with the signal length and can handle EEG signals of lengths exceeding 40000. In the comparison of inference speed, Single-task EEGMamba has no obvious advantage when the sequence length is less than 5000. However, as the sequence length increases, the inference speed of Transformer-based model decreases sharply, while that of Single-task EEGMamba decreases gently.

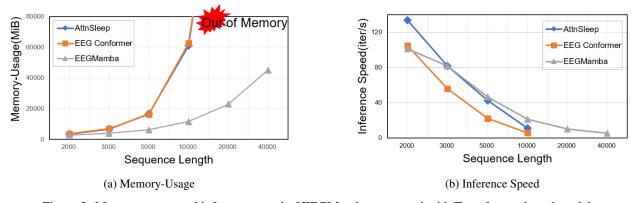


Figure 3: Memory-usage and inference speed of EEGMamba compared with Transformer-based models.

To summarize, compared with the previous classification networks, Single-task EEGMamba achieves better performance, lower memory usage and faster inference speed when dealing with long sequences.

## 4.2 EEGMamba for Multi-task EEG Classification

## 4.2.1 MoE in Multi-Task Classification

We build EEGManba according to the structure described in Figure 1. Table 2 shows the accuracy of EEGMamba on different datasets compared with EEGNet Lawhern et al. [2018], Attnsleep Eldele et al. [2021], EEG Conformer Song et al. [2022], and single-task EEGMamba. It is worth noting that all classification networks, except for EEGMamba, are trained on a single dataset. Single datasets typically have consistency in data distribution, features, and labels, which allows the model to better adapt and optimize for the specific patterns and characteristics of that dataset, thereby improving accuracy. Even though, the performance of EEGMamba surpasses that of all classification networks except for Single-task EEGMamba. This reflects the ability of the proposed model to simultaneously process diverse EEG data.

Table 2: Accuracy of EEGMamba compared with other classification models on different datasets. Bold fonts indicate the highest accuracy, and red fonts indicate the second highest accuracy.

| Classification       | Universal | Epilepsy Detection |         | Sleep Stages Cl | assification | Emotion | Recognition | Motor Imagery |           |  |
|----------------------|-----------|--------------------|---------|-----------------|--------------|---------|-------------|---------------|-----------|--|
| Network              | Model     | Bonn               | CHB-MIT | SleepEDF-20     | SHHS         | DEAP    | SEED        | Shu           | BCI-IV-2a |  |
| EEGNet               | X         | 0.8700             | 0.9927  | 0.8269          | 0.8310       | 0.6172  | 0.8241      | 0.6301        | 0.5758    |  |
| AttnSleep            | X         | 0.9300             | 0.9947  | 0.8402          | 0.8320       | 0.6055  | 0.6562      | 0.6505        | 0.4695    |  |
| EEG Conformer        | X         | 0.9200             | 0.9936  | 0.8331          | 0.8247       | 0.6094  | 0.6469      | 0.6510        | 0.5453    |  |
| Single-task EEGMamba | X         | 0.9600             | 0.9969  | 0.8689          | 0.8720       | 0.6680  | 0.7104      | 0.6751        | 0.5827    |  |
| EEGMamba             | ✓         | 0.9300             | 0.9973  | 0.8500          | 0.8540       | 0.5234  | 0.7469      | 0.6430        | 0.4626    |  |

To obtain the corresponding results presented in Table 2, our EEGMamba only needs to be trained and set the input channels and the number of classes once. However, other classification networks need to be trained multiple times, requiring manual reconfiguration of data length, number of channels, and number of classes each time, which is very inconvenient.

Furthermore, we explore the role of MoE in EEGMamba through a series of experiments. In Table 3, we show the results of no-MoE, adding MoE in each Mamba block and adding MoE after all Mamba blocks (EEGMamba). It is evident that when performing multi-task classification, MoE can effectively enhance the model's capability. We also investigate the appropriate placement and number of MoEs. In the "MoE Each Mamba" configuration, we add one MoE for each Mamba block, resulting in a total of 8 MoE modules in the entire model. Although the model has become more complex, we have not observed significant performance improvements.

