# A Multi-feature Fusion Transformer Neural Network for Motor Imagery EEG Signal Classification

Zhangfang Hu, Lingxiao He, Haoze Wu,

*Abstract*—**In recent years, the classification method of motor imagery(MI) electroencephalography(EEG) signals based on deep learning(DL) has become more and more mature in the field of brain-computer interface(BIC). However, most of the studies tend to use a single feature or two associated features when dealing with motor imagery EEG signal classification, while ignoring other features, resulting in poor classification performance. Therefore, this paper proposes a neural network feature fusion algorithm called Multi-feature Fusion Transformer(M-FFT). The network is built based on convolutional neural network (CNN) and Transformer. This method uses CNN and wavelet transform to extract the time-frequency and space-time features, and uses the converter to fuse the three feature domains contained in the two features to establish the information interaction between the three feature domains contained in the two features. Then, the global feature pooling is used to output the feature vector, and finally the softmax function is used to classify the feature vector. In the training process, we use cross entropy as the loss function. Finally, on the brain-computer interface competition IV data set 2a, the average classification accuracy is 85.66%, and the average kappa value is 0.833. The experimental results validate the algorithm performance.**

*Index Terms*—**motor imagery, brain-computer interface, Transformer, multi-feature fusion, CNN**

## I. Introduction

THE BIC (Brain-Computer Interface) serves as a communication channel bridging the gap between humans and their external environment. It operates autonomously, free from reliance on human neural pathways and muscular activity. Its main signal is a bioelectrical signal composed of EEG signals generated by brain nerve activity. In the realm of biomedical research, the neuroimaging methods used to record EEG signals primarily involve techniques like electrocorticography (EcoG), magnetoencephalography (MEG), and conventional electroencephalography (EEG). [1] EEG is favored because of its simple acquisition, short time consuming and low acquisition cost. Methods for acquiring EEG signals can be broadly categorized into three main types: invasive, partially invasive, and non-invasive approaches. [2]. The non-invasive brain-computer interface, while it may offer a lower signal quality compared to the invasive counterpart, presents substantial advantages in contrast to the associated risks of implanting EEG sensors directly into the brain.

Human movement is commonly associated with $\mu$-rhythms (8-13 Hz) and $\beta$-rhythms (17-30 Hz) in the EEG signal. When subjects are at rest or inert, they show a significant increase in the amplitude of these rhythms, a physiological phenomenon known as event-related synchronization (ERS) [3]–[5]. When the subjects begin to carry out motor imagery, the above specific rhythms will show a significant decrease in activity, which is called event-related desynchronization (ERD). Numerous quantitative techniques exist for evaluating ERD/ERS, encompassing classical frequency band power analysis, trial variance assessment, autoregressive modeling, spectrum decomposition, time spectrum evolution analysis, task-related energy fluctuation analysis, Kolmogorov-chaitin complexity, and Fourier spectrum entropy. Notably, Kolmogorov-chaitin complexity and fourier spectral entropy represent nonlinear metrics, while the remaining methods hinge on frequency band power entropy. This array of physiological attributes and associated quantification methodologies forms the underpinning for extracting EEG data.

Motor imagery involves the psychological simulation of motor intention, which activates related neural pathways in the absence of physical movement execution. This stimulation changes the brain waves. By capturing and analyzing these brain waves, information can be deciphered. These decoded data can achieve human-computer interaction through sensors and control circuits. Applications include assistive prostheses for the disabled, mechanical remote control, and personnel status assessment. [6]. Hence, enhancing the accuracy of motor imagery EEG signal classification tasks holds significant importance.

Machine learning has become widely utilized for the classification of EEG signal data. Conventional approaches typically encompass three key components: preprocessing, feature extraction, and classification.

Preprocessing typically involves several steps. These include channel selection, signal filtering, and artifact removal. Channel selection aims to pick the electrode with the most relevant motion imagery features from a set of EEG electrodes. Choosing fewer EEG channels has multiple advantages. It not only reduces system complexity, computation time, and equipment cost but can also enhance system performance. Additionally, it mitigates the risk of overfitting,

Manuscript received on June 16, 2023; revised on October 14, 2023. This work was supported in part by the Youth Fund Program of the National Natural Science Foundation of China (Grant No. 61703067), the Chongqing Basic Science and Frontier Technology Research Program (Grant No. Cstc2017jcyjAX0212), and the Science and Technology Research Program of Chongqing Municipal Education Commission (KJ1704072).

Zhangfang Hu is a Professor at the Key Laboratory of Optical Information Sensing and Technology, School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China(e-mail: 3565207151@qq.com)

Lingxiao He is a graduate student of the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (corresponding author phone: 198-238-24589; e-mail: s220431028@stu.cqupt.edu.cn)

Haoze Wu is a graduate student of the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail:617971700@qq.com)

which may occur with uncorrelated channels. However, Hassanpour's study pointed out that the number and location of the electrodes chosen must be very appropriate [7]. Mwatavelu conducted a study on the optimal EEG channel of motor imagery, and concluded that different subjects had different optimal electrodes for judging motor imagery [8].This aspect contributes to the limited generalization of current EEG signal training. The primary purpose of signal filtering is to isolate the frequency band that holds the most pertinent information for the classification of motor imagery EEG signals. For motor imagery, the ERD/ERS of sensorimotor rhythms occurs mainly in the $\mu$ (8-12 Hz) and $\beta$ (18-26 Hz) frequency bands. Pfurtscheller found that the alpha and beta frequency bands are extremely important for motor imagery, with 10-12 Hz and 18-26 Hz being the most suitable bands for discrimination, and bilateral asynchrony was found in the 10-14 Hz band [9]. In most of the current studies, the bands 8-12 Hz and 18-26 Hz have been used, but some studies have also pointed out that it is sufficient to use the band 6-35 Hz, because the main purpose of filtering by frequency is to eliminate low-frequency artifacts and high-frequency noise, and most of the low-frequency artifacts are below 6 Hz, while high-frequency noise is also mainly concentrated above 35 Hz [10]. The most common methods for artifact removal are Independent Component Analysis (ICA) proposed by Tayeb [11] and Common Average Reference (CAR) proposed by Yang [12].

In motor imagery EEG classification research, feature extraction primarily focuses on temporal, spatial, and frequency features. Typically, either one type of feature or a combination of two associated features is extracted. However, there is a lack of mature research on simultaneous extraction and processing of multiple types of features. This constraint leads to partial information extraction from EEG signals and diminishes the universality of EEG signal feature extraction networks. Traditional feature extraction methods mainly use the Common Spatial Patterns (CSP) proposed by Luo [13] and various algorithms developed from it, such as Filter Bank Common Spatial Pattern (FBCSP), which can also achieve better good accuracy. However, since it can only handle simple constraints, but not multi-constraint space and multivariate problems. The propositional logic approach may not be able to solve the constraint satisfaction problem completely because it may not be able to handle uncertainty and ambiguity constraints. Also it has poor generality in other fields. So feature extraction in the field of deep learning has gradually replaced the traditional feature extraction in recent years.

