Motor Imagery EEG Classification Based on a Weighted Multi-branch Structure Suitable for Multisubject Data

Huiyang Wang, Jiuchuan Jiang[+], John Q. Gan, and Haixian Wang*, *Senior Member, IEEE*

***Abstract*— *Objective*: Electroencephalogram (EEG) signal recognition based on deep learning technology requires the support of sufficient data. However, training data scarcity usually occurs in subject-specific motor imagery tasks unless multisubject data can be used to enlarge training data. Unfortunately, because of the large discrepancies between data distributions from different subjects, model performance could only be improved marginally or even worsened by simply training on multisubject data. *Method*: This paper proposes a novel weighted multi-branch (WMB) structure for handling multisubject data to solve the problem, in which each branch is responsible for fitting a pair of source-target subject data and adaptive weights are used to integrate all branches or select branches with the largest weights to make the final decision. The proposed WMB structure was applied to six well-known deep learning models (EEGNet, Shallow ConvNet, Deep ConvNet, ResNet, MSFBCNN, and EEG_TCNet) and comprehensive experiments were conducted on EEG datasets BCICIV-2a, BCICIV-2b, high gamma dataset (HGD) and two supplementary datasets. *Result*: Superior results against the state-of-the-art models have demonstrated the efficacy of the proposed method in subject-specific motor imagery EEG classification. For example, the proposed WMB_EEGNet achieved classification accuracies of 84.14%, 90.23%, and 97.81% on BCICIV-2a, BCICIV-2b and HGD, respectively. *Conclusion*: It is clear that the proposed WMB structure is capable to make good use of multisubject data with large distribution discrepancies for subject-specific EEG classification.**

***Index Terms*— EEG decoding, Motor Imagery, Brain-computer interfaces, Deep Learning, Data Distribution**

I. Introduction

Electroencephalogram (EEG) signal decoding is a crucial task in a brain-computer interface (BCI) system, which translates observed brain activities into meaningful information to communicate between the brain and external environments [1]. Motor imagery-based EEG has gained popularity from clinical to industrial applications due to its low clinical risk, low cost, convenience, and no need for stimulus targets. Applications include controlling a wheelchair for the disabled [2], remotely controlling a robot to work [3], and playing computer games for entertainment [4]. EEG-based BCIs have also been developed in the motor recovery field for stroke [5][6]. However, the low signal-to-noise ratio (SNR) of EEG signals, high complexity of brain cognitive processing procedures, and high variance among different subjects limit the capability of motor imagery BCIs to decode mental activities.

Existing work on motor imagery classification can be grouped into classical machine learning and deep learning methods. Classical machine learning methods involve multistage tasks. First, EEG signals must be preprocessed, such as removing artifacts [7] and bandpass filtering [8]. Second, as a crucial step in the classification of EEG signals, handcrafted feature extraction is used to obtain concise information and reduce data dimensionality, such as power spectral density (PSD) [9], entropy feature sets [10], autoregressive (AR) models [11], common spatial pattern (CSP) [12] and its variants including filter band CSP (FBCSP) [13], and multi-time and multi-band CSP [14]. Finally, these features are fed into classifiers such as support vector machine (SVM) [15], linear discriminant analysis (LDA) [16], k-nearest neighbours (KNN) [17], and random forest [18]. However, the performance of these multistage methods is mainly determined by handcrafted features that heavily rely on human selection, which may ignore underlying information from the EEG signals [19]. Feature extraction and classifier optimisation are separated, preventing the situation in which the two stages may promote each other and result in suboptimal classification performance.

Deep learning methods have achieved great success in image classification, speech recognition, object localisation, and so forth. Thus, many studies have attempted to design a deep learning model suitable for EEG signal decoding [20][21], which has outperformed classical machine learning methods. However, the improvements by deep learning are marginal and far below our expectations. Deep learning can shine in other fields but has stagnated in EEG signal recognition because the scarcity of training data results in heavy overfitting problems [22]. Thus, some studies fully used data from different subjects for motor imagery tasks to mitigate the issues of training data scarcity [23][24]. Unfortunately, data from different subjects have large distribution discrepancies, resulting in poor common domain-invariant representations for multisource data. In other words, the information from other subjects may not be helpful for subject-specific tasks.

This work was supported by the National Natural Science Foundation of China under Grant 92270113 and the Key Research and Development Plan (Industry Foresight and Common Key Technology) of Jiangsu Province, China under Grant BE2022157 ([+] Joint first author: Jiuchuan Jiang; * Corresponding author: Haixian Wang).

Huiyang Wang and Haixian Wang are with the Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Science & Medical Engineering, Southeast University, Nanjing 210096, Jiangsu, PR China (e-mail: stickovercarrot@foxmail.com; hxwang@seu.edu.cn).

Jiuchuan Jiang is with the School of Information Engineering, Nanjing University of Finance and Economics, Nanjing 210003, Jiangsu, PR China (e-mail: jcjiang@nufe.edu.cn).

John Q. Gan is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK (e-mail: jqgan@essex.ac.uk).

Furthermore, transfer learning technology cannot be avoided when dealing with data distribution discrepancies. It directly supervises intermediate feature distribution by maximum mean discrepancy (MMD) [25] or uses generative adversarial network (GAN) to confuse models to distinguish the source of samples [26]. However, the performance of the above methods is just passable and still has ample space to develop.

This paper proposes a novel weighted multi-branch (WMB) structure that efficiently fits multisubject features to help improve the classification accuracy of subject-specific tasks for the first time (the implementation code is made publicly available[1]). The core idea of our method is to transform the fitting of multisubject data into the fitting of multiple pairs of source-target data, i.e., a branch is responsible for fitting a couple of source-target data. Concretely, any off-the-shelf model consists of multiple layers of convolutional neural networks (CNNs), which can be separated into base convolution to extract shallow features of all subjects (the sources and the target) and branch convolution to extract deep features of source subject and target subject pairs. The extracted deep features are fed into multi-branch classifiers to obtain a list of class probability distribution vectors. Furthermore, a method for adaptive weighting is proposed to optimise each branch's contributions by adaptively increasing the weights of classifiers with small classification error rates during training. The final decision is based on the weighted sum of each branch's class probability distribution vectors. This study employs three public EEG-based motor imagery datasets to validate our proposed methods in a subject-specific manner. The experimental results have demonstrated the superiority of our proposed methods over different start-of-the-art methods. In addition, insights into the source-target feature distribution affecting the classification results are also analysed and discussed in detail. The following are the main contributions of this paper:

1) This work designs a novel weighted multi-branch structure for recognising EEG-based motor imagery tasks for the first time, which solves the problem that a single-branch structure has difficulty fitting features of multiple subjects, resulting in failure to improve subject-specific tasks.

2) This work proposes an adaptive weighting method for combining class probability distribution vectors, which plays a role in optimising the contributions of each branch. During the model inference time, the branches with the most significant weights can be reserved to reduce the number of parameters and computational complexity of the model to achieve a balance between performance and compactness.

3) It has been demonstrated by visualising that the ineffectiveness of multisource EEG data for subject-specific tasks is attributed to the problem that a single-branch structure cannot solve the feature distribution discrepancy of multiple subjects.

4) Comprehensive experiments using six widely used EEG deep learning models have demonstrated the superior performance of our proposed methods on three EEG-based motor imagery datasets and two supplementary datasets.

[1] https://github.com/stickOverCarrot/WMB_EEGNet

II. Related Work

In previous studies on EEG classification, many classical deep learning models or structures for computer vision (CV) or natural language processing (NLP) were applied to classify motor imagery tasks. Schirrmeister et al. [19] designed a feature extraction module based on a residual structure to obtain results as good as FBCSP. Lawhern et al. [27] proposed a novel network called EEGNet based on depth-wise separable convolution with approximately 2000 trainable parameters, creating a balance between performance and compactness. Tang and Zhang [28] proposed a channel-projection mixed-scale CNN, which refers to the densely connected structure, to improve the transmission of EEG features and reduce trainable parameters. Li et al. [29] proposed fusing the CNN and long short-term memory (LSTM) units in parallel to extract temporal and spatial features of EEG signals at the same time. Ayca et al. [30] used an image processing technique based on GoogLeNet to control a robot manipulator. Lun et al. [31] proposed a GCNs_Net with the aid of graph neural networks (GNN) and achieved an excellent capacity for decoding EEG signals.

However, the problem is the scarcity of EEG-based motor imagery datasets, which hinders the performance of the above excellent models. For example, the BCICIV2a dataset [32], widely used in motor imagery tasks as a benchmark, has no more than 300 training samples for each subject, which may result in a heavy overfitting problem. Data augmentation may be a promising method to solve this problem. Tang and Zhang [28] added noise to spectral images' amplitudes, called amplitude-perturbation data augmentation, to improve deep learning model robustness. Mattioli et al. [33] used the synthetic minority oversampling technique (SMOTE) to create synthetic data mainly for minority classes, achieving better classification results. Yang et al. [34] proposed a data augmentation method based on the circular translation strategy without losing any information about the original samples or introducing any extra noise. Lu et al. [35] created a set of crops using a sliding window to expand the dataset. The above methods can be summarised as changing original samples to generate 'new samples'. However, these 'new samples' may not work due to noise or overlapping information. Thus, as mentioned in the introduction section, how to fully use data from source subjects to improve the performance of models for a target subject is more potential and practical.

