A Frequency-Selective Multi-Window Attention Model for EEG Motor Imagery Classification

Yixin Wang, Tianwei Shi*, Xintong Ye

Abstract—**This paper proposes a hybrid model that integrates Convolutional Neural Networks (CNN) and Transformer architectures, leveraging both temporal and time-frequency features to enhance EEG signal classification performance in motor imagery tasks. The model innovatively fuses raw EEG signals with time-frequency features extracted through Continuous Wavelet Transform (CWT) via a weighted fusion strategy, thereby effectively capturing dynamic variations in both the temporal and frequency domains. Subsequently, the fused features are processed by a Transformer framework that employs a windowed multi-head attention mechanism with frequency band – adaptive window sizes, which significantly improves frequency-domain modeling. Finally, classification is achieved using a Gated Feed-Forward Network (GFFN), which adaptively integrates multi-layer features via gating mechanisms, thereby enhancing feature selection and representation.**

Experimental results indicate that on the BCI 2a competition dataset, FAMW achieves an average classification accuracy improvement of 11.41%, 10.32%, 9.19%, and 6.44% over CNN-ELM, EEGNet, Deep ConvNet, and Conformer, respectively. FAMW shows improvements of 10.69%, 10.26%, 9.49%, and 5.66% on self-collected datasets compared to corresponding methods. These findings validate that the proposed model significantly enhances classification accuracy while demonstrating good generalization ability and robustness.

Index Terms—**EEG Signal Classification， Motor Imagery， Frequency-Selective Multi-Window Attention， Transformer**

I. INTRODUCTION

Brain-computer Interface (BCI) technology can directly decode Brain activity signals [1] to realize the interaction between people and devices, and has shown wide application prospects in the fields of medical rehabilitation, neuroscience research and auxiliary device control. In recent years, a variety of BCI paradigms based on electroencephalography (EEG) have been developed. Steady-State Visual Evoked Potential (SSVEP) [2], Event-Related Potential (event-related Potential, ERP) [3], emotional decoding [4], and Motor Imagery (MI) [5]-[9]. Among them, SSVEP generates periodic EEG signals through visual stimulation, which has the advantages of a strong signal and high transmission rate, but depending on visual stimulation, it may cause visual fatigue and is not suitable for visually impaired users. Based on event-related potentials induced by specific stimuli, ERP is easy to use because of the apparent signal characteristics and the need for complex training, Still, the transmission rate is low, the real-time is insufficient, and it is easy to be disturbed by noise. In contrast, MI-based BCI does not require external stimulation, can significantly reduce user fatigue, is suitable for long-term use, and has outstanding advantages in the field of medical rehabilitation and smart device control. MI provides a new means for stroke patients to reconstruct motor function in combination with rehabilitation robots or electrical stimulation devices by decoding the user's motor intention [10]. Provide innovative treatment for patients with movement disorders such as Parkinson's disease, and improve movement ability through EEG-driven devices [11]; For the control of smart wheelchairs and prosthetics to enhance the quality of life and independence of people with paralysis or amputations [12]; In addition, it has demonstrated the potential to enhance the naturalness of human-computer interaction in screen cursor control and virtual reality applications [13]. There are two main approaches for MI-based BCI decoding: traditional machine learning and deep learning.

Traditional machine learning methods usually involve two different steps of: feature extraction and feature classification. Conventional methods typically rely on hand-designed features or shallow machine-learning techniques. Such as common space mode (CSP) [14], continuous wavelet transform(CWT) [15], wavelet transform (WT) [16], and short-time Fourier transform (STFT) [17]. These methods use different mathematical techniques to extract features from EEG signals for classification. These methods can effectively describe some key aspects of EEG signals through the extracted features and realize the decoding of motor imagination tasks to a certain extent. Traditional classification methods usually include random forest (RF) [18], support vector machine (SVM) [19], linear discriminant Analysis (LDA) [20], etc. Some methods rely on manually selected features for classification, often requiring domain experts to select the appropriate features based on experience. This feature selection method is effective to a certain extent, but in high-dimensional and complex data, manually designed features often cannot fully cover the diversity of EEG signals, resulting in some

Manuscript received March 21, 2025; revised May 13, 2025.

This work was supported by the Liaoning Science and Technology Department's 'Jiebang, Guashuai' (Unveiling the List and Appointing the Leader) Project (2024-241).

This work was supported by the Liaoning Provincial Department of Education University Basic Scientific Research Business Fee Project (LJ242410146070).

Yixin Wang is a graduate student at the School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China (phone: +86-185-2433-5776; e-mail: 2177746309@qq.com).

Tianwei Shi* is a Professor at the School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China (Corresponding author to provide phone: +86-139-9805-3962; e-mail: 1552872449@qq.com).

Xintong Ye is a graduate student at the School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China (phone: +86-176-4124-2463; e-mail: 1925575202@qq.com).

potentially valuable signal information can not be effectively captured, thus affecting the classification performance. In addition, traditional classification methods are difficult to adapt to the dynamic changes and nonlinear characteristics of EEG signals, which limits their application in more complex tasks.