The main applications of classifiers in the field of motion imagery are decision tree classification, rule-based classification, K-neighborhood, plain Bayesian classifier, neural network classification, and support vector machines. The most frequently employed classification methods include neural network classification and support vector machines. Mainly due to the emergence of neural networks in recent years in various fields have been widely used and have achieved significant results. Deep learning algorithms have demonstrated remarkable effectiveness in image recognition, and they have also found widespread application in EEG signal processing, yielding highly commendable results. So using neural networks for classification has become the mainstream method in the field of EEG signal processing nowadays. The current deep learning neural networks have also derived numerous branches after a period of development such as CNN [14] and its variants such as attention-based CNN [15], residual-based CNN [16], Dense Net [17], 3D-CNNs [18], multi-branch CNNs [19], etc., RNNs [20] and their variants such as GRU [21], LSTM [22], Transformer [23], etc. Applications in the field of EEG signal processing are still mainly based on CNN, RNN and their variants [10]. Presently, the latest Transformer and its derivative networks have made substantial advancements across various domains, particularly in natural language processing. However, the application and study of this network in the realm of EEG signal processing have not been widely explored.

To address the issues of poor robustness and low classification accuracy resulting from single-feature information extraction and to enhance overall accuracy, this study introduces a novel hybrid feature extraction and processing network called Multi-feature Fusion Transformer (M-FFT). This network separately assesses the spatio-temporal domain and the time-frequency domain and subsequently performs multi-feature fusion. Finally, the classification of motor imagery EEG signals is achieved.

An innovative M-FFT decoding network is proposed, and its contributions can be summarized as the following two points:

A multi-dimensional feature extraction approach is proposed, which provides a direction for feature extraction in more feature domains in the future, and the in-depth representation and comprehensive extraction of relevant features of motor imagery EEG signals are investigated.

This paper introduces a multi-dimensional feature processing approach based on the Transformer model. Firstly, it leverages Transformer, a proven technology in natural language processing, and applies it to motor imagery EEG signal processing. An effective methodology for integrating Transformer into the field of motion picture EEG signal processing is described in this paper.Additionally, the enhancements made to Transformer in this study enable it to facilitate fusion among different feature domains. The improved Transformer can allocate weights and facilitate information interaction across various feature domains. This advancement paves the way for incorporating more features in the future.This research successfully achieves multi-information fusion feature extraction and classification by integrating it with the multi-information feature extraction network. This addresses issues related to the robustness of single-information feature extraction and contributes to improved classification performance. It successfully solves the problem of poor robustness caused by single information feature extraction, and further improves the classification accuracy.

## II. Methods

### A. Data Pre-processing

*1) Data Filtering and Artifact Removal Processing:* Studies have shown that the most relevant part of the EEG signal for motor imagery is mainly contained in 6Hz-35Hz [10], so the acquired EEG waves were filtered, and the Butterworth third-order filter was used to filter the raw data to obtain the

motor imagery EEG signal from 6Hz-35Hz, and the EOG channel was removed.

Since the data recorded by EEG also includes a series of other uncorrelated signals such as electrooculogram signals, this creates enormous difficulties in analyzing and processing EEG signals.EEG signals. To address this challenge, this study utilizes the ICA approach to remove EEG artifacts from filtered signals.

*2) Signal Enhancement:* Since the data collection of EEG is more difficult, time-consuming, and has high ethical and safety risks, the data collection of EEG signals is very little, the database is small, and it cannot meet the data demand for large and deep networks with high risk of overfitting, so data enhancement using certain means becomes necessary. We use a sliding window to enhance the data [25], a window of length 3S is designed, and then the sliding starts sequentially at 0.1S intervals, assuming that the original EEG signal is a given segment with E electrodes and T sampling points, namely $\mathrm{E}^i \in \mathbb{R}^{E*T}$ , Thus, a set of window EEG data is cut out, $\mathrm{T}_s = 0S$ representing the starting point of the first sliding window. $T_e = 3S$ signifies the conclusion of this window. $T_0 = 0.1S$ representing the sliding time interval, this paper slides 10 times, thus turning one EEG data into 10 EEG data,$S^i = [S^i_{T_s,T_e}, S^i_{T_s+T_0,T_e+T_0}, ..., S^i_{T_s+10T_0,T_e+10T_0}]$, $i \in (1, 288)$ , This allows more data sources to be available during the training process.

### B. Multi-feature Fusion Transformer Network Overview

In this paper, we first input the processed EEG data into a parallel feature preprocessing network, and extract the time-space features(TSF) and time-frequency features(TFF) of the data by adding the Temporal-frequency Block (T-F Block) and the Temporal-spatial Block (T-S Block), respectively. The two features are input to the Self-attention Block (S-A Block) module for processing, and each feature is given attention internally to do preliminary feature processing and network learning, The TSF data and TFF data that pass the self-attention block, as well as the TSF data and TFF data that do not pass the self-attention block, are then fed into the proposed parallel feature fusion transformer (FF Transformer) module. Then, the TSF data and TFF data that pass the self-attention block, and the TSF data and TFF data that do not pass the self-attention block, are subjected to further feature extraction, information fusion, and network learning in the parallel feature fusion transformer (FF Transformer) module.The structure of the complex of the proposed multi-feature fused network is shown in Figure 1.

### C. Multi-feature Extraction

*1) Temporal-frequency Block:* Time-frequency analysis is highly interpretable, given that neural oscillations are an inherent characteristic of the brain, time-frequency analysis offers a more direct insight into the neurophysiological mechanisms responsible for the processes reflected in EEG data. Therefore, this paper adopts the time-frequency signal as one of the signals for multi-feature fusion.

To extract the time-frequency signal, we use wavelet transform [26] in this paper, but using wavelet transform will increase the dimensionality of the data, which will cause 2 problems: firstly, the increase in dimensionality will be detrimental to the weight deployment within the source later, because Transfromer is good at dealing with the two-dimensional matrix composed of word vectors, while the three-dimensional tensor is not conducive to the addition of attention. Secondly, it will cause a large difference in the dimensionality with the spatio-temporal features, which is not conducive to the later information interoperability and fusion as well as the weight deployment between two sources. To solve these two problems, a CNN is constructed in this paper to optimize the spatio-temporal features after wavelet transform. The processed data are then subjected to Position Embedding (PE), and finally turned into a form of data that is easy to be processed by the Transformer, as shown in Figure 2.

The following will elaborate the specific wavelet transform, CNN and Position Embedding implementation steps, the specific network details are shown in Table I.

