frontiers in Neuroscience

ORIGINAL RESEARCH
published: 09 May 2018
doi: 10.3389/fnins.2018.00272

Improving Generalization Based on l_1-Norm Regularization for EEG-Based Motor Imagery Classification

Yuwei Zhao[1†], Jiuqi Han[1†], Yushu Chen[1], Hongji Sun[1], Jiayun Chen[1,2], Ang Ke[1,3], Yao Han[1,4], Peng Zhang[1,3], Yi Zhang[1], Jin Zhou[1] and Changyong Wang[1*]*

[1] Department of Neural Engineering and Biological Interdisciplinary Studies, Institute of Military Cognition and Brain Sciences, Academy of Military Medical Sciences, Beijing, China, [2] College of Life Science and Technology, Huazhong Agricultural University, Wuhan, China, [3] Neural Interface & Rehabilitation Technology Research Center, Huazhong University of Science and Technology, Wuhan, China, [4] Stem Cell and Tissue Engineering Lab, Beijing Institute of Transfusion Medicine, Beijing, China

OPEN ACCESS

Edited by:
Ioan Opris,
University of Miami, United States

Reviewed by:
Dezhong Yao,
University of Electronic Science and Technology of China, China
Zhong Yin,
University of Shanghai for Science and Technology, China

***Correspondence:**
Jin Zhou
sisun819@yahoo.com
Changyong Wang
wcy2000_te@yahoo.com

[†]These authors have contributed equally to this work.

Specialty section:
This article was submitted to Neural Technology, a section of the journal Frontiers in Neuroscience

Received: 05 January 2018
Accepted: 09 April 2018
Published: 09 May 2018

Citation:
Zhao Y, Han J, Chen Y, Sun H, Chen J, Ke A, Han Y, Zhang P, Zhang Y, Zhou J and Wang C (2018) Improving Generalization Based on l_1-Norm Regularization for EEG-Based Motor Imagery Classification. Front. Neurosci. 12:272. doi: 10.3389/fnins.2018.00272

Multichannel electroencephalography (EEG) is widely used in typical brain-computer interface (BCI) systems. In general, a number of parameters are essential for a EEG classification algorithm due to redundant features involved in EEG signals. However, the generalization of the EEG method is often adversely affected by the model complexity, considerably coherent with its number of undetermined parameters, further leading to heavy overfitting. To decrease the complexity and improve the generalization of EEG method, we present a novel l_1-norm-based approach to combine the decision value obtained from each EEG channel directly. By extracting the information from different channels on independent frequency bands (FB) with l1-norm regularization, the method proposed fits the training data with much less parameters compared to common spatial pattern (CSP) methods in order to reduce overfitting. Moreover, an effective and efficient solution to minimize the optimization object is proposed. The experimental results on dataset IVa of BCI competition III and dataset I of BCI competition IV show that, the proposed method contributes to high classification accuracy and increases generalization performance for the classification of MI EEG. As the training set ratio decreases from 80 to 20%, the average classification accuracy on the two datasets changes from 85.86 and 86.13% to 84.81 and 76.59%, respectively. The classification performance and generalization of the proposed method contribute to the practical application of MI based BCI systems.

Keywords: motor imagery, electroencephalography (EEG), classification, l_1-norm regularization, generalization

1. INTRODUCTION

Noninvasive brain-computer interface (BCI) based on electroencephalography (EEG) has attracted an increasing interest in recent decades owing to its significant potential in practical applications (Wolpaw et al., 2002; Nicolas-Alonso and Gomez-Gil, 2012). For example, motor imagery EEG (MI-EEG) offers users direct control of different devices such as a wheelchair, quadcopter, or robotic arm (Graimann et al., 2008; Lafleur et al., 2013; Meng et al., 2016) through the modulation of thought without external stimuli. Typical MI-EEG data is composed of multichannel signals

recorded from several electrodes placed on the scalp corresponding to the motor-relevant cortex (Blankertz et al.,, 2008). In order to achieve high classification accuracy, merging of signals from scalp spatial districts is required to suppress the data noise cause by imperfect conductivity of human tissues.

Apparently, as a mirror of the total brain activity in specific regions, multichannel EEG signals interact with each other intrinsically. This interaction is believed to originate from the fundamental mechanism of the information processing within the brain, such as the distributed and co-related function of different cerebral cortex (Baillet et al., 2001). Thus, a specific brain activity is typically mirrored by more than one site on the scalp, leading to considerably redundant information involved in multichannel EEG signals. Moreover, informative EEG features such as task relevant and event-related potentials are likely mixed with blurred features and submerged into the raw data owing to the artifacts and merging effects of the conductive scalp and skull (Pfurtscheller et al., 2006). Due to the insufficient EEG data for classifier training, the complexity of classification algorithms may increase with redundant features involved in EEG signals, adversely affecting their generalization.

In past decades, numerous literatures, such as the common spatial patterns (CSP) (Müllergerking et al., 1999; Ramoser et al., 2000), have focused on this research area. CSP usually amplifies the class disparation in spatial domain by covariance analysis. However, it ignores the task related differences across local regions in frequency domain, which is also important in processing rhythmic activities such as motor imagery EEG. Besides, it may bring in relatively large number of undetermined parameters (the dimensions are the number of output channels multiply by the number of input channels), leading to complicated models, and therefore vulnerable to overfitting especially when the training samples are insufficient.

To avoid this limitation and reduce overfitting during EEG classification, we propose a novel framework named COL (Channel optimization based on l_1-norm). For the sake of mitigating generalization error caused by overfitting, we introduce a sparse l1-norm regularization to solve the optimal weights of channels during combination of each channel's decision value, in which the sparse optimal weights are solved by minimizing the least square error between the predicted labels and the real labels. The optimized model has only a few feature parameters, that is, the channel number, the upper/lower frequency band, and the weight. Benefited from extracting the information from different channels on independent frequency bands with L1-norm regularization, the algorithms proposed fits the training data with much less parameters compared to CSP methods, which enables it to reduce overfitting.

Experimental results on real world datasets demonstrate the effectiveness and robustness of the proposed method, validating its generalization in practical applications.

Our main contributions are highlighted in the following:

- We provide a simple but effective model to reduce overfitting in EEG classification by reducing the number of undetermined parameters.
- We introduce an effective and efficient iterative solution to train the model.
- We demonstrate the superiority of the generalization of our methods on real world datasets.

The remainder of this paper is organized as follows. In section 2, we overview related works. We formulate the proposed method and provide an efficient solution and complexity analysis in section 3. A description of the datasets, the details of the experimental setups, experimental results and discussion are presented in section 4, followed by our conclusions in section 5.

2. RELATED WORKS

Significant efforts have been made in the classification of motor imagery EEG signals. A key point to promote the accuracy of classification algorithms is to prevent overfitting during EEG classification. Here we give a brief review of existing methods for EEG classification from two strategies, and some efforts to reduce overfitting.

Because of the characteristics of different regions of the brain, a number of researches have attempted to process signals from different channels independently. An approach was presented to determine the contribution of different bandwidths of the EEG signal in different recording sites using the multiple kernel learning (MKL) method in Schrouff et al. (2016). Channel-frequency map (CFM) was proposed as a tool to develop data-driven frequency band selection methods for parallel EEG processing in Suk and Lee (2011). Genetic algorithm was utilized to identify individually optimized brain areas and frequency ranges based on a predefined chromosome simultaneously in Lee et al. (2012). Popular deep learning was also introduced in this area. For example, deep belief network (DBN) was employed to reveal the critical frequency bands for emotion recognition (Zheng et al., 2015). Support vector machine (SVM) was considered as a useful method to solve small sample and nonlinear classification problems (Boser et al., 1992). SVM was applied in the feature optimization and classification of MI-EEG (Chatterjee and Bandyopadhyay, 2016; Ma et al., 2016), resulting in a speedup of classification while loss in generalization remained acceptable (Xu et al., 2010). Hybrid spatial finite impulse response (FIR) filters of high-order and data-driven were channel-specifically designed to complement broadband CSP filtering in Yu et al. (2013). In this manner, they facilitate the study of the specific properties of the channels. Nevertheless, their disregard of the interaction among channels likely submerged significant data into irrelevant and redundant signals, negatively influencing the classification performance. Another disadvantage of this approach is the significant computational burden related to the enormous volume of signals.

There have also been several researches that have attempted to address the combination of multichannel EEG data. Well-known CSP methods combined signals from multiple channels by amplifying the class disparity in the spatial domain by covariance analysis (Blankertz et al., 2007; Li et al., 2013). Improved CSPs, such as common spatio-spectral pattern (CSSP)

(Lemm et al., 2005), iterative spatio-spectral pattern learning (ISSPL) (Wu et al., 2008), and filter bank common spatial pattern (FBCSP) (Kai et al., 2012) were introduced to optimize the combination of multichannel signals by designing novel spectral weight coefficient evaluation. Another spatial filtering algorithm called discriminative spatial patten (DSP) solved single trial EEG classification by maximizing the between-class separation (Duda et al., 2001; Hoffmann et al., 2006). CSP and DSP were combined to more efficient feature extraction and classification of single trial EEG during finger movement tasks (Liao et al., 2007). In addition to these methods, there are numerous researches focusing on subset selection of EEG channels. Based on grouped automatic relevance determination, group-sparse Bayesian linear discriminant analysis (gsBLDA) was presented to select EEG channels (Yu et al., 2015). The Separability & Correlation (SEPCOR) approach was designed to automatically search for an optimal EEG channel subset with minimum correlation and maximum class separation (Shri and Sriraam, 2016; Student and Sriraam, 2017). Sequential floating forward selection (SFFS) performed a loop of channel selection continuously by iteratively adding and eliminating EEG channels (Pudil et al., 1994; Meng et al., 2011). By considering adjacent channels as one feature according to their distribution on the cerebral cortex, an improved SFFS (ISFFS) was proposed to remove task-irrelevant and redundant channels with low computational burden (Qiu et al., 2016). In order to reduce overfitting, L1 norm regularization was applied in constructing spatial filters for its competence to achieve sparse solution (Silva et al., 2004; Donoho, 2006; Farquhar et al., 2006). Sparse common spatial pattern (SCSP) was applied to optimally select the least number of channels while containing high performance in classification, with l1/l2 norm as the regularization term (Arvaneh et al., 2011). By combining L1 norm based Eigen decomposition into CSP, a L1-norm based CSP was proposed to effectively improve the robustness of BCI system to EEG outliers and achieved higher classification accuracy than the conventional CSP (Li et al., 2013). A modified CSP with l1 sparse weighting method was developed for EEG trial selection, and successfully rejected low-quality trials in a sparsity-aware way (Tomida et al., 2015). These approaches are effective in determining the informative subset or combination weights of channels based on shallow features extracted from voltage signals. However, the CSP in EEG classification generates a spatial filter matrix that generally contains too many parameters, and therefore vulnerable to be overfitting especially when insufficiency training data is available. A model that requires few number of parameters while utilizing the features right related to task will be potential for EEG classification.