In recent years, the rise of deep learning technology has provided a new solution for EEG signal processing. Convolutional neural networks (CNNS) have become an essential deep learning structure in brain-computer interfaces (BCI) based on motor imagination (MI) because of their powerful representation learning ability. Several studies have investigated the effect of different configurations of CNN parameters on MI classification performance, such as convolution mode, kernel size, number of cores, and network depth. For example, Schmidmeister et al. proposed two different CNN-based architectures, Shallow ConvNet and Deep ConvNet, for end-to-end classification of MI tasks, and found that the depth of CNN significantly impacts its performance [21]. Lawhern et al. introduced separable convolution operations into CNN and developed EEGNet, a general BCI classification framework successfully applied to classify multiple tasks [22]. Hermosilla et al. tried shallow CNN and adjusted the number and size of cores further to improve the MI classification performance [23]. However, CNNS have some limitations, especially when it comes to capturing long-range dependencies and global information. Due to the limitations of convolutional kernels, CNNS are often unable to effectively model long time series relationships in signals, which is a challenge for processing complex timing and cross-band dynamic changes in EEG signals. To address these issues, Transformer architecture has been introduced to process and classify EEG signals. With its self-attention mechanism, Transformer is able to capture long distance dependencies rather than being limited to local receptive fields, so it can better handle long time series and global information in EEG signals. Overcome the shortcomings of CNN in long time series data. Combined with the advantages of CNN and Transformer, time rate features can be extracted from EEG signals more comprehensively, thus improving classification performance and accuracy. However, when processing EEG signals, the model often difficult to fully capture the complex changes of time-frequency characteristics and time characteristics at the same time and may ignore the critical role of frequency band information, which limits its performance in complex tasks.

To address these issues, we proposed a CNN-Transformer model that integrates time-frequency features to enhance EEG classification performance in motor imagery tasks. First, the model combines time-frequency features extracted from Continuous Wavelet Transform (CWT) with original EEG signals. The weighted fusion strategy is used to make full use of the complementary information of the two feature domains, to enhance the ability of the model to represent multi-dimensional features. Second, in order to map the time-frequency features and original EEG signals to the feature space more suitable for Transformer model processing, the representation capability of time series data is enhanced by Embedding. Then, in order to capture multi-scale time features more efficiently, the Windowed Multi-Head Attention mechanism is adopted. It can dynamically adjust the attention window size on the time frequency segment to accurately model long-range dependencies and local features in time series data. Next, GFFN is used to adaptively weight the features of different network layers, effectively reduce the interference of redundant information, and further improve the feature extraction effect. Finally, the fused features pass through the fully connected layer, and the Softmax classifier calculates each category's probability distribution to complete the final classification of the motor imagery task.

II. Method

This paper proposes a Cnn-Transformer model with time-time-frequency features is proposed to significantly improve the classification performance of EEG signals in motor imagination tasks. Firstly, the original EEG signal is weighted with the time-frequency features extracted by continuous wavelet transform (CWT) to make full use of the complementary information of the two feature domains, thereby enhancing the model's ability to represent multi-dimensional features. Then, in order to map the fused

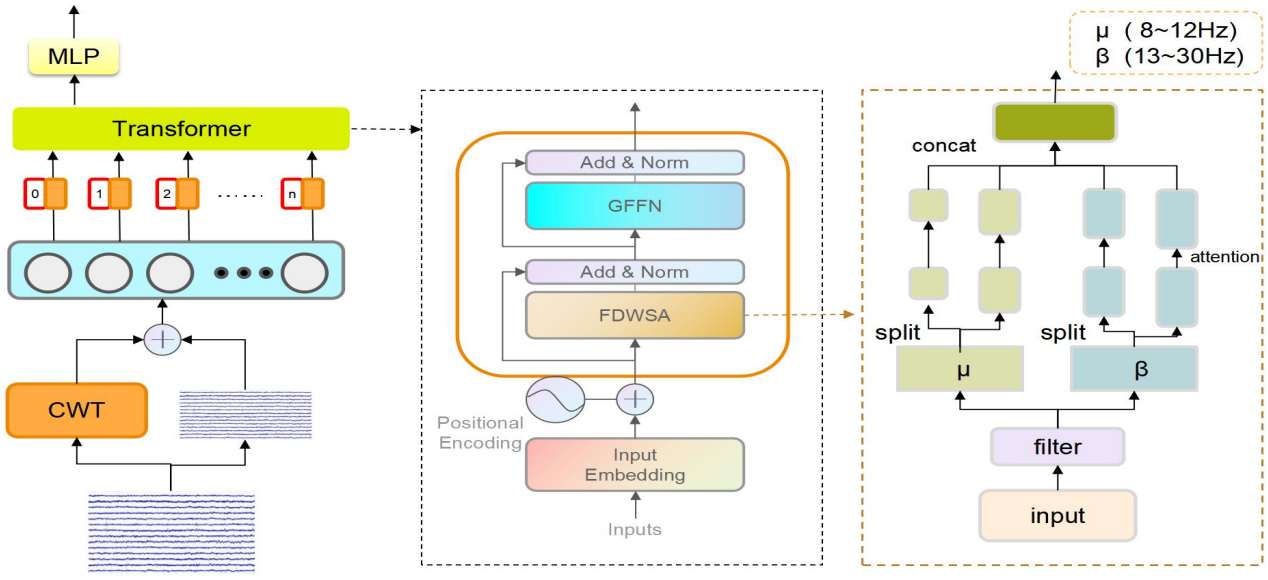

Fig. 1. Figure 1 Model network structure diagram

features to the feature space suitable for Transformer processing, the method of Embedding is adopted, thus enhancing the representation capability of time series data. Subsequently, the model captures multi-scale temporal features efficiently by Windowed Multi-Head Attention. This mechanism can dynamically adjust the attention window size over time frequency segments to accurately model long-range dependencies and local features in time series data. In order to further improve the feature extraction effect, a gated feed-forward network (GFFN) is used to adaptively weight the features of different network layers and reduce the interference of redundant information. Finally, the fused features are input to Softmax classifier through the fully connected layer, and the probability distribution of each category is calculated to complete the final classification of the motor imagery task. The complete network structure is shown in Fig. 1.

