frontiers in Neuroscience

ORIGINAL RESEARCH
published: 22 October 2021
doi: 10.3389/fnins.2021.760979

A Generalizable and Discriminative Learning Method for Deep EEG-Based Motor Imagery Classification

Xiuyu Huang[1]*, *Nan Zhou*[2,1] *and Kup-Sze Choi*[1]*

[1] *Centre for Smart Health, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China,* [2] *College of Control Engineering, Chengdu University of Information Technology, Chengdu, China*

Convolutional neural networks (CNNs) have been widely applied to the motor imagery (MI) classification field, significantly improving the state-of-the-art (SoA) performance in terms of classification accuracy. Although innovative model structures are thoroughly explored, little attention was drawn toward the objective function. In most of the available CNNs in the MI area, the standard cross-entropy loss is usually performed as the objective function, which only ensures deep feature separability. Corresponding to the limitation of current objective functions, a new loss function with a combination of smoothed cross-entropy (with label smoothing) and center loss is proposed as the supervision signal for the model in the MI recognition task. Specifically, the smoothed cross-entropy is calculated by the entropy between the predicted labels and the one-hot hard labels regularized by a noise of uniform distribution. The center loss learns a deep feature center for each class and minimizes the distance between deep features and their corresponding centers. The proposed loss tries to optimize the model in two learning objectives, preventing overconfident predictions and increasing deep feature discriminative capacity (interclass separability and intraclass invariant), which guarantee the effectiveness of MI recognition models. We conduct extensive experiments on two well-known benchmarks (BCI competition IV-2a and IV-2b) to evaluate our method. The result indicates that the proposed approach achieves better performance than other SoA models on both datasets. The proposed learning scheme offers a more robust optimization for the CNN model in the MI classification task, simultaneously decreasing the risk of overfitting and increasing the discriminative power of deeply learned features.

Keywords: electroencephalogram, motor imagery, convolutional neural networks, label smoothing, center loss

Edited by:
Yuanpeng Zhang,
Nantong University, China

Reviewed by:
Hongru Zhao,
Soochow University, China
Yufeng Yao,
Changshu Institute of Technology, China

***Correspondence:**
Xiuyu Huang
xiuyu.huang@connect.polyu.hk
Kup-Sze Choi
thomasks.choi@polyu.edu.hk

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *19 August 2021*
Accepted: *06 September 2021*
Published: *22 October 2021*

Citation:
Huang X, Zhou N and Choi K-S (2021) A Generalizable and Discriminative Learning Method for Deep EEG-Based Motor Imagery Classification. Front. Neurosci. 15:760979. doi: 10.3389/fnins.2021.760979

1. INTRODUCTION

Brain–computer interface (BCI) has been raising interest from the research community. It provides an important way for the disabled to interact with the outside world without using any muscular movements (Wolpaw et al., 2002). This technology aims to recognize the user intentions based on the distinct patterns of neural events. Motor imagery (MI) is one of the crucial topics in the area of BCI, referring to a cognitive procedure of the motion imagination such as lifting left or right leg, without any actual moving actions (Ahn and Jun, 2015). The most popular technology to signalize such cognitive procedures is the electroencephalogram (EEG), being noninvasive and relatively easy to set up (Ahn and Jun, 2015; Ni et al., 2020; Zhang et al., 2020). The principle of the EEG-based MI-BCI system is to match the type of motion imagination and its corresponding EEG signals. Such matching systems have been practiced in a variety of applications, including speller (Rezeika et al., 2018), wheelchair (Kaufmann et al., 2014), and prosthesis (Vidaurre et al., 2016).

Accurate classification of EEG-MI pattern is one of the most decisive factors to the BCI performance but remains a significant challenge due to the low signal-to-noise ratio (SNR) characteristics of the EEG signal (Goldenholz et al., 2009; Zhang et al., 2019). Convolutional neural networks (CNNs) have been widely explored and achieved great success in the MI recognition area (Bashivan et al., 2015; Roy et al., 2019). It significantly pushes the boundary of the state-of-the-art (SoA) in classification accuracy compared to the conventional methods such as band power analysis (Martinez-Leon et al., 2015), independent component analysis (ICA) (Lee et al., 1999), and common spatial filter (CSP) (Ramoser et al., 2000). The most common framework of CNN is to perform feature generation and label prediction, learning deep features from raw EEG data by the CNN pipeline, then making label predictions based on the learned features (see **Figure 1A**).

The training of CNNs in the MI classification task is mainly guided by minimizing the cross-entropy (Hertz et al., 2018). This objective function is "greedy" and encourages the largest possible logit gaps, making the model less adaptive, and sometimes overconfident to its predictions (Szegedy et al., 2016). The model learns to assign a full probability to the ground-truth label for each training example, even though some noisy data are mixed in the training set (Müller et al., 2019). This overfitting phenomenon incredibly easily occurs when the training sample size is small. Coincidentally, the EEG data have a low SNR and contains much noise. In addition, the MI-BCI system is usually designated as subject dependent, so it usually has limited training data. To reduce the risk of overfitting (overconfidence) issue, we adopt a label smoothing technique introduced by Szegedy et al. (2016) in the training scheme of the MI classification. It computes a modified cross-entropy, called smoothed cross-entropy, not by using "hard" one-hot encoded labels such as [0, 0, 1, 0] from training data, but a weighted mixture of these hard labels with the uniform distribution (**Figure 1B**). Label smoothing alternatively encourages small logit gaps and prevent overconfident predictions. This technology has successfully increased the performance of CNN models across multiple tasks, including image classification (Szegedy et al., 2016), speech recognition (Chorowski and Jaitly, 2016), and machine translation (Vaswani et al., 2017). It is expected to benefit the model training in MI classification by tackling the overfitting problem, leading to a generalizable and adaptive CNN model.

In addition to ensuring the model's generalizability, we also aim to increase its discriminative power. As shown in the common framework of CNN (**Figure 1A**), the last fully connected layer acts as a linear classifier, and the cross-entropy only encourages the separability (Hertz et al., 2018) but does not guarantee the high discriminative characteristics, where features have both a large inter-class difference and a tight intra-class variation (**Figure 1C**). Therefore, the resulting features generated by the model trained via the cross-entropy are not sufficiently effective for the MI classification. To enhance the discriminative capacity of the deep features, we apply a center loss (Wen et al., 2016) for the model training. Specifically, the center of deep features is computed by their means in each class and updated across every epoch. The distances between deep features with their corresponding class centers are minimized at each training iteration. The parameters are optimized by jointly minimizing the cross-entropy and center loss. Intuitively, the cross-entropy forces the deep features from different classes to stay apart, and the center loss pulls the features belonging to the same class toward their centers. With joint supervision, we can concurrently enlarge the inter-class difference and reduce the intra-class variation so as to improve the discriminative power of deep features.