3. PROPOSED METHODS

In this section, we introduce notations used throughout this paper and present the concrete formulation of the proposed method. We then provide a simple yet effective algorithm to solve this problem, followed by an analysis of its computational complexity.

3.1. Notations

In this document, scalars, matrices, vectors, sets, and functions are denoted as small, boldface capital, boldface lowercase, fraktur capital, and script capital letters, respectively. $\mathbf{x}^T$, $\mathbf{X}^T$, $\mathbf{x}_i$, $\mathbf{X}_i$, $\mathbf{X}_{ij}$, $\mathbf{X}_{(i,:)}$ and $\mathbf{X}_{(:,j)}$ indicate the transpose of vector $\mathbf{x}$, the transpose of matrix $\mathbf{X}$, the i-th element of $\mathbf{x}$, the i-th sample of the variable $\mathbf{X}$, the element of $\mathbf{X}$ occurring in the i-th row and j-th column, the i-th row of $\mathbf{X}$ and the j-th column of $\mathbf{X}$ respectively. Moreover, $||\mathbf{x}||_1$ is the l_1-norm of $\mathbf{x}$, $||\mathbf{X}||_1$ and $||\mathbf{X}||_2$ are the m_1-norm and m_2-norm of matrix $\mathbf{X}$. See the Appendix section for definitions of norm terms.

3.2. Problem Formulation

In order to reduce the model complexity and the number of undecided parameters, we firstly generate features (rough dichotomous probabilities) of each single channel. Then, features of all channels are selected by sorting each channel according to their Fisher Criterion scores (F-score) (Müller et al., 2004; Duda et al., 2012), denoting distance between the class means in relation to the intra-class variances.

Afterwards, we introduce a l_1-norm regularized sparse least square regression to directly minimize the error between predicted and ground-truth labels, contributing to a simplified model with optimized parameters. The optimized model has much fewer parameters than most existing models.

Assume that we have recorded EEG signals of N trials, and let $\mathfrak{X} = \{\mathbf{X}_i\}_{i=1}^N$ be the set of EEG signal corresponding to the i-th trial of motor imagery. Specifically, we represent the segment of EEG signals as matrix $\mathbf{X}^{M\times C}$, where M and C are the number of sampled time points and channels in a trial respectively. The class indicator vector can be denoted as $\tilde{\mathbf{y}} \in \{0,1\}^N$, where $\tilde{\mathbf{y}}_i = 0$ and $\tilde{\mathbf{y}}_i = 1$ indicate that the i-th trial is left/right hand and foot motor imagery, respectively.

3.2.1. Extracting and Selecting Features From Each Channel

Suppose that all channels are independent from each other, we take a signal vector $\mathbf{x}^{M\times 1}$ of one channel as an example to define the feature extraction and selection method.

First, we remove the average values channel-wise by applying the common average reference (CAR), which is widely used in EEG preprocessing (Offner, 1950; Wu and Ge, 2013), that is

$$\mathbf{x} \leftarrow \mathbf{x} - \bar{\mathbf{x}}, \tag{1}$$

where $\bar{\mathbf{x}}$ is the average over all values of $\mathbf{x}$ at each channel.

Then, $\mathbf{x}$ is preprocessed by a band-pass filter. The upper bound f_{max} and lower bound f_{min} of the filter is chosen from the frequency list $\mathfrak{F} = \left\{f_0\theta^{n_f} | n_f \in \{0,1,...,N_f\}\right\}$, where f_0 is the base frequency, θ is a constant scaling factor, n_f is the selected power coefficient, and N_f is the number of candidate element frequency bands. With the band pass filtered signals, the envelope data is obtained using a discrete-time Hilbert transform, whose complex magnitude, denoted as $\hat{\mathbf{x}}$, is used to replace $\mathbf{x}$ for further feature extraction.

Afterwards, we extract its feature as

$$\gamma = \log(\frac{1}{M}||\hat{\mathbf{x}}||_2^2). \tag{2}$$

Next, we determine f_{max} and f_{min} by maximizing the F-score (Müller et al., 2004; Duda et al., 2012), which are determined by

$$F\text{-}score = \frac{(\overline{\gamma^+} - \overline{\gamma^-})^2}{\overline{(\gamma^+ - \overline{\gamma^+})^2} + \overline{(\gamma^- - \overline{\gamma^-})^2}}, \tag{3}$$

where γ^+ and γ^- denote the features of trials labeled "1" or "0," respectively.

Lastly, the predicted label, the main feature from signal $\mathbf{x}$ of the current channel, can be obtained by

$$p = \mathcal{S}\left(\frac{\delta(\overline{\gamma^+} - \overline{\gamma^-})}{\sqrt{\Delta_\gamma}}\left(\gamma - \frac{\overline{\gamma^+} + \overline{\gamma^-}}{2}\right)\right), \tag{4}$$

where $\mathcal{S}(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, $\delta(x)$ is the sign function, and $\Delta(x)$ is the variance of x.

3.2.2. Defining the Object of Our Simplified Model

With each element gained from the above section channel-by-channel and trial-by-trial, we could get the feature matrix $\mathbf{P}$. Thus, we define the final decision value as $\mathbf{p} = \mathbf{P}\mathbf{w} + \mathbf{1}^T b \in [0,1]^N$, where $\mathbf{P}_{ij}^{N\times C}$, $\mathbf{w}^{C\times 1}$ and $\mathbf{1}^{1\times N}$, b is the predicted label gained in the i-th trial by signals in the j-th channel, the weights of C channels, a vector of all elements "1" and the bias of all trials. Inspired by the least absolute shrinkage and selection operator (lasso) (Tibshirani, 2011) , the object can be written as

$$\min \mathcal{L} : \min \frac{1}{2}||\mathbf{p} - \tilde{\mathbf{y}}||_2^2 + \alpha\mathcal{R}(\mathbf{w}), \tag{5}$$

or

$$\min \mathcal{L} : \min \frac{1}{2}||\mathbf{P}\mathbf{w} + \mathbf{1}^T b - \tilde{\mathbf{y}}||_2^2 + \alpha\mathcal{R}(\mathbf{w}). \tag{6}$$

where α is a balance parameter proportional to N and $\mathcal{R}(\mathbf{w})$ is a regularization term on $\mathbf{w}$.

Clearly, $\mathbf{P}$ is obtained through the above section. Thereby, the undetermined variables in Equation (6) are $\mathbf{w}$ and b[1], if we properly define functions $\mathcal{R}(\mathbf{w})$.

Recalling that there is frequently redundant and irrelevant channels in practical MI-BCI, we can define $\mathcal{R}(\mathbf{w})$ as a sparsity metric on $\mathbf{w}$, such as its l_1-norm. Therefore, the optimization problem we must solve can be expressed as

$$\begin{aligned} &\min \mathcal{L}(\mathbf{w}, b) : \\ &\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{P}\mathbf{w} + \mathbf{1}^T b - \tilde{\mathbf{y}}||_2^2 + \alpha||\mathbf{w}||_1. \end{aligned} \tag{7}$$

3.3. Solution to the Formulation

Intuitively, an iterative multiplicative updating procedure is designed to solve Equation (7). In each step, we first fix $\mathbf{w}$ to determine the optimal b, and then solve $\mathbf{w}$ by fixing b.

[1]In our proposed method, only $\mathbf{w}$ and b need to be trained. That means that only $C + 1$ parameters need to be trained, much fewer than existing approaches, including CSP. Thus, intuitively, the generalization of our method is better and it is probable to train our models with relatively small number of labeled samples. The experimental results in section4 demonstrate this hypothesis.

For Equation (7), we redefine the object of b as

$$\begin{aligned} &\min \mathcal{G}(b) : \\ &\min_b \frac{1}{2}||\mathbf{P}\mathbf{w} + \mathbf{1}^T b - \tilde{\mathbf{y}}||_2^2. \end{aligned} \tag{8}$$

It should be noted that the inequality $||\mathbf{A}||_2^2 + ||\mathbf{B}||_2^2 \geq \frac{1}{2}||\mathbf{A}+\mathbf{B}||_2^2$ holds. Thus, we can determine that $||\mathbf{P}\mathbf{w} + \mathbf{1}^T b_1 - \tilde{\mathbf{y}}||_2^2 + ||\mathbf{P}\mathbf{w} + \mathbf{1}^T b_2 - \tilde{\mathbf{y}}||_2^2 \geq 2||\mathbf{P}\mathbf{w} + \mathbf{1}^T\frac{b_1+b_2}{2} - \tilde{\mathbf{y}}||_2^2$, which indicates that $\mathcal{G}(\mathbf{b})$ is convex in terms of $\mathbf{b}$. Therefore, we have

$$\begin{aligned} \frac{\partial \mathcal{G}(b)}{\partial b} &= \mathbf{1}(\mathbf{P}\mathbf{w} - \tilde{\mathbf{y}}) + \mathbf{1}\mathbf{1}^T b \\ &= \mathbf{1}(\mathbf{P}\mathbf{w} - \tilde{\mathbf{y}}) + Nb. \end{aligned} \tag{9}$$

Then, by setting $\frac{\partial \mathcal{G}(b)}{\partial b} = 0$, we obtain the optimal b as

$$b \leftarrow \frac{\mathbf{1}(\mathbf{P}\mathbf{w} - \tilde{\mathbf{y}})}{N}. \tag{10}$$

Similarly, with b fixed as in Equation (10), we redefine the object of $\mathbf{w}$ as

$$\begin{aligned} &\min \mathcal{H}(\mathbf{w}) : \\ &\min_{\mathbf{w}} \frac{1}{2}||\mathbf{P}\mathbf{w} + \frac{\mathbf{1}^T\mathbf{1}(\mathbf{P}\mathbf{w} - \tilde{\mathbf{y}})}{N} - \tilde{\mathbf{y}}||_2^2 + \alpha||\mathbf{w}||_1 \\ &\Leftrightarrow \min_{\mathbf{w}} \frac{1}{2}||(\mathbf{P} + \frac{\mathbf{1}^T\mathbf{1}\mathbf{P}}{N})\mathbf{w} - (\mathbf{I}_N + \frac{\mathbf{1}^T\mathbf{1}}{N})\tilde{\mathbf{y}}||_2^2 + \alpha||\mathbf{w}||_1. \end{aligned} \tag{11}$$

where $\Leftrightarrow$, $\mathbf{I}_N$ denotes equivalence and the N-by-N identity matrix.