In previous EEG studies, models with a multi-branch structure are usually adapted to multi-band and multi-scale signals. Amin et al. [36] proposed multi-branch CNNs for motor imagery classification. Each branch has a different network depth, with EEG signals of various frequency bands, which extracts domain-specific knowledge to build class-discriminative generic features to improve EEG decoding. Jia et al. [37] proposed multi-branch CNNs composed of five parallel inception networks to solve the problems that the best convolution scale varies with different subjects but differs from time to time. Li et al. [29] used CNN and LSTM as two branches for extracting multi-view features. These multi-branch models branch from the first layer and may bring too much computational burden because of the high dimensionality of EEG signals. Wu et al. [38] and Mouad et al. [39] designed a multi-branch model that branches from the middle layer, but the heavy

computational burden still exists. Models with a multi-branch structure were also used in non-EEG studies. YOLOv3 [40] has three branches composed of feature pyramid networks (FPN) in the last layer for obtaining feature maps with different resolutions, which is helpful for object detection with different sizes. Wang et al. [41] designed a model branching from the last few layers for person re-identification. Each branch is responsible for mining discriminative information with various granularities. Due to the dimensionality reduction by the previous multi-layer networks, the subsequent multi-branch networks have less computational burden.

Inspired by the above models, this paper proposes a multi-branch structure with branching from the last few layers to avoid a heavy computational burden. The focus is on the relationship between multi-branch and multi-subject, and the multi-branch structure is used to eliminate the problem of data distribution discrepancies.

III. Methods

A. Definitions and Notations

An EEG dataset for a specific subject can be defined as $S=\{(x_i,y_i),i=1,2,\cdots,M\}$, where $x_i \in R^{C\times T}$ represents a raw sample from the *i-th* trial and has T discretised time points and C electrodes, M is the total number of trials, and $y_i \in \{1,2,\cdots,K\}$ is the corresponding label of x_i, where K is the total number of classes. For a motor imagery task, it is ideal to find a perfect mapping that automatically assigns the correct label to x_i. The decoder model can be mathematically formalised as

$$\tilde{y}=f(x_i;\theta) \tag{1}$$

where $\tilde{y}$ is the predicted label corresponding to x_i and θ represents all of the trainable parameters in the decoder model. If the decoder model is based on a neural network, then Formula (1) can be rewritten as

$$h^l=\sigma\left(h^{l-1}*W^l+b^l\right) \tag{2}$$

$$\tilde{y}=softmax\left(h^L\times W^*+b^*\right) \tag{3}$$

where h^l represents the output from the *l-th* convolution layer whose weights and bias are defined by W^l and b^l, respectively; $h^0=x$ and σ include other common operations such as batch normalisation, linear or nonlinear activation, and pooling; L is the total number of convolution layers; W^* and b^* are the weights and bias of the fully connected layer, respectively; * represents the convolution operation; and × represents matrix multiplication. Therefore, $\theta=\{W^l,W^*,b^l,b^*;\ l=1,2,\cdots,L\}$. The softmax function activates the final output, a vector with K dimensions representing the probability distribution for K classes. In the supervised classification task for motor imagery EEG signal decoding, the learning process minimises the cross-entropy loss L, as shown in Formula (4), to optimise the decoder model

$$L=\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{K}-\delta(j==y_i)\log\tilde{y}_j+\lambda|\theta|^2 \tag{4}$$

where B is the batch size, δ represents the signal function, $|\bullet|^2$ represents the regularisation component, and λ is the trade-off regularisation weight.

B. Detailed Architecture of the Proposed Approach

The overall weighted multi-branch structure is shown in Fig. 1, which contains four components: base convolution, branch convolution, branch classifier, and adaptive weights.

Base convolution: EEG signals from a target subject and all source subjects are concatenated batch-wise as the input to the base convolution to obtain shallow target features f_T^s and a list of shallow source features $f_{S,1}^s, f_{S,2}^s,\cdots, f_{S,\mathrm{n}}^s$, respectively. Multiply accumulate operations (MACCs) of many EEG deep learning models are mainly determined by temporal convolution (the first layer of convolution). Thus, computational complexity increases significantly if the base convolution adopts a multi-branch structure.

Branch convolution: This component adopts a multi-branch structure, and the number of branches is consistent with the number of source subjects. f_T^s and $f_{S,i}^s$ are concatenated batch-wise as the input to the i-th branch to obtain the i-th target subject's deep features $f_{T,i}^d$ and the i-th source subject's deep features $f_{S,i}^d$, respectively. Finally, $f_{T,1}^d, f_{T,2}^d,\cdots, f_{T,\mathrm{n}}^d$ and $f_{S,1}^d, f_{S,2}^d,\cdots, f_{S,\mathrm{n}}^d$ are obtained through forwarding propagation. Each branch can exclusively learn the common feature representation of a pair of target subject and source subject. The deep feature distribution of the target subject in each branch is different, equivalent to obtaining a multi-view feature representation for the target subject.

Branch classifier: This component adopts a multi-branch structure in which classifiers and branches of branch convolution are in one-to-one correspondence. $f_{T,i}^d$ and $f_{S,i}^d$ are input to the *i-th* branch of the branch classifier and are activated by the softmax function to obtain the *i-th* class probability distribution vector $p_{T,i}$ of the target subject and the class probability distribution vector $p_{S,i}$ of the *i-th* source subject, respectively. The deep learning model is trained by minimising the cross-entropy loss, with Formula (4) rewritten as follows:

$$L_S=\frac{1}{nB}\sum_{i=1}^{n}\sum_{j=1}^{B}\sum_{k=1}^{K}-\delta\left(k==y_{S,i}^j\right)\log p_{S,i}^{j,k} \tag{5}$$

$$L_T=\frac{1}{nB}\sum_{i=1}^{n}\sum_{j=1}^{B}\sum_{k=1}^{K}-\delta\left(k==y_{T,i}^j\right)\log p_{T,i}^{j,k} \tag{6}$$

where n is the number of branches, B is the batch size, K is the number of classes, $y_{S,i}^j$ and $y_{T,i}^j$ are the labels, and δ represents the signal function.

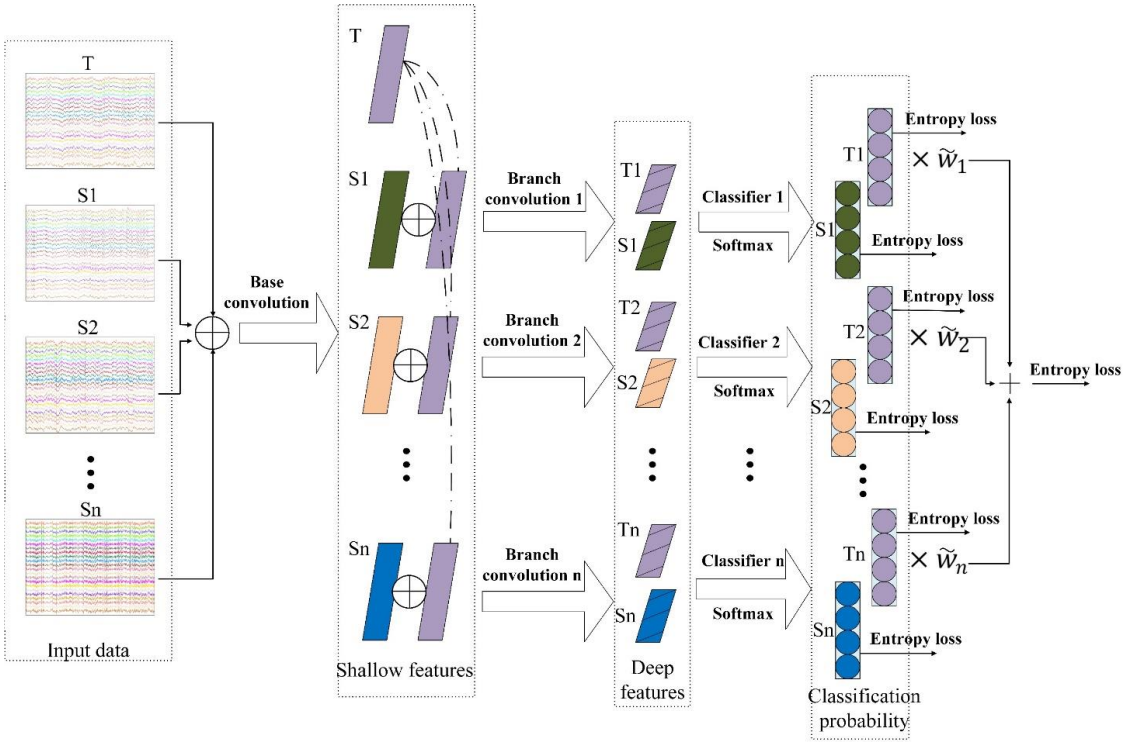

Fig. 1. Weighted multi-branch structure. $\oplus$ represents batchwise concatenation, and + represents elementwise addition; $\tilde{w}_1, \tilde{w}_2, \cdots, \tilde{w}_n$ are adaptive weights; T represents target subject (specific subject) and S1, S2, ⋯, Sn represent source subject.