TABLE I: Implementation details of the proposed Temporal-frequency Block project.

| Layer Type | Filters | Kernel | Stride | Output shape (channel × height × width) |
|---|---|---|---|---|
| Input | | | | 22 × 100 × 750 |
| Conv-1 | 10 | 100 × 10 | 1 × 10 | 220 × 1 × 75 |
| recombine | | | | 22 × 10 × 75 |
| Max Pooling | 1 | 10 × 3 | 1 × 3 | 22 × 1 × 25 |
| Position Embedding | | | | 22 × 1 × 25 |

The Fourier transform employs trigonometric functions as its basis, covering the entire time domain of a signal. However, it is unable to capture localized signal features. Although the "windowing" technique can enhance the Short-time Fourier Transform (STFT) to extract local features, STFT's fixed and uniform window width limits its effectiveness.To address this limitation, this paper employs wavelet transform, which allows for variable window widths. This approach facilitates more effective extraction of local signal features. The formula of wavelet transform is as follows:

$$WT(\alpha,\tau) = \frac{1}{\sqrt{\alpha}} \int_{-\infty}^{\infty} f(t) * \psi(\frac{t-\tau}{\alpha}) dt \qquad (1)$$

$f(t)$ is the signal to be analyzed, $t$ is the time, the result of wavelet transform is $WT(\alpha,\tau)$ , $\alpha$ represents the scale factor of wavelet transform, $\tau$ represents the translation of wavelet transform, $\psi$ is the wavelet transform function, this paper chooses ComplexGaussian wavelet, the scale is chosen as 100.

CNN is a feedforward neural network characterized by convolutional computation and a deep structure. It can learn the feature representation of the input data, which is very suitable for processing image data.CNN uses local perceptual field and shared weight strategy to extract local information in the feature map, this paper uses CNN mainly for data downscaling and simple extraction of features, mainly including two parts of convolution and pooling, in the convolution part, especially for temporal information using a convolution kernel of size 100 to process the sequence,ten convolution kernels are employed to extract temporal and frequency

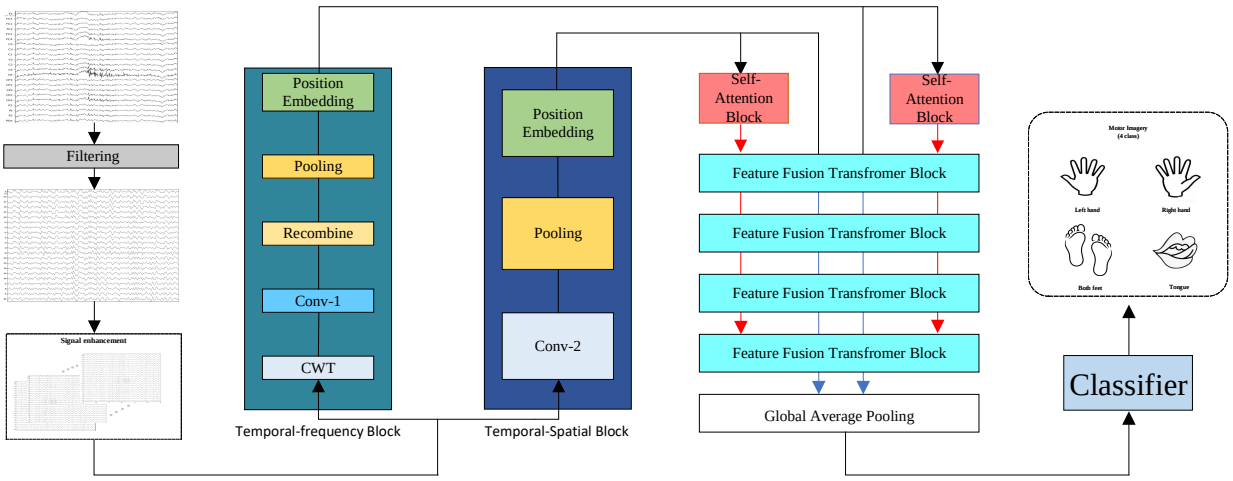


Fig. 1: Multi-feature Fusion Transformer Network structure

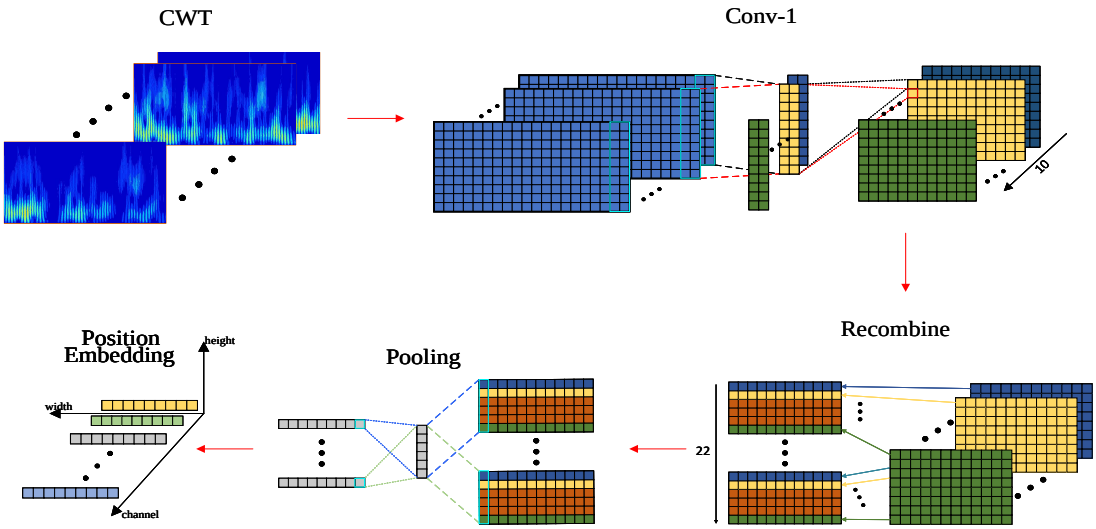


Fig. 2: Temporal-frequency Block structure diagram

features from the same channel, ensuring the thoroughness of feature extraction. The output of the convolutional layer is calculated as follows:

$$x_j^l = f(\sum_{i \in M_j} x_i^{l-1} * \omega_{ij}^l + b_j^l) \tag{2}$$

Where $x_i^{l-1}$ is the region corresponding to the $i$ th convolution kernel of the th layer, $l$ is the th feature map of the th layer, $x_j^l$ is the feature input map, $\omega$ is the weight matrix of the convolution kernel, $b$ is the bias, $f$ is the activation function, $*$ is the convolution operation.

Upon the completion of the convolution process, the next steps involve batch normalization, followed by a nonlinear transformation facilitated by the activation function. Since the data sets output from the 10 convolution kernels are used to dope the channel groups, the output features are reorganized and the convolution data of the respective channels will be reorganized to finally ensure that each group is the information data of the same channel. In the pooling part, since the number of dimensions after extraction is still high, this paper takes max pooling for further dimensionality reduction of EEG features.

$$x_j^l = down_{\max}(x_j^{l-1}) \tag{3}$$

Where $down_{\max}$ is the max pooling function and $x_j^l$ is the output feature map of the pooling layer.