In this paper, we propose a novel training scheme for CNN-based model in the MI classification by using a combined loss with smoothed cross-entropy and center loss. The main contributions are as follows:

1. To our best knowledge, although structures of the CNN model are heavily investigated, this is the first attempt to use the proposed loss to help supervise training in the context of MI classification. With joint supervision of the smoothed cross-entropy and the center loss, both generalizable and discriminative model can be obtained for robust MI recognition.
2. We present extensive experiments on two famous MI public datasets, called BCI-competition IV-2a and IV-2b. Our new approach achieves superior performance compared to other SoA methods.
3. We also conduct an ablation study to demonstrate the effectiveness of the label smoothing and the center loss.

The remainder of the paper is organized as follows. In section 2, conventional and deep learning methods on MI classification are introduced. Section 3 describes the proposed approach. Sections 4 and 5 present the experiment result and analysis. Section 6 concludes the current study.

2. RELATED WORKS

A sophisticated feature extractor is the key to success in conventional methods for the MI classification task. One of the most frequently and widely used approaches is the common spatial pattern (CSP) (Pfurtscheller and Neuper, 2001; Yu et al., 2019). It tries to generate optimal spatial filters that have minimum or maximum variance between classes in a particular frequency band. The features used in the winner algorithm of the BCI competition IV are based on the filter bank CSP (FBCSP) (Ang et al., 2008) that finds a set of optimal spatial filters in multiple frequency bands. The Naive Bayes Parzen Window classifier using these features achieved an outstanding classification performance with an accuracy of 67.75% on the dataset IV-2a (Ang et al., 2008). After the competition, a novel method based on the support vector machine (SVM) with Riemannian covariance achieved a better performance (75.74%)

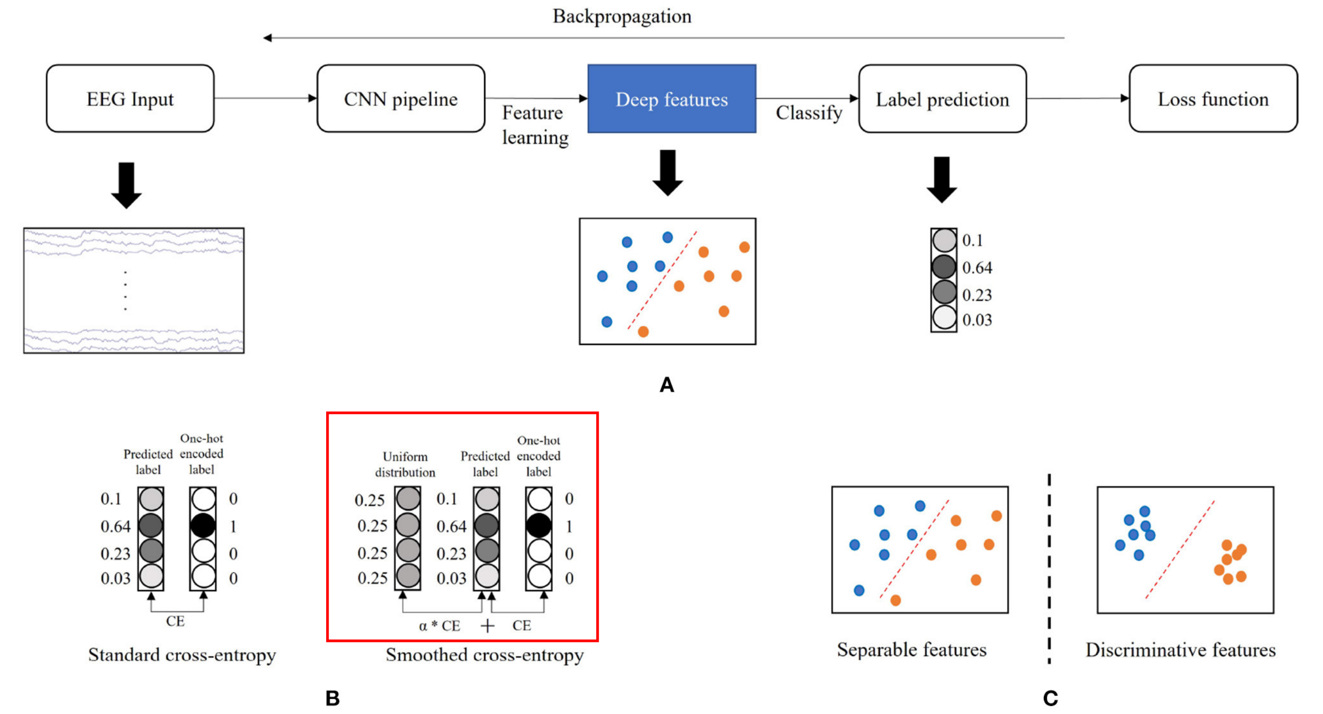

FIGURE 1 | Illustration of the concepts in the study. **(A)** The common framework of convolutional neural network (CNN). **(B)** Label smoothing. **(C)** Deep feature distributions. The CE denotes the cross-entropy calculation. The α denotes the weight of the CE between predicted label and uniform distribution.

for the same database (Hersche et al., 2018). In addition to these vector-based methods, matrix-form strategies such as the logistics regression classifier with a nuclear norm regularization (Zhou and Li, 2014), the rank-k SVM (Lal et al., 2004), and the support matrix machine (SMM) (Zheng et al., 2018) were also developed by multiple research groups. The leading edge of these methods is to directly process the 2-D MI EEG data on a matrix basis instead of stacking features as a vector input to a classifier, which preserves the informative structural patterns.

Deep learning (DL) models were also exploited to tackle the MI classification challenge. For instance, the multilayer perceptron (MLP) was proposed to generate nonlinear patterns from CSP features and also to substitute the SVM as a classifier for MI recognition (Kumar et al., 2016). Similarly, a channel-wise convolution with channel mixing (C2CM) was introduced to classify the spatial-temporal features generated by FBCSP (Sakhavi et al., 2018). Bashivan et al. (2015) converted the EEG waves into spectral topographies via short-time Fourier transform (STFT). These topographies were then fitted into CNNs for further transformation and classification. Tabar and Halici (2016) also used the STFT approach to extract spatial-temporal images as the feature input to the CNN-SAE model for classification. These feature input (FI) models still require complex feature generation from raw EEG data prior to the DL modeling. Several research groups investigate raw signal input (RSI) models to provide an end-to-end scheme for MI recognition to address this limitation. For example, the two most well-known RSI networks, EEGNet (Lawhern et al., 2018), and ConvNet (Schirrmeister et al., 2017), achieved competitive classification performance without using any pre-processing techniques. In general, although model architectures were heavily investigated, the neural network learning process did not receive too much attention from the MI community. Rather than figuring out a more sophisticated architecture, we propose a potentially efficient objective function for both generalizable and discriminative learning in the CNN-based model.