Adaptive weights: A list of the target's class probability distribution vectors $p_{T,1}, p_{T,2}, \cdots, p_{T,n}$ are obtained from the branch classifiers. A simple method to make the final prediction is to average the vectors, i.e., $\frac{1}{n}\sum_{i=1}^{n} p_{T,i}$. However, this method has two drawbacks: On the one hand, the contribution of each branch to the EEG signal classification is different so that the average is not optimal; on the other hand, if there are too many source subjects, there will be too many branches, which may significantly increase the number of parameters and computational complexity of the model. It goes against the real-time performance and compactness of portable BCI systems. However, assigning a weight to each branch can solve the above problems. Our first thought is to use an additional validation set to test the classification accuracy of each branch and determine the weight values according to their accuracy. However, it is too expensive to separate part of the data to assess weights because the data amount is usually small. This paper proposes an adaptive weighting method for combing the class probability distribution vectors. The weight of each branch can be adjusted by gradient descent along with all model parameters. Given n weights $w_1, w_2, \cdots, w_n$, the softmax function is used to make the sum of all the weights equal 1:

$$\tilde{w}_i = \frac{\exp(w_i)}{\sum_{j=1}^{n}\exp(w_j)} \tag{7}$$

The final prediction of the target object is obtained by weighted summation as follows:

$$p_T = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i p_{T,i} \tag{8}$$

Again cross-entropy loss is used here, with Formula (4) rewritten as follows:

$$\hat{L}_T = \frac{1}{B}\sum_{j=1}^{B}\sum_{k=1}^{K} -\delta\left(k == y_T^j\right)\log p_T^{j,k} \tag{9}$$

Therefore, updating w_i can be described by the following Formula:

$$w_i \leftarrow w_i - \eta \frac{\partial \hat{L}_T}{\partial p_T}\left(\frac{\partial p_T}{\partial \tilde{w}_i}\frac{\partial \tilde{w}_i}{\partial w_i} + \sum_{j=1, j\neq i}^{n}\frac{\partial p_T}{\partial \tilde{w}_j}\frac{\partial \tilde{w}_j}{w_i}\right) \tag{10}$$

where η represents the learning rate. Adding the regularisation term and combining it with Formulae (5) and (6), the final total loss is

$$L_{all} = L_S + L_T + \hat{L}_T + \lambda\left|\theta\right|^2 \tag{11}$$

In the reference phase, either Formula (8) can be directly used for prediction, or the branches with the most significant weights can be selected first to reduce the running complexity of the model. When a branch with the largest weight is reserved, the number of parameters and running time are consistent with the original model (baseline).

C. The Proposed Model WMB_EEGNet

EEGNet [27], a widely used EEG classification model, is adopted as the backbone for various experiments in this paper, and EEGNet based on the weighted multi-branch structure is named as WMB_EEGNet. The WMB_EEGNet implementation details are shown in the **supplementary Table VII**. Specifically, EEGNet is composed of three convolutional layers (temporal_conv, depthwise_conv, and separable_conv) and a fully connected layer (classifier). In WMB_EEGNet, the temporal_conv and depthwise_conv are combined as base convolution for extracting common shallow features of all subjects, and the separable_conv is used as branch convolution to extract

common deep features of a pair of target subject and source subject. Since the separable_conv in each branch is updated by different source subject data, resulting in its different weights, which leads to extract multi-view feature representation for the same target subject data. The fully connected layer is used as a branch classifier to output the class probability distribution vector. Each fully connected layer independently outputs a class probability distribution vector for the corresponding feature. Combining the class probability distribution vectors with adaptive weights results in the final decision.

The parameters $F1$ and $F2$ in WMB_EEGNet control the number of filters in base convolution and branch convolution, respectively, and they will affect the classification performance of WMB_EEGNet. In the default experimental setting, $F2 = 2 \times F1$, and $F1$ and $F2$ are to 16 and 32, respectively. For a fair comparison, $F1$ and $F2$ are also set to 16 and 32 for EEGNet, respectively. The following are the variants of WMB-EEGNet with non-default settings:

WMB_EEGNet $(\times a)$: The values of $F1$ and $F2$ are a times of their default values, e.g., $a = 1.5$ means that $F1$ and $F2$ are 24 and 48, respectively.

WMB_EEGNet (variable): The optimal parameter values for different target subjects may differ. Some previous studies have found optimal subject-specific network parameters to obtain the best classification results [42][43]. For example, EEG_TCNet [43] needs to consider kernel size, drop rate, number of filters, and other hyperparameters. In the experiments in this paper, only the value of $F1$ is optimised for different target subjects.

WMB_EEGNet_*b*: In the inference phase, only the branches with weights ranked in the top *b* will be reserved to reduce the number of parameters and computational complexity.

Apart from EEGNet, we also applied WMB to other baseline models, including Shallow ConvNet [19], Deep ConvNet [19], ResNet [19], MSFBCNN [38], and EEG_TCNet [43] to demonstrate the universality of our proposed method.

D. *Datasets and Implementation Details*

BCICIV-2a Dataset [32], BCICIV-2b Dataset [44] and High-Gamma Dataset (HGD) [19] are used as the primary materials. The datasets details, implementation details and evaluation metric refer to **Section A of supplementary document**. Moreover, Upper Limb Movement Dataset (ULMD) [45] and P300 Dataset are supplementary, and the datasets details and relevant experiments are shown in **Section E of supplementary document**.

IV. Results

A. *WMB_EEGNet* vs. *State-of-the-Art Models*

Deep learning is an end-to-end technique that outperforms traditional methods relying on mining handcrafted features. To make a fair comparison, ten state-of-the-art models with various structures and almost the same input data strategies are compared with WMB_EEGNet in this paper to embody the superior performance of WMB_EEGNet. On the one hand, some models, including ResNet, Shallow ConvNet, Deep ConvNet, EEGNet, EEG_TCNet, and FBCNet with source code publicly available, are directly reproduced. On the other hand, for some models without open source code, including SCCNet, MSFBCNN, Incep-EEGNet, and CNN+LSTM, we carefully followed the detailed implementation details described in the original papers to reproduce them. Luckily, all reproduced results are no worse than those reported in the original papers. Those models are generally described as follows: 1) ResNet, composed of 20 residual network blocks, has the deepest structure among these models. 2) Shallow ConvNet contains only three layers (two convolution layers and one fully connected layer), and its advantage derives from using fewer parameters in exchange for generalisability. 3) Deep ConvNet is a deeper version of shallow ConvNet, and it has more trainable parameters but fewer MMACs. 4) EEGNet takes advantage of depth-wise separable convolution to sharply decrease the number of trainable parameters while retaining the competitive EEG decoding results. 5) SCCNet contains only two convolution layers, with the first being the spatial convolution layer and the second the temporal convolution layer, which is in the reverse order of the two convolution layers of Shallow ConvNet. 6) MSFBCNN adopts a parallel network structure where the filter bank is multi-scale to mine optimised features along the temporal space of raw EEG signals. 7) EEG_TCNet combines EEG_Net with a temporal convolutional network (TCN) capable of exponentially extending the receptive field size while increasing the number of parameters linearly without suffering from exploding or vanishing gradient issues, which contrasts with other time-series networks such as LSTM networks. 8) Incep-EEGNet adopts a parallel network structure based on the inception block to extract information along the temporal space of features from the former layer. 9) FBCNet creates a multi-view representation of the broad-band EEG data wherein each view represents a narrow-band localised EEG. It utilises the temporal variance operation, which represents the spectral power (ERD/ERS) in the given time series, instead of temporal convolution to extract the temporal information effectively. 10) CNN+LSTM is no longer a novel combination, while its parallel structure may be more suitable to decode EEG signals than serial structure. Besides, FBCSP is used as a traditional handcrafted feature extraction method to compare with our proposed method, and SVM and LDA are used as classifiers, respectively.