The final output data is passed through PE, with additional position information, to complete the feature extraction and preliminary processing in the time-frequency domain. The data are encoded and the input sequence is linearly projected into a new sequence of dimension , PE with the same dimensions as the input embedding, and the two are summed to obtain the final data.PE is given by equation (4)(5).

$$PE(pos, 2i) = \sin(pos/10000^{2i/d_{\text{model}}}) \tag{4}$$

$$PE(pos, 2i+l) = \cos(pos/10000^{2i/d_{\text{model}}}) \tag{5}$$

Where $pos$ is the position index of each sequence data, $d_{\text{model}}$ is the feature embedding dimension, $i$ is the feature dimension, $\sin / \cos$ takes values in different dimensions from $2\pi$ to $10000 \times 2\pi$ .

*2) Temporal-spatial Block:* With the development of brain neuroscience, the study of the functions of different brain regions has become more and more in-depth, and the information in the spatio-temporal domain has become more and more important. Recently, more and more researchers have started to try to extract feature information in the spatio-temporal domain, and have achieved better classification results [10]. Therefore, this paper adopts the spatio-temporal signal as one of the signals for multi-feature fusion. In this study, the preprocessed EEG signal is fed into the spatio-temporal module, where specific convolution kernels are employed to downscale the 3D features, facilitating the extraction of spatio-temporal characteristics. Simultaneously, structural adjustments to the data are made through techniques like maximum pooling to align it with the structural prerequisites for interaction with time-frequency domain signals, thus enhancing the extraction of spatio-temporal features. TheTemporal-spatial Block structure diagram is shown in Figure 3.

The method of extracting spatio-temporal signals is roughly the same as the method of extracting time-frequency signals are extracted using CNN, which will not be discussed in detail in this paper, and the details related to the extraction of time-frequency features with different convolutional kernel sizes, etc., are shown in Table II

TABLE II: Implementation details of the proposed Temporal-frequency Block project.

| Layer Type | Filters | Kernel | Stride | Output shape (channel × height × width) |
|---|---|---|---|---|
| Input | | | | 22 × 750 |
| Conv-1 | 22 | 22 × 10 | 1 × 10 | 22 × 1 × 75 |
| Max Pooling | 1 | 1 × 3 | 1 × 3 | 22 × 1 × 25 |
| Position Embedding | | | | 22 × 1 × 25 |

*3) Self-attention Block:* As the overall neural network model becomes larger and larger, the computational power of the computer begins to dry up. At the same time, the over fitting and gradient disappearance of the local model become more and more obvious with the large model. [23] Therefore, introduces the mechanism of attention to alleviate the above problems.The mechanism of attention is to allot different kinds of weights to the information entered into the computer, depending on the information entered and the importance of the information entered, so as to screen out more useful information for transmission to the next level model, which effectively avoids the over fitting and gradient disappearance caused by the large model. The self attention module is shown in Figure 4. we do the self-attention operation on the time-frequency feature signal and the space-time feature signal firstly, and perform the internal attention weight allocation between the space-time feature signal and the time-frequency feature signal, and then perform the weight allocation between the feature domains afterward. The calculation formula is as follows:

$$Attention(Q, K, V) = \omega(QK^T)V \tag{6}$$

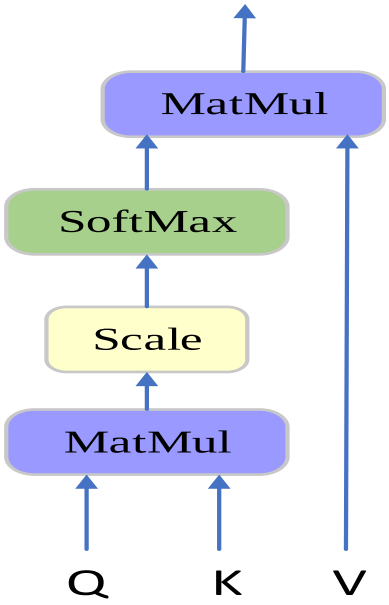


Fig. 4: Self-attention module structure

where $Attention(Q, K, V)$ is the value of the obtained attention, $Q, K, V$ represents Query Vector, Key Vector, Value Vector, respectively, $\omega(\bullet)$ and represents the activation function $Soft\max(\frac{\bullet}{\sqrt{d_k}})$ .

### D. Multi-feature Domain Encode

Most of the current studies use simple splicing in feature fusion, and do not pay attention to the different effects of different features on the classification results, and there is no information interaction and fusion between different features, which leads to no significant improvement in the robustness and accuracy of classification results of many networks after multi-feature extraction. In this paper, we propose Multi-feature domain Encoder to solve this problem, which consists of two self-attentive and Feature Fusion Transformer. Firstly, the TFF and TSF are connected to the self-attention module, and then the output data of TFF and TSF that are not connected to the attention module and the output data of TFF and TSF that are connected to the self-attention block are connected to the feature fusion Transformer to realize the information interaction and weight deployment of different features [27]. The feature fusion Transformer module is shown in Figure 5.

These four signal taps are named Inattentional Temporal-frequency Signal (ITF), Inattentional Temporal-spatial Signal (ITS), Attentional Temporal-spatial Signal (ATF), and Attentional Temporal-spatial Signal (ATS). When performing cross-attention, make sure that the query vector and key vector come from one feature and the value vector comes from another feature. For example, in this paper, we use query vectors and key vectors from the inattentive temporal-frequency feature and value vectors from the inattentive feature.