3. METHOD

This section first introduces the notations and definitions used in this work and describes the CNN architecture. Then, the novel proposed loss is presented in detail.

3.1. Definition and Notations

Assuming that the DL model input is on a per-trial basis, where the continuous EEG is segmented into labeled trials, we define the segmented trials of a subject as $\{[x_i, y_i]\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^{E\times T}$ represents ith of EEG trials recorded by E electrodes and T sampling time points. $y_i \in \mathbb{R}^{M}$ denotes the corresponding i^{th} labels of M classes. Let the ground truth distribution p over labels $p(y|x_i)$, and $\sum_{y=1}^{M} p(y|x_i) = 1$. We also define a CNN-based model with θ that predicts label distribution $q_\theta(y|x_i)$, and certainly $\sum_{y=1}^{M} q_\theta(y|x_i) = 1$. We are

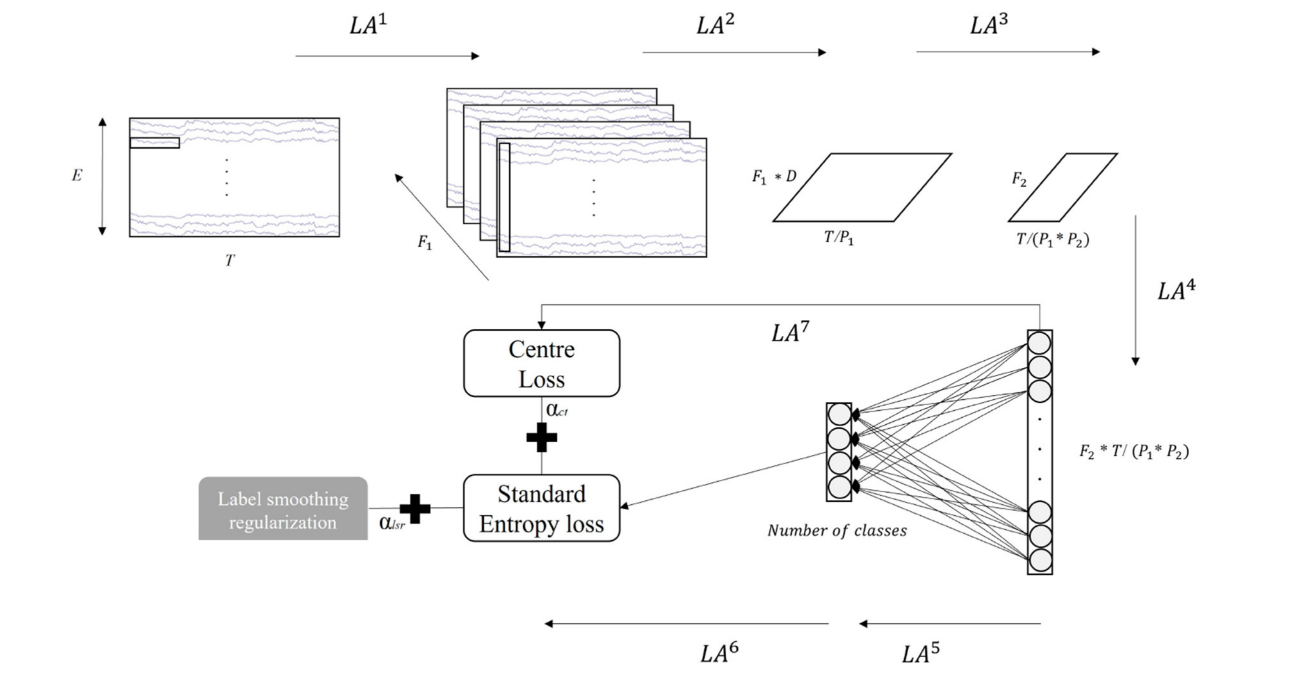

FIGURE 2 | Illustration of the convolutional neural network (CNN)-based model. T = number of the timestamps, E = number of electricodes, F_1 = number of filters in the temporal CNN, D = number of depthwise convolution output channels, F_2 = number of spatial filters, P_1 and P_2 are sizes of the average pooling kernel, and α_{lsr} and α_{ct} are weights of label smoothing regularization and center loss.

motivated to adopt the label smoothing technology and center loss to improve the generalizable and discriminative power of the CNN-based model.

3.2. Network Architecture

In the current study, we inherit the CNN architecture of the EEGNet (Lawhern et al., 2018) but make two modifications, where the kernel size of the first temporal CNN filter (LA^1) is decreased to attain temporal information above 8 Hz, as the alpha (8–12 Hz) and beta (12.5–30 Hz) band contain most relevant information of the motor imagery task (Wierzgała et al., 2018). The illustration of the network is displayed in **Figure 2**, and the details of each layer are presented in **Table 1**. The model begins with a 2D-CNN directly linked to the raw EEG data with a kernel size in $(K_1, 1)$ to capture temporal patterns in each electrode. A depthwise convolution layer with a kernel size of $(1, E)$ is followed and utilized for spatial feature extraction. The separableConv2D with a kernel size in $(K_2, 1)$ is then performed to gain deeper and more abstract temporal information across all electrodes. As shown in **Table 1**, it is noted that batch normalization (Ioffe and Szegedy, 2015), exponential linear unit (ELU) (Clevert et al., 2015) activation, and average pooling are sequentially followed after some of these convolutions for covariate shift avoidance (Bickel et al., 2009), nonlinear transformation, and dimension reduction, respectively. The deep feature generated by the CNN pipeline is then flattened as a vector (nodes) by a flatten layer. The vectors of each training batch are used to compute the center loss. The dense layer is subsequently connected to these nodes and acts as a classifier. The *softmax* function finally performs the estimation of the probability for each MI class. The cross-entropy between the probability estimation and the smoothed label represents the classification loss (standard cross-entropy + label smoothing regularization). The weighted sum of the classification and center losses supervises the training of the entire network.