We exploit the gradient descent method to optimize $\mathbf{w}$ with positive initialization values.

With Equation (11), we have

$$\begin{aligned} \nabla_{\mathbf{w}} &= \frac{\partial \mathcal{H}(\mathbf{w})}{\partial \mathbf{w}} \\ &= (\mathbf{P} + \frac{\mathbf{1}^T\mathbf{1}\mathbf{P}}{N})^T(\mathbf{P} + \frac{\mathbf{1}^T\mathbf{1}\mathbf{P}}{N})\mathbf{w} - (\mathbf{P} + \frac{\mathbf{1}^T\mathbf{1}\mathbf{P}}{N})^T(\mathbf{I} + \frac{\mathbf{1}^T\mathbf{1}}{N})\tilde{\mathbf{y}} \\ &+ \alpha\delta(\mathbf{w}) \\ &= (\frac{3(\mathbf{1}\mathbf{P})^T(\mathbf{1}\mathbf{P})}{N} + \mathbf{P}^T\mathbf{P})\mathbf{w} - (\mathbf{P}^T + \frac{3\mathbf{P}^T\mathbf{1}^T\mathbf{1}}{N})\tilde{\mathbf{y}} + \alpha\delta(\mathbf{w}). \end{aligned} \tag{12}$$

where $\delta(\mathbf{w})$ is the sign function. Therefore, we have the update rule

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\nabla_{\mathbf{w}}, \tag{13}$$

where $\eta > 0$ is the step length determined by Algorithm 1, with which the objective value is minimized along the negative gradient direction.

With positive initialization values, $\mathbf{w}$ decreases gradually toward zero. Once any element of $\mathbf{w}$ reaches zero, signals from the corresponding channel are removed, and the updating in terms of this channel is terminated.

Thus, with a new trial, we can first preprocess the EEG data using band-pass filtering and average removing according to the

Algorithm 1 Algorithm to determine η in Equation (13)

Input:

Current $\mathbf{w}$ and the corresponding value of the object function $\mathcal{H}(\mathbf{w})$, as well as the gradient $\nabla_{\mathbf{w}}$.

Output:

Step length η.

Initialize the linear step length $\eta_l = \frac{\mathcal{H}(\mathbf{w})}{||\nabla_{\mathbf{w}}||_2^2}$;
Compute the maximum step length under the nonnegative constraint $\boldsymbol{\eta}_{nn}$ by dividing $\mathbf{w}$ by $\nabla_{\mathbf{w}}$ element-wise;
Preserve all positive elements of $\boldsymbol{\eta}_{nn}$ as $\boldsymbol{\eta}_{nn}^{+}$;
Set the maximum step length $\eta_m \leftarrow \min\{\eta_l, \boldsymbol{\eta}_{nn}^{+}\}$;
while $\mathcal{H}(\mathbf{w} - \eta_m \nabla_{\mathbf{w}}) \geq \mathcal{H}(\mathbf{w})$ **do**
  Set $\eta_m \leftarrow \eta_m/2$;
end while
Set $\tilde{\eta} \leftarrow \eta_m$ and $\tilde{\mathcal{H}} \leftarrow \mathcal{H}(\mathbf{w} - \eta_m \nabla_{\mathbf{w}})$;
if $\mathcal{H}(\mathbf{w} - \frac{\eta_m}{2} \nabla_{\mathbf{w}}) < \tilde{\mathcal{H}}$ **then**
  Update $\tilde{\eta} = \frac{\eta_m}{2}$ and $\tilde{\mathcal{H}} = \mathcal{H}(\mathbf{w} - \frac{\eta_m}{2} \nabla_{\mathbf{w}})$;
end if
Compute the parabolic approximation parameters using

$$\begin{pmatrix} a_\eta \\ b_\eta \end{pmatrix} = \begin{pmatrix} \frac{\eta_m^2}{4} & \frac{\eta_m}{2} \\ \eta_m^2 & \eta_m \end{pmatrix}^{-1} \begin{pmatrix} \mathcal{H}(\mathbf{w} - \frac{\eta_m}{2} \nabla_{\mathbf{w}}) - \mathcal{H}(\mathbf{w}) \\ \mathcal{H}(\mathbf{w} - \eta_m \nabla_{\mathbf{w}}) - \mathcal{H}(\mathbf{w}) \end{pmatrix} \quad (14)$$

Set $\eta \leftarrow \min\{-\frac{b_\eta}{2a_\eta}, \eta_m\}$;
if $a_\eta \leq 0$ **then**
  Set $\eta = \eta_m$;
end if
if $\mathcal{H}(\mathbf{w} - \eta \nabla_{\mathbf{w}}) > \tilde{\mathcal{H}}$ **then**
  Set $\eta = \tilde{\eta}$.
end if

channel-specific f_{min} and f_{max}. Then, features are extracted using Equation (2), followed by obtaining the decision value channel-wise using Equation (4). Then, the combined predicted label p is computed with learnt $\mathbf{w}$ and b.

It should be noted that the combined predicted label can exceed $[0, 1]$, hence, we define another sigmoid function to normalize this as

$$p_{normal} = \mathcal{S}\big(\beta(p - 0.5)\big), \quad (15)$$

where β is a constant and we fix $\beta = 4$ to set the derivative on $p = 0.5$ as "1".

3.4. Flowchart of Algorithm

Based on the above analysis, we summarize the detailed optimization algorithm of COL in Algorithm 2.

3.5. Complexity Analysis

In this subsection, we analyze the time complexity of Algorithm 2. For all channels, searching for f_{min} and f_{max} requires $O(MNCN_f^2)$. Computing $\mathbf{P}$ requires $O(MNC)$ time. Further, in each iteration for optimizing $\mathbf{w}$ and b, updating $\mathbf{w}$ requires $O(NC^2)$ and b requires $O(NC)$ time. Thus, it requires $O(TNC^2)$ to update $\mathbf{w}$ and b. Therefore, the overall cost for Algorithm 2 is $O(MNCN_f^2 + TNC^2)$, where T is the number of iterations.

Algorithm 2 Algorithm to Solve COL

Input:

N EEG data matrices $X_1^{M\times C}$, $X_2^{M\times C}$,..., and $X_N^{M\times C}$; balance parameter α; number of candidate element frequency bands N_f; base frequency f_0 and constant scaling factor θ.

Output:

Weight vector of C channels $\mathbf{w}$; bias of N trials b.

for $c = 1, 2, ..., C$ **do**
  Remove average values from EEG signals channel-wise using Eq.(1);
  Initialize $n_{f_{min}} = 0$ and $n_{f_{max}} = 1$;
  Set $f_{min} = f_0\theta^{n_{f_{min}}}$ and $f_{max} = f_0\theta^{n_{f_{max}}}$;
  Filter signals by band pass defined by f_{min} and f_{max};
  Obtain the envelope data using a discrete-time Hilbert transform;
  Extract features using Eq.(2);
  Compute F-Score using Eq.(3), save it as the temporal best FS_{best}, store f_{min} and f_{max};
  for $n_{f_{min}} = 1, 2, 3, ..., N_f - 1$ **do**
    for $n_{f_{max}} = n_{f_{min}}, n_{f_{min}} + 1, ..., N_f$ **do**
      Set $f_{min} = f_0\theta^{n_{f_{min}}}$ and $f_{max} = f_0\theta^{n_{f_{max}}}$;
      Filter signals by band pass defined by f_{min} and f_{max};
      Extract features using Eq.(2);
      Compute F-Score using Eq.(3) as FS;
      if $FS > FS_{best}$ **then**
        Store f_{min} and f_{max};
        Update $FS_{best} \leftarrow FS$;
      end if
    end for
  end for
  Filter signals by band pass defined by f_{min} and f_{max};
end for
Obtain the decision matrix $\mathbf{P}$ element-wise using Eq.(4);
Initialize $\mathbf{w}$ and b.
repeat
  Fix $\mathbf{w}$, and update b using Eq.(10);
  Fix b, and update $\mathbf{w}$ using Eq.(13);
until Convergence criterion satisfied.

4. EXPERIMENTS

We compared the proposed COL with several classical state-of-the-art methods in terms of the classification and generalization performance. The experimental setups and results are presented in this section.

4.1. Experimental Setups

4.1.1. DataSets

In our experiments, we selected two public real world datasets. A brief description of these datasets is provided below and their statistics are summarized in **Table 1**.

TABLE 1 | Statistics of the 2 datasets.

Datasets	No. of channels	Sampled frequency (Hz)	No. of Subjects	No. of trials per class
DS1	118	100	5	140
DS2	59	100	4	100

DS1: Dataset IVa from BCI Competition III is a public dataset provided by the Berlin BCI group Fraunhofer FIRST (Intelligent Data Analysis Group) and Campus Benjamin Franklin of the Charité University (Neurophysics Group). This public dataset is recorded from five healthy subjects during right hand and right foot motor imageries. The EEG recordings consist of 118 channels at positions of the extended international 10/20-system. We chose a version of the data that was downsampled at 100 Hz for analysis. In the experiments, subjects performed three motor imageries for 3.5 s after visual cues for left hand, right hand, or right foot. After the duration of motor imagery, a resting interval with random length of 1.75–2.25 s was inserted for relaxation. The dataset provided only EEG trials for right hand and right foot imagery. For each subject, the dataset consisted of signals of 140 trials per class.