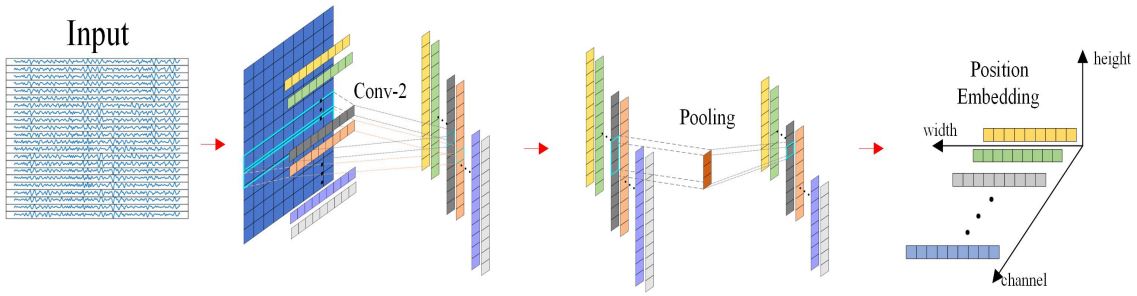


Fig. 3: Temporal-spatial Block structure diagram

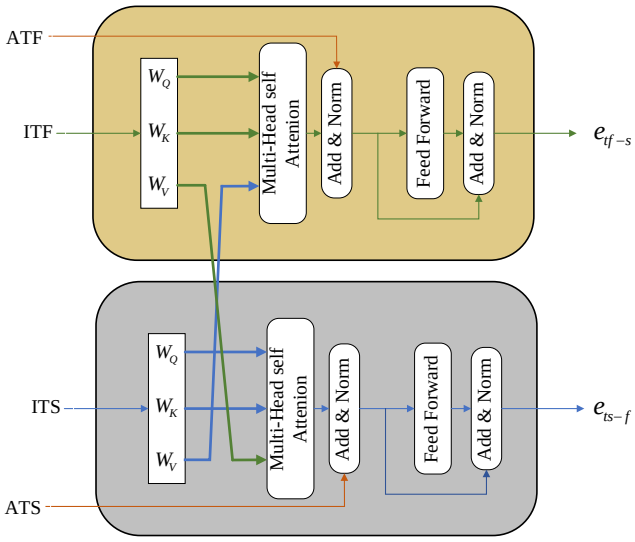


Fig. 5: Feature Fusion Transformer Block structure diagram

This approach can effectively deploy different information features and make different features to be fused. The fusion of cross-attention [27] and self-attention [23] using residual networks not only prevents the gradient from disappearing, but also allows further fusion and weighting between different features. For example, in this paper, the signal after cross-attention of the interrogation vector and key vector of the no-attention time-frequency feature and the value vector of the no-attention feature is connected with the residuals of the self-attention time-frequency signal. The attention mechanism is calculated as follows:

$$Attention(\mathrm{Q}_{itf},\mathrm{K}_{itf},V_{its}) = Soft\max(\frac{Q_{itf}K_{itf}^{T}}{\sqrt{d_k}})V_{its} \tag{7}$$

$$Attention(\mathrm{Q}_{its},\mathrm{K}_{its},V_{itf}) = Soft\max(\frac{Q_{its}K_{its}^{T}}{\sqrt{d_k}})V_{itf} \tag{8}$$

$\mathrm{Q}_{itf}$ , $\mathrm{K}_{itf}$ and $\mathrm{V}_{its}$ denote the inattentive time-frequency signals as the query vector, key vector, and value vector income in the attention module, spectively. $\mathrm{Q}_{its}$ , $\mathrm{K}_{its}$ and $\mathrm{V}_{itf}$denote the inattentive time-space signals as the query vector, key vector, and value vector inputs in the attention module, separately. $Soft\max(.)$ denotes the softmax function. $d_k$ denotes the dimensionality of $\mathrm{K}_{itf}$ , $\mathrm{K}_{its}$ respectively.

### E. Classifier

After the multi-feature fusion process two-way high-level features are obtained, denoted as $e_{tf-s}$ , $e_{ts-f}$ respectively. Then these two-way features are spliced, and the spliced features are denoted as $e_{Mf}$ , and the spliced vector is subjected to global average pooling to reduce the number of parameters and prevent overfitting, and the pooling produces a one-dimensional vector of length 22. The outcome is subsequently normalized. Global average pooling calculates the average value across all pixels within each channel of the output feature map. This process yields a feature vector with dimensions equal to the number of categories, which is then directly fed into the softmax layer, as depicted in Figure 6. [28] Finally, the result is normalized. The final output is the multi-domain features of the fused time-space-frequency signal, so this paper uses the loss function for it to learn parameters. The final loss function equation in follows:

$$L_{cls_{Mf}} = CEL(F(GAP(e_{Mf})), y) \tag{9}$$

where $CEL(.)$ is the cross-entropy loss function, $F(.)$ is the classifier, $GAP(.)$ is the global average pooling, and $y$ is the signal label.

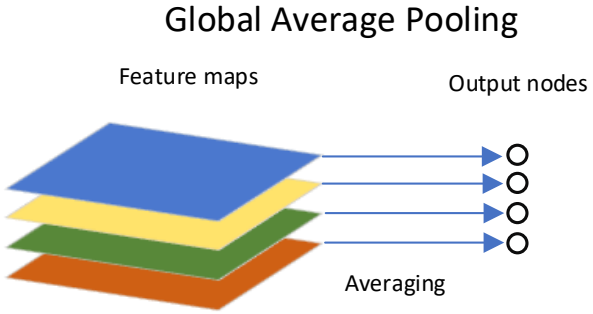


Fig. 6: Global Average Pooled Block structure

## III. Experiment and Results

### A. Experimental Dataset

The data were collected from the 2008 Brain-computer Interface Competition IV 2a have nine subjects who were asked to perform four different categories of motor imagery (left hand, right hand, foot, and tongue) during the experiment,

and each subject performed the experiment on two different days to obtain two sessions, which were used for training and testing the classifier. Each session could be subdivided into 6 runs, each run containing 48 trials, of which 12 were left hand motor imagery, 12 were right hand motor imagery, 12 were foot motor imagery, and 12 were tongue motor imagery, and the order was randomly distributed. 6 runs were followed by a rest period for the subjects.

The procedure of each experiment was that there was a monitor in front of the subject and a fixed cross would appear on the monitor at the beginning of the experiment, accompanied by a short cue tone. After two seconds, an arrow cueing up, down, left, and right (corresponding to tongue, foot, left hand, and right hand) appeared on the display for approximately 1.25 S. The subject was prompted to imagine the corresponding movement. Each subject was asked to imagine this process only until the cue arrows disappeared from the screen, which took about 6S. Then the screen went black again, with an average time of about 8S each time. The process is shown in Figure 7. [24]

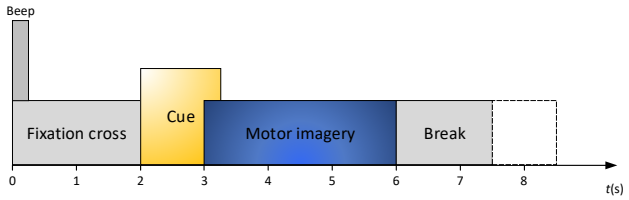


Fig. 7: Flow chart of motion imagery acquisitionm

### B. Experimental Setup

The approach in this paper is implemented using Tensorflow and Keras in python with an Inter(R) Core(TM) i7-9750H (2.60GHz) CPU and four NVIDIA GPUs (RTX 2080). In this paper, only EEG channel data is used to classify the above dataset, and three electrooculogram channel (EOG channel) data are discarded in the experiments of this paper. In this paper, four feature fusion Transformers are used for the experiments, and the model in this paper is trained with a stochastic gradient descent optimizer with a learning rate of 0.001, and the network is iterated 350 times with the weight decay ratio and momentum set to 0.001 and 0.9, respectively. In addition, if no improvement is observed in the training set after 30 iterations, the training is terminated early to avoid overfitting. In this study, classification accuracy and kappa coefficient were used as the evaluation criteria of the model. kappa coefficient is a measure of classification accuracy, and its expression is given by:

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{10}$$

The formula $p_0$ is the total of the number of properly categorised samples in each of the classes split by the total number of samples.That is, the aggregate categorisation accuracy,, and $p_e$ is the random guess accuracy.

### C. Ablation Experiments

The following tests were carried out using the filtering module, the time-frequency feature extraction module, the presence of the spatio-temporal feature extraction module and different numbers of feature fusion Transformer modules.