Tables I-III present the results of WMB_EEGNet on BCICIV-2a, BCICIV-2b, and HGD, respectively, compared with ten state-of-the-art deep learning models and two traditional machine learning methods. All the deep learning models have surpassed FBCSP on the three datasets, showing the advantages of deep learning methods. It can be seen from Table II that WMB_EEGNet outperformed other models in terms of classification accuracy and k-score on BCICIV-2a dataset. WMB_EEGNet achieved the highest average accuracy of 84.14% and the highest average k-score of 0.79, at least 2.31% and 0.03 higher than the other models, respectively. WMB_EEGNet achieved the best results for Subjects 1-3, 6, and 8, with the accuracies improved by at least 3.18%, 1.15%, 1.04%, 3.13%, and 3.82%, respectively. For the standard deviation of the two indices, WMB_EEGNet is second only by CNN+LSTM, proving its stability and robustness when classifying different subjects. A similar trend was observed on other

TABLE I
RESULTS ON THE BCICIV2A DATASET

Year	Subject		1	2	3	4	5	6	7	8	9	Mean	Std	p-value
2012	FBCSP+SVM* [11]	Acc%	65.62	56.94	65.28	62.15	25	47.92	75.35	64.24	62.15	58.29	13.65	<0.001
		k-scores	0.54	0.43	0.54	0.5	0	0.31	0.67	0.52	0.5	0.44	0.18	
2012	FBCSP+LDA* [11]	Acc%	68.4	48.96	65.97	58.68	26.04	43.75	72.22	68.4	52.08	56.06	14.1	<0.001
		k-scores	0.58	0.32	0.55	0.45	0.01	0.25	0.63	0.58	0.36	0.41	0.19	
2017	Shallow ConvNet* [19]	Acc%	84.38	62.15	90.28	73.96	70.49	56.94	91.67	86.11	81.94	76.48	12.38	0.002
		k-scores	0.79	0.49	0.87	0.65	0.60	0.43	0.88	0.81	0.75	0.70	0.16	-
2017	Deep ConvNet* [19]	Acc%	77.08	55.90	90.97	74.31	**78.82**	68.40	90.97	81.25	82.29	77.77	10.95	0.004
		k-scores	0.76	0.39	0.79	0.71	0.69	0.55	0.89	0.78	0.79	0.71	0.16	-
2017	ResNet*[19]	Acc%	77.45	56.60	86.46	61.11	54.51	51.74	83.68	75.69	75.69	69.21	13.25	0.002
		k-scores	0.71	0.41	0.83	0.48	0.41	0.34	0.78	0.69	0.67	0.59	0.18	
2018	EEGNet* [27]	Acc%	81.25	66.32	96.18	72.92	72.92	60.76	87.14	87.85	**88.89**	79.36	11.78	0.004
		k-scores	0.75	0.55	0.95	0.64	0.64	0.48	0.83	0.84	**0.85**	0.73	0.16	-
2019	SCCNet* [23]	Acc%	86.11	67.01	90.62	75.00	57.29	58.68	84.03	84.38	80.90	76.00	12.28	0.002
		k-scores	0.81	0.56	0.88	0.67	0.43	0.45	0.79	0.79	0.75	0.68	0.16	
2019	MSFBCNN* [38]	acc%	84.03	61.11	93.4	76.39	73.61	60.07	82.29	86.11	80.90	77.55	11.14	0.002
		k-scores	0.79	0.48	0.91	0.69	0.65	0.47	0.76	0.81	0.75	0.70	0.15	-
2020	EEG_TCNet [43]	Acc%	85.77	65.02	94.51	64.91	75.36	61.4	87.36	83.76	78.03	77.35	11.58	0.002
		k-scores	0.81	0.53	0.93	0.53	0.67	0.49	0.83	0.78	0.71	0.70	0.15	-
2020	Incep-EEGNet [39]	acc%	86.11	57.29	93.75	77.78	57.99	60.42	**95.83**	84.72	82.64	77.39	15.14	0.010
		k-scores	0.81	0.43	0.92	0.7	0.44	0.47	**0.94**	0.8	0.77	0.7	0.2	
2021	FBCNet* [46]	Acc%	88.89	64.24	95.49	**84.72**	75.35	62.15	94.10	84.38	87.15	81.83	12.09	0.049
		k-scores	0.85	0.52	0.94	**0.80**	0.67	0.50	0.92	0.79	0.83	0.76	0.16	
2022	CNN+LSTM (parallel)* [29]	Acc%	81.25	64.93	89.58	71.18	74.31	67.01	77.43	85.07	84.72	77.28	**8.57**	0.002
		k-scores	0.75	0.53	0.86	0.62	0.66	0.56	0.70	0.80	0.80	0.70	**0.11**	
2023	WMB_EEGNet	Acc%	**92.01**	**68.06**	**97.22**	76.74	78.47	**70.14**	**94.44**	**91.67**	88.54	**84.14**	10.94	-
		k-scores	**0.89**	**0.57**	**0.96**	0.69	**0.71**	**0.60**	**0.93**	**0.89**	**0.85**	**0.79**	0.15	-

The best results are marked in bold. * Reproduced.

TABLE II
RESULTS ON THE BCICIV2B DATASET

Year	Subject		1	2	3	4	5	6	7	8	9	Mean	Std	p-value
2012	FBCSP+SVM*	Acc%	60.94	56.43	55.94	95	78.44	78.75	78.12	88.44	75.94	74.22	12.99	<0.001
		k-scores	0.22	0.13	0.12	0.9	0.57	0.57	0.56	0.77	0.52	0.48	0.26	
2012	FBCSP+LDA*	Acc%	60	54.29	50.94	94.38	80.31	81.56	75.62	88.44	77.5	73.67	14.34	<0.001
		k-scores	0.2	0.09	0.02	0.89	0.61	0.63	0.51	0.77	0.55	0.47	0.29	
2017	Shallow ConvNet*	Acc%	76.25	66.07	78.75	97.19	98.75	86.88	88.75	93.12	85.31	85.67	10.55	0.002
		k-scores	0.53	0.32	0.57	0.94	0.97	0.74	0.78	0.86	0.71	0.71	0.21	
2017	Deep ConvNet*	Acc%	78.75	72.14	84.06	97.5	98.75	85.31	93.44	95	88.75	88.19	8.97	0.006
		k-scores	0.57	0.44	0.68	0.95	0.97	0.71	0.87	0.9	0.78	0.76	0.18	
2017	ResNet*	Acc%	77.19	65.36	70	97.81	96.25	85.94	85	92.5	85	83.89	11.23	0.006
		k-scores	0.54	0.31	0.4	0.96	0.93	0.72	0.7	0.85	0.7	0.68	0.23	
2018	EEGNet*	Acc%	80	73.21	83.12	97.81	97.5	**90.94**	**94.06**	93.12	89.06	88.76	8.37	0.014
		k-scores	0.6	0.46	0.66	0.96	0.95	**0.82**	**0.88**	0.86	0.78	0.78	0.17	-
2019	SCCNet*	Acc%	76.88	70	81.88	95.94	99.06	84.69	91.56	91.88	80.62	85.83	9.5	0.002
		k-scores	0.54	0.4	0.64	0.92	0.98	0.69	0.83	0.84	0.61	0.72	0.19	
2019	MSFBCNN*	acc%	77.5	66.79	78.75	98.12	98.44	88.12	89.69	93.75	88.44	86.62	10.47	0.004
		k-scores	0.55	0.34	0.57	0.96	0.97	0.76	0.79	0.88	0.77	0.73	0.21	
2020	EEG_TCNet*	Acc%	80.31	**75**	80.62	**98.44**	99.06	87.81	92.5	95	85.31	88.23	8.59	0.025
		k-scores	0.61	**0.5**	0.61	**0.97**	0.98	0.76	0.85	0.9	0.71	0.76	0.17	
2020	Incep-EEGNet	acc%	78.12	61.07	64.06	95	98.12	87.5	81.56	91.56	83.44	82.27	12.88	0.002
		k-scores	0.56	0.22	0.28	0.9	0.96	0.75	0.63	0.83	0.67	0.65	0.26	
2021	FBCNet*	Acc%	79.06	60.36	70	96.88	95.94	89.06	82.19	92.81	88.75	83.89	12.33	0.002
		k-scores	0.58	0.21	0.4	0.94	0.92	0.78	0.64	0.86	0.78	0.68	0.25	
2022	CNN+LSTM (parallel)*	Acc%	76.25	72.14	82.5	98.12	91.25	75.62	90.31	**95.31**	85.62	85.24	9.23	0.010
		k-scores	0.53	0.44	0.65	0.96	0.82	0.51	0.81	**0.91**	0.71	0.7	0.18	
2023	WMB_EEGNet	Acc%	**84.69**	73.93	**84.38**	97.81	**100**	90.31	**94.06**	95	**91.88**	**90.23**	**8.09**	-
		k-scores	**0.69**	0.48	**0.69**	0.96	**1**	0.81	**0.88**	0.9	**0.84**	**0.8**	**0.16**	-

The best results are marked in bold. * Reproduced.

datasets. From Table II, it can be seen that WMB_EEGNet achieved the best results on BCICIV-2b dataset, with the highest average accuracy of 90.23%, highest average k-score of 0.80, and lowest standard deviation of both indices among all models. To further demonstrate the adaptability of the proposed WMB_EEGNet, the performance of WMB_EEGNet was compared with other models on dataset HGD. Table III shows that WMB_EEGNet is superior to the start-of-the-art models in terms of classification accuracy and k-score on HGD, which achieved the best classification accuracy for eight subjects, and the accuracies for six subjects reached 100%. Furthermore, WMB_EEGNet can be considered as an enhanced version of EEGNet. Compared with EEGNet, the average accuracies of WMB_EEGNet increases by 4.78%, 1.47%, and 3.53%, respectively, and the average k-scores increases by 0.06, and 0.02 0.05, on the three datasets. Additionally, the standard deviation of the two indices decreases.