Table 3: The effect of different MoE usage on Multi-task EEGMamba.

|                             | ъ.                  | no-M                    | oE EEGM                 | Iamba   | MoE t                   | for Each N              | <b>I</b> amba           | EEGMamba             |                      |                      |  |
|-----------------------------|---------------------|-------------------------|-------------------------|---|-------------------------|-------------------------|-------------------------|----------------------|----------------------|----------------------|--|
| Task                        | Dataset             | ACC                     | AUC                     | amba<br>F1<br>0.8837 0<br>0.9762 <b>0</b><br>0.7580 0<br>0.7450 0<br>0.4488 0<br>0.7410 <b>0</b><br>0.6514 <b>0</b> | ACC                     | AUC                     | F1                      | ACC                  | AUC                  | F1                   |  |
| Epilepsy Detection          | Bonn<br>CHB-MIT     | 0.8700<br>0.9972        | 0.9850<br>0.9971        |   | 0.8900<br><b>0.9978</b> | 0.9851<br><b>0.9989</b> | 0.9045<br><b>0.9810</b> | <b>0.9300</b> 0.9973 | <b>0.9917</b> 0.9958 | <b>0.9417</b> 0.9768 |  |
| Sleep Stages Classification | SleepEDF-20<br>SHHS | 0.8348<br>0.8477        | 0.9415<br>0.9364        |   | 0.8399<br>0.8483        | 0.9471<br>0.9454        | 0.7621<br>0.7350        | 0.8500<br>0.8540     | 0.9503<br>0.9512     | 0.7784<br>0.7524     |  |
| Emotion Recognition         | DEAP<br>SEED        | 0.4961<br>0.7427        | 0.4808<br>0.8813        |   | 0.5469<br>0.7485        | 0.4935<br>0.8828        | 0.5031<br><b>0.7478</b> | 0.5234<br>0.7469     | 0.5277<br>0.8894     | <b>0.5127</b> 0.7448 |  |
| Motor Imagery               | Shu<br>BCI-IV-2a    | 0.6514<br><b>0.4744</b> | 0.7181<br><b>0.7146</b> |   | <b>0.6735</b> 0.4380    | <b>0.7277</b> 0.6904    | <b>0.6729</b> 0.4392    | 0.6430<br>0.4626     | 0.7104<br>0.7051     | 0.6430<br>0.4634     |  |

## 4.2.2 Contribution of Task-aware MoE

We explore the role of designed task-aware MoE in practical application. We calculate the probability that each expert will be activated in different tasks with and without task-aware MoE, as shown in Figure 4. When using task-aware MoE, the model shows an obvious expert selection preference for a given task, suggesting that each expert has task it is good at. However, when task-aware MoE is replaced by ordinary MoE, the step diagram is almost a straight line (Figure 4b), indicating that the difference in activation probability between experts is very subtle for the same task, which is contrary to our expectation of assigning different tasks to different experts for processing.

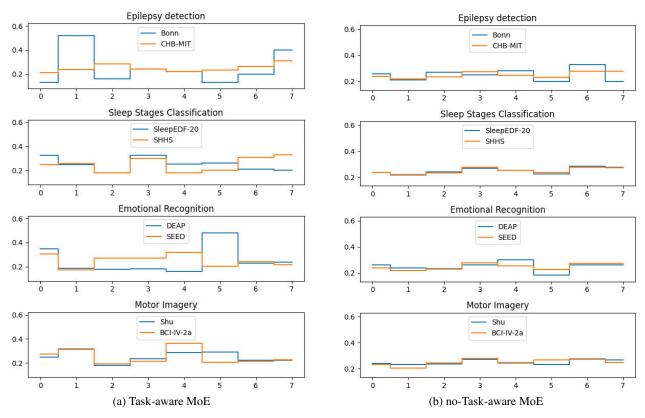


Figure 4: Expert activation probability with task-aware MoE and no-task-aware MoE.

## 4.3 Ablation Study

We compared the model performance using only task-aware gating, only universal expert, and both applied simultaneously, as shown in Table 4. Obviously, removing either of the task-aware gating and the universal expert will lead to a decrease in model performance compared to the task-aware MoE with universal expert. This may be due to the fact that the model after removing the modules cannot simultaneously capture the commonalities and differences between EEG data of different tasks. Therefore, the proposed task-aware MoE and universal experts can effectively enhance the model's performance.