A. Data Acquisition

BCI Competition IV 2a [24] : 9 subjects completed four different motor imagination tasks (left hand (Class 1), right hand (class 2), foot (class 3) and tongue (class 4)). The experiment was divided into 6 rounds, each consisting of 48 tests. Participants sat relaxed in a comfortable armchair and looked at a 21-inch LCD monitor during the experiment. Each test starts with a fixed "+" symbol appearing in the center of the display, prompting the sound (t = 0 s). After two seconds (t = 2 s), the display prompts left, right, down, and up symbols corresponding to the motor imagination task of the left hand, right hand, feet, and tongue, respectively (lasting about 1.25 seconds). During this time, participants completed the motor imagination process until the "+" symbol on the screen disappeared (t = 6 s). After the test is over, the monitor briefly turns black and goes to rest until the next test begins.

Ustl MI Dateset: The Ustl MI dataset consisted of 12 healthy subjects (age: 21.4±3.3 years, numbers 1-12) who completed four motor imagination (MI) tasks in the left hand (class 1), right hand (class 2), foot (class 3), and tongue (class 4). All subjects had no history of neurological diseases and were not taking drugs during the study. The experiment was in accordance with the Declaration of Helsinki, and participants read and signed informed consent forms before participating.

A neuroscan (NuAmps) electrode cap was used in the experiment. The electrodes were arranged according to the international 10-20 system, and bilateral mastoids were used as reference electrodes. EEG signals were collected from multiple scalp locations (FP1, FP2, F7, F3, FZ, F4, F8, FT7, FC3, FCZ, FC4, FT8, T3, C3, CZ, C4, T4, TP7, CP3, CPZ, CP4, TP8, T5, P3, PZ, P4, T6, O1, OZ, and O2). It is stored with a 500 Hz sampling rate and 32-bit accuracy. During the experiment, the subjects were required to remain still and avoid any significant movements or sounds.

The experimental device is a DELL XPS 8940 microserver with an i7-11700 CPU, RTX 3060Ti graphics card, and 32 GB RAM. Each participant completed 10 sets of experiments on the same day, each set 5 minutes apart, and repeated each of the four tasks three times. The timing of the experiment is shown in Fig. 2., covering both single and continuous MI tasks.

B. Author ListData Preprocessing

First, the collected original MI task EEG signals are filtered by 50 Hz notch to remove power interference. Secondly, three types of electrode signals [' EOG-left ', 'EOG-central', 'EOG-right'] are selected as bad signal channels. Then, the ICA method is used to isolate and eliminate the eye artefact. Finally, for the data extracted in each training period (epoch), the z-score standard score is used to eliminate the differences between different measurement scales, enhance the comparability between data features, and make the data distribution more uniform, thus improving the numerical stability and efficiency in the algorithm training process. The formula for the z-score is as follows.

$$z = \frac{(x - \mu)}{\sigma} \tag{1}$$

Where x is the raw data, μ is the mean data, and σ is the standard deviation of the data.

C. Feature Extraction

The processed EEG signal is taken as input, and the EEG signal is $X = \{x_i\}_{i=1}^{N}$, where $x_i \in R^{(C \times T)}$, N represents the total number of training samples, C represents the number of EEG channels, and T represents the number of samples included in each trial.

Time-frequency feature extraction: For each channel x_i, continuous wavelet transform (CWT) is used to convert the time data into the corresponding spectrum [25]. The order represents the time domain EEG data of the first channel, referred to as $s(t)$. The continuous wavelet transform is carried out by the following equation:

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int s(t) \psi \left(\frac{t-b}{a} \right) dt \tag{2}$$

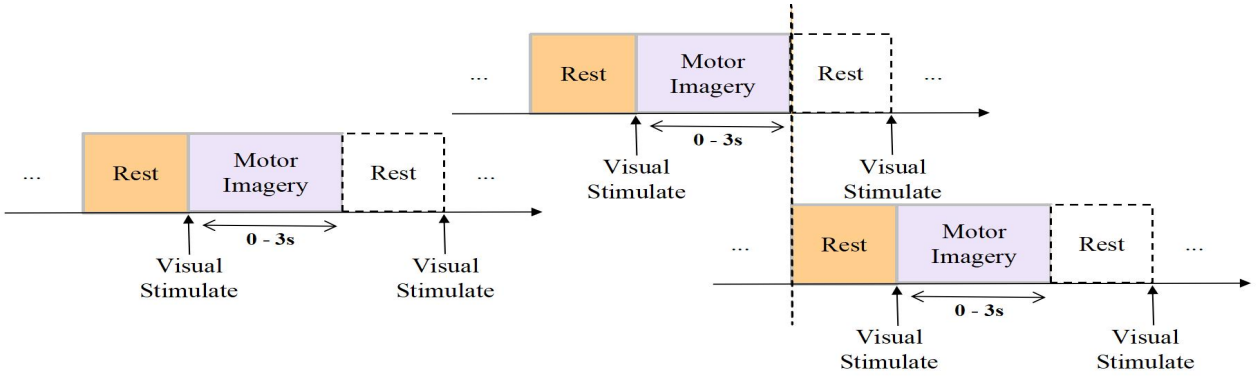

Fig. 2. Signal acquisition timing for single and continuous MI tasks

Where, ψ represents the wavelet generating function, ψ is complex conjugate, a and b represents the extended and shifted variables. In this study, Morlet wavelet [25] is selected as the wavelet generating function, which is defined as:

$$\psi(t) = exp(-(\frac{\beta^2 t^2}{2}))cos(\pi t) \quad (3)$$

Where, β is the parameter that balances the time resolution and frequency resolution of the Morlet wavelet.