TABLE III
RESULTS ON HGD

Year	Subject		1	2	3	4	5	6	7	8	9	10	11	12	13	14	Mean	Std	p-value
2012	FBCSP+ SVM*	Acc%	61.9	68.8	93.1	89.4	75.6	68.8	66.0	78.8	50.6	65	61.9	81.9	76.1	57.5	71.1	11.6	<0.001
		k-scores	0.49	0.58	0.91	0.86	0.68	0.58	0.55	0.72	0.34	0.53	0.49	0.76	0.68	0.43	0.61	0.15	
2012	FBCSP+ LDA*	Acc%	67.5	64.4	89.4	89.4	78.1	70.6	69.2	77.5	49.4	71.3	63.1	88.1	78.6	59.4	72.6	11.4	<0.001
		k-scores	0.57	0.53	0.86	0.86	0.71	0.61	0.59	0.7	0.32	0.62	0.51	0.84	0.71	0.46	0.63	0.15	
2017	Shallow ConvNet*	Acc%	95.6	97.5	**100**	98.8	98.8	98.8	94.3	98.1	97.5	91.4	**98.8**	96.9	96.2	79.4	95.9	5.2	0.001
		k-scores	0.94	0.97	**1**	0.98	0.98	0.98	0.92	0.97	0.97	0.88	**0.98**	0.96	0.95	0.72	0.94	0.07	
2017	Deep ConvNet*	Acc%	97.50	97.5	97.5	99.4	99.4	97.5	95.6	93.1	96.9	92.5	97.5	98.1	**98.7**	76.9	95.6	97.5	0.002
		k-scores	0.97	0.97	0.97	0.99	0.99	0.97	0.94	0.91	0.96	0.90	0.97	0.97	**0.98**	0.69	0.94	0.97	
2017	ResNet*	Acc%	91.9	90	95.6	98.1	91.9	93.1	93.7	96.3	96.9	93.8	96.3	94.4	89.9	83.8	93.3	3.7	<0.001
		k-scores	0.89	0.87	0.94	0.97	0.89	0.91	0.92	0.95	0.96	0.92	0.95	0.93	0.87	0.78	0.91	0.05	
2018	EEGNet*	Acc%	95.0	96.9	96.9	99.4	94.4	98.1	95.0	95.0	97.5	94.4	85.6	**99.4**	96.2	76.3	94.3	6.2	<0.001
		k-scores	0.93	0.96	0.96	0.99	0.93	0.97	0.93	0.93	0.97	0.93	0.81	**0.99**	0.95	0.68	0.92	0.08	
2019	SCCNet*	Acc%	94.4	96.9	98.8	97.5	96.3	91.9	95.6	95.6	97.5	91.9	81.3	96.3	95.0	70.0	92.8	7.8	<0.001
		k-scores	0.93	0.96	0.98	0.97	0.95	0.89	0.94	0.94	0.97	0.89	0.75	0.95	0.93	0.6	0.9	0.1	
2019	MSFBCNN*	acc%	96.3	97.5	**100**	99.4	96.9	98.1	91.2	95.0	96.9	92.5	99.4	98.1	96.9	**91.3**	96.4	**2.9**	0.045
		k-scores	0.95	0.97	**1**	0.99	0.96	0.97	0.88	0.93	0.96	0.90	0.99	0.97	0.96	**0.88**	0.95	**0.04**	
2020	EEG_TCNet*	Acc%	95.0	94.4	99.4	98.8	96.3	95.0	94.3	93.8	96.9	93.8	83.8	97.5	95.0	80.6	93.9	5.3	<0.001
		k-scores	0.93	0.93	0.99	0.98	0.95	0.93	0.92	0.92	0.96	0.92	0.78	0.97	0.93	0.74	0.92	0.07	
2020	Incep-EEGNet	acc%	98.8	93.1	98.1	**100**	**100**	98.1	**96.9**	98.8	99.4	95	98.1	**99.4**	98.1	66.3	95.7	8.7	0.016
		k-scores	0.98	0.91	0.97	**1**	**1**	0.97	**0.96**	0.98	0.99	0.93	0.97	**0.99**	0.97	0.55	0.94	0.12	
2021	FBCNet*	Acc%	94.4	90.6	99.4	98.8	97.5	95.6	89.9	98.1	93.1	94.3	83.8	98.1	96.9	76.9	93.4	6.4	<0.001
		k-scores	0.93	0.88	0.99	0.98	0.97	0.94	0.87	0.97	0.91	0.93	0.78	0.97	0.96	0.69	0.91	0.09	
2022	CNN+LSTM (parallel)*	Acc%	94.4	92.5	95.6	97.5	93.8	90.0	91.2	93.1	91.9	95.6	78.8	96.3	95.0	71.9	91.2	7.2	<0.001
		k-scores	0.93	0.9	0.94	0.97	0.92	0.87	0.88	0.91	0.89	0.94	0.72	0.95	0.93	0.62	0.88	0.1	
2023	WMB_ EEGNet	Acc%	**100**	**98.8**	**100**	99.4	**100**	**100**	**96.9**	**100**	**100**	**96.9**	98.1	98.1	98.1	83.8	**97.8**	4.38	
		k-scores	**1**	**0.98**	**1**	0.99	**1**	**1**	**0.96**	**1**	**1**	**0.96**	0.97	0.97	0.97	0.78	**0.97**	0.06	

The best results are marked in bold. * Reproduced.

TABLE IV
RESULTS WITH OPTIMAL $F1$ ON THE BCICIV2A DATASET

Subject	1	2	3	4	5	6	7	8	9	mean
$F1$	24	24	16	32	24	32	48	32	32	-
WMB_EEGNet	93.06	71.53	97.22	85.07	84.72	72.22	97.22	92.01	90.28	87.04
WMB_EEGNet_1	92.7	71.53	97.22	84.72	84.38	71.18	97.22	92.01	90.28	86.80

The relationship between the loss and iterations on the three datasets is shown in **supplementary Fig. 4**. It can be seen that WMB_EEGNet and EEGNet almost had the same loss curve, indicating that WMB did not change the convergence trend of the baseline model.

B. Preserve One Branch Only

The branches are weighted for two reasons. On the one hand, weights can play a role in optimising the contributions of every branch. On the other hand, branches can be pruned according to the weight values to reduce the inference complexity. This section considers the situation when only one branch with the maximum weight is preserved. When WMB_EEGNet finishes training and only retains one branch, i.e., its special case WMB_EEGNet_1, whose MMACs and the number of parameters are consistent with EEGNet. Another issue is that the optimal hyperparameters for each subject are different and $F1$ greatly influences the model's performance among all hyperparameters (e.g., dropout rate, learning rate, and kernel size). Hence, we tried to find an optimal subject-specific value of $F1$ for WMB_EEGNet_1, leading to the variable model WMB_EEGNet_1 $(\times a)$.

The results of WMB_EEGNet_1 $(\times a)$ on the BCICIV2a dataset, with a = 1, 1.5, 2, 3, are shown in **supplementary Fig. 5**. It can be seen that even though only one branch with the largest weight of WMB_EEGNet is preserved after training, the classification accuracy is still superior. This means that the branch with the maximum weight dominates the classification decision. Hence, the classification accuracy is still superior when only one branch with the maximum weight is preserved. Except for the branch with the maximum weight, the classification accuracies of the other branches are lower than the baseline, and the corresponding weight ratios are very low. This indicates that branches with low weights make limited contributions to the final decision. However, decent results can still be obtained if branches with the 2nd to 8th largest weight are retained, demonstrating the superiority of a multi-branch structure.

Another obvious result is that the optimal values of $F1$ for different subjects are different. The more intuitive comparison results are listed in Table IV. For example, when $F1$ of WMB_EEGNet_1 is 24, Subject 1 achieved the highest accuracy of 92.7%. The optimal value of $F1$ for Subject 7 is 48, and the corresponding accuracy is 97.22%. Moreover, the difference in subject-level accuracy between WMB_EEGNet and WMB_EEGNet_1 is slight. The mean accuracy of WMB_EEGNet_1 is only 0.24% lower than that of WMB_EEGNet. Consequently, it makes sense to prune branches according to their weights. However, not all cases are extreme. When the ratio of the maximum weight to the sum of all weights is too small, the branch with the maximum weight may fail to achieve good classification performance. More branches may need to be preserved to maintain the superior classification performance in this case. The dataset and structure of a model may affect the distribution of weights. (Please refer to **Section B of supplementary document** for the related research.)