DS2: Dataset I from BCI Competition IV is a public dataset provided by the Berlin BCI group Fraunhofer FIRST (Intelligent Data Analysis Group) and Campus Benjamin Franklin of the Charité University (Neurophysics Group). This public dataset is recorded from four healthy subjects during two classes of motor imagery selected from three classes: left hand, right hand, and foot (side chosen by the subject; optionally also both feet). In the experiment, the data was continuous signals of 59 EEG channels and visual cues pointing left, right or down were presented for a period of 4.0 s during which the subject was instructed to perform the cued motor imagery task. These periods were interleaved with 2.0 s of blank screen and 2.0 s with a fixation cross displayed in the center of the screen. The dataset provided only EEG trials for left hand and foot imagery. For each subject, the dataset consisted of signals of 100 trials per class.

Since the band-pass filtering is involved in the proposed method, the pre-processing in the experiment is mainly two data normalization. The first one is removing the direct component in each trial, and the second one is normalizing all channels with mean zero by subtracting the mean values of each channel.

4.1.2. Baselines

To validate the effectiveness of the proposed COL, we compared it with the following feature selection and optimization methods.

1. **All channels**: Signals of all available channels are used for EEG classification;
2. **3C channels**: Signals of C3, Cz, and C4 are used for EEG classification;
3. **gsBLDA** (Yu et al., 2015): Signals of channels selected based on group Bayesian linear discriminant analysis are used for EEG classification;
4. **MRCS** (Zhang et al., 2016): Signals of channels selected by combining ReliefF and SVM are used for EEG classification;
5. **CSTI** (Yang et al., 2016): A subject-specific channel selection method based on a criterion, called F score, to realize the parameterization of both time segment and channel positions;
6. **NSGA-II** (Kee et al., 2015): Signals of channels selected by a multi-objective genetic algorithm, i.e., NSGA-II, are used for EEG classification.

TABLE 2 | OFB of channels with non-zero weights on subject ay of DS1.

Channel number	Channel location	f_{min}	f_{max}
60	CCP5	18.92	23.08
52	C3	8.54	34.35
70	CP3	10.42	28.16
87	P7	15.51	34.35
32	FT7	10.42	12.71
53	C1	10.42	34.35
9	AF8	8.54	12.71
85	PCP8	28.16	34.35
72	CPz	23.08	34.35
77	TP10	10.42	12.71

In order to ensure that the performance comparison mainly focused on the feature selection and optimization abilities of different algorithms, we used the same training and testing partitions for all methods when performing the cross-validation. We used autoregression analysis (AR) (Pfurtscheller et al., 1998) for EEG feature extraction after channel optimization by these comparison algorithms, except for the CSTI which contained the feature extraction procedure in its framework. Then linear discrimination analysis (LDA) was used for classification.

4.1.3. Parameter Settings

It is universally acknowledged that the performance of the majority of methods depends on their parameters. Therefore, we set the parameters used in our experiments in advance. As EEG is recorded continuously, it is necessary to choose a time interval to cut signals into a specific duration. In this work, we selected different durations of data from 3.5 s of continuous motor imagery for data processing and classification for DS1 and 4 s for DS2, namely 0–2, 0–2.5, 0.5–2.5, 0.5–3, 1–3, 1–3.5, 1.5–3.5, 1.5–4, and 2–4 s. Except for the results on different time intervals, we used signals in the time interval of 0.5-3 s for each trial on DS1, and that of 0–4 s for each trial on DS2. Moreover, α, N_f, f_0, and θ were set to 0.01, 9, 7, and 1.22, respectively. Further, we defined two convergence criteria in Algorithm 2; the iteration terminated if either of them was satisfied. The first one is $\mathcal{H}(\mathbf{w} - \eta\nabla_{\mathbf{w}}) > 0.9999\mathcal{H}(\mathbf{w})$, indicating the updating of $\mathbf{w}$ is almost stopped. The second was that the number of iteration met the maximum iterations, which was set to 1,000 in this work. All trials of the DS1 were used to perform cross validation. Only the calibration data of DS2 were used, since these data were provided with complete marker information. Except for the experiments on different sizes of training sets, eighty percent of the data were used for training data, the remaining part were used for testing

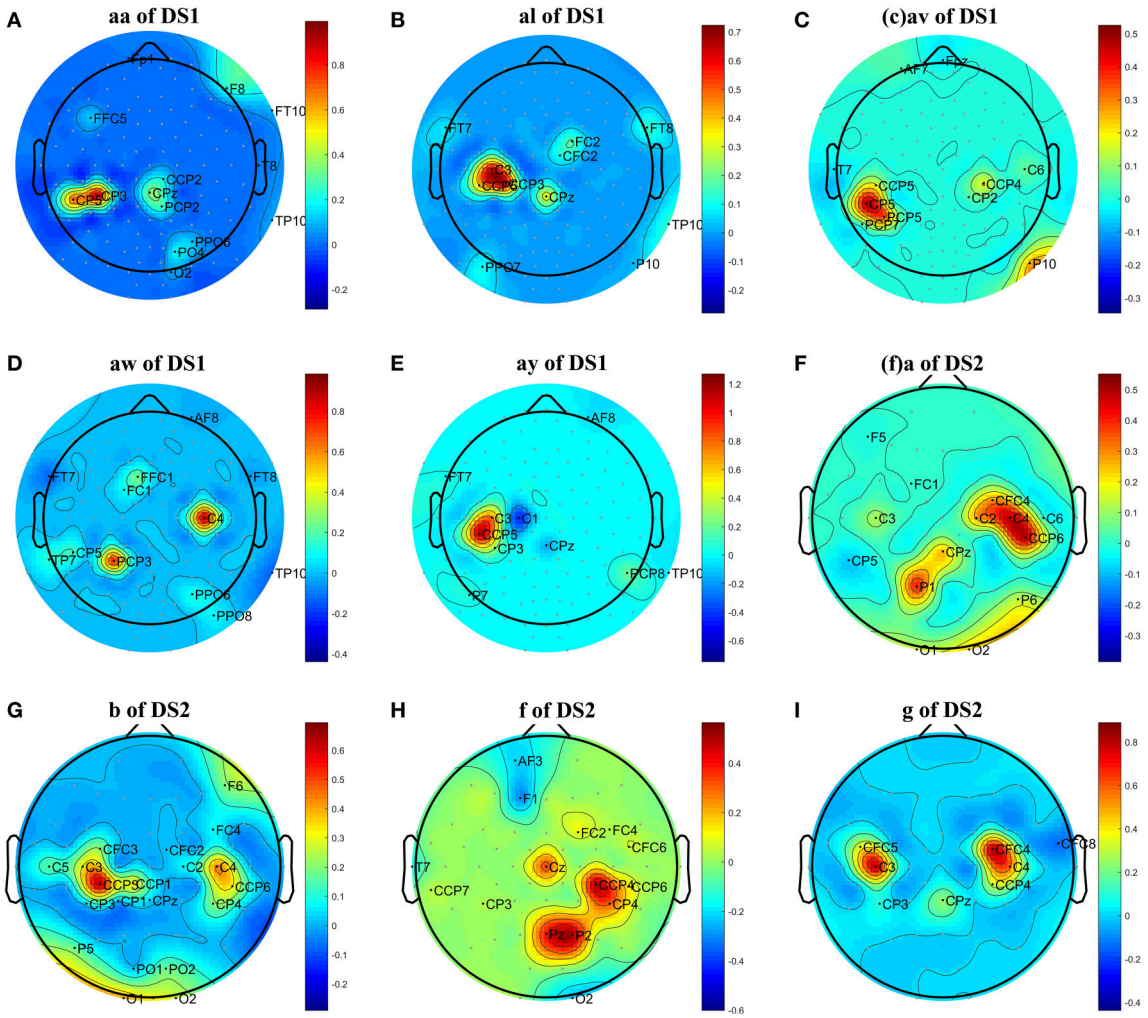

FIGURE 1 | Topographical map of optimal channels and weights on subjects: **(A)** aw of DS1, **(B)** av of DS1, **(C)** al of DS1, **(D)** aa of DS1, **(E)** ay of DS1, **(F)** a of DS2, **(G)** b of DS2, **(H)** f of DS2, **(I)** g of DS2.

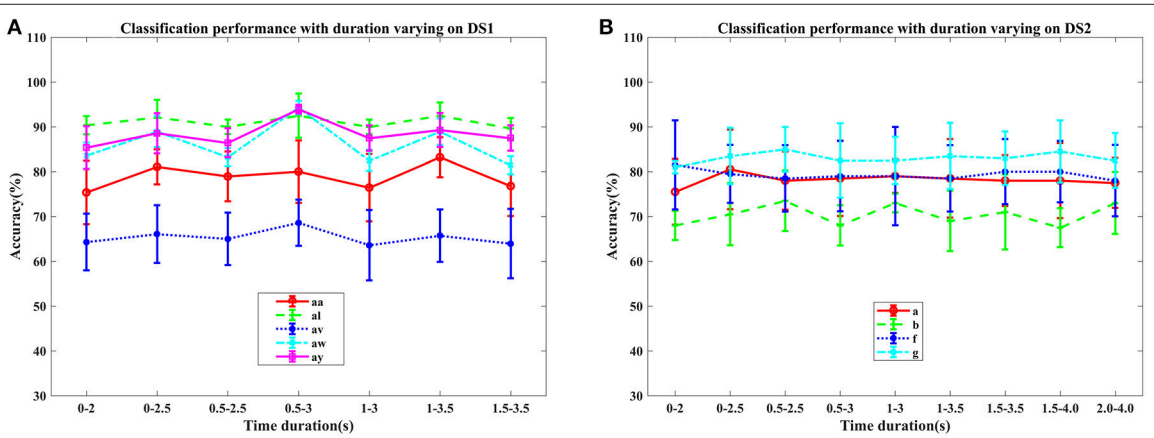

FIGURE 2 | Average classification accuracy (%) and standard deviation of COL with signal time duration varying on nine subjects: **(A)** aa, al, av, aw, and ay of DS1, and **(B)** a, b, f, and g of DS2. Each line type represents a subject's mean and std. of classification accuracy at different time durations.