| Table 4: Module ablation study in task-aware MoE with universal expert. |                     |                       |                      |                  |                  |                         |                  |                               |                      |                  |  |
|---|---------------------|-----------------------|----------------------|------------------|------------------|-------------------------|------------------|-------------------------------|----------------------|------------------|--|
|   | _                   | only Universal Expert |                      |                  | only Ta          | ask-aware               | Gating           | Task-aware + Universal Expert |                      |                  |  |
| Task  | Dataset             | ACC                   | AUC                  | F1               | ACC              | AUC                     | F1               | ACC                           | AUC                  | F1               |  |
| Epilepsy Detection  | Bonn<br>CHB-MIT     | 0.8700<br>0.9960      | 0.9813<br>0.9956     | 0.8806<br>0.9641 | 0.9200<br>0.9953 | 0.9770<br>0.9924        | 0.9287<br>0.9609 | 0.9300<br>0.9973              | 0.9917<br>0.9958     | 0.9417<br>0.9768 |  |
| Sleep Stages Classification   | SleepEDF-20<br>SHHS | 0.8350<br>0.8491      | 0.9417<br>0.9418     | 0.7525<br>0.7279 | 0.8447<br>0.8461 | 0.9412<br><b>0.9579</b> | 0.7721<br>0.7225 | 0.8500<br>0.8540              | <b>0.9503</b> 0.9512 | 0.7784<br>0.7524 |  |
| Emotion Recognition   | DEAP<br>SEED        | 0.5898<br>0.7543      | 0.6028<br>0.8924     | 0.5557<br>0.7543 | 0.5547<br>0.6393 | 0.5066<br>0.8021        | 0.5164<br>0.6390 | 0.5234<br>0.7469              | 0.5277<br>0.8894     | 0.5127<br>0.7448 |  |
| Motor Imagery   | Shu<br>BCI-IV-2a    | 0.6555<br>0.3927      | <b>0.7201</b> 0.6607 | 0.6555<br>0.3915 | 0.6606<br>0.4931 | 0.7165<br><b>0.7248</b> | 0.6598<br>0.4875 | 0.6430<br>0.4626              | 0.7104<br>0.7051     | 0.6430<br>0.4634 |  |

Table 4: Module ablation study in task-aware MoE with universal expert

# 5 Conclusion

In this paper, we propose EEGMamba, which utilizes ST-Adaptive module to adaptively extract features of EEG data with different lengths and channel numbers. We introduce bidirectional State Space Models SSM to achieve high accuracy and fast inference speed when processing long-term EEG datasets. We design a task-aware Mixture of Experts

(MoE) and an EEG universal expert, allowing the model to process multiple tasks simultaneously and better learn the commonalities among EEG signals from different tasks. We evaluate our model on eight publicly available EEG datasets across four tasks, and experimental results demonstrate the superior performance of our proposed model in multi-task classification scenarios.