The output of the CWT is a complex-valued matrix, where each element represents the wavelet coefficient at a specific time and frequency. The amplitude of the wavelet coefficient indicates the intensity of EEG activity at that time and frequency. The EEG signal is bandpass filtered to remove the interference of irrelevant frequency bands, and then the frequency range of the spectral graph is aligned with the filtered EEG signal passband to reduce the influence of noise on the spectral graph. In order to generate the time-frequency graph, firstly, the amplitude of the wavelet coefficients is squared. It is then plotted as a function of time and frequency. The CWT of each channel's EEG signal produces an 800×600 time-spectral image. In this model, the spectrum images corresponding to each channel are spliced together in the width direction, and the spliced images are downsampled to 224×224.

Feature Fusion: In the model, the input of the fully connected layer is the weighted average of the time frequency S and the time network output feature F . If η, $0 \le \eta \le 1$, represents the feature weight, the weighted fusion of time-frequency and time network output features can be expressed as:

$$G = \eta \cdot S + (1-\eta) \cdot F \quad (4)$$

By Concatenation, the time-domain feature and time-frequency feature are fused. This method can retain the information of different feature sources and learn the appropriate combination pattern through the neural network. The fused feature is a high-dimensional eigenvector of EEG signal that integrates time-frequency domain and time-frequency domain.

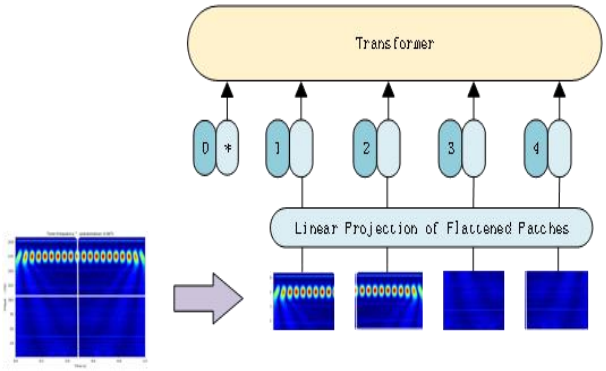

Fig. 3. Structure diagram of feature embedding module

Feature Embedding Module: Feature embedding module is the starting point of the whole Transformer model. Its primary function is to map the input multimodal data, which integrates the time-domain features of the original EEG signal and the time-frequency features of continuous wavelet transform (CWT), to the feature space suitable for Transformer processing. The module mainly includes Patch Split, Convolutional Embedding and Patch Embedding. The structure flow is shown in Fig. 3.

Patch Split: Since EEG signals are high-dimensional time series data, direct input into the Transformer will result in high computational complexity. Therefore, the Patch segmentation strategy was adopted to divide the time-time-frequency features into patches (fragments) of equal length according to the time axis, and set patch_size as the size of the sliding window. Each Patch contains multiple EEG sampling points and corresponding CWT spectrum information, which can capture local time characteristics and frequency information.

Convolutional Embedding: After Patch partitioning, local features are extracted using one-dimensional convolution (Conv1D) and projected into a fixed dimensional embedding space (embedding_dim). To fit the Transformer structure:

$$E_i = f\left(W \cdot P_i + b\right) \quad (5)$$

Where, P_i is the i th Patch, W and b are the weights and biases of the convolution kernel, and $f(\cdot)$ represents the nonlinear activation function.

Patch Embedding: All Patch embedding vectors are combined to form the final embedding representation. Assuming that each Patch is embedding_dim, the final output is (batch_size, patch_num, embedding_dim). Where, patch_num is the number of patches divided, and embedding_dim is the dimension of the embedding vector.

$$E_i = \left[E_1, E_2, \cdots, E_N\right] \quad (6)$$

Where, E_i is the embedding matrix of the output, and N is the total number of patches.

The feature embedding module uses CNN technology to convert the high-dimensional long sequence signal into the input sequence of fixed format through segmentation and embedding. This module uses local convolution operations to extract local time features, which lays a foundation for subsequent windowing multi-head attention mechanisms and classification tasks.

D. Feature Classification

This study is based on transformer architecture, which classifies EEG signals of motor imagery tasks by combining time domain, frequency domain and time-frequency characteristics. Transformer effectively captures temporal dependencies in the input sequence through its Multi-Head Self-Attention mechanism, which is particularly important for dynamic brain electrical activity in EEG signals. Each attentional head focuses on a different part of the input sequence, and by weighting and splicing the output of multiple heads, the model can synthesize different timing information to improve classification. In this study, a dynamic window size strategy is introduced to adjust the window size dynamically according to the characteristic changes of the signal. This method enables the model to

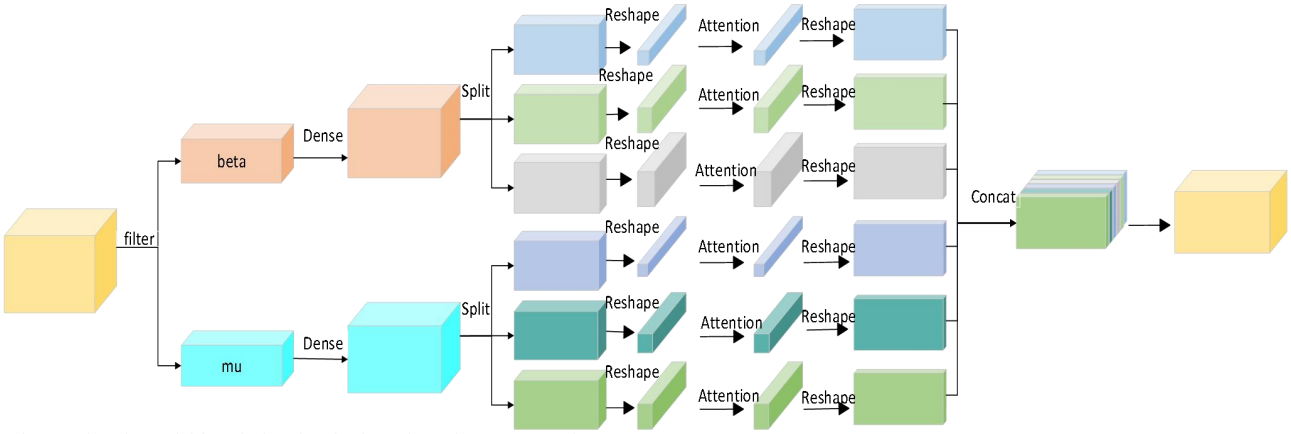

Fig. 4. Signal acquisition timing for single and continuous MI tasks

adjust the frequency range adaptively under different tasks so as to better capture the key information in EEG signals.