*1) Different Modules:* The outcome is given in Table III. The No filter (NF) experiment is designed to remove the filter module and keep all the modules except the filter module, and input the data into the network without filtering and artifact removal. NO Temporal-frequency Feature (NTF) extraction experiment, i.e., the Temporal-frequency Feature extraction module is deleted while all other modules except the Temporal-frequency Feature extraction module are retained, and the Temporal-frequency Feature extraction module outcome is connected to the original Temporal-frequency Feature extraction module outcome, and the data are processed without extracting temporal-frequency features. The data can be processed without extracting the time-frequency features. NO Temporal-spatial Feature (NTS) extraction experiment, i.e., delete the TSF extraction module while keeping all other modules except the TSF extraction module, connect the TFS extraction module outcome to the output of the original Temporal-spatial Feature extraction module, and process the data without extracting temporal features. data are processed. Also the full module experimental data are included in the table for easy comparison.

Finally, as shown by the results in Table 4, both the removal of the filter module, the TFF extraction module, and the TSF extraction module cause a decrease in the classification accuracy, from this, it is obvious that the decrease in accuracy is especially obvious when the time-frequency module is removed, which indicates that the information provided by the time-frequency signal features is more important in this method, while the information provided by the spatio-temporal signal is less, and the minimum and maximum difference accuracy of the no-filter module experiment and the full module experiment is smaller than that of the no-time-frequency module experiment and the full module experiment. The minimum and maximum difference accuracy is smaller than that of the no-time-frequency module experiment and the no-time-space module experiment. It shows that the multi-feature fusion can improve the system robustness.

*2) Different Number of Feature Fusion Transformer Modules:* In this paper, we address the influence of the number of feature fusion Transformer modules on the classification accuracy, and the results are shown in Table IV. In this paper, we compare 0-6 feature fusion Transformer modules, when the number of feature fusion Transformer modules is 0, i.e., the two data are directly pooled globally averaged, and then the output is performed by classifier, and the The classification accuracy is recorded in the table, and it can be concluded that increasing the module depth can improve the classification accuracy, but the network performance decreases slightly when the number of modules exceeds 4. This indicates that too many feature fusion Transformer modules will cause learning redundancy, which leads to overfitting. Therefore, 4 feature fusion Transformer modules are chosen as the frame depth.

### D. Experimental Results and Analysis

*1) Experimental Results:* Figure 8 shows the training process of the M-FFT model when the window size is 3S, the

TABLE III: Comparison of tests under different modules.

| Methods | Subject | | | | | | | | | Avg acc |
|---|---|---|---|---|---|---|---|---|---|---|
| | A01 | A02 | A03 | A04 | A05 | A06 | A07 | A08 | A09 | |
| NF+TF+TS | 62.64 | 66.49 | 80.77 | 67.94 | 69.42 | 63.08 | 77.25 | 72.25 | 73.51 | 70.37 |
| F+NTF+TS | 56.73 | 40.68 | 70.43 | 59.38 | 54.51 | 49.31 | 79.75 | 61.49 | 56.60 | 58.76 |
| F+TF+NTS | 77.45 | 70.78 | 85.63 | 73.30 | 71.92 | 68.75 | 80.64 | 81.47 | 74.85 | 73.75 |
| F+TF+TS | 83.32 | 75.83 | 91.65 | 80.54 | 85.71 | 79.53 | 93.67 | 91.59 | 89.13 | 85.66 |

TABLE IV: Comparison of experiments with different numbers of feature fusion Transformer modules.

| Number | Subject | | | | | | | | | Avg acc |
|---|---|---|---|---|---|---|---|---|---|---|
| | A01 | A02 | A03 | A04 | A05 | A06 | A07 | A08 | A09 | |
| 0 | 64.64 | 55.43 | 73.62 | 60.56 | 75.52 | 69.54 | 62.65 | 70.42 | 68.59 | 66.77 |
| 1 | 82.64 | 73.42 | 89.34 | 78.39 | 82.84 | 76.35 | 90.21 | 88.75 | 87.42 | 83.26 |
| 2 | 84.54 | 74.29 | 88.82 | 78.73 | 83.24 | 78.36 | 89.46 | 89.37 | 88.79 | 83.96 |
| 3 | 83.17 | 73.69 | 90.21 | 79.68 | 84.31 | 78.58 | 91.77 | 90.63 | 90.33 | 84.70 |
| 4 | 83.32 | 75.83 | 91.65 | 80.54 | 85.71 | 79.53 | 93.67 | 91.59 | 89.13 | 85.66 |
| 5 | 82.78 | 73.21 | 89.86 | 80.21 | 84.92 | 78.74 | 92.39 | 91.13 | 88.49 | 84.64 |
| 6 | 82.67 | 74.26 | 90.38 | 79.87 | 85.38 | 78.42 | 91.93 | 90.32 | 88.23 | 84.60 |

transverse scale is the number of epochs and the portrait scale is the classification accuracy. After 350 iterations of training, the final training accuracy and test accuracy converge to a stable level. This results in an average classification accuracy of 85.66% on dataset 2a of the BCI Competition IV competition.

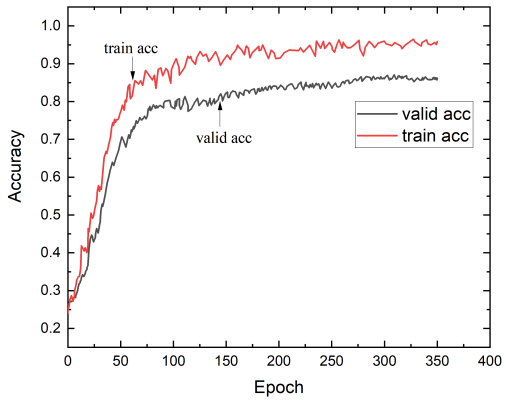


Fig. 8: Neural network training accuracy

*2) Experimental Analysis:* To assess the effectiveness of the recommended methodology, the methods in this paper were compared with existing state-of-the-art methods, including a method using traditional machine learning, FBCSP [29], two traditional CNN methods without feature fusion, EEGNet [30] and ConvNet [31], a method using multiple feature domains, SPCNN [32], and a domain adaptive method CCSP [33], aiming to show the differences between different methods for the same topic.

We evaluate the proposed model in this paper on dataset 2a of BCI Competition IV and objectively demonstrate the performance of the recommended methodology in this paper.as set out in the table V, The average accuracy and kappa values of the classification of the above models and and the proposed in this paper are modelled are included. The classification accuracies are listed in terms of accuracy for different subjects, and the average accuracy is given at the end.

The data were plotted by different subjects into bar graphs as shown in Figure 9.

The following analysis can be made from Figure 8: For the four-classification motion imagery classification task, the M-FFT method proposed in this paper is 2.01% and 0.199 higher in mean accuracy and kappa coefficient, respectively, than the second best classification accuracy method. it obtains the best accuracy on four subjects (including A02, A04, A06, and A07), especially on subjects A02 and A06 The second highest precision was 4.31% and 4.89% higher than the second highest precision, respectively, while better results were also obtained on other subjects. Since the EEG signal is a non-stationary random signal and its background noise is also strong, the FBCSP showed a slight difference in the results using the traditional method compared with those using the deep learning method, while the CCSP showed no significant advantage using the domain adaptive method. eegNet and ConvNet using the traditional CNN, although improved compared with the traditional method, did not have a significant advantage due to their The SPCNN uses multi-feature domain extraction to improve the classification performance, but the latter is better than the former in terms of final classification accuracy because it does not pay attention to inter-feature weighting compared to M-FFT feature fusion.