## References

- Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Mahboobeh Jafari, Parisa Moridian, Roohallah Alizadehsani, Maryam Panahiazar, Fahime Khozeimeh, Assef Zare, Hossein Hosseini-Nejad, et al. Epileptic seizures detection using deep learning techniques: A review. *International journal of environmental research and public health*, 18(11): 5780, 2021.
- Mahboobeh Jafari, Afshin Shoeibi, Marjane Khodatars, Sara Bagherzadeh, Ahmad Shalbaf, David López García, Juan M Gorriz, and U Rajendra Acharya. Emotion recognition in eeg signals using deep learning methods: A review. *Computers in Biology and Medicine*, page 107450, 2023.
- Hamdi Altaheri, Ghulam Muhammad, Mansour Alsulaiman, Syed Umar Amin, Ghadir Ali Altuwaijri, Wadood Abdul, Mohamed A Bencherif, and Mohammed Faisal. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. *Neural Computing and Applications*, 35(20):14681–14722, 2023.
- Tellakula Ramya Sri, Jahnavi Madala, Sai Lokesh Duddukuru, Rupasri Reddipalli, Phani Kumar Polasi, et al. A systematic review on deep learning models for sleep stage classification. In 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), pages 1505–1511. IEEE, 2022.
- Xun Chen, Chang Li, Aiping Liu, Martin J McKeown, Ruobing Qian, and Z Jane Wang. Toward open-world electroencephalogram decoding via deep learning: A comprehensive survey. *IEEE Signal Processing Magazine*, 39 (2):117–134, 2022.
- Siavash Sakhavi, Cuntai Guan, and Shuicheng Yan. Learning temporal information for brain-computer interface using convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 29(11):5619–5629, 2018.
- Punnawish Thuwajit, Phurin Rangpong, Phattarapong Sawangjai, Phairot Autthasan, Rattanaphon Chaisaen, Nannapas Banluesombatkul, Puttaranun Boonchit, Nattasate Tatsaringkansakul, Thapanun Sudhawiyangkul, and Theerawit Wilaiprasitporn. Eegwavenet: Multiscale cnn-based spatiotemporal feature extraction for eeg seizure detection. *IEEE transactions on industrial informatics*, 18(8):5547–5557, 2021.
- Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.
- Yang Dai, Xiuli Li, Shanshan Liang, Lukang Wang, Qingtian Duan, Hui Yang, Chunqing Zhang, Xiaowei Chen, Longhui Li, Xingyi Li, et al. Multichannelsleepnet: A transformer-based model for automatic sleep stage classification with psg. *IEEE Journal of Biomedical and Health Informatics*, 2023.
- Jin Xie, Jie Zhang, Jiayao Sun, Zheng Ma, Liuni Qin, Guanglin Li, Huihui Zhou, and Yang Zhan. A transformer-based approach combining deep learning network and spatial-temporal information for raw eeg classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:2126–2136, 2022.
- Zhe Wang, Yongxiong Wang, Chuanfei Hu, Zhong Yin, and Yu Song. Transformers for eeg-based emotion recognition: A hierarchical spatial information learning model. *IEEE Sensors Journal*, 22(5):4359–4368, 2022.
- Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint* arXiv:2312.00752, 2023.
- Alison O'Shea, Gordon Lightbody, Geraldine Boylan, and Andriy Temko. Neonatal seizure detection from raw multi-channel eeg using a fully convolutional architecture. *Neural Networks*, 123:12–25, 2020.
- Huy Phan, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467, 2022.
- Mona Algarni, Faisal Saeed, Tawfik Al-Hadhrami, Fahad Ghabban, and Mohammed Al-Sarem. Deep learning-based approach for emotion recognition using electroencephalography (eeg) signals using bi-directional long short-term memory (bi-lstm). *Sensors*, 22(8):2976, 2022.
- Phairot Autthasan, Rattanaphon Chaisaen, Thapanun Sudhawiyangkul, Phurin Rangpong, Suktipol Kiatthaveephong, Nat Dilokthanakul, Gun Bhakdisongkhram, Huy Phan, Cuntai Guan, and Theerawit Wilaiprasitporn. Min2net: End-to-end multi-task learning for subject-independent motor imagery eeg classification. *IEEE Transactions on Biomedical Engineering*, 69(6):2105–2118, 2021.
- Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

- Hangyu Zhu, Laishuan Wang, Ning Shen, Yonglin Wu, Shu Feng, Yan Xu, Chen Chen, and Wei Chen. Ms-hnn: Multi-scale hierarchical neural network with squeeze and excitation block for neonatal sleep staging using a single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:2195–2204, 2023a.
- Hangyu Zhu, Wei Zhou, Cong Fu, Yonglin Wu, Ning Shen, Feng Shu, Huan Yu, Wei Chen, and Chen Chen. Masksleepnet: A cross-modality adaptation neural network for heterogeneous signals processing in sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 27(5):2353–2364, 2023b.
- Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2016.
- Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*, 2024.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* preprint arXiv:2010.11929, 2020.
- William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.
- Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*, 2023.
- Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.
- Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.
- Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.
- Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.
- Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.
- Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In 2013 6th international IEEE/EMBS conference on neural engineering (NER), pages 81–84. IEEE, 2013.
- Jun Ma, Banghua Yang, Wenzheng Qiu, Yunzhe Li, Shouwei Gao, and Xinxing Xia. A large eeg dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, 9(1):531, 2022.
- Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, 16:1–6, 2008.
- Pedro Fonseca, Niek Den Teuling, Xi Long, and Ronald M Aarts. Cardiorespiratory sleep stage detection using conditional random fields. *IEEE journal of biomedical and health informatics*, 21(4):956–966, 2016.
- Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021.
- Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.
- Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11): 1998–2008, 2017.