C. MACCs and #Parameters

TABLE V
COMPARISON WITH THE START-OF-THE-ART MODELS IN TERMS OF FIVE METRICS ON THE BCICIV2A DATASET

Method	#Parameters	MACCs	Time(s/inference)	Acc%	Information density
Resnet* [19]	1.41 M	1.59 G	46.15	69.21	0.49
Incep-EEGNet	190.98 k	97.63 M	**2.83**	74.07	3.88
CP-MixedNet [28]	839.81 k	186.24 M	5.41	74.60	0.89
SCCNet*	13.40k	6.64M	**0.19**	76.00	56.72
Shallow ConvNet*	47.36 k	64.09 M	**1.86**	76.18	16.09
CNN+LSTM(serial)* [35]	452.22 k	131.85 M	3.83	76.95	1.7
CNN+LSTM(parallel)*	290.92 k	54.88 M	**1.59**	77.28	2.66
EEG_TCNet(x0.5)	4.27 k	6.80 M	**0.2**	77.35	181.15
EEG_TCNet*	10.97k	15.01M	**0.44**	77.16	70.34
MSFBCNN*	158.56 k	315.25 M	9.15	77.55	4.89
Deep ConvNet*	284.48 k	38.10 M	**1.11**	77.77	2.73
EEGNet(x0.5)*	2.63 k	13.1M	**0.38**	76.51	290.91
EEGNet*	5.60 k	27.33 M	**0.79**	79.36	141.71
CNN++	220.00 k	18.20 M	**0.53**	81.10	3.69
MMCNN [37]	648.45 k	119.00 M	3.45	81.40	1.26
FBCNet*	63.65k	65.13 M	**1.89**	81.83	12.86
TPCT [47]	7.78 M	1.73 G	50.22	88.87	0.11
WMB_EEGNet_1(x0.5)	2.63 k	13.1M	**0.38**	79.25	**301.33**
WMB_EEGNet_1	5.60 k	27.33 M	**0.79**	83.68	149.43
WMB_EEGNet_1(x1.5)	9.72 k	41.11 M	**1.19**	84.41	86.84
WMB_EEGNet_1(x2)	13.25 k	54.95 M	**1.6**	85.75	64.72
WMB_EEGNet_1(x3)	22.95 k	82.86 M	**2.41**	85.10	37.08
DFFN(variable) [42]	1.07 M	132 M	3.83	79.71	0.74
EEG_TCNet(variable)	20.5 k	12.1 M	**0.35**	83.84	40.9
WMB_EEGNet_1(variable)	22.95 k	82.86 M	**2.41**	86.80	37.82

Results with inference time less than 3 s are marked in bold. Information density = 10^6*accuracy (between 0 and 1)/#parameters, which is used to evaluate the balance between accuracy and the number of parameters.

Compared to baseline models, models with multi-branch structure increase the MACCs and #parameters. The improvement of model classification performance at the cost of low compactness is sometimes infeasible. For real-time and privacy, some motor imagery tasks need to be carried out on a portable device with a low-power microcontroller unit (MCU) rather than on a cloud server.

Fortunately, we can still solve this problem by pruning branches. According to the study in the previous sections, it is possible only to preserve the branch with the maximum weight to achieve superior classification performance but keep MMACs and the number of parameters consistent with the baseline. Table V compares the state-of-the-art models in terms of the metrics on the BCICIV2a dataset. It is also shown in **supplementary Fig. 6**. For the variable models, we followed [43] to report the maximum number of parameters and MACCs, as they pose a hard requirement when models are transplanted to a MCU.*1) Fixed model*. With an increased magnification of $F1$, the MACCs and number of parameters of WMB_EEGNet_1 increase, while WMB_EEGNet_1 (x2) achieves the highest accuracy of 85.75% among the five models tested with different $F1$. It is 3.12% below the TPCT (A advanced deep learning model using the information of electrode locations achieves the highest accuracy on the BCICIV2a dataset so far), whose accuracy is 88.87%. However, the highest accuracy of TPCT comes at the cost of low compactness. The MACCs and number of WMB_EEGNet_1 (x2) parameters are 31.5 and 587.2 times lower than those of TPCT. For example, referring to the throughput of 34.45 MMAC/s of an ARM M7 processor [48], TPCT needs to take 50.22 s per inference, but WMB_EEGNet_1 (x2) takes only 1.60 s per inference. If the minimum real-time requirement is 3 s per inference, TPCT does not satisfy the condition, but our method does. In other words, WMB_EEGNet_1 (x2) achieved the highest accuracy under the conditions of real-time inference. WMB_EEGNet_1(x0.5) has the highest information density, and its accuracy is nearly 80%. w.*2) Variable model.* EEG_TCNet (variable) is WMB_EEGNet_1 (variable) 's main opponent among the three variable models. Regarding the number of parameters and MACCs, WMB_EEGNet_1 (variable) is higher than EEG_TCNet (variable). Although the inference time of EEG_TCNet (variable) is 6.9 times faster than WMB_EEGNet_1 (variable), WMB_EEGNet_1(variable) still meets the real-time condition and achieves the highest accuracy, 2.96% higher than EEG_TCNet (variable). WMB_EEGNet_1 (variable) 's information density is only 3.08 lower than that of EEG_TCNet (variable).

Table VI
COMPUTATIONAL COST ON THE BCICIV-2A DATASET

	Times	#Operations
EEGNet	2.01s	4.7x10^10
WMB_EEGNet	2.17s	4.5x10^11

We provided the training time per epoch of EEGNet and WMB_EEGNet during model training on an RTX 3080 GPU. In addition, we directly counted The number of operations (additions and multiplications) in EEGNet and WMB_EEGNet per epoch and the total number of forward and backward passes during training can be estimated as follows:

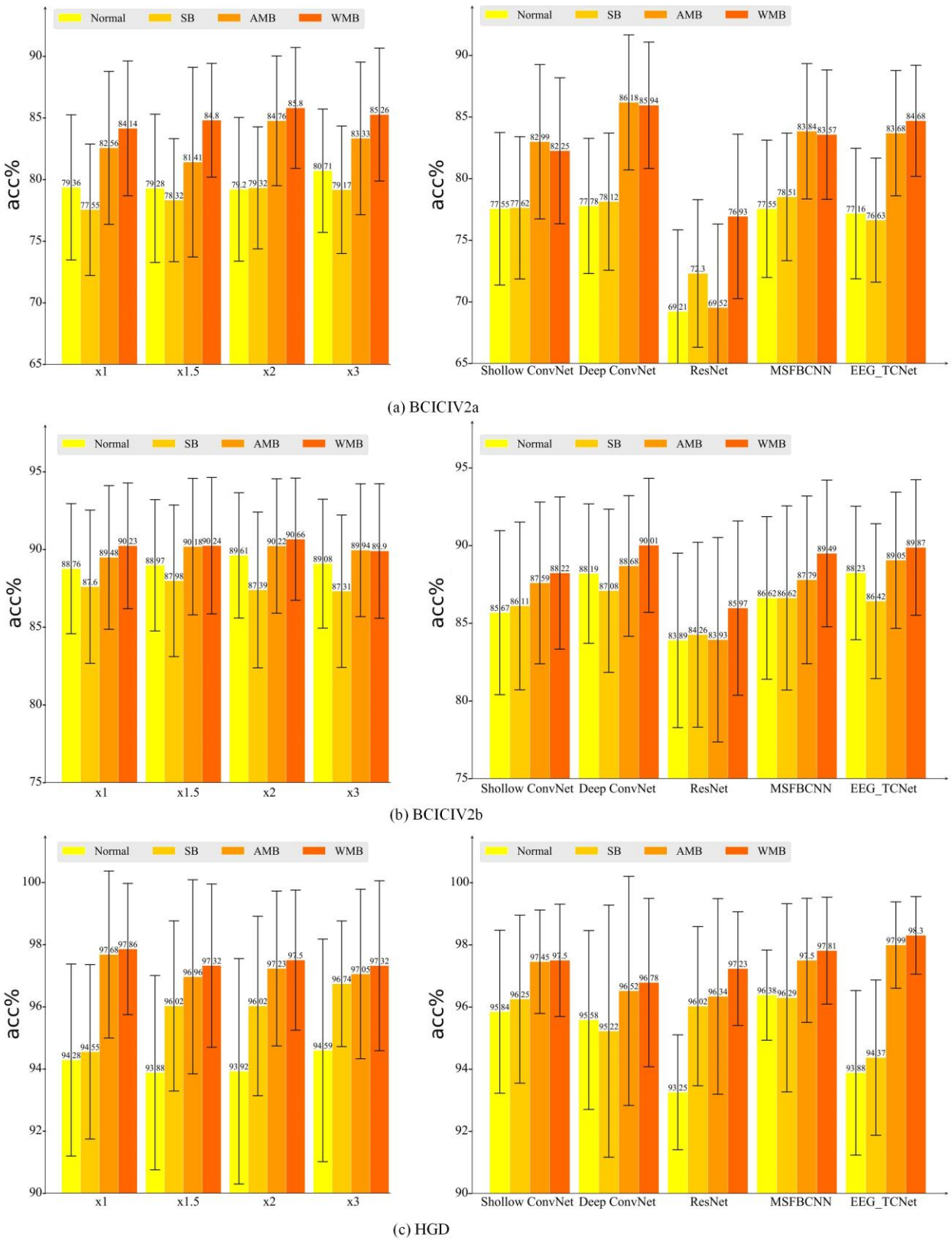

Fig. 2. Results of EEGNet (with four different $F1$), Shallow ConvNet, Deep ConvNet, ResNet, MSFBCNN, and EEG_TCNet under four conditions – normal, SB, AMB, and WMB, respectively. (a) Results on BCICIV2a dataset. (b) Results on BCICIV2b dataset. (c) Results on HGD.

$$\begin{aligned} &\#Operations = MACCs \times 2 \times 3 \\ &(\text{for forward and backward pass}) \times N_e \end{aligned} \tag{12}$$

where N_e is the number of examples in a dataset. The training time per epoch of EEGNet and WMB_EEGNet during model training on an RTX 3080 GPU was recorded.