3.3. Proposed Loss

To ensure both generability and discriminative power of the CNN-based model in the context of MI classification, we propose a combined loss that jointly optimizes the classification loss (e.g., cross-entropy computed by smoothed labels) and the center loss. Most models in previous studies performed the standard cross-entropy for the objective function defined as

$$\begin{aligned} L_{cl} &= \sum_{i=1}^{n} H_i(p, q_\theta) \\ &= -\sum_{i=1}^{n}\sum_{y=1}^{M} p(y|x_i) log\ q_\theta(y|x_i) \end{aligned} \tag{1}$$

标准交叉熵

Given the $p(y|x_i)$ is one-hot encoded in classification task where

$$p(y|x_i) = \begin{cases} 1, & y = y_i; \\ 0, & otherwise. \end{cases} \tag{2}$$

TABLE 1 | Architecture setting.

Layer	Function	Filter	Kernel	Output shape
	Input			$T, E, 1$
LA^1	Conv2d	F_1	$(K_1, 1)$	T, C, F_1
	BatchNorm			
	DepthwiseConv2d	$F_1 * D$	$(1, E)$	
	BatchNorm			$T, 1, F_1 * D$
LA^2	ELU Activation			
	Average pooling		$(P_1, 1)$	$T/P_1, 1, F_1 * D$
	SparableConv2d	F_2	$(K_2, 1)$	
	BatchNorm			$T/P_1, 1, F_2$
LA^3	ELU Activation			
	Average pooling		$(P_2, 1)$	$T/(P_1 * P_2), 1, F_2$
LA^4	Flatten			$(T * F_2)/(P_1 * P_2)$
LA^5	Fully connected			*number of classes*
LA^6	Softmax (CEL)			
LA^7	Lambda (CL)			1

(1) T = number of the timestamps, E = number of electricodes, K_1 = kernel size of the first CNN, D = number of depthwise convolution output channels, F_1 = number of temporal filters, F_2 = number of spatial filters, K_2 = size of th kernel in the spatial filer, and P_1 and P_2 are sizes of average pooling kernels.

(2) CEL stands for cross-entropy computed by smoothed labels. CL stands for center loss. Lamda layer is a self-customize layer for the calculation of the center loss.

TABLE 2 | Hyperparameter settings.

Hyperparameter	CNN for IV-2a	CNN for IV-2b
T	1,000	1,000
E	22	3
F_1	8	8
F_2	16	16
K_1	32	32
K_2	16	16
D	2	2
P_1	8	8
P_2	8	8

We can further reduce (1) as

$$L_{cl} = -\sum_{i=1}^{n} log\ q_\theta(y_i|x_i) \tag{3}$$

For each training sample i, the $q_\theta(y_i|x_i)$ is usually calculated by the *softmax* function as follows:

$$q_\theta(y_i|x_i) = \frac{exp(z_{y_i})}{\sum_{j=1}^{M} exp(z_j)} \tag{4}$$

Here, z_j is the logit value or unnormalized log-probability for each label j. By using the one-hot ground-truth label, minimizing the objective function L_{cl} is equivalent to do the log-likelihood maximum. The maximum is not achievable with finite data, so it can only be estimated in the case when $z_{y_i} >> z_j$ for all $j \neq y_i$ (e.g., the logit of the ground-truth label is much larger than all other logits) over the training dataset (Szegedy et al., 2016). In such a case, the model learns to classify every training sample correctly with a confidence of nearly 1, which is the signal of overfitting. This phenomenon relatively easily occurs in the scenario that the MI EEG task often only contains a small sample size of training data.

We adopt the label smoothing mechanism where a noise distribution $u(y|x)$ is added to the one-hot ground truth label to prevent the model from having overconfidence and to reduce the risk of overfitting. Then, the new ground truth label distribution is $p'(y|x_i) = (1-\varepsilon)p(y|x_i) + \varepsilon u(y|x_i)$, where ε is a weight factor, $\varepsilon \in [0, 1]$. By replacing $p(y|x_i)$ with $p'(y|x_i)$ in (1) and (2), the new classification loss L'_{cl} computed by smoothed labels can be written as

$$\begin{aligned} L'_{cl} &= \sum_{i=1}^{n}\sum_{y=1}^{M} p'(y|x_i) log\ q_\theta(y|x_i) \\ &= -\sum_{i=1}^{n}\sum_{y=1}^{M} [(1-\varepsilon)p(y|x_i) + \varepsilon u(y|x_i)]\ q_\theta(y|x_i) \\ &= (1-\varepsilon)\sum_{i=1}^{n} H_i(p, q_\theta) + \varepsilon \sum_{i=1}^{n} H_i(u, q_\theta) \end{aligned} \tag{5}$$

The first half of (5) is the weighted standard cross-entropy L_{cl}. Let the second half is weighted loss of label-smoothing regularization (L_{lsr}), which penalizes the deviation of predicted label distribution p from noise distribution u with a relative weight $\varepsilon/(1-\varepsilon)$. We set the noise as the uniform distribution $u(y|x) = 1/M$ (Szegedy et al., 2016). Then, the $H_i(u, q_\theta)$ measures the dissimilarity between predicted label distribution p to uniform. Therefore, L_{lsr} heavily penalizes overconfident predictions and prevents poor generalization during the training.

Let the weight of L_{cl} is fixed as 1, and the relative weight $\varepsilon/(1-\varepsilon)$ of L_{lsr} is redefined as α_{lsr}. Therefore, L'_{cl} can be elaborated as

$$L'_{cl} = L_{cl} + \alpha_{lsr} L_{lsr} \tag{6}$$

In addition to the generalizability, we also try to ensure the discriminative ability of the deep learned features extracted by the CNN pipeline. Intuitively, simultaneously maximizing the inter-class distance and minimizing the intra-class variation is the fundamental strategy to keep features of different classes divisible. The cross-entropy minimization only assures to enlarge the inter-class distance, so we further employ a center loss to achieve intra-class variation reduction. We follow the equation proposed by Wen et al. (2016), and the center loss is defined as

$$L_{ct} = \frac{1}{2}\sum_{i=1}^{n} ||x_i - c_{y_i}||_2^2 \tag{7}$$

where c_{y_i} denotes the y_i center of the feature extracted by the CNN pipeline. Minimizing the distance between each deeply learned feature and its class center naturally decreases the intra-class variation. Finally, the objective function L is to jointly optimize the classification loss L'_{cl} and L_{ct}, defined as

$$L = L'_{cl} + \alpha_{ct} L_{ct} \tag{8}$$

where α_{ct} is the weight of center loss. Based on (6), L can be re-defined as,

$$\boxed{L = L_{cl} + \alpha_{lsr} L_{lsr} + \alpha_{ct} L_{ct}} \tag{9}$$

4. EXPERIMENTS

The BCI competition IV-2a and IV-2b datasets are used to evaluate the proposed approach. These two datasets are publicly available. The people involved in the datasets have obtained ethic approval. Users can download the data for free for research and publish relevant articles, so the ethical review and approval were waived for the current study.