- Zhenghua Chen, Min Wu, Wei Cui, Chengyu Liu, and Xiaoli Li. An attention based cnn-lstm approach for sleep-wake detection with heterogeneous sensors. *IEEE Journal of Biomedical and Health Informatics*, 25(9):3270–3277, 2020.
- Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv* preprint arXiv:2111.00396, 2021.
- Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. arXiv preprint arXiv:2208.04933, 2022.
- Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.
- Suparerk Janjarasjitt. Epileptic seizure classifications of single-channel scalp eeg data using wavelet-based features and svm. *Medical & biological engineering & computing*, 55(10):1743–1761, 2017.
- Edward A Wolpert. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. *Archives of General Psychiatry*, 20(2):246–247, 1969.
- Phan Huy, Fernando Andreotti, Navin Cooray, Oliver Y Chen, and Maarten De Vos. 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, 2019.
- Muhammad Khateeb, Syed Muhammad Anwar, and Majdi Alnowami. Multi-domain feature fusion for emotion classification using deap dataset. *IEEE Access*, 9:12134–12142, 2021.
- Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

# A Appendix

## A.1 Related Works

## A.1.1 EEG Classification

The development of deep learning has greatly advanced EEG classification tasks. CNNs are a classic type of neural network with mature applications in EEG classification. Schirrmeister et al. Schirrmeister et al. [2017]proposed a shallow convolutional network with both spatiotemporal convolutional layers to decode task-related information from raw EEG signals. Similarly, Lawhern et al. Lawhern et al. [2018] introduced EEGNet, a classic EEG classification network based on depthwise separable convolution, which has demonstrated stable and robust performance in various EEG classification tasks. Recurrent Neural Networks (RNNs) are proposed to capture temporal dependencies in time-series EEG signals. Supratak et al. Supratak et al. [2017] used the RNN architecture for sleep stage classification. Chen et al. Chen et al. [2020] used CNN and Long Short Term Memory (LSTM) networks for sleep stage classification. EEG classification networks based on Transformers have also made significant progress. Eldele et al. [2021] introduced attention mechanisms into EEG classification networks for classifying sleep stages. Song et al. Song et al. [2022] proposed EEG Conformer, a EEG classification network based on spatiotemporal convolution and Transformers. EEG Conformer effectively extracts local and global features from EEG signals, and it performs well in tasks such as motor imagery and emotion recognition.

## A.1.2 State Space Model

A state-space model is a mathematical model that represents a physical system as a set of input, output, and state variables related by a first-order differential equation. Gu et al. Gu et al. [2021] proposed the Structured State-Space Sequence Model (S4) to model long-term dependencies. Smith et al. Smith et al. [2022] introduced a new S5 layer by incorporating Multiple Input Multiple Output (MIMO) SSM and efficient parallel scanning within the S4 layer. Fu Fu et al. [2022] designed a new SSM layer, H3, which further narrowed the performance gap between SSM and Transformers. Recently, Gu et al. Gu and Dao [2023] proposed a data-dependent SSM structure and built a universal language model backbone network: Mamba. Its selective mechanism and hardware-aware design allow it to maintain computational efficiency and excellent performance while scaling to billions of parameters.

# A.1.3 Mixture of Experts

The Mixture of Experts model was first introduced by Jacobs et al. [1991], which controls a system composed of different networks called experts through a supervisory program, with each expert responsible for handling a specific subset of training samples. Shazeer et al. [2016] introduced the concept of sparsity into MoE and applied it to LSTM models for translation tasks. With the development of large language models, Fedus et al. [2024] extensively investigated the stability issues of MoE models during training and fine-tuning processes, and built a MoE model with 16 trillion parameters and 2048 experts. Recently, Xue et al. Fedus et al. [2022] proposed OpenMOE, which further explores the details of MoE using the power of the open-source community, thereby promoting the development of MoE.

#### A.2 Dataset

## A.2.1 Bonn

The Bonn dataset is composed of EEG data from 5 healthy individuals and 5 patients with epilepsy, totaling 5 subsets, which are F, S, N, Z, and O. The Bonn dataset is a single-channel dataset, where each subset contains 100 data segments. The duration of each data segment is 23.6 seconds with a sampling frequency of 173.61Hz. During the data cutting process, noise signals such as myogenic and ocular artifacts have been removed. The data Z and O are scalp EEGs, collected from 5 healthy individuals, forming the control group. The segments in Z are EEGs when the subjects have their eyes open, and the segments in O are EEGs when the subjects have their eyes closed. The data N, F, and S are intracranial EEGs, collected from 5 patients who have been diagnosed preoperatively. N and F are collected during the interictal phase of epilepsy, and S is collected during the ictal phase. To facilitate model processing, we have truncated the length of this dataset to 4,096 signal points.