Frequency-Selection Different Window based MultiScale: Traditional Transformer's self-attention Traditional Transformer's self-attention mechanism faces the problem of quadratic increase in computational complexity when processing high-resolution inputs, especially for time-frequency graph data, where high resolution will significantly increase the computational burden. In order to effectively capture multi-scale information and reduce computing costs, this method innovatively applies Different Window-based multi-scale self-attention mechanism with different frequencies to extract local and global features, which is especially suitable for the task of frequency division processing. It combines multiple techniques such as multi-head attention, band selection, and windowing to capture timing and frequency domain features in EEG data. Figure 4 shows the multi-head attention structure of band selection windowing.

The input data is the time-frequency graph data which combines the time-frequency domain features of the original EEG signal with the time-frequency domain features transformed by continuous wavelet. In the process of data preprocessing, the frequency range of filter retention is [8 Hz,30 Hz]. It includes mu wave [8Hz,12Hz], beta wave [13Hz,30Hz]. It extracts a specific frequency band from the input tensor $X \in R^{B\times T\times F}$ through the frequency index. Where B is the batch size, T is the time step (sequence length), and F is the frequency dimension. According to the frequency dimension F, data $X_{mu} \in R^{B\times T\times F_{mu}}$ and $X_{beta} \in R^{B\times T\times F_{beta}}$ of the corresponding frequency band are extracted.

Window division: Input features X_{mu} and X_{beta} of each frequency band are divided into multiple Windows according to time dimension T. The window size can be set according to the characteristics of the frequency band: mu waves are suitable for capturing small Windows of fine-grained time-local features, and beta waves are suitable for capturing large Windows of features over a longer time range.

Assuming the window size is W, the number of Windows in the time dimension is:

$$N_w = \left\lfloor \frac{T}{W} \right\rfloor \tag{7}$$

The feature of window partition is $X_{window} \in R^{B\times N_W\times T\times F_{mu}}$

Apply a self-attention mechanism to each window to extract local features within the window:

$$Attention(Q,K,V) = Soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

Where Q, K, V are queries, keys, and values obtained by linear transformations. Is the attention scaling factor, equal to the dimension K. For different frequency bands, different attention heads and window sizes are defined independently, and local features are processed separately.

Cross-window feature aggregation: The features of each window are integrated through global pooling or other aggregation methods to restore the global feature dimension:

$$X_{pooled} = \frac{1}{W}\sum_{i-1}^{W} X_{window} \tag{9}$$

After cross-window aggregation, the resulting feature is $X_{aggregated} \in R^{B\times N_W\times F_{band}}$

Band fusion: The features of multiple frequency bands are spliced together to form the final global multi-scale feature representation:

$$X_{final} = Concat\left(X_{mu\_aggregate}, X_{beta\_agregate}\right) \tag{10}$$

Gated Feed-Forward Network (GFFN) : Gated Feed-Forward Network (GFFN) is a dynamic feature optimization module that selects key features through a gating mechanism and nonlinear mapping. It is an improved feed-forward network, which integrates the gating mechanism to control the flow of feature information dynamically. Unlike traditional feed-forward networks, GFFN can more flexibly select important features and suppress redundant information, thus improving the performance of the model, especially in complex time-frequency feature extraction and classification tasks. By weighting and fusing features, the GFFN module helps the model focus more effectively on key features and suppress noise and irrelevant information, especially for high-dimensional complex data such as EEG signals.

Nonlinear feature extraction: Input feature X is used to extract higher-order features through the first layer of a fully connected network W_1:

$$X_{dense} = RELU(W_1 \cdot X + b_1) \tag{11}$$

Gating mechanism: Generates gating weights

$$Gate = \sigma\left(W_g \cdot X_{dense} + b_g\right) \tag{12}$$

Where σ is the activation function , W_g and b_g are the gated weights and biases.

Feature fusion: Gated weights are used to dynamically weight feature output:

$$Y = Gate \odot X_{dense} + \left(1 - Gate\right) \odot \left(W_2 \cdot X + b_2\right) \tag{13}$$

Where $\odot$ represents multiplication-by-element W_g and b_g are the weights and biases of the second fully connected layer.

The processed 3D features are first converted to 2D feature representations. The aim is to compress the original time dimension and feature dimension into a flat vector representation that can be used as input to the fully connected layer. The flattened vector is mapped through the fully connected layer to a smaller hidden layer feature space.

In the hidden layer, ReLU activation function is used to enhance the learning ability of the model for nonlinear features, and the complex relationship between input features is mined. The output of the hidden layer is further mapped to the dimension space corresponding to the number of categories. With the Softmax activation function, the scores for each category are converted into a probability distribution. Where, each element represents the probability that the sample belongs to the corresponding category. Finally, the model outputs the prediction based on the maximum probability.

E. Training

In the training process, the Adam algorithm is used as the optimizer and cross-entropy loss function is used as the loss measure. The optimizer formula is as follows:

$$L = -\sum y \times \log\left(y\_pred\right) \tag{14}$$

Where L is the value of the loss function, y is the One-Hot encoding of the real label, and y_pred is the predicted output value.