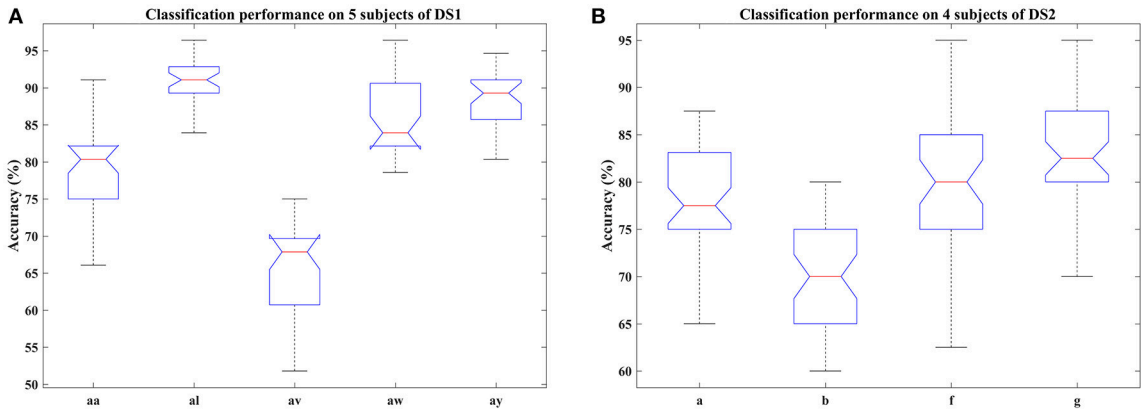

FIGURE 3 | Classification performance of COL on nine subjects: **(A)** aa, al, av, aw, and ay of DS1, **(B)** a, b, f, and g of DS2. Blue box indicates the range from 25 and 75% of the accuracy distribution, red line marks the median of the accuracy, notches in the box indicate the variability of the median between samples at the 5% significance level, extended whisker is the furthest observations in the range of 1.5 times the interquartile range away from the top or bottom of the box.

TABLE 3 | The mean classification accuracy and standard deviation (%) of 5 cross-validation with the number of corresponding channels (listed in the bracket) of COL and baselines on 5 subjects of DS1.

Method	Subject					Mean	Significant
	aa	al	av	aw	ay		
BS1	71.43 ± 4.37	83.57 ± 5.56	65.71 ± 8.41	76.07 ± 4.11	78.57 ± 3.34	75.07	****
BS2	69.29 ± 5.84	91.07 ± 2.19	58.93 ± 3.79	79.29 ± 2.99	85.36 ± 6.61	76.79	****
gsBLDA	70.36(76) ± 4.11	87.86(70) ± 3.19	68.21(101) ± 6.61	80.00(118) ± 4.62	77.50(97) ± 6.87	76.79	****
MRCS	75.00(94) ± 1.26	84.64(89) ± 5.45	66.07(115) ± 7.14	77.50(94) ± 2.71	79.64(86) ± 7.21	76.57	****
CSTI	74.64(8) ± 4.62	**94.64(20)** ± 1.26	**72.50(18)** ± 8.80	81.79(16) ± 4.07	92.86(12) ± 6.92	83.29	ns
NSGA-II	75.36(55) ± 6.11	85.36(55) ± 4.96	65.00(51) ± 3.70	76.43(71) ± 4.62	77.50(64) ± 8.43	75.93	****
L2-norm	73.57(72.4) ± 2.93	87.86(74) ± 3.43	61.79(77) ± 4.66	83.57(76.6) ± 4.45	81.79(75.2) ± 4.62	77.71	****
Proposed	**80.00(13.4)** ± 6.96	92.50(12.2) ± 4.96	68.57(12) ± 5.14	**94.29(11.2)** ± 1.49	**93.93(9.8)** ± 0.98	**85.86**	–

*The best performance is highlighted in bold. Performance of BS1 and BS2 indicates classification accuracy with all channels and 3C channels, respectively. Performance of L2-norm indicates classification accuracy with L2 norm replacing L1 norm as the regularization term. ****$p < 0.0001$, ordinary one-way ANOVA with Dunnett's multiple comparisons test.*

the data in each fold. Then, five-fold cross validation was repeated five times and the accuracies were averaged to obtain the mean result of the five-fold cross validation.

4.2. Results

To examine the effectiveness of the proposed COL, the classification on the two-class MI-EEG experimental results are given and analyzed in this section. We first present the optimal frequency bands with nonzero weights of the channels subject-specifically in **Table 2** and **Figure 1**. The performance of the proposed COL on different signal time duration and subjects is displayed in **Figures 2**, **3**. We discuss the sensitivity of the proposed COL on different sizes of training sets in **Figure 5**. Further comparison and analysis of these results with several channel selection algorithms in literature (Kee et al., 2015; Yu et al., 2015; Yang et al., 2016; Zhang et al., 2016) are provided in **Tables 3**, **5**, **9**, **10**.

4.2.1. Results on OFB and Weights of Channels

We first analyzed the OFB and weights obtained by the proposed COL in this subsection. The results were presented in **Table 2** and **Figure 1**. **Table 2** listed the selected channels with different weights on subject ay, as well as their locations on the scalp and optimal f_{min}, f_{max}. In **Figure 1**, the weights of all channels were

TABLE 4 | The mean F1 score and standard deviation (%) of 5 cross-validation of COL and baselines on 5 subjects of DS1.

Method	Subject					Mean
	aa	al	av	aw	ay	
BS1	70.33 ± 4.77	82.81 ± 5.81	65.70 ± 8.35	76.24 ± 4.30	78.12 ± 3.12	74.64
BS2	68.32 ± 5.69	90.71 ± 2.25	58.42 ± 3.17	77.36 ± 3.91	84.30 ± 7.87	75.82
gsBLDA	67.24 ± 7.30	87.07 ± 3.90	66.34 ± 8.72	79.23 ± 5.04	75.44 ± 8.22	75.06
MRCS	74.40 ± 1.59	83.74 ± 6.12	66.77 ± 6.33	77.59 ± 2.19	78.48 ± 8.19	76.20
CSTI	73.47 ± 4.35	**94.42** ± 1.31	**73.47** ± 9.43	81.91 ± 3.45	93.09 ± 6.55	83.27
NSGA-II	73.92 ± 6.37	83.99 ± 5.84	63.52 ± 5.47	77.01 ± 3.70	75.73 ± 10.28	74.83
L2-norm	71.86 ± 3.89	87.45 ± 4.14	59.63 ± 4.66	83.19 ± 4.17	81.29 ± 4.84	76.69
Proposed	**79.38** ± 7.63	92.52 ± 4.81	68.33 ± 5.15	**94.27** ± 1.57	**93.81** ± 1.01	**85.66**

The best performance is highlighted in bold.Performance of BS1 and BS2 indicates classification accuracy with all channels and 3C channels, respectively. Performance of L2-norm indicates classification accuracy with L2 norm replacing L1 norm as the regularization term.

reported in the topographical map, where pseudo-red regions were selected channels with large positive weights. Among the total number of 118 channels in DS1 and 59 channels in DS2, the COL selected a dozen of optimal channels for feature classification, leading to a simplified model.

Table 2 demonstrated the capability of the proposed COL to determine the OFB for each channel independently. It was beneficial for exploring and exploiting sensitive frequencies of different cerebral regions during motor imagery tasks. Moreover, an interesting observation from this table was that these OFBs were all approximately among μ and β rhythm, which was believed to be closely related to motor imagery.

From **Figure 1**, we have three main observations. First, the significant channels selected by COL exhibited a physiologically interpretable topography, where the regions near the mid-central vertex and left hemisphere were pivotal to discriminating the foot and right-hand imagery. Secondly, the weights obtained by COL were heavily concentrated on one or two regions, and signals from the majority of the channels were discarded, which contributed to reduce the computational costs and overfitting. Thirdly, significant regions for subject av of DS1 and subject b of DS2 were marginally farther from the vertex than other subjects, not totally focusing on the sensorimotor cortex, which could adversely influence the MI classification accuracy.

4.2.2. Performance on Different Signal Time Duration and Subjects

In this section, we examined the classification performance of COL on different signal time durations and subjects through five-fold cross-validation. The accuracies with respect to the selection of time interval and subjects were displayed in **Figures 2**, **3**. **Figure 2** plotted the mean accuracies and standard deviation with different signal time durations. In **Figure 3**, the distribution of the classification accuracies on the five subjects were displayed in box figures, where the tops and bottoms of the blue boxes were the 25th and 75th percentiles of the samples of the accuracy distribution, the red line marked the median of the accuracy, notches in the boxes indicated the variability of the median between samples at the 5% significance level, the extended whisker was the furthest observations in the range of 1.5 times the interquartile range away from the top or bottom of the box.

From these two figures, we can make the following observations. For DS1, higher accuracies were frequently achieved by 2.5 s intervals, instead of 2 s intervals, indicating that more data was beneficial for improving the performance. Specifically, the interval of 0.5–3 s provided the best classification accuracies, approximately. For DS2, the proposed COL got a more stable classification accuracy across different time durations. Accuracy distributions differed among all nine subjects and the proposed COL could determine a relatively stable classification accuracy for the majority of the subjects except for av of DS1 and b of DS2.

We also compared the proposed COL with several classical and state-of-the-art methods; the results were presented in **Tables 3**, **5**. Each table summarized the mean accuracies of the different methods. The standard deviation and the number of selected channels were also reported. And the F1 score of these methods were also presented for further evaluating the classification accuracy in **Tables 4**, **6**. The best values were highlighted in bold.