## A.2.2 CHB-MIT

The CHB-MIT scalp EEG database is collected by the Children's Hospital Boston, which contains 24 cases of 23 patients with intractable seizures. The first 23 cases are from 22 patients (17 females, aged 1.5-19 years; 5 males, aged 3-22 years). For the last case, there is no clear gender or age record. the Children's Hospital Boston evaluated the potential conditions for surgical intervention in all epilepsy patients after discontinuing medication for a period of time, and monitored the patients for several days. The original EEG record was obtained using 256 Hz sampling rate with 16-bit resolution from electrodes placed according to the international 10-20 EEG electrode positions and nomenclatureJanjarasjitt [2017]. Given that the number of available channels varies among different patients, we select 23 common channels and discarded data from less than 23 channels. Due to the varying duration of the original data ranging from tens of minutes to several hours, we have truncated it into 4-second segments for easy classification.

## A.2.3 SleepEDF-20

SleepEDF-20 includes Polysomnography (PSG) records from each subject for two consecutive days and nights. The recording of subject 13 on the second night was lost due to a failing cassette or laserdisc. Sleep experts use R&K rules Wolpert [1969] to visually determine signal characteristics and label each 30 second period in the dataset as one of eight stages W, N1, N2, N3, N4, REM, MOVEMENT, UNKNOWN. Similar to previous work Huy et al. [2019], N3 and N4 were merged into N3. In addition, the stages of "MOVEMENT" and "UNKNOWN" have also been removed. In our experiment, Fpz-Cz EEG with a sampling rate of 100Hz was adopted for sleep staging.

## A.2.4 SHHS

Sleep Heart Health Study (SHHS) is a multi-center cohort study on the cardiovascular and other consequences associated with sleep apnea. The research subjects suffer from various diseases, including lung disease, cardiovascular disease, and coronary heart disease. To reduce the impact of these diseases, we referred to the research method of Fonseca et al. [2016] and selected subjects who were considered to have regular sleep patterns (such as those with apnea hypopnea index (AHI) less than 5). Finally, we select 329 for the experiment out of 6441 participants. It is worth noting that we chose the C4-A1 channel with a sampling rate of 125 Hz.

## A.2.5 DEAP

In the DEAP dataset, movies are used as emotional inducers in experiments. This dataset contains data from over 32 participants aged between 19 and 37, half of whom are females. Participants sit one meter away from the screen. The device records EEG signals at a sampling rate of 512Hz. 40 selected music video clips were used to trigger emotions. At the end of each video, participants were asked to evaluate their level of arousal, valence, preference, and dominance. The self-assessment scale ranges from 1 to 9. The scores of the subjects are divided into two categories (low or high) based on a stable threshold of 4.5. During the preprocessing process, the EEG signal is downsampled to 128Hz and a bandpass filter with a cutoff frequency of 4-45Hz is applied. In this paper, we use the same channel selection as Khateeb et al. [2021], which includes four electrodes: FP1, FP2, F3, and C4.

## **A.2.6 SEED**

The SEED dataset collects EEG data from 15 participants while watching emotional movies. It contains a total of 45 experiments. The EEG data is collected by 62 channels based on the international 10-20 system and a sampling rate of 1000Hz. During the preprocessing process, the data is downsampled to 200Hz and subjected to a bandpass filter ranging from 0 to 75Hz. The extraction of EEG sections was based on the duration of each movie. Within each subject's data

file, there are 16 arrays, with 15 of these arrays containing 15 preprocessed segments of EEG data from the experiment. The label array includes corresponding emotional labels, where 1 for positive, 2 for negative, and 3 for neutral emotions.

## A.2.7 Shu

The motor imagery dataset experiment consists of three phases. The first phase (0-2 seconds) is the resting preparation period, during which subjects can rest, perform minor physical activities, and blink. The second phase (2-4 seconds) is the cue phase, where an animation of left or right hand movement appears on the monitor, indicating the upcoming task. The third phase (4-8 seconds) is the MI (Motor Imagery) phase, during which subjects perform the hand movement MI task as prompted, and EEG signals are recorded. Each session consists of 100 trials, with five sessions conducted for each subject every 2 to 3 days, resulting in a total of 500 trials per subject.

## A.2.8 BCI-IV-2a

The BCI-IV-2a dataset includes EEG signals obtained from trials involving 9 subjects. This experiment includes four different motor imagery tasks: left hand, right hand, foot, and tongue. Each participant participated in two training sessions, with six sessions per session. In each run, there were 48 trials, a total of 288 trials (12 trials per MI task, a total of 72 trials per task). A set of 25 Ag/AgCl electrodes were used in the experiment, of which 22 were dedicated to recording EEG signals, while the remaining three electrodes recorded eye movement signals (not used in our experiment). All recorded signals are processed through a bandpass filter of 0.5 to 100Hz and a 50Hz notch filter. The sampling frequency is set to 250Hz.