4.1. Dataset Description

The BCI competition IV-2a (Tangermann et al., 2012) was recorded from 9 healthy individuals (A01-A09) by 22 EEG and 3 EOG channels in a sample rate of 250 Hz. The cue-based paradigm is used during the data collection. It consists of four MI classes, including the imagination motions of the left hand, right hand, tongue, and both feet. Two separate sessions were implemented for each subject. Each session comprises a total of 4*72 (a single MI class) = 288 trials. For fair comparisons with other approaches, the same data division scheme as that in the competition was used in our experiment. The first section is for model training and the second for model testing. Only the 4-s temporal segment (from the start of the cue until the end of the MI) in each trial is used in our model. Given the 250 Hz sample rate, our experiment's training and testing data are on a 1,000-sample series basis.

The BCI competition IV-2b (Tangermann et al., 2012) was also collected from nine healthy people (B01–B09) but only with 3 EEG channels (C3, Cz, and C4) attached to the frontal cortex. The dataset comprises two MI classes, including left-hand and right-hand movement imagination based on a cue-based BCI paradigm. Five independent sessions were recorded for each individual. We also keep using the same data division as that in the competition. The first three sessions are for training, and the remaining two are for evaluation. The 4-s temporal interval, from the starting point of the cue until the end of the MI, is used as a trial in our experiment. Given the recording frequency of 250 Hz, each training or testing trial is also on a 1,000-point basis.

4.2. Experimental Setup

Our approach is performed on a Tesla V100-SXM2 GPU running on Google online platform (Colab). The CNN network and the proposed loss function are implemented by Keras. The model is trained with Adam (Kingma and Ba, 2014) optimizer using a learning rate of 0.001, mini-batch size of 64, and 750 epochs. According to the result of the ablation study stated in section 5.2, the loss weights α_{lsr} and α_{ct} are set as 0.5 and 0.5, respectively. Other hyperparameters of the model architecture are shown in **Table 2**.

To evaluate the effectiveness of the proposed approach, we compare our strategy against existing SoA methods, including two conventional approaches [a vector-based method, e.g., the competition winner algorithm FBCSP (Ang et al., 2008), and a matrix-based method, SMM (Zheng et al., 2018)], two compact well-known DNN methods [EEGNet (Lawhern et al., 2018) and shallow ConNet (Schirrmeister et al., 2017)], and one more DNN methods (DRDA, Zhao et al., 2020) with complex architecture. The evaluation metric is the classification accuracy (*acc*).

5. RESULTS

5.1. Comparison With State-Of-The-Art Methods

The comparisons between the proposed methods and other models on the BCI competition IV-2a and IV-2b datasets are shown in **Tables 3**, **4**, respectively. The classification accuracy of each subject and the average accuracy are reported in a subject-dependent basis (e.g., training and testing data are from the same subject) as the same as the competition data division scheme. The model that has the best performance for each subject is highlighted in boldface. **Tables 3**, **4** clearly show that the proposed strategy has the best classification accuracy for nearly all subjects on both datasets with a maximum of 14.16% (subject A05) better than the second-best on IV-2a and of 11.43% (B02) better than second-best on IV-2b. On the average level, the classification average accuracies of our method have improvements of around 5.33 and 3.54% on IV-2a and IV-2b compared to other SoAs. We conduct paired t-tests between our approach and other SoA strategies to verify if the improvements are statistically significant. The p values obtained from the tests are indicated in **Table 5**. We can see that all p values are <0.05, which advises that the performance improvements of our method against others are statistically significant. In addition, it also can be seen that the corresponding standard deviations (SDs) of our method are 10.07 and 8.54%, which are both the smallest SD on the respective datasets. This result suggests that our method is a more robust classifier in a subject-independent manner than other models. All these results, as mentioned earlier, demonstrate that the CNN model trained using the proposed loss provides a more accurate and stable classification outcome for the MI recognition task.

5.2. Ablation Result Analysis

Ablation studies are carried out to study the contributions of the label smoothing technique and center loss to the CNN modeling. The hyperparameter α_{lsr} controls the degree of the smoothness on the label, and α_{ct} dominates the intra-class variations of the deep features. They are both significant. Therefore, two experiments on dataset IV-2a are explored to investigate the sensitiveness of these two hyperparameters.

5.2.1. Label-Smoothing Regularization

In the first experiment, we fix the α_{ct} as 0, where no center loss is applied, and vary α_{lsr} from 0 to 1 (inclusive) to train different

TABLE 3 | Classification accuracies (%) obtained with the dataset BCI competition IV-2a.

Methods	**Subject**									**Average (SD)**
	A01	**A02**	**A03**	**A04**	**A05**	**A06**	**A07**	**A08**	**A09**	
FBCSP	76.00	56.50	81.25	61.00	55.00	45.25	82.75	81.25	70.75	67.75 (12.94)
SMM	81.94	59.38	81.60	62.85	59.03	49.36	86.11	77.78	78.47	70.72 (12.35)
EEGNet	85.76	61.46	88.54	67.01	55.90	52.08	89.58	83.33	**86.81**	74.50 (14.36)
ConNet	76.39	55.21	89.24	74.65	56.94	54.17	**92.71**	77.08	76.39	72.53 (13.42)
DRDA	83.19	55.14	87.43	**75.28**	62.29	57.15	86.18	83.61	82.00	74.75 (12.22)
Ours	**89.32**	**66.78**	**94.14**	74.56	**76.45**	**62.33**	86.28	**85.61**	85.23	**80.08 (10.07)**

Highest values are highlighted in boldface.

TABLE 4 | Classification accuracies (%) obtained with the dataset BCI competition IV-2b.