III. Experiment and Results

A. Experimental Settings

The experimental environment, MI task EEG signal preprocessing and classification model were built using python, mne library and tensorflow library respectively, and were run on a Dell laptop with Intel i5-9300H CPU and 16 GB memory. The experimental data of a single subject was run for 200 epochs, the batch size was set to 32, and the learning rate was set to 0.005. The data were divided into the training set and the test set according to the ratio of 8:2, and the training set was divided by the cross-validation of 50%. The training set is used for model training and parameter adjustment. The test set is not involved in model training and is only used to evaluate model performance.

Select Accuracy as the evaluation parameter. Among them, TP represents the number of true positives, FP represents the number of false positives, TN represents the number of true negatives, and FN represents the number of false negatives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

B. Experimental Results

To verify the validity of the model, four advanced deep learning architectures were selected for comparative analysis, and each model was re-evaluated on two datasets.

EEGNet[25] designed a compact and efficient convolutional neural network (CNN) to classify EEG signals by introducing dependency and depthwise separable convolutions.

Deep ConvNet [26] combines temporal and spatial filtering layers, and enhances the model's ability to extract temporal features by introducing multiple convolution and maximum pooling blocks in the temporal filtering layer.

Cnn-elm [27] (Convolutional Neural Network - Extreme Learning Machine) combines convolutional neural network (CNN) and extreme learning machine (ELM). First, CNNS extract features from input data, such as images or EEG signals, and automatically learn spatial or temporal patterns through convolutional layers. Then, the features extracted by CNN are input into ELM for classification or regression.

Conformer [28] combines the convolutional neural network (CNN) and Transformer structure to improve sequence data's processing efficiency and performance. At its core, it combines convolution operations and self-attention mechanisms to full exploit both.

The results of the accuracy comparison between the FAMW algorithm and other comparison methods on the dataset of BCI Competition IV 2a are shown in Table Ⅰ. The highest and average classification accuracy reached 92.3% and 90.8%, respectively.

The results of accuracy comparison between FAMW algorithm and other comparison methods on self-collected data sets are shown in Table Ⅱ. The highest and average classification accuracy reached 91.62% and 90.94%, respectively.

It is recommended that footnotes be avoided (except for the unnumbered footnote with the receipt date on the first page). Instead, integrate the footnote information into the text and the reference part.

As can be seen from Table 1 and Table 2, the highest classification accuracy of the CNN-ELM algorithm is 81.52% and 82.48%, and the average classification accuracy is 79.40% and 80.25%, respectively, in the data set of BCI Competition IV 2a and self-collected data set. The highest classification accuracy of the EEG NET algorithm was 82.62% and 83.62%, and the average classification accuracy was 80.48% and 80.68%, respectively. The highest classification accuracy of the Deep ConvNet algorithm was 84.1% and 83.45%, and the average accuracy was 81.61% and 81.45%, respectively. The highest classification accuracy of the Conformer algorithm is 89.95% and 90.23%, and the average accuracy is 84.36% and 85.28%, respectively. The proposed FAMW algorithm outperforms the other four methods on the BCI Competition IV 2a dataset, with

maximum accuracy improvements of 10.78%, 9.68%, 8.2%, and 2.35%, respectively. The average classification accuracy was improved by 11.41%, 10.32%, 9.19% and 6.44%, respectively. The highest accuracy on self-collected data sets was increased by 9.69%, 8.55%, 8.72% and 1.94%, respectively, and the average classification accuracy was increased by 10.69%%, 10.26%, 9.49 and 5.66%, respectively.

As shown in Figures 5 and 6, a box plot shows the accuracy of different models in BCI Competition IV 2a and self-collected data sets. The accuracy distribution of CNN-ELM is relatively concentrated, and the overall accuracy is low. EEG NET and Deep ConvNet were slightly more accurate, showing a narrower distribution of nearly 82% and 84%, respectively. Conformer had a wider range of accuracy, but the median improved significantly to nearly 88%. The FAMW model performed best, with an accuracy of nearly 92% and a smaller box, indicating a more stable performance. Because FAMW multi-layer feature extraction enhances the ability to represent complex signals, windowing multi-point attention mechanism and gated feed-forward network (GFFN) effectively reduce the interference of redundant information, ensure the adaptability of the model to different time scales, and improve the model's capture of long-range dependence and local features. Enhanced robustness and stability in multiple tasks and data sets.

TABLE I
ACCURACY (%) COMPARISON RESULTS ON THE BCI COMPETITION IV 2A DATA SET

Subject	CNN-ELM	EEG NET	Deep ConvNet	Conformer	FAMW（Ours）
1	81.52	80.55	79.73	86.93	92.3
2	78.20	82.12	78.26	85.41	91.13
3	79.86	80.91	83.33	83.33	90.97
4	79.73	79.69	83.3	80.55	89.68
5	80.76	81.3	80.5	81.59	90.97
6	79.54	78.16	84.1	85.19	91.67
7	80.33	80.35	83.2	85.06	90.27
8	81.37	82.62	80.21	89.95	89.58
9	78.89	77.97	80.13	84.72	91.67
10	78.76	80.12	81.22	86.11	89.58
11	77.25	80.6	82.96	82.63	91.3
12	76.54	81.41	82.41	80.9	90.48
AVG	79.39	80.48	81.61	84.36	90.8

TABLE II
ACCURACY (%) COMPARISON RESULTS ON THE SELF-COLLECTED DATA SET

Subject	CNN ELM	EEG NET	Deep Conv Net	Conformer	FAMW（Ours）
1	80.45	80.12	83.12	85.4	91.25
2	78.24	82.36	80.19	90.23	89.94
3	78.46	80.91	79.89	85.71	91.62
4	81.23	80.55	82.67	84.02	91.62
5	80.58	80.23	83.43	82.80	91.62
6	79.63	79.95	80.31	85.41	92.17
7	81.56	80.35	78.96	83.33	90.62
8	79.54	83.62	81.08	84.72	90.50
9	78.85	77.97	81.98	86.45	91.17
10	81.28	80.12	79.07	84.72	89.26
11	82.48	80.6	83.32	84.02	90.50
12	80.78	81.41	83.45	86.59	91.06
AVG	80.25	80.68	81.45	85.28	90.94