## A.3 Experimental Related Supplements

## A.3.1 Load Balance and Model stability in MoE

Training an MoE typically encounters two issues: (1) Load imbalance: the gating network tends to select only a few experts. (2) Training instability: excessively large gating values for a few experts lead to an unstable training process. To address these issues, we incorporate balance loss  $L_b$  Shazeer et al. [2016] and router z-loss  $L_z$  Zoph et al. [2022] as auxiliary losses for the model to mitigate load imbalance and training instability, as shown in Equation (11), where B represents the batch size.

$$L_{b} = \frac{Std(e(t_{cls}^{*}))}{Mean(e(t_{cls}^{*}))}$$

$$L_{z} = \frac{1}{B} \sum_{i=1}^{B} (log(exp(t_{cls}^{*})))^{2}$$

$$L_{aux} = L_{b} + L_{z}$$
(11)

## A.3.2 Visualization of Features Extracted by Bidirectional Mamba

Figure 5 shows t-distributed stochastic neighbor embedding (t-SNE) plots of features extracted by Single-task EEG-Mamba from different datasets. The plot exhibits distinct distances between features of different classes and small distances within the same class, indicating the successful extraction of features from different classes by EEGMamba. This may indicate its comprehensive performance superiority across different datasets.

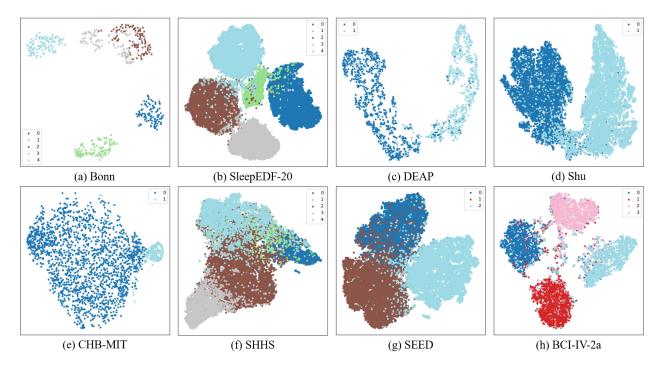


Figure 5: Visualization results of feature extracted by EEGMamba on different datasets.

## A.3.3 Bidirectional Mamba Ablation Study

We analyze the role of each module in EEGMamba through a series of ablation analysis, as shown in Table 5. When using Mamba, the accuracy decreases on datasets except CHB-MIT and SleepEDF-20. This indicates that bidirectional modeling can better capture the dependency relationship between EEG sequences. We also attempt to combine single-directional causal convolution with bidirectional SSM, and the results are similar to single-directional modeling. Overall, the combination of bidirectional SSM and bidirectional causal convolution leads to better performance.

Table 5: Module ablation study in bidirectional Mamba. I represents the combination of single-directional causal convolution and single-directional SSM. II represents the combination of single-directional causal convolution and bidirectional SSM. III represents the combination of bidirectional causal convolution and bidirectional SSM used by EEGMamba.

| m 1                         |                     | I                        |                         |                         |                         | II               |                         | III                     |                         |                         |  |
|-----------------------------|---------------------|--------------------------|-------------------------|-------------------------|-------------------------|------------------|-------------------------|-------------------------|-------------------------|-------------------------|--|
| Task                        | Dataset             | ACC                      | AUC                     | F1                      | ACC                     | AUC              | F1                      | ACC                     | AUC                     | F1                      |  |
| Epilepsy Detection          | Bonn<br>CHB-MIT     | 93.00%<br><b>99.70</b> % | 0.9914<br><b>0.9987</b> | 0.9345<br><b>0.9748</b> | 95.00%<br>99.68%        | 0.9923<br>0.9985 | 0.9564<br>0.9728        | <b>96.00%</b> 99.69%    | <b>0.9931</b> 0.9986    | <b>0.9644</b> 0.9734    |  |
| Sleep Stages Classification | SleepEDF-20<br>SHHS | <b>87.07%</b> 86.96%     | <b>0.9753</b> 0.9696    | <b>0.8141</b> 0.7698    | 86.56%<br>86.85%        | 0.9742<br>0.9700 | 0.8060<br>0.7720        | 86.89%<br><b>87.20%</b> | 0.9742<br><b>0.9714</b> | 0.8117<br><b>0.7753</b> |  |
| Emotion Recognition         | DEAP<br>SEED        | 66.02%<br>70.42%         | 0.6751<br>0.8728        | 0.6448<br>0.7026        | 66.41%<br>70.57%        | 0.6771<br>0.8729 | 0.6426<br>0.7065        | 66.80%<br>71.04%        | 0.6897<br>0.8788        | 0.6529<br>0.7108        |  |
| Motor Imagery               | Shu<br>BCI-IV-2a    | 67.14%<br>56.79%         | 0.7295<br>0.8106        | 0.6714<br>0.5678        | 66.85%<br><b>58.27%</b> | 0.7339<br>0.8155 | 0.6682<br><b>0.5827</b> | 67.51%<br>58.27%        | 0.7310<br>0.8199        | <b>0.6749</b> 0.5802    |  |