Table VI shows the computational cost on the BCICIV-2a dataset. The #operations of WMB_EEGNet is almost 10 times more than that of EEGNet, but the training time per epoch only increases by 0.16s.

D. WMB vs. AMB vs. SB

Section A shows the superior performance of WMB_EEGNet. In this section, we not only applied WMB to more baseline models, including Shallow ConvNet, Deep ConvNet ResNet, MSFBCNN, and EEG_TCNet to demostrate the universality of our proposed method, but comparied WMB with averaging multi-branch (AMB) and single branch (SB) method to show the importance of weighting in optimising the contributions of every branch. For AMB, the final prediction is to average all the multi-branch's output. For SB, data from all subjects are fed into a baseline model.

Below, we briefly introduce how those baseline models are transformed into the multi-branch structure (WMB or AMB). a) Shallow ConvNet: branch from the last convolution layer, (b) Deep ConvNet: branch from the third last convolution layer, (c) ResNet: branch from the last residual block, (d) MSFBCNN: branch from the last convolution layer, and (e) EEG_TCNet: branch from the temporal convolutional network (TCN).

Fig. 2 shows performance comparison among WMB, AMB, and SB based on six baseline models in classification accuracy on the three datasets. SB does not always improve the performance of the baseline models and may even be counterproductive. For example, SB_EEGNet achieved the worst classification accuracy among the four conditions on BCICIV2a and BCICIV2b datasets, and so did SB_EEG_TCNet. However, SB achieved no poor performance on HGD (the slight reduction on SB_Deep ConvNet and SB_MSFBCNN could be ignored). The reason may be that the data distribution from different subjects in HGD varies slightly only so that the models with a single branch can fit the data well. According to Table V, the number of parameters of EEGNet and EEG_TCNet is 5.6k and 10.97k, respectively. The parameters of the other baseline models are at least four times more than that of EEGNet and EEG_TCNet. When the data distribution from different subjects varies obviously, the SB-based models with fewer parameters may be difficult to fit. Thus, data from other subjects is a burden for EEGNet and EEG_TCNet and worsens the classification results. It is worth noting that SB_ResNet is much better than ResNet on the three datasets because of its vast number of parameters and very deep structure, so it is free of the risk of underfitting.

AMB and WMB performed much better on the three datasets compared with SB. In dealing with data from different subjects, multi-branch is more suitable than single branch and guaranteed to surpass the baseline models consistently.The results on the BCICIV2a dataset show the accuracies of AMB_Deep ConvNet and WMB_Deep ConvNet are 86.18% and 85.94%, respectively, approximately 8% higher than that of Deep ConvNet. Compared with AMB, WMB is more advantageous. First, WMB_EEGNet almost always outperformed AMB_EEGNet regardless of the dataset and $F1$ value. Second, except for AMB_Shallow ConvNet, AMB_Deep ConvNet, and AMB_MSFBCNN ConvNet on the BCICIV2a dataset, the other models based on WMB surpassed those based on AMB on the three datasets. It is a remarkable fact that AMB is at risk of not working, as seen on AMB_ResNet on the BCICIV2a and BCICIV2b datasets. The reason is that some branches in AMB_ResNet have a poor ability to decode EEG features, so they adversely impact classification results. However, this problem can be solved by weighing the branches. Giving higher weights to branches with better feature mining ability and lowering the weights of weaker branches helps to produce superior performance.

In addition, the study on branch depth and the comparative results with fine-tuning methods are presented in **Sections C and D of supplementary document**.

V. Discussions

A. Analysis of Data Distribution

EEG data distribution varies with different subjects. Thus, subject-specific tasks usually do not consider data from other subjects; otherwise, it may backfire. As seen in Fig. 2, the classification accuracies of many models based on SB are lower than normal. However, examples also can be found where the classification accuracy of SB is higher than normal (Fig. 2), indicating that data information from other subjects may be helpful for subject-specific tasks. In the final analysis, the critical link is how the positive effect of multisource data can overwhelm the adverse impact.

WMB was designed to solve the problem that a single-branch structure has difficulty extracting the common domain-invariance representations for multisource data. The core idea of WMB is to transform multisource data fitting into multiple pairs of source-target data fitting, i.e., a branch is responsible for fitting a pair of source-target data. The weighted sum of all branches' decisions is the final decision. To further explore the differences in feature distribution produced by SB and WMB, **supplementary Fig. 7** visualises the feature distribution with the aid of t-distributed stochastic neighbour embedding (T-SNE) [49] from the last convolution layers. Except for 8-5, the target features and source features in the right T-SNE maps can be separated almost linearly, showing a clear distribution gap between the target features and the source features produced by SB_EEGNet. However, the target and source features in the left T-SNE maps are almost fused, demonstrating the powerful fitting ability of WMB. We argue that only when the distribution distance between the source and target features is short enough can the subject-specific task absorb the source data's knowledge. Otherwise, source data may become a burden to a subject-specific task.

From another perspective, the input target data into each branch are the same, while each branch's output feature distribution differs. Thus, we can obtain the multi-representation features for the target data. As seen in **supplementary Fig. 8**, the feature distributions of the eight branches are different from each other. Each branch maps the target data to features with various distributions, whose processes are both influenced by the source data and the initialised weights of each branch.

B. Statistical Significance Test

Following the conventions in previous EEG studies [21][50], a Wilcoxon signed-rank test is used to evaluate the proposed method in terms of statistical significance for performance improvement. Tables I-III list partial results for WMB_EEGNet. It is demonstrated that on the three datasets, our proposed WMB_EEGNet significantly outperformed all the state-of-the-art models and the baseline EEGNet ($p< 0.05$). More detailed results are summarised in **supplementary Table VIII**. On the one hand, the performance improvement of our proposed

method compared to the corresponding baselines is significant ($p< 0.05$, WMB_Shallow ConvNet vs. Shallow ConvNet, WMB_Deep ConvNet vs. Deep ConvNet, WMB_ResNet vs. ResNet, WMB_MSFBCNN vs. MSFBCNN, and WMB_EEGTCNeT vs. EEG_TCNet) regardless of the dataset. On the other hand, our proposed method significantly outperformed most state-of-the-art models ($p<0.05$). Note that WMB_ResNet performed poorly on the BCICIV2a and BCICIV2b datasets while showing advanced performance on HGD. The reason may be that WMB_ResNet has deeper CNNs than other models, which needs more trainable data for studying, and the amount of data in HGD is larger than that in the other two datasets.

C. Role of Weights

The adaptive weights for multi-branch decision fusion are another important contribution of this paper. Multiple branches solve the problem that the models have difficulty fitting distributions from multisubject features, while the adaptive weights play a role in optimising the contributions of every branch. The use of weights is motivated by the Adaptive Boosting Algorithm. The output of the other learning algorithms ('weak learners') is combined into a weighted sum representing the boosted classifier's final output. Similarly, each branch corresponds to a weak classifier, and the weighted sum of each branch's decision is the final decision. During training, the weight of the weak classifier with a low classification error rate is adaptively increased, so it plays a more prominent decisive role in the final classification decision. In addition, there are more complex implementation methods for weighting, e.g., weight by attention mechanism [51][52]. However, the attention mechanism inevitably increases the number of training parameters and running memory and makes models more complex, which goes against real-time and compactness.

D. Balance Between Performance and Compactness

Multiple branches inevitably increase the number of parameters and model complexity, including a longer inference time and more memory. As seen in **supplementary Table IX**, the number of parameters and MACCs of baseline models based on WMB or AMB increases with increased branches. Interestingly, the growth rate of MACCs is lower and much smaller than the number of parameters. Because the computational burden is concentrated in the head for most EEG deep learning models, it is a nice property for multi-branch structure, so it may not bring too much extra computational burden. Figs. 4-6 also show that as long as the baseline is real-time, the baseline based on WMB is real time, except for Deep ConvNet. Hence, the multi-branch structure may not affect real-time. However, the increase in the number of parameters is considerable. For example, it increases approximately four times for WMB_Shallow ConvNet. Pruning branches according to the weights can reduce the number of parameters but may affect performance (**supplementary Table IX, supplementary Figs. 1-3**). Therefore, it is essential to set the number of preserved branches correctly. After training, keeping half of the branches reserves over 50% weight. Consequently, we argue that the decision is dominated by half of the branches with the maximum weights. Additionally, we observed that the preserved branches between 20% and 50% led to near-optimal performance and were at least superior to the baselines (Figs. 1-3). Thus, it is feasible to balance performance and compactness by pruning branches according to weights. In addition, experiments were also conducted to investigate the effect of the number of source subjects on the performance of WMB models by training models using only a pair of source and target subjects, as well as by training models after randomly removing one source subject. A preliminary argument is that a decrease in the number of source subjects will reduce the performance of the model, and too few source subjects (i.e., only keeping one source subject) will even degrade the results compared to the baseline model.