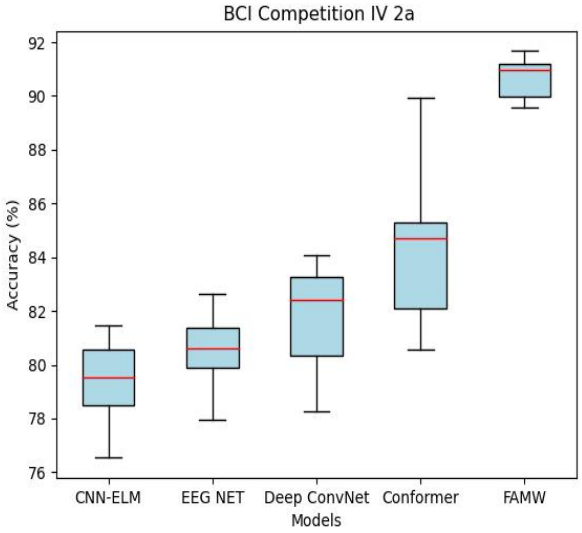

Fig. 5. Results on the BCI Competition IV 2a

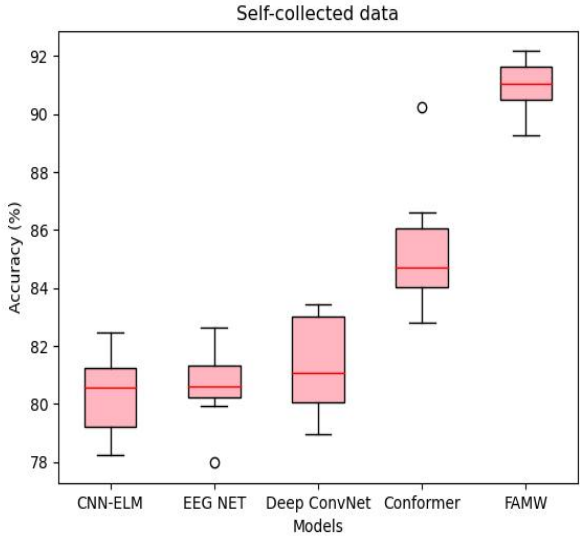

Fig. 6. Results on the Self-collected data

C. *Ablation experiment*

The aim of this experiment is to remove different modules in FAMW algorithm gradually. The effects of Feature Fusion, Frequency Band Selection Windowed Multi-Head Attention and global feature fusion network (GFFN) on the performance of the algorithm are analyzed. Specific ablation experiment modules are as follows:

TABLE III
THE COMPARISON RESULTS OF ABLATION EXPERIMENTS COMPLETED ON THE DATASET OF BCI COMPETITION IV 2A

FWM	FF	GFFN	ACC(%)
			82.80
√			88.54
√	√		89.95
√	√	√	92.3

(1) FF (Feature Fusion Ablation): time-frequency and time-domain feature fusion modules in FAMW algorithm are removed, and time-domain information is used separately.

(2) FWM (Frequency Band Selection Windowed Multi-Head Attention Removal) : Remove the multi-head attention module of frequency band selection and windowed multi-head attention removal, and use the ordinary multi-head self-attention.

(3) Global Feature Fusion Network Removal (GFFN) : Remove the Global Feature Fusion Network (GFFN) module.The comparison results are shown in Table III.

IV. CONCLUSION

This paper proposes a Cnn-Transformer model with time-time-frequency features is proposed to significantly improve the classification performance of EEG signals in motor imagination tasks. By weighting the original EEG signals and the time-frequency features extracted by continuous wavelet transform (CWT), The model fully leverages the complementary information of time-domain and frequency-domain features to enhance multi-dimensional feature representation.The feature embedding method can effectively map the fused features to the feature space suitable for Transformer processing, and further enhance the representation capability of time series data.

The Windowed Multi-Head Attention mechanism is excellent at capturing multi-scale time features, and can dynamically adjust the attention window size over time frequency segments to accurately model long-range dependencies and local features in time series data. By introducing gated feed-forward network (GFFN), the model adaptively weights the features of different network layers, reduces the interference of redundant information, and further improves the effect of feature extraction. Finally, the fused features are input to the Softmax classifier through the fully connected layer to complete the final classification of the motor imagery task. The experimental results show that compared with other methods, the proposed model significantly improves classification accuracy and demonstrates superior performance in MI EEG signal classification.

Future studies will further optimize the model to explore different feature fusion strategies and finer attention mechanisms to improve classification accuracy. At the same time, more EEG data and different task scenarios will be combined to verify the generalization ability and robustness of the model and promote its wide application in brain-computer interface (BCI) applications.

REFERENCES

[1] M. Zabcikova, Z. Koudelkova, R. Jasek, and J. J. Lorenzo Navarro, "Recent advances and current trends in brain-computer interface research and their applications," Int. J. Dev. Neurosci., vol. 82, no. 2, pp. 107–123, 2022.

[2] C. M. Wong, B. Wang, Z. Wang, K. F. Lao, A. Rosa, and F. Wan, "Spatial filtering in SSVEP-based BCIs: Unified framework and new improvements," IEEE Trans. Biomed. Eng., vol. 67, no. 11, pp. 3057–3072, Nov. 2020.

[3] B. Abibullaev and A. Zollanvari, "A systematic deep learning model selection for P300-based brain–computer interfaces," IEEE Trans. Syst. Man, Cybern. Syst., vol. 52, no. 5, pp. 2744–2756, May 2022.