Methods	**Subject**									**Average (SD)**
	B01	**B02**	**B03**	**B04**	**B05**	**B06**	**B07**	**B08**	**B09**	
FBCSP	70.00	60.36	60.94	**97.50**	93.12	80.63	78.13	92.50	86.88	80.01 (13.06)
SMM	67.81	51.79	53.44	93.31	82.81	74.69	72.19	82.50	75.62	72.68 (12.77)
EEGNet	68.44	57.86	61.25	90.63	80.94	63.13	84.38	93.13	83.13	75.88 (12.57)
ConNet	76.56	50.00	51.56	96.88	93.13	85.31	83.75	91.56	85.62	79.37 (16.27)
DRDA	81.37	62.86	63.63	95.94	93.56	88.19	85.00	95.25	**90.00**	83.98 (11.94)
Ours	**83.33**	**74.29**	**72.65**	96.09	**95.97**	**88.84**	**92.24**	**96.09**	88.16	**87.52 (8.54)**

Highest values are highlighted in boldface.

models. The verification accuracy for each subject and the averaged accuracy across these models are displayed in **Table 6**. From the average accuracy column, it is clear that only using the standard cross-entropy (in the case of $\alpha_{lsr} = 0$) for the model training is not an excellent choice. Proper selection of α_{lsr} can improve the CNN-based model's verification accuracy on the MI recognition. Second, we also observe that the model performance remains relatively stable across different values of α_{lsr} in a range of [0.25, 1]. This phenomenon suggests that different levels of smoothness on labels may have a similar effect on the model performance in the MI area. Finally, it can be seen that the model has the most remarkable improvements on subjects A02 and A05, who have low predicting accuracy, by using the smoothed cross-entropy compared to using the standard one. We further visualize the training and testing loss during the optimization of these two subjects in **Figure 3**. It is recognized that the models using standard cross-entropy suffer from an overfitting issue where the training loss decreases at the beginning and flattens gradually, but the testing loss decreases at the beginning while increases after several epochs. On the contrary, the models using smoothed cross-entropy have a good learning curve. Both training and testing errors decrease at the beginning and then flatten until the end of optimization. Together with the verification accuracy improvement, this finding suggests that the label smoothing technique can degrade the influence of the overfitting on the CNN model in the MI classification.

TABLE 5 | Paired *t*-test (*p*-values) between our method and others.

Model	**IV-2a**	**IV-2b**
FBCSP	0.0002	0.0058
SMM	0.0005	0.0001
EEGNet	0.0443	0.0013
ConNet	0.0168	0.0228
DRDA	0.0120	0.0479

5.2.2. Center Loss

In the second experiment, we fix the α_{lsr} as 0.5 (performing the best in the first experiment) and vary α_{ct} in a range from 0 to 1 (inclusive) to train different models. The performances of these models are displayed in **Table 7**. When the α_{ct} is larger than zero, the center loss is activated, and the performance improves across almost all subjects and in the averaged level. This result suggests that the involvement of the center loss increases the discriminative power of the model. For a clear illustration and intuition, the principal component analysis (PCA) (Jolliffe and Cadima, 2016) is further employed to convert the high dimensional features of the second last layer in the model for subject A07 (randomly selected) into 2-D vectors. The distributions of these vectors are shown in **Figure 4**. It is clear that, without the center loss (**Figure 4A**), the deep features within each class are dispersive and have a larger intra-class variation. Alternatively, with the center loss in the joint supervision (**Figure 4B**), the features have both a compact intraclass distance and a clear interclass boundary. These results suggest that the center loss is beneficial to the discriminative feature learning for MI classification modeling.

TABLE 6 | Classification accuracies (%) of models with different α_{lsr} values on the BCI competition IV-2a.

α_{lsr}	A01	A02	A03	A04	A05	A06	A07	A0	A09	Average
0	84.34	52.65	**93.04**	66.67	52.90	57.00	87.73	81.55	80.3	72.91
0.25	80.78	**64.66**	92.67	67.54	69.57	**60.00**	85.92	80.07	78.79	75.58
0.5	**84.34**	61.48	90.84	**72.37**	**73.19**	59.53	85.56	**82.66**	**81.44**	**76.82**
0.75	82.92	63.60	90.11	66.23	72.46	53.49	85.92	82.29	81.44	75.39
1	81.14	62.19	91.94	64.04	69.20	58.14	**89.53**	80.07	82.58	75.43

The α_{ct} is fixed as 0. The best accuracy for each subject is highlighted in boldface.

TABLE 7 | Classification accuracies (%) of models with different α_{ct} values on the BCI competition IV-2a.

α_{ct}	A01	A02	A03	A04	A05	A06	A07	A08	A09	Average
0	84.34	61.48	90.84	72.37	73.19	59.53	85.56	82.66	81.44	76.82
0.25	87.54	64.31	94.14	73.25	74.64	**63.26**	86.64	80.07	81.82	78.41
0.5	**89.32**	**66.78**	**94.14**	**74.56**	**76.45**	62.33	86.28	**85.61**	**85.23**	**80.08**
0.75	89.32	66.43	93.77	73.68	71.38	60.47	89.17	84.5 0	85.23	79.33
1	87.9	59.72	93.77	71.49	74.64	57.67	**89.89**	83.39	81.82	77.81

The α_{lsr} is fixed as 0.5. The best accuracy for each subject is highlighted in boldface.