# A.3.4 The Original Results of Single-task Mamba Comparison Experiment

Table 6 records the original results of the Single-task EEGMamba comparison experiment, and its visualization results are shown in Figure 2.

Table 6: The original results of Single-task Mamba comparison experiment.

|                             |                    | $\mathcal{C}$           |                         |                         | 0                |                  |                  | 1                |                  |                  |                      |                      |                      |
|-----------------------------|--------------------|-------------------------|-------------------------|-------------------------|------------------|------------------|------------------|------------------|------------------|------------------|----------------------|----------------------|----------------------|
|                             |                    | EEGNet                  |                         |                         | AttnSleep        |                  |                  | EEG Conformer    |                  |                  | EEGMamba             |                      |                      |
| Task                        | Dataset            | ACC                     | AUC                     | F1                      | ACC              | AUC              | F1               | ACC              | AUC              | F1               | ACC                  | AUC                  | F1                   |
| Epilepsy Detection          | Bonn<br>CHB-MIT    | 0.8700<br>0.9927        | 0.9662<br>0.9909        | 0.8729<br>0.9373        | 0.9300<br>0.9947 | 0.9896<br>0.9931 | 0.9301<br>0.9554 | 0.9200<br>0.9936 | 0.9913<br>0.9903 | 0.9205<br>0.9451 | 0.9600<br>0.9969     | 0.9931<br>0.9986     | 0.9644<br>0.9734     |
| Sleep Stages Classification | SleepEDF20<br>SHHS | 0.8269<br>0.8310        | 0.9545<br>0.9474        | 0.7518<br>0.7130        | 0.8402<br>0.8320 | 0.9585<br>0.9510 | 0.7680<br>0.7201 | 0.8331<br>0.8247 | 0.9561<br>0.9454 | 0.7575<br>0.6907 | 0.8689<br>0.8720     | 0.9742<br>0.9714     | 0.8117<br>0.7753     |
| Emotion Recognition         | DEAP<br>SEED       | 0.6172<br><b>0.8241</b> | 0.5898<br><b>0.9444</b> | 0.5483<br><b>0.8241</b> | 0.6055<br>0.6562 | 0.5757<br>0.8429 | 0.5825<br>0.6584 | 0.6094<br>0.6469 | 0.5533<br>0.8245 | 0.5455<br>0.6479 | <b>0.6680</b> 0.7104 | <b>0.6897</b> 0.8788 | <b>0.6529</b> 0.7108 |
| Motor Imagery               | Shu<br>BCI-IV-2a   | 0.6301<br>0.5758        | 0.6725<br>0.8026        | 0.6301<br>0.5751        | 0.6505<br>0.4695 | 0.7008<br>0.7283 | 0.6505<br>0.4698 | 0.6510<br>0.5453 | 0.7067<br>0.7855 | 0.6509<br>0.5452 | 0.6751<br>0.5827     | 0.7310<br>0.8199     | 0.6749<br>0.5802     |

## A.4 Limitations

Although the current experimental results show that EEGMamba can be well applied to EEG multi-task classification, it still has some limitations. On the one hand, this paper only covers four kinds of EEG tasks to verify the performance of EEGMamba, which is only a small part of the tasks that EEG can accomplish. Therefore, EEGMamba is still far from a universal EEG classification model. On the other hand, it should be extended to other one-dimensional time signals besides EEG to prove the universality of the model in one-dimensional time signals.