The experimental signals measured by different subjects are influenced by a series of factors such as their genes and growth environment, which makes the classification accuracy vary greatly when discriminating between subjects. For example, when using traditional methods such as FBCSP for motor imagery classification, the difference between subjects with the highest and lowest classification accuracy is 37.5%.The difference between the Subjects who had the highest classification accuracy and subjects who had

TABLE V: Comparison of experiments with different numbers of Feature Fusion Transformer modules.

| Methods | Subject | | | | | | | | | Avg acc | (kappa) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | A01 | A02 | A03 | A04 | A05 | A06 | A07 | A08 | A09 | | |
| EEGNet [30] | 83.68 | 63.89 | 90.97 | 64.24 | 59.72 | 52.08 | 87.85 | 82.29 | 86.81 | 74.61 | 0.661 |
| CCSP [33] | 84.72 | 52.78 | 80.90 | 59.38 | 54.51 | 49.31 | 88.54 | 71.88 | 56.60 | 66.51 | 0.553 |
| FBCSP[ [29] | 76.00 | 56.50 | 81.25 | 61.00 | 55.00 | 45.25 | 82.75 | 81.25 | 70.75 | 67.75 | 0.570 |
| ConvNet [31] | 76.39 | 55.21 | 89.24 | 74.65 | 56.94 | 54.17 | 92.71 | 77.08 | 76.39 | 72.53 | 0.634 |
| SPCNN [32] | 80.63 | 71.52 | 92.64 | 75.40 | 89.70 | 74.64 | 93.66 | 93.04 | 91.25 | 83.65 | - |
| M-FFT | 83.32 | 75.83 | 91.65 | 80.54 | 85.71 | 79.53 | 93.67 | 91.59 | 89.13 | 85.66 | 0.833 |

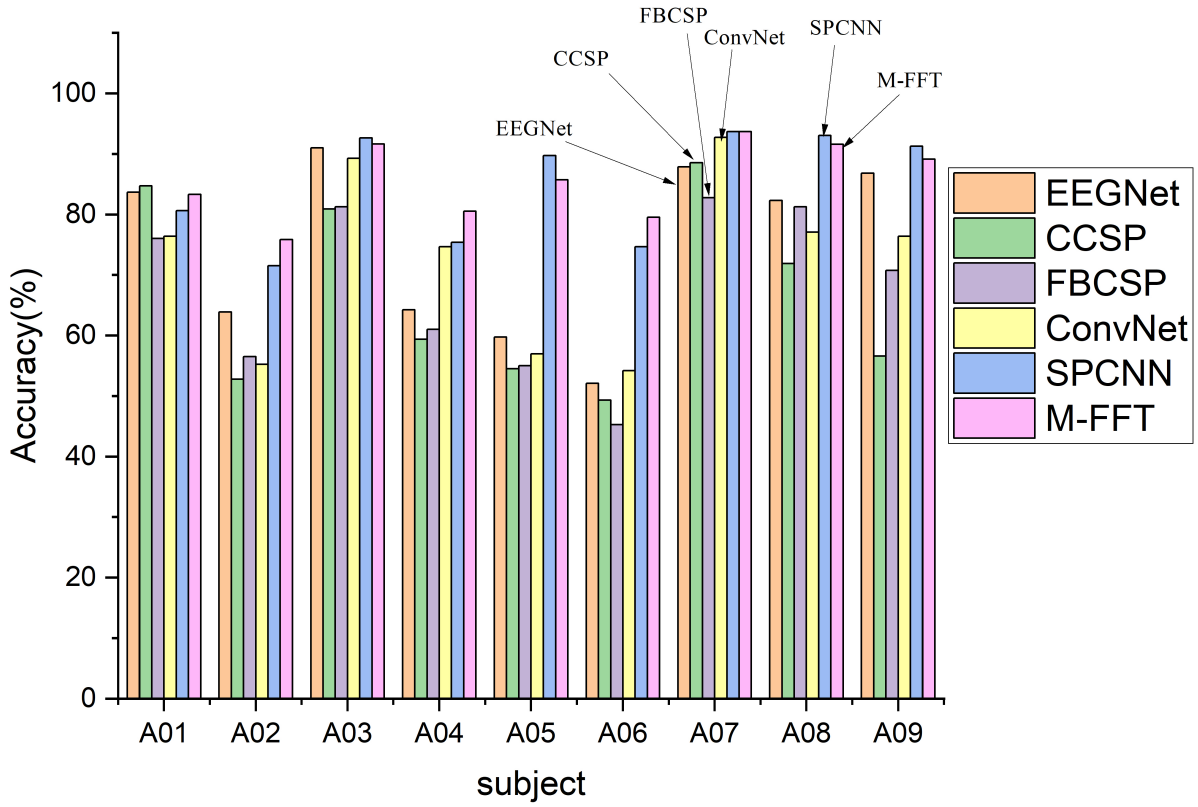


Fig. 9: Average classification accuracy per subject

the lowest classification accuracy was 35.77% using CNN conventional methods such as EEGNet.While using multi-feature fusion networks such as SPCNN and M-FFT, there was a difference of 22.14% and 17.83% between the Subjects who had the highest classification accuracy and subjects who had the lowest classification accuracy, respectively.Therefore, the multi-feature fusion method obviously helps to improve the robustness of the system.

## IV. Conclusion

In this paper, a deep learning classification model M-FFT for decoding four classified motor imagery EEG signals is proposed. First, the filtering block is used to filter high and low frequency noise and remove artifacts using ICA, second, the TF Block is used to extract TFF of EEG signals, which combines temporal and spatial information, and the Temporal-spatial Block is used to extract spatial and temporal features of EEG signals. Temporal-spatial Block is used to extract the spatio-temporal features of EEG signals, which combines temporal and spatial information. The Multi-feature domain Encoder model is designed to learn various aspects of EEG signals, including frequency features, spatial location information, and temporal characteristics. It achieves effective fusion of time-frequency and spatio-temporal features within the feature image, thereby enhancing the recognition performance for Motor Imagery EEG (MI-EEG) data.Following feature extraction, global average pooling is employed, and the resultant pooled features are subsequently used for classification. Finally, the classification performance of the proposed method is compared to other state-of-the-art techniques. The experimental results demonstrate that the method yields higher average classification accuracy and average kappa value when applied to the Brain-Computer Interface Competition IV dataset 2a, showcasing

its effectiveness and robustness. Future work aims to implement the deep learning network model proposed in this study in an operational brain-computer interface control system to validate its effectiveness and robustness and explore its real-time performance.