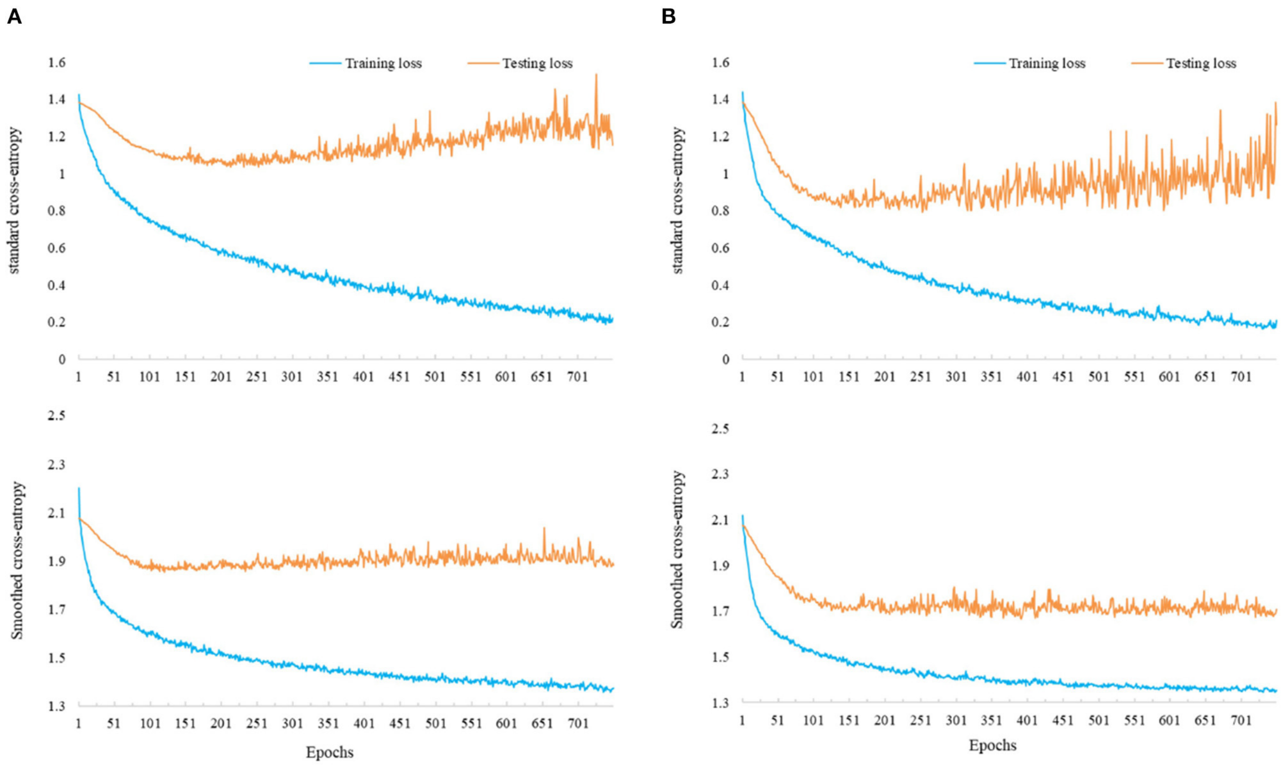

FIGURE 3 | Training and testing losses of the models using standard cross-entropy (top) and smoothed cross-entropy (bottom). **(A)** Learning curve of the model on A02. **(B)** Learning curve of the model on A05.

6. CONCLUSION

This paper proposes a new deep learning scheme for the CNN-based model in the MI classification. By jointly combining the smoothed cross-entropy with center loss, the robustness and discriminative power of the model can be highly enhanced for the classification. Extensive and systematic experiments are conducted to validate our strategy on two well-known

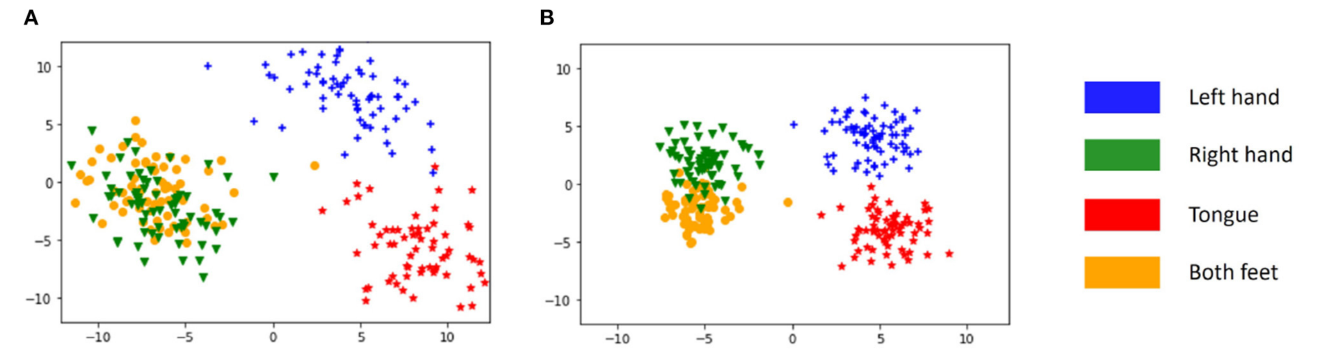

FIGURE 4 | Visualization of deep feature distributions of subject A07 on dataset IV-2a under two conditions. **(A)** Model with $\alpha_{lsr} = 0.5$ and $\alpha_{ct} = 0$. **(B)** Model with $\alpha_{lsr} = 0.5$ and $\alpha_{ct} = 0.5$.

benchmarks. Several suggestions have been made based on experimental findings. First, the label smoothing technique can degrade the overfitting issue caused by the scarcity and low SNR of the EEG data on the CNN model training. Second, the center loss along with the cross-entropy efficiently decreases the intra-class variance and thus increases the discriminative ability of the deep features by pulling them toward their corresponding latent class centers. It reduces the negative impact of the non-stationary characteristics of the EEG data on the MI classification task. Finally, the proposed loss offers a robust and discriminative training scheme for CNN-based modeling in the MI area. This phenomenon uncovers the fact that, in addition to sophisticated model structure development, implementing an efficient loss function for the learning guidance is also beneficial to model performance in MI recognition. This research has thus encouraged more attempts on the objective function innovation for the deep learning model in the MI field. It can be an interesting alternative for overcoming the bottleneck performance to the model architecture that has been heavily investigated.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://www.bbci.de/competition/iv/.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

XH conceptualized the study, performed the majority of the experiments and analyses, made the figures, and wrote the first draft of the manuscript. NZ and K-SC performed some experiments, updated the figures, performed the statistics, and edited the manuscript. All authors approved the submitted version.

FUNDING

The work in this paper was supported in part by the Hong Kong Innovation and Technology Fund (MRP/015/18), the Hong Kong Research Grants Council (PolyU 152006/19E), and National Nature Science Foundation of China (Nos. 61802036 and 11901063).

REFERENCES

Ahn, M., and Jun, S. C. (2015). Performance variation in motor imagery brain-computer interface: a brief review. *J. Neurosci. Methods* 243, 103–110. doi: 10.1016/j.jneumeth.2015.01.033

Ang, K. K., Chin, Z. Y., Zhang, H., and Guan, C. (2008). "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (Hong Kong), 2390–2397.

Bashivan, P., Rish, I., Yeasin, M., and Codella, N. (2015). Learning representations from EEG with deep recurrent-convolutional neural networks. *arXiv [Preprint]. arXiv:1511.06448.*

Bickel, S., Brückner, M., and Scheffer, T. (2009). Discriminative learning under covariate shift. *J. Mach. Learn. Res.* 10, 2137–2155. doi: 10.1145/1577069.1755858

Chorowski, J., and Jaitly, N. (2016). Towards better decoding and language model integration in sequence to sequence models. *arXiv [Preprint]. arXiv:1612.02695.* doi: 10.21437/Interspeech.2017-343

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv [Preprint]. arXiv:1511.07289.*

Goldenholz, D. M., Ahlfors, S. P., Hämäläinen, M. S., Sharon, D., Ishitobi, M., Vaina, L. M., et al. (2009). Mapping the signal-to-noise-ratios of cortical sources in magnetoencephalography and electroencephalography. *Hum. Brain Mapp.* 30, 1077–1086. doi: 10.1002/hbm.20571

Hersche, M., Rellstab, T., Schiavone, P. D., Cavigelli, L., Benini, L., and Rahimi, A. (2018). "Fast and accurate multiclass inference for mi-bcis using large multiscale temporal and spectral features," in *2018 26th European Signal Processing Conference (EUSIPCO)* (Rome), 1690–1694. doi: 10.23919/EUSIPCO.2018.8553378

Hertz, J., Krogh, A., and Palmer, R. G. (2018). *Introduction to the Theory of Neural Computation*. Redwood City: CRC Press. doi: 10.1201/9780429499661

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning* (Lille), 448–456.