Tables 3, **4** indicated that COL achieved the best performance compared with two baselines and four state-of-the-art methods. For the two excellent subjects aw and ay, the proposed COL maintained a relatively stable classification accuracy, indicated by the small standard deviations. Notably, the number of selected channels of COL outperformed the other methods in the vast majority of cases. One-way ANOVA with Dunnett's multiple comparisons test showed that the improvement of COL was significant ($p < 0.0001$) except for CSTI. With regard to the other methods, the proposed COL not only achieved superior accuracy but also preserved fewer channels, demonstrating the superior performance of the proposed method over the state-of-the-art methods. The proposed COL was not the best discriminating right hand and foot imagery on subject al, and av, but the best at classifying the signals of aa, aw, and ay.

Tables 5, **6** illustrated that COL was more than 10% superior to the comparison algorithms on performance. Specifically, COL could discriminate motor imagery of subject b and g with an improvement of 11 and 15% in accuracy over the best baselines, respectively. One-way ANOVA with Dunnett's multiple comparisons test showed that the improvement of COL was significant ($p < 0.0001$). The results apparently verified the positive effect of the proposed channel-optimization strategy on MI-EEG classification. The average number of selected channels of COL maintained the same level in DS1, contributing to a simplified model with optimized parameters.

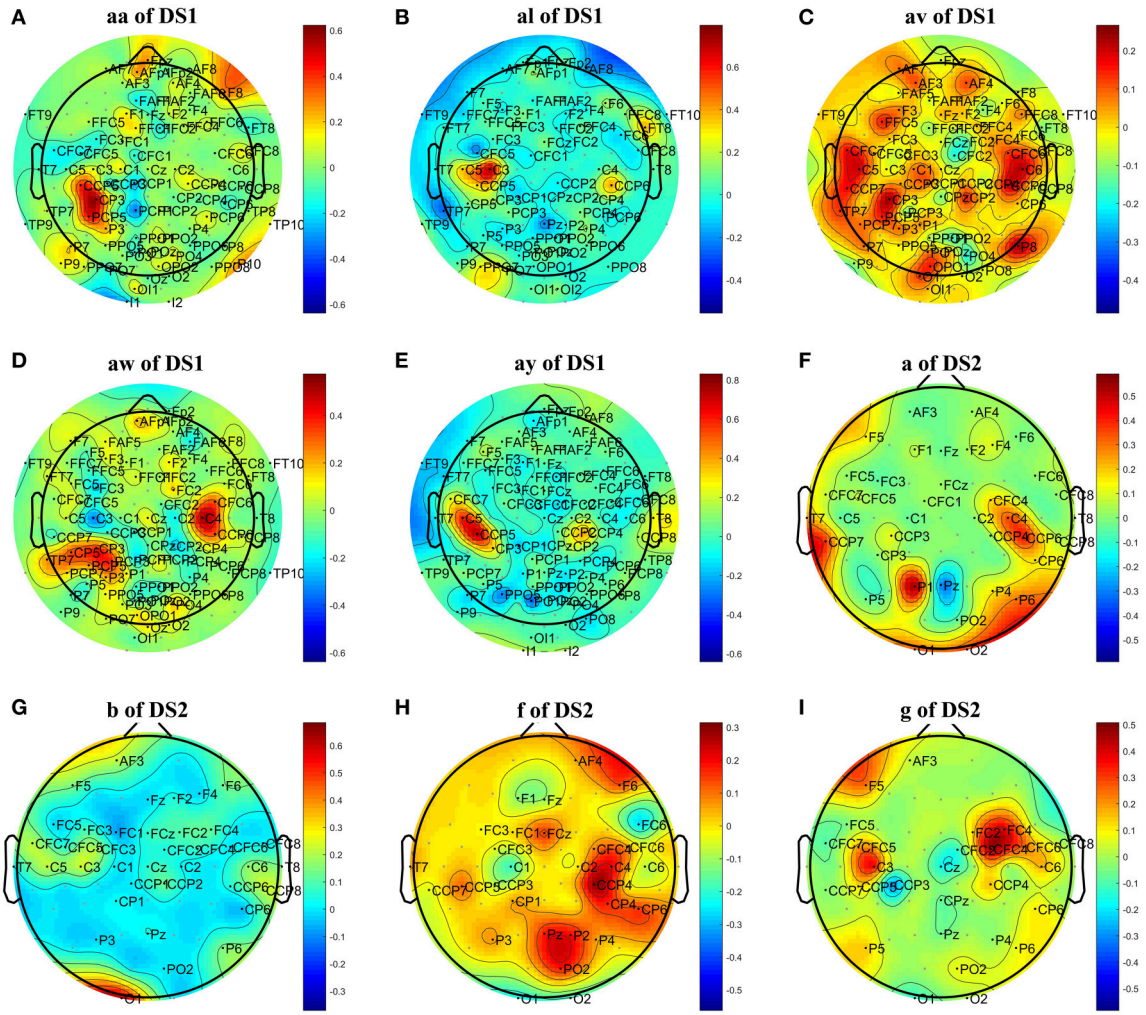

FIGURE 4 | Topographical map of optimal channels and weights with L2 norm regularization term on subjects: **(A)** aw of DS1, **(B)** av of DS1, **(C)** al of DS1, **(D)** aa of DS1, **(E)** ay of DS1, **(F)** a of DS2, **(G)** b of DS2, **(H)** f of DS2, **(I)** g of DS2.

To further investigate the effect of L1 norm on the simplification and classification accuracy of COL, we replaced the L1 norm by a similar L2 norm regularization term and compared its influence on COL's performance. The L1 norm and L2 norm are two different sparse strategies. In comparison, the L1 norm can achieve sparser optimization, and the optimization process is more robust and less susceptible to interference from signal changes, noise, and other factors (Li et al., 2013; Peterson et al., 2015). **Tables 3**, **5** showed that the optimization based on L2 norm did not result in a simplified classification model. The number of optimal channels was 3–6 times higher than that based on L1 norm. The topographical maps of optimal channels and weights with L2 norm regularization term were also plotted in **Figure 4**, showing that many redundant channels were selected by the L2 norm regularization. As a result, the classification accuracy obtained by L2 norm was lower than the L1 norm. These results proved that L1 norm played an important role in the optimization of the COL model.

Considering EEG reference may have a fundamental impact on the result, we performed EEG zero-reference by means of the reference electrode standardization technique (REST, Yao, 2001, 2017), which can be found at www.neuro.uestc.edu.cn/rest (Dong et al., 2017). Comparison of the classification results before and after REST processing were provided in **Tables 7**, **8**. We found that the classification performance didn't change significantly after REST processing, except for subject ay of DS1 and subject g of DS2, who received a decrease of approximate to 10% in classification accuracy. The number of optimal channels maintained the same level as before. We infer that the L1 norm-based sparse channel optimization and combination procedure may be insensitive to the detailed characteristics of the signal. As a result, the EEG signal obtained by the simple CAR preprocessing or the more accurate zero-reference preprocessing does not cause apparent difference in the classification performance.

4.2.3. Results on Different Sizes of Training Sets

To further verify the generalization of the proposed COL, we plotted the mean classification accuracy with the ratio of the number of training samples to all samples varying from 0.2 to 0.8 with step 0.2 in **Figure 5**. It appears that the performance of the proposed COL was relatively stable over a large range of the ratio on DS1. Even using only 20% of the samples for training, the accuracy was >90% for subject aw. Furthermore, this result implied that the proposed approach could leverage the supervision information of small numbers of training sets to predict labels of a considerably larger quantity of testing samples. This favorable generalization based on the small sample training complemented the rapid implementation and application practically of the proposed COL. For subject b and g of DS2, an obvious improvement of approximately 20% with training samples increasing from 20 to 80 could be observed. It clearly indicated that COL was capable of leveraging supervised labels to acquire more appropriate undetermined parameters.

The generalization of COL was compared with other mentioned methods above by evaluating the classification performance under small training sets (20% samples for training), as shown in **Tables 9**, **10**. Compared with **Table 3**, results in **Table 9** illustrate that COL was relatively stable with small training sets on DS1, while the other methods indicated a decrease approximate to 9% when training sets became small. The generalization of COL on DS2 did not outperform the others, possibly due to the big drop of classification accuracy of subject b and g with small training sets.

TABLE 5 | The mean classification accuracy and standard deviation (%) of 5 cross-validation with the number of corresponding channels (listed in the bracket) of COL and baselines on 4 subjects of DS2.

Method	Subject				Mean	Significant
	a	b	f	g		
BS1	74.50 ± 3.26	67.00 ± 4.81	62.00 ± 4.81	72.00 ± 5.97	68.88	****
BS2	69.00 ± 6.02	61.00 ± 3.79	65.00 ± 5.59	77.50 ± 3.95	68.13	****
gsBLDA	74.00 (59) ± 6.52	68.50 (46) ± 1.37	68.00 (3) ± 8.91	74.00 (7) ± 8.02	71.13	****
MRCS	74.50 (59) ± 3.26	67.00 (59) ± 4.81	63.00 (58) ± 4.11	76.00 (36) ± 2.85	70.38	****
CSTI	82.50 (21) ± 8.10	69.00 (9) ± 7.20	76.00 (16) ± 5.18	73.50 (19) ± 4.54	75.25	****
NSGA-II	73.50 (38) ± 5.18	65.00 (34) ± 10.75	67.50 (36) ± 1.77	70.50 (8) ± 8.55	69.13	****
L2-norm	84.00 (35.2) ± 8.77	66.50 (33.2) ± 7.42	71.50 (30.8) ± 4.54	74.00 (28.6) ± 14.85	74.00	****
Proposed	**87.50 (13.2)** ± 9.19	**80.00 (19.8)** ± 7.50	**84.50 (12)** ± 5.70	**92.50 (10.4)** ± 3.54	**86.13**	–

*The best performance is highlighted in bold. Performance of BS1 and BS2 indicates classification accuracy with all channels and 3C channels, respectively. Performance of L2-norm indicates classification accuracy with L2 norm replacing L1 norm as the regularization term. ****$p < 0.0001$, ordinary one-way ANOVA with Dunnett's multiple comparisons test.*

4.3. Discussion

Simplified and well-generalized classification models are essential for the practical application of MI based BCI. Many efforts have been done in the feature selection and optimization of MI-EEG based on CSP, DSP, and SVM et al. However, due to the insufficient EEG data for model training in practical applications, the more training parameters a classifier requires, the more likely it tends to be overfitting, which reduces its practical value. In this study, we establish a L1 norm based channel optimization algorithm for MI-EEG