## REFERENCES


[1] Yang, J.; Gao, S.; Shen, T. A Two-Branch CNN Fusing Temporal and Frequency Features for Motor Imagery EEG Decoding. Entropy 2022, 24, 376.

[2] Zhang, K.; Xu, G.; Zheng, X.; Li, H.; Zhang, S.; Yu, Y.; Liang, R. Application of Transfer Learning in EEG Decoding Based on Brain-Computer Interfaces: A Review. Sensors 2020, 20, 6321.

[3] McFarland, D.J.; Miner, L.A.; Vaughan, T.M.; Wolpaw, J.R. Mu and beta rhythm topographies during motor imagery and actual movements. Brain Topogr. 2000, 12, 177–186.

[4] Shahid, S.; Sinha, R.K.; Prasad, G. Mu and beta rhythm modulations in motor imagery related post-stroke EEG: A study under BCI framework for post-stroke rehabilitation. BMC Neurosci. 2010, 11, P127.

[5] Djemal, R.; Bazyed, A.G.; Belwafi, K.; Gannouni, S.; Kaaniche, W. Three-Class EEG-Based Motor Imagery Classification Using Phase-Space Reconstruction Technique. Brain Sci. 2016, 6, 36.

[6] Li, H.; Ding, M.; Zhang, R.; Xiu, C. Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network. Biomed. Signal Process Control 2021, 72, 103342.

[7] Hassanpour A, Moradikia M, Adeli H. A novel end-to-end deep learning scheme for classifying multi-class motor imagery electroencephalography signals. Expert Systems, 2019, 36(6): e12494.

[8] Mwata-Velu, T.Y.; Avina-Cervantes, J.G.; Ruiz-Pinales, J.; Garcia-Calva, T.A.; González-Barbosa, E.A.; Hurtado-Ramos, J.B.; González-Barbosa, J.J. Improving Motor Imagery EEG Classification Based on Channel Selection Using a Deep Learning Architecture. Mathematics 2022, 10, 2302.

[9] Pfurtscheller, G.; Neuper, C.; Flotzinger, D.; Pregenzer, M. EEG-based discrimination between imagination of right and left hand movement. Electroencephalogr. Clin. Neurophysiol. 1997, 103, 642–651.

[10] Altaheri, H.; Muhammad, G.; Alsulaiman, M.; Amin, S.; Altuwaijri, G.; Abdul, W.; Bencherif, M.; Faisal, M. Deep Learning Techniques for Classification of Electroencephalogram (EEG) Motor Imagery (MI) Signals: A Review. Neural Comput. Appl. 2021, 1–42.

[11] Tayeb, Z.; Fedjaev, J.; Ghaboosi, N.; Richter, C.; Everding, L.; Qu, X.; Wu, Y.; Cheng, G.; Conradt, J. Validating deep neural networks for online decoding of motor imagery movements from EEG signals. Sensors 2019, 19, 210.

[12] Yang, J.; Ma, Z.; Wang, J.; Fu, Y. A Novel Deep Learning Scheme for Motor Imagery EEG Decoding Based on Spatial Representation Fusion. IEEE Access 2020, 8, 202100–202110.

[13] Luo, T.; Chao, F. Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network. BMC Bioinf. 2018, 19, 344.

[14] Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proc. IEEE 1998, 86, 2278–2324.

[15] Fan, C.C.; Yang, H.; Hou, Z.G.; Ni, Z.L.; Chen, S.; Fang, Z. Bilinear neural network with 3-D attention for brain decoding of motor imagery movements from the human EEG. Cogn. Neurodyn. 2021, 15, 181–189.

[16] Lee, M.H.; Kwon, O.Y.; Kim, Y.J.; Kim, H.K.; Lee, Y.E.; Williamson, J.; Fazli, S.; Lee, S.W. EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy. GigaScience 2019, 8, giz002.

[17] Alwasiti, H.; Yusoff, M.Z.; Raza, K. Motor imagery classification for brain computer interface using deep metric learning. IEEE Access 2020, 8, 109949–109963.

[18] Zhao, X.; Zhang, H.; Zhu, G.; You, F.; Kuang, S.; Sun, L. A Multi-Branch 3D Convolutional Neural Network for EEG-Based Motor Imagery Classification. IEEE Trans. Neural Syst. Rehabil. Eng. 2019, 27, 2164–2177.

[19] Liu, T.; Yang, D. A Densely Connected Multi-Branch 3D Convolutional Neural Network for Motor Imagery EEG Decoding. Brain Sci. 2021, 11, 197.

[20] Elman, J.L. Finding structure in time. Cogn. Sci. 1990, 14, 179–211.

[21] Luo, T.; Chao, F. Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network. BMC Bioinf. 2018, 19, 344.

[22] Tayeb, Z.; Fedjaev, J.; Ghaboosi, N.; Richter, C.; Everding, L.; Qu, X.; Wu, Y.; Cheng, G.; Conradt, J. Validating deep neural networks for online decoding of motor imagery movements from EEG signals. Sensors 2019, 19, 210.

[23] Ashish, V.; Noam, S.; Niki, P.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. Adv. Neural Inf. Process. Syst. 2017, 30, 5998–6008.

[24] Brunner, C.; Leeb, R.; Muller-Putz, G.; Schlogl, A.; Pfurtscheller, G. BCI Competition 2008—Graz Data Set A; Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces); Graz University of Technology: Graz, Austria, 2008; Volume 16, pp. 1–6.

[25] Li H, Ding M, Zhang R. Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network. Biomedical signal processing and control, 2022, 72: 103342.

[26] Malan, N.; Sharma, S. Motor imagery EEG spectral-spatial feature optimization using dual-tree complex wavelet and neighbourhood component analysis. IRBM 2022, 43, 198–209.

[27] Zhang D, Li H, Xie J. MI-CAT: A Transformer-Based Domain Adaptation Network for Motor. Available at SSRN 4331172.

[28] Lin, M.; Chen, Q.; Yan, S. Network In Network. arXiv 2013, arXiv:1312.4400.

[29] Lawhern, V.J.; Solon, A.J.; Waytowich, N.R.; Gordon, S.M.; Hung, C.P.; Lance, B.J. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. J. Neural Eng. 2018, 15, 056013.

[30] Kang, H.; Nam, Y.; Choi, S. Composite common spatial pattern for subject-to-subject transfer. IEEE Signal Process. Lett. 2009, 16, 683–686.

[31] Ang, K.; Chin, Z.Y.; Zhang, H.; Guan, C. Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface. In Proceedings of the IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008.

[32] Schirrmeister, R.T.; Springenberg, J.T.; Fiederer, L.D.J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; Ball, T. Deep learning with convolutional neural networks for EEG decoding and visualization. Hum. Brain Mapp. 2017, 38, 5391–5420.

[33] Zhao, X.; Liu, D.; Ma, L.; Liu, Q.; Chen, K.; Xie, S.; Ai, Q. Deep CNN model based on serial-parallel structure optimization for four-class motor imagery EEG classification. Biomed. Signal Process. Control. 2021, 72, 103338.