Jolliffe, I. T., and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A* 374:20150202. doi: 10.1098/rsta.2015.0202

Kaufmann, T., Herweg, A., and Kübler, A. (2014). Toward brain-computer interface based wheelchair control utilizing tactually-evoked event-related potentials. *J. Neuroeng. Rehabil*. 11, 1–17. doi: 10.1186/1743-0003-11-7

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint]. arXiv:1412.6980*.

Kumar, S., Sharma, A., Mamun, K., and Tsunoda, T. (2016). "A deep learning approach for motor imagery EEG signal classification," in *2016 3rd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)* (Nadi), 34–39. doi: 10.1109/APWC-on-CSE.2016.017

Lal, T. N., Schroder, M., Hinterberger, T., Weston, J., Bogdan, M., Birbaumer, N., et al. (2004). Support vector channel selection in BCI. *IEEE Trans. Biomed. Eng*. 51, 1003–1010. doi: 10.1109/TBME.2004.827827

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). Eegnet: a compact convolutional neural network for EEG-based brain-computer interfaces. *J. Neural Eng*. 15:056013. doi: 10.1088/1741-2552/aace8c

Lee, T.-W., Girolami, M., and Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Comput*. 11, 417–441. doi: 10.1162/089976699300016719

Martinez-Leon, J.-A., Cano-Izquierdo, J.-M., and Ibarrola, J. (2015). Feature selection applying statistical and neurofuzzy methods to EEG-based BCI. *Comput. Intell. Neurosci*. 2015:781207. doi: 10.1155/2015/781207

Müller, R., Kornblith, S., and Hinton, G. (2019). When does label smoothing help? *arXiv [Preprint]. arXiv:1906.02629*.

Ni, T., Gu, X., and Jiang, Y. (2020). Transfer discriminative dictionary learning with label consistency for classification of EEG signals of epilepsy. *J. Ambient Intell. Human. Comput*. 1–12. doi: 10.1007/s12652-020-02620-9

Pfurtscheller, G., and Neuper, C. (2001). Motor imagery and direct brain-computer communication. *Proc. IEEE* 89, 1123–1134. doi: 10.1109/5.939829

Ramoser, H., Muller-Gerking, J., and Pfurtscheller, G. (2000). Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans. Rehabil. Eng*. 8, 441–446. doi: 10.1109/86.895946

Rezeika, A., Benda, M., Stawicki, P., Gembler, F., Saboor, A., and Volosyak, I. (2018). Brain-computer interface spellers: a review. *Brain Sci*. 8:57. doi: 10.3390/brainsci8040057

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng*. 16:051001. doi: 10.1088/1741-2552/ab260c

Sakhavi, S., Guan, C., and Yan, S. (2018). Learning temporal information for brain-computer interface using convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst*. 29, 5619–5629. doi: 10.1109/TNNLS.2018.2789927

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp*. 38, 5391–5420. doi: 10.1002/hbm.23730

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas), 2818–2826. doi: 10.1109/CVPR.2016.308

Tabar, Y. R., and Halici, U. (2016). A novel deep learning approach for classification of EEG motor imagery signals. *J. Neural Eng*. 14:016003. doi: 10.1088/1741-2560/14/1/016003

Tangermann, M., Müller, K.-R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., et al. (2012). Review of the BCI competition IV. *Front. Neurosci*. 6:55. doi: 10.3389/fnins.2012.00055

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Long Beach), 5998–6008.

Vidaurre, C., Klauer, C., Schauer, T., Ramos-Murguialday, A., and Müller, K.-R. (2016). Eeg-based bci for the linear control of an upper-limb neuroprosthesis. *Med. Eng. Phys*. 38, 1195–1204. doi: 10.1016/j.medengphy.2016.06.010

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision* (Amsterdam: Springer), 499–515. doi: 10.1007/978-3-319-46478-7_31

Wierzgała, P., Zapała, D., Wojcik, G. M., and Masiak, J. (2018). Most popular signal processing methods in motor-imagery BCI: a review and meta-analysis. *Front. Neuroinformatics* 12:78. doi: 10.3389/fninf.2018.00078

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clin. Neurophysiol*. 113, 767–791. doi: 10.1016/S1388-2457(02)00057-3

Yu, H., Lu, H., Wang, S., Xia, K., Jiang, Y., and Qian, P. (2019). A general common spatial patterns for EEG analysis with applications to vigilance detection. *IEEE Access* 7, 111102–111114. doi: 10.1109/ACCESS.2019.2934519

Zhang, Y., Li, X., Zhu, J., Wu, C., and Wu, Q. (2019). Epileptic EEG signals recognition using a deep view-reduction tsk fuzzy system with high interpretability. *IEEE Access* 7, 137344–137354. doi: 10.1109/ACCESS.2019.2942641

Zhang, Y., Zhou, Z., Bai, H., Liu, W., and Wang, L. (2020). Seizure classification from EEG signals using an online selective transfer tsk fuzzy classifier with joint distribution adaption and manifold regularization. *Front. Neurosci*. 14:496. doi: 10.3389/fnins.2020.00496

Zhao, H., Zheng, Q., Ma, K., Li, H., and Zheng, Y. (2020). Deep representation-based domain adaptation for nonstationary EEG classification. *IEEE Trans. Neural Netw. Learn. Syst*. 32, 535–545. doi: 10.1109/TNNLS.2020.3010780

Zheng, Q., Zhu, F., Qin, J., Chen, B., and Heng, P.-A. (2018). Sparse support matrix machine. *Pattern Recogn*. 76, 715–726. doi: 10.1016/j.patcog.2017.10.003

Zhou, H., and Li, L. (2014). Regularized matrix regression. *J. R. Stat. Soc. Ser. B* 76, 463–483. doi: 10.1111/rssb.12031

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.