VI. Conclusion

This paper proposes a novel motor imagery EEG decoding method based on a weighted multi-branch structure suitable for multisubject data. It can handle the problem that different subject data has large distribution discrepancy, enhance subject-specific model learning, and produce better classification accuracy. In the proposed architecture for EEG classification, pruning branches according to weight values can obtain a slim model and balance its performance and compactness. The proposed method achieved significantly higher classification accuracy on six widely used EEG deep learning models across three public motor imagery datasets and two supplementary datasets. Moreover, with the analysis and visualisation of data distribution, it can be concluded that the critical link for multisubject training is how the positive effect of multisource data (helpful information) can overwhelm the adverse impact (large distribution discrepancy). The extensive experimental results have demonstrated that the efficient fitting of multisubject features drove improved performance by WMB. Nevertheless, it is still a challenging problem to determine the optimal number of branches with many subjects. Future work will focus on this problem and apply the proposed method to more subjects.

References

[1] D. Zapała et al., "The effects of handedness on sensorimotor rhythm desynchronization and motor-imagery BCI control," *Scientific Reports*, vol. 10, 1, pp. 1-11, 2020.

[2] R. Zhang et al., "Control of a wheelchair in an indoor environment based on a brain-computer interface and automated navigation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 1, pp. 128-139, 2015.

[3] B. Xu et al., "Motor imagery based continuous teleoperation robot control with tactile feedback," *Electronics*, vol. 9 no. 1, pp. 174, 2020.

[4] M. Li et al., "The MindGomoku: an online P300 BCI game based on Bayesian deep learning," *Sensors*, vol. 21, no. 5, pp. 1613, 2021.

[5] R. Mane et al., "BCI for stroke rehabilitation: motor and beyond," *Journal of Neural Engineering*, vol. 17, nol. 4, pp. 041001, 2020.

[6] R. R. Lu et al., "Motor imagery based brain-computer interface control of continuous passive motion for wrist extension recovery in chronic stroke patients," *Neuroscience Letters*, vol. 718, pp. 134727, 2020.

[7] B. Somers et al., "A generic EEG artifact removal algorithm based on the multi-channel Wiener filter," *Journal of Neural Engineering*, vol. 15, no. 3, pp. 036007, 2018.

[8] M. Sharma et al., "An accurate sleep stages classification system using a new class of optimally time-frequency localized three-band wavelet filter bank," *Computers in Biology and Medicine*, vol. 98, pp. 58-75, 2018.

[9] F. Rohit et al., "Real-time drowsiness detection using wearable, lightweight brain sensing headbands," *IET Intelligent Transport Systems*, vol. 11, no. 5, pp. 255-263, 2017.

[10] J. Hu et al., "Noise robustness analysis of performance for EEG-based driver fatigue detection using different entropy feature sets," *Entropy*, vol. 19, no. 8, pp. 385, 2017.

[11] Y. Zhang et al., "Classification of EEG signals based on autoregressive model and wavelet packet decomposition," *Neural Processing Letters*, vol. 45, no. 2, pp. 365-378, 2017.

[12] B. Blankertz et al., "The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, no. 2, pp. 539-550, 2007.

[13] K. K. Ang et al., "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Frontiers in Neuroscience*, vol. 6, no. 39, pp. 2012.

[14] J. Yang et al., "Multi-time and multi-band CSP motor imagery EEG feature classification algorithm," *Applied Sciences*, vol. 11, no. 21, pp. 10294, 2021.

[15] J. Li et al., "Multisource transfer learning for cross-subject EEG emotion recognition," *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3281-3293, 2019.

[16] Y. Zhang et al., "Boosting-LDA algorithm with multi-domain feature fusion for motor imagery EEG decoding," *Biomedical Signal Processing and Control*, vol. 70, pp. 102983, 2021.

[17] L. H. Chew et al., "Aesthetic preference recognition of 3D shapes using EEG," *Cognitive Neurodynamics*, vol. 10, no. 2, pp. 165-173, 2019.

[18] L. Fraiwan et al., "Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier," *Computer Methods and Programs in Biomedicine*, vol. 108, no. 1, pp. 10-19, 2012.

[19] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391-5420, 2017.

[20] B. Accou et al., "Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network," in *IEEE 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1175-1179.

[21] R. Zhang et al., "Hybrid deep neural network using transfer learning for EEG motor imagery decoding," *Biomedical Signal Processing and Control*, vol. 63, pp. 102144, 2021.

[22] Y. Roy et al., "Deep learning-based electroencephalography analysis: a systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, pp. 051001, 2019.

[23] C. S. Wei et al., "Spatial component-wise convolutional network (SCCNet) for motor-imagery EEG classification," In I*EEE 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2019, pp. 328-331.

[24] X. Liu et al., "Multi-scale space-time-frequency feature-guided multitask learning CNN for motor imagery EEG classification," *Journal of Neural Engineering*, vol. 18, no. 2, pp. 026003, 2021.

[25] Y. Zhu et al., "Multi-representation adaptation network for cross-domain image classification," *Neural Networks*, vol. 119, pp. 214-221.

[26] X.Tang et al., "Conditional adversarial domain adaptation neural network for motor imagery EEG decoding," *Entropy*, vol. 22, no. 1, pp. 96, 2020.

[27] V. J. Lawhern et al. "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, pp. 056013, 2018.

[28] Y. Li et al. "A channel-projection mixed-scale convolutional neural network for motor imagery EEG decoding," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 6, pp. 1170-1180, 2019.

[29] H. Li et al., Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network. *Biomedical Signal Processing and Control*, vol. 72, pp. 103342, 2022.

[30] A. Ak et al., "Motor imagery EEG signal classification using image processing technique over GoogLeNet deep learning algorithm for controlling the robot manipulator," *Biomedical Signal Processing and Control*, vol. 72, pp. 103295, 2022.

[31] Y. Hou et al., "GCNs-net: a graph convolutional neural network approach for decoding time-resolved EEG motor imagery signals," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[32] C. Brunner et al., "BCI Competition 2008–Graz data set A," *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, vol. 16, pp. 1-6, 2008.

[33] F. Mattioli et al., "A 1D CNN for high accuracy classification and transfer learning in motor imagery EEG-based brain-computer interface," *Journal of Neural Engineering*, vol. 18, no. 6, pp. 066053, 2022.

[34] L. Yang et al., "Motor imagery EEG decoding method based on a discriminative feature learning strategy," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 368-379, 2021.

[35] P. Lu et al., (2019, October). "Combined CNN and LSTM for motor imagery classification," in *IEEE 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2019, pp. 1-6.

[36] S. U. Amin et al., "Deep learning for EEG motor imagery classification based on multi-layer CNNs feature fusion," *Future Generation computer systems*, vol. 101, pp. 542-554, 2019.

[37] Z. Jia et al., (2020, September). "MMCNN: a multi-branch multi-scale convolutional neural network for motor imagery classification," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (MLKDD)*, 2020, pp. 736-751.

[38] H. Wu et al., "A parallel multi-scale filter bank convolutional neural networks for motor imagery EEG classification," *Frontiers in Neuroscience*, vol. 13, pp. 1275, 2019.

[39] M. Riyad et al., "Incep-EEGNet: a convent for motor imagery decoding," in *International Conference on Image and Signal Processing (ISP)*, 2020, pp. 103-111.

[40] J. Redmon et al., "Yolov3: An incremental improvement," 2018, arXiv:1804.02767.

[41] G. Wang et al., "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia (ACM)*, 2018, pp. 274-282.

[42] D. Li et al., "Densely feature fusion based on convolutional neural networks for motor imagery EEG classification," *IEEE Access*, vol. 7, pp. 132720-132730, 2019.

[43] T. M. Ingolfsson et al. "EEG-TCNet: an accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 2958-2965.

[44] C. Brunner et al., "BCI competition 2008–Graz data set A," *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, vol. 16, pp. 1-6, 2008.

[45] P. Ofner et al., "Upper limb movements can be decoded from the time-domain of low-frequency EEG," *PloS one*, vol. 12, no. 8, pp. e0182578.

[46] R. Mane et al., "FBCNet: a multi-view convolutional neural network for brain-computer interface," 2021, arXiv:2104.01233.

[47] M. A. Li et al., "A novel MI-EEG imaging with the location information of electrodes," *IEEE Access*, vol. 8, pp. 3197-3211, 2019.

[48] X. Wang et al., "An accurate EEGNet-based motor-imagery brain–computer interface for low-power edge computing," in *IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2020, pp. 1-6.

[49] L. Van der Maaten et al., "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 10, no. 11, 2008.

[50] F. Li et al., "A novel simplified convolutional neural network classification algorithm of motor imagery EEG signals based on deep learning," *Applied Sciences*, vol. 10, no. 5, pp. 1605, 2020.

[51] D. Zhang et al., "Motor imagery classification via temporal attention cues of graph embedded EEG signals," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 9, pp. 2570-2579, 2020.

[52] X. Liu et al., "Parallel spatial-temporal self-attention CNN-based motor imagery classification for BCI," *Frontiers in Neuroscience*, vol. 14, pp. 587520, 2020.