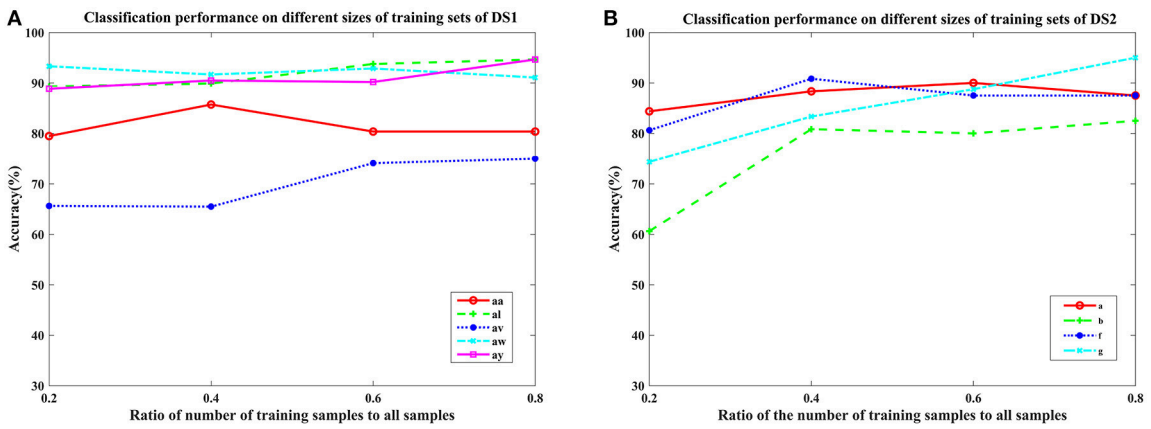

FIGURE 5 | Average classification accuracy of COL with the ratio of the number of training samples to all samples varying from 0.2 to 0.8 with step 0.2 on nine subjects: **(A)** aa, al, av, aw, and ay of DS1, and **(B)** a, b, f, and g of DS2. Each line type represents a subject's classification accuracy at different training samples.

TABLE 6 | The mean F1 score and standard deviation (%) of 5 cross-validation of COL and baselines on 4 subjects of DS2.

Method	Subject				Mean
	a	b	f	g	
BS1	74.09 ± 2.60	67.95 ± 4.34	60.77 ± 8.73	71.51 ± 7.26	68.58
BS2	66.70 ± 5.44	62.08 ± 3.37	62.65 ± 8.19	76.57 ± 4.63	67.00
gsBLDA	73.52 ± 6.30	68.63 ± 1.34	68.41 ± 9.49	73.58 ± 8.33	71.04
MRCS	74.09 ± 2.60	67.95 ± 4.34	61.95 ± 7.91	75.56 ± 4.12	69.89
CSTI	82.05 ± 8.82	68.46 ± 6.98	75.57 ± 6.25	74.47 ± 3.65	75.14
NSGA-II	72.66 ± 5.66	64.50 ± 10.76	66.02 ± 5.66	67.97 ± 12.10	67.79
L2-norm	83.87 ± 8.74	62.60 ± 11.46	71.36 ± 3.73	73.28 ± 14.81	72.78
Proposed	**87.00** ± 10.12	**79.20** ± 7.71	**84.49** ± 5.66	**92.38** ± 3.61	**85.77**

The best performance is highlighted in bold. Performance of BS1 and BS2 indicates classification accuracy with all channels and 3C channels, respectively. Performance of L2-norm indicates classification accuracy with L2 norm replacing L1 norm as the regularization term.

TABLE 7 | The mean classification accuracy and standard deviation (%) of 5 cross-validation with the number of corresponding channels (listed in the bracket) of COL and baselines on 5 subjects of DS1 after using REST for zero-reference.

Method	Subject					Mean
	aa	al	av	aw	ay	
BS1	77.50 ± 3.48	85.00 ± 5.14	60.71 ± 4.19	81.43 ± 4.30	78.21 ± 3.43	76.57
BS2	64.64 ± 9.14	88.93 ± 5.11	66.79 ± 4.30	77.50 ± 6.63	82.14 ± 4.37	76.00
gsBLDA	77.50(118) ± 3.48	87.50(97) ± 4.55	63.57(85) ± 4.82	83.21(95) ± 3.70	80.00(105) ± 2.33	78.36
MRCS	77.50(118) ± 3.48	85.71(112) ± 3.99	63.93(108) ± 5.84	82.50(100) ± 6.24	78.21(118) ± 3.43	77.57
CSTI	73.21(8) ± 4.55	88.57(31) ± 4.48	**68.93(12)** ± 7.64	75.00(19) ± 3.79	59.64(93) ± 7.32	73.07
NSGA-II	76.43(55) ± 4.62	84.29(59) ± 4.07	64.29(35) ± 6.31	85.00(68) ± 4.30	78.93(64) ± 5.70	77.79
Proposed	**77.86(15.8)** ±5.30	**91.43(14.6)** ±2.65	68.57(11.4) ±5.30	**86.43(15)** ±2.71	**83.57(7.6)** ±4.26	**81.57**

The best performance is highlighted in bold. Performance of BS1 and BS2 indicates classification accuracy with all channels and 3C channels, respectively.

TABLE 8 | The mean classification accuracy and standard deviation (%) of 5 cross-validation with the number of corresponding channels (listed in the bracket) of COL and baselines on 4 subjects of DS2 after using REST for zero-reference.

Method	Subject				Mean
	a	b	f	g	
BS1	70.50 ± 8.55	67.00 ± 8.18	68.50 ± 5.48	63.50 ± 6.75	67.38
BS2	64.00 ± 3.35	54.50 ± 6.71	62.00 ± 10.22	75.00 ± 8.84	63.88
gsBLDA	71.50(55) ± 10.40	68.00(58) ± 7.79	68.50(59) ± 5.48	66.00(57) ± 8.59	68.50
MRCS	72.00(55) ± 6.47	67.50(58) ± 7.29	70.50(43) ± 4.11	72.00(3) ± 3.71	70.50
CSTI	78.00(12) ± 5.70	66.50(32) ± 7.42	75.00(14) ± 7.07	75.00(10) ± 4.68	73.63
NSGA-II	72.50(27) ± 10.0	68.00(33) ± 4.47	68.00(27) ± 6.22	73.00(9) ± 6.47	70.38
Proposed	**86.00(13.4)** ±3.79	**77.00(17.8)** ±6.94	**87.00(12.0)** ±6.47	**82.50(11.0)** ±7.91	**83.13**

The best performance is highlighted in bold. Performance of BS1 and BS2 indicates classification accuracy with all channels and 3C channels, respectively.

TABLE 9 | The mean classification accuracy and standard deviation (%) of 5 cross-validation (20% samples for training) with the number of corresponding channels (listed in the bracket) of COL and baselines on 5 subjects of DS1.

Method	Subject					Mean
	aa	al	av	aw	ay	
BS1	66.43 ± 3.90	74.73 ± 1.21	55.80 ± 2.42	70.63 ± 2.63	66.43 ± 2.31	66.80
BS2	61.96 ± 3.75	75.18 ± 7.10	54.82 ± 5.45	70.89 ± 3.32	74.29 ± 3.48	67.43
gsBLDA	64.64 (106) ± 5.26	80.18 (81) ± 2.76	56.70 (96) ± 5.28	73.75 (95) ± 4.65	68.84 (86) ± 5.01	68.82
MRCS	66.87 (115) ± 5.61	76.16 (71) ± 1.08	57.68 (49) ± 2.15	73.12 (75) ± 3.21	69.46 (33) ± 1.94	68.66
CSTI	64.73 (1) ± 13.52	89.64 (7) ± 5.92	59.82 (89) ± 3.09	68.30 (104) ± 2.53	85.80 (9) ± 5.50	73.66
NSGA-II	69.37 (47) ± 4.20	77.86 (39) ± 2.29	58.39 (45) ± 2.31	72.50 (60) ± 3.44	71.25 (34) ± 1.47	69.87
Proposed	**80.43 (19)** ± 3.34	**90.88 (14.8)** ± 3.81	**68.43 (17.2)** ± 5.29	**94.13 (15.6)** ± 2.15	**93.18 (14.2)** ± 2.00	**84.81**

The best performance is highlighted in bold. Performance of BS1 and BS2 indicates classification accuracy with all channels and 3C channels, respectively.

classification. Compared with commonly used CSP methods, the parameters required for COL training are greatly reduced, contributing to a simplified and generalized classification model.

It is shown that the optimal channel distribution of the COL exhibit a physiologically interpretable topography, and the optimal frequency bands are mostly distributed

TABLE 10 | The mean classification accuracy and standard deviation (%) of 5 cross-validation (20% samples for training) with the number of corresponding channels (listed in the bracket) of COL and baselines on 4 subjects of DS2.

Method	Subject				Mean
	a	b	f	g	
BS1	63.38 ± 2.75	55.37 ± 2.36	59.00 ± 4.71	62.00 ± 1.35	59.94
BS2	57.00 ± 5.44	52.75 ± 3.47	57.50 ± 5.43	63.00 ± 4.89	57.56
gsBLDA	61.12(43) ± 3.96	57.37(36) ± 6.00	57.87(48) ± 2.88	63.88(54) ± 1.68	60.06
MRCS	63.38(59) ± 2.75	55.75(31) ± 3.84	60.25(24) ± 2.24	63.25(50) ± 2.94	60.66
CSTI	67.62(55) ± 3.99	58.25(48) ± 5.56	62.38(1) ± 3.01	71.62(58) ± 3.82	64.97
NSGA-II	64.87(24) ± 2.23	57.62(23) ± 3.96	60.25(20) ± 3.58	63.50(35) ± 2.82	61.56
Proposed	**81.75(14.2)** ± 4.04	**64.25(20.4)** ± 11.49	**78.88(15)** ± 8.79	**81.5(15.4)** ± 6.50	**76.59**

The best performance is highlighted in bold. Performance of BS1 and BS2 indicates classification accuracy with all channels and 3C channels, respectively.

around the μ and β rhythm, which is believed to be closely related to motor imagery. In addition, results on BCI competition datasets show that the COL maintains relatively high classification accuracy and F1 score with sparse features, indicating its good potential in practical applications. Further tests under small ratio of training samples show that the COL has good generalization performance, especially on DS1. As the training set ratio decreased from 80 to 20%, the average classification accuracy on DS1 changed from 85.86 to 84.81%, maintaining relatively high classification accuracy. The generalization of the COL algorithm benefits from the simplified model design and efficient extraction of motor-related features.