[4] W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, F. Wan, and X. Chen, "EEG-based emotion recognition via channel-wise attention and self-attention," IEEE Trans. Affect. Comput., vol. 14, no. 1, pp. 382–393, 2020.

[5] B. Wang et al., "Common spatial pattern reformulated for regularizations in brain-computer interfaces," IEEE Trans. Cybern., vol. 51, no. 10, pp. 5008–5020, Oct. 2021.

[6] A. Al-Saegh, S. A. Dawwd, and J. M. Abdul-Jabbar, "Deep learning for motor imagery EEG-based classification: A review," Biomed. Signal Process. Control, vol. 63, Jan. 2021, Art. no. 102172.

[7] H. Altaheri et al., "Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review," Neural Comput. Appl., vol. 35, no. 20, pp. 14681–14722, Jul. 2023.

[8] P. Chen, Z. Gao, M. Yin, J. Wu, K. Ma, and C. Grebogi, "Multiattention adaptation network for motor imagery recognition," IEEE Trans. Syst. Man, Cybern. Syst., vol. 52, no. 8, pp. 5127–5139, Aug. 2022.

[9] J. Fumanal-Idocin, Y.-K. Wang, C.-T. Lin, J. Fernández, J. A. Sanz, and H. Bustince, "Motor-imagery-based brain–computer interface using signal derivation and aggregation functions," IEEE Trans. Cybern., vol. 52, no. 8, pp. 7944–7955, Aug. 2022.

[10] R. Foong et al., "Assessment of the efficacy of EEG-based MI-BCI with visual feedback and EEG correlates of mental fatigue for upper-limb stroke rehabilitation," IEEE Trans. Biomed. Eng., vol. 67, no. 3, pp. 786–795, 2019.

[11] V. Zotev, A. Mayeli, M. Misaki, and J. Bodurka, "Emotion self-regulation training in major depressive disorder using simultaneous real-time fMRI and EEG neurofeedback," NeuroImage: Clinical, vol. 27, p. 102331, 2020.

[12] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Representations, 2021.

[13] K. K. Ang and C. Guan, "EEG-based strategies to detect motor imagery for control and rehabilitation," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 4, pp. 392–401, Apr. 2017.

[14] P. Autthasan et al., "Min2net: End-to-end multi-task learning for subject-independent motor imagery EEG classification," IEEE Trans. Biomed. Eng., vol. 69, no. 6, pp. 2105–2118, 2022.

[15] O. Rioul and P. Duhamel, "Fast algorithms for discrete and continuous wavelet transforms," IEEE Trans. Inf. Theory, vol. 38, no. 2, pp. 569–586, 1992.

[16] Y. Wang, K. C. Veluvolu, and M. Lee, "Time-frequency analysis of band-limited EEG with BMFLC and Kalman filter for BCI applications," J. NeuroEng. Rehabil., vol. 10, no. 1, pp. 1–16, Dec. 2013.

[17] K. Samiee, P. Kovács, and M. Gabbouj, "Epileptic seizure classification of EEG time-series using rational discrete short-time Fourier transform," IEEE Trans. Biomed. Eng., vol. 62, no. 2, pp. 541–552, Feb. 2015.

[18] M. Bentlemsan, E.-T. Zemouri, D. Bouchaffra, B. Yahya-Zoubir, and K. Ferroudji, "Random forest and filter bank common spatial patterns for EEG-based motor imagery classification," in Proc. 5th Int. Conf. Intell. Syst. Modeling Simulation, Jan. 2014, pp. 235–238.

[19] L. Quoc Thang and C. Temiyasathit, "Increase performance of four-class classification for motor-imagery based brain-computer interface," in Proc. Int. Conf. Comput., Inf. Telecommun. Syst. (CITS), Jul. 2014, pp. 1–5.

[20] P. Gaur, R. B. Pachori, H. Wang, and G. Prasad, "An empirical mode decomposition based filtering method for classification of motor-imagery EEG signals for enhancing brain-computer interface," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2015, pp. 1–7.

[21] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," Hum. Brain Mapp., vol. 38, no. 11, pp. 5391–5420, Nov. 2017.

[22] V. J. Lawhern et al., "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," J. Neural Eng., vol. 15, no. 5, Oct. 2018, Art. no. 056013.

[23] D. Milanés Hermosilla et al., "Shallow convolutional network excels for classifying motor imagery EEG in BCI applications," IEEE Access, vol. 9, pp. 98275–98286, 2021.

[24] C. Brunner et al., "BCI competition 2008–Graz data set A," Inst. Knowl. Discovery (Lab. Brain-Comput. Interfaces), Graz Univ. Technol., Graz, Austria, Tech. Rep., vol. 16, pp. 1–6, 2008.

[25] J. Lin and L. Qu, "Feature extraction based on Morlet wavelet and its application for mechanical fault diagnosis," J. Sound Vib., vol. 234, no. 1, pp. 135–148, 2000.
[26] O. Rioul and P. Duhamel, "Fast algorithms for discrete and continuous wavelet transforms," IEEE Transactions on Information Theory, vol. 38, pp. 569–586, 1992.
[27] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," Journal of Neural Engineering, vol. 15, no. 5, p. 056013, Oct. 2018.
[28] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," Human Brain Mapping, vol. 38, no. 11, pp. 5391–5420, 2017.
[29] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG Conformer: Convolutional transformer for EEG decoding and visualization," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 710–719, 2022.
[30] Y. R. Yuvaraj, A. Baranwal, A. A. Prince, et al., "Emotion recognition from spatio-temporal representation of EEG signals via 3D-CNN with ensemble learning techniques," Brain Sciences, vol. 13, no. 4, p. 685, 2023.