We further compared the classification performance under different reference conditions. The REST, an accurate zero-reference method, is applied to the multichannel EEG recordings. However, the classification performance is not significantly improved after REST processing compared to the simple CAR reference. We infer that the L1 norm-based sparse channel optimization and combination procedure may be insensitive to the detailed characteristics of the signal. As a result, the EEG signal obtained by the simple CAR preprocessing or the more accurate zero-reference preprocessing does not cause apparent difference in the classification performance.

There are still some limitations in this study. Firstly, the current study lacks theoretical and experimental proof of the convergence of COL. Secondly, this study only utilizes binary classification to evaluate the classification and generalization performance of the COL algorithm. In the future work, we will prove the convergence of COL theoretically and experimentally. More attention will be paid to flexible EEG processing and classification methods for improved channel-specific prediction accuracy. Further, novel embedding and interaction metrics for signals from multi-channels are also of great interest. And multi-classification problem will also be considered for improving the effectiveness of the COL algorithm in practical applications.

5. CONCLUSION

We introduced a novel method to optimize the features extracted from multichannel EEG by integrating inter-channel and intra-channel factors on motor imagery signal processing in this paper. Specifically, an l_1-norm-based sparse regularized linear least square regression was introduced to learn a compact and accurate representation of MI-EEG. By maximizing the F-score of the EEG classification channel-specifically, COL discretely determines the optimal frequency bands for each channel independently. Simultaneously, by virtue of the sparse regularization term on channel weights, redundant and uninformative channels are discarded, while significant and task-relevant channels preserved. Subsequently, we designed an iterative algorithm to efficiently solve the constrained optimization problem and analyze its computational complexity. Experimental results on real world EEG datasets not only validated the effectiveness and efficiency of the proposed method compared with state-of-the-art methods but also provided convincing evidence of its feasible application in practical BCI systems.

AUTHOR CONTRIBUTIONS

YZ and JH processed and analyzed the data, and wrote the manuscript; YC developed the l1-norm regularized method; HS, JC, and AK helped to acquire and interpret the data; YH and PZ helped data analysis; YZ helped in data interpretation and manuscript edit; JZ and CW supervised development of work, helped in manuscript edit and evaluation.

ACKNOWLEDGMENTS

The authors thank the reviewers for their valuable comments and thank the editors for their fruitful work. This study is financially supported by International Cooperation and Exchange of the National Natural Science Foundation of China (No. 31320103914), General Program of National Natural Science Foundation of China (No. 31370987), National Natural Science Funds for Outstanding Young Scholar (No. 81622027), Beijing NOVA Program of China (No. 2016B615), and National Key Research and Development Program of China (No. 2017YFA0106100).

REFERENCES

Arvaneh, M., Guan, C., Ang, K. K., and Quek, C. (2011). Optimizing the channel selection and classification accuracy in EEG-based BCI. *IEEE Trans. Biomed. Eng.* 58, 1865–1873. doi: 10.1109/TBME.2011.213114

Baillet, S., Mosher, J. C., and Leahy, R. M. (2001). Electromagnetic brain mapping. *Signal Process. Mag. IEEE* 18, 14–30. doi: 10.1109/79.962275

Blankertz, B., Losch, F., Krauledat, M., Dornhege, G., Curio, G., and Müller, K. R. (2008). The berlin brain-computer interface: accurate performance from first-session in BCI-naïve subjects. *IEEE Trans. Biomed. Eng.* 55, 2452–2462. doi: 10.1109/TBME.2008.923152

Blankertz, B., Tomioka, R., Lemm, S., and Kawanabe, M. (2007). Optimizing spatial filters for robust EEG single-trial analysis. *Signal Process. Mag. IEEE* 25, 41–56. doi: 10.1109/MSP.2008.4408441

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers," in *The Workshop on Computational Learning Theory* (Pittsburgh: PA), 144–152.

Chatterjee, R., and Bandyopadhyay, T. (2016). "EEG based motor imagery classification using SVM and MLP," in *International Conference on Computational Intelligence and Networks*, 84–89.

Dong, L., Li, F., Liu, Q., Wen, X., Lai, Y., Xu, P., et al. (2017). Matlab toolboxes for reference electrode standardization technique (rest) of scalp EEG. *Front. Neurosci.* 11:601. doi: 10.3389/fnins.2017.00601

Donoho, D. L. (2006). For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Commun. Pure Appl. Math.* 59, 797–829. doi: 10.1002/cpa.20131

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. New York, NY: John Wiley & Sons, 1–654.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Patten Classification, 2nd Edn)*. New York, NY: John Wiley & Sons.

Farquhar, J., Hill, N. J., Lal, T. N., and Schölkopf, B. (2006). "Regularised CSP for sensor selection in BCI," in *Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course* (Graz).

Graimann, B., Allison, B., Mandel, C., Lüth, T., Valbuena, D., and Graeser, A. (2008). "Non-invasive brain-computer interfaces for semi-autonomous assistive devices," in *Robust Intelligent Systems*, ed A. Schuster (London: Springer), 113–138.

Hoffmann, U., Vesin, J. M., and Ebrahimi, T. (2006). "Spatial filters for the classification of event-related potentials," in *Esann 2006, European Symposium on Artificial Neural Networks* (Bruges: EPFL scientific publications), 47–52.

Kai, K. A., Zheng, Y. C., Wang, C., Guan, C., and Zhang, H. (2012). Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Front. Neurosci.* 6:39. doi: 10.3389/fnins.2012.00039

Kee, C.-Y., Ponnambalam, S., and Loo, C.-K. (2015). Multi-objective genetic algorithm as channel selection method for p300 and motor imagery dataset. *Neurocomputing* 161, 120–131. doi: 10.1016/j.neucom.2015.02.057

Lafleur, K., Cassady, K., Doud, A., Shades, K., Rogin, E., and He, B. (2013). Quadcopter control in three-dimensional space using a noninvasive motor imagery based brain-computer interface. *J. Neural Eng.* 10, 1–29. doi: 10.1088/1741-2560/10/4/046003

Lee, C., Jung, J., Kwon, G., and Kim, L. (2012). "Individual optimization of EEG channel and frequency ranges by means of genetic algorithm," in *International Conference of the IEEE Engineering in Medicine and Biology Society* (San Diego, CA), 5290–5293.

Lemm, S., Blankertz, B., Curio, G., and Müller, K. R. (2005). Spatio-spectral filters for improving the classification of single trial EEG. *IEEE Trans. Biomed. Eng.* 52:1541. doi: 10.1109/TBME.2005.851521

Li, P., Xu, P., Zhang, R., Guo, L., and Yao, D. (2013). L1 norm based common spatial patterns decomposition for scalp EEG BCI. *Biomed. Eng. Online* 12, 1–12. doi: 10.1186/1475-925X-12-77

Liao, X., Yao, D., Wu, D., and Li, C. (2007). Combining spatial filters for the classification of single-trial EEG in a finger movement task. *IEEE Trans. Biomed. Eng.* 54, 821–831. doi: 10.1109/TBME.2006.889206

Ma, Y., Ding, X., She, Q., Luo, Z., Potter, T., and Zhang, Y. (2016). Classification of motor imagery EEG signals with support vector machines and particle swarm optimization. *Comput. Math. Methods Med.* 2016, 1–8. doi: 10.1155/2016/4941235

Meng, J., Zhang, S., Bekyo, A., Olsoe, J., Baxter, B., and He, B. (2016). Noninvasive electroencephalogram based control of a robotic arm for reach and grasp tasks. *Sci. Rep.* 6:38565. doi: 10.1038/srep38565

Meng, L., Jin, J., and Wang, X. (2011). "A comparison of three electrode channels selection methods applied to SSVEP BCI," in *International Conference on Biomedical Engineering and Informatics* (Shanghai), 584–587.

Müller, K. R., Krauledat, M., Dornhege, G., Curio, G., and Blankertz, B. (2004). Machine learning techniques for brain-computer interfaces. *Biomed. Eng. Biomed. Tech.* 49, 11–22. doi: 10.13109/9783666351419.11

Müllergerking, J., Pfurtscheller, G., and Flyvbjerg, H. (1999). Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clin. Neurophysiol.* 110, 787–798.

Nicolas-Alonso, L. F., and Gomez-Gil, J. (2012). Brain computer interfaces, a review. *Sensors* 12, 1211–1279. doi: 10.3390/s120201211

Offner, F. F. (1950). The EEG as potential mapping: the value of the average monopolar reference. *Electroencephalogr. Clin. Neurophysiol.* 2, 213–214.

Peterson, V., Rufiner, H. L., and Spies, R. (2015). "L1-norm regularization for sparse representation and p300 wave detection in brain-computer interfaces," in *V Congreso de Matemática Aplicada, Computacional e Industrial* (Tandil).

Pfurtscheller, G., Brunner, C., Schlögl, A., and Silva, F. H. (2006). Mu rhythm (de)synchronization and eeg single-trial classification of different motor imagery tasks. *Neuroimage* 31, 153–159. doi: 10.1016/j.neuroimage.2005.12.003

Pfurtscheller, G., Neuper, C., Schlögl, A., and Lugger, K. (1998). Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters. *IEEE Trans. Rehabil. Eng.* 6, 316–325.

Pudil, P., Novovičová, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recogn. Lett.* 15, 1119–1125.

Qiu, Z., Jin, J., Lam, H.-K., Yu, Z., Xingyu, W., and Andrzej, C. (2016). Improved SFFS method for channel selection in motor imagery based bci. *Neurocomputing* 207, 519–527. doi: 10.1016/j.neucom.2016.05.035

Ramoser, H., Müller-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

Schrouff, J., Mourão Miranda, J., Phillips, C., and Parvizi, J. (2016). Decoding intracranial EEG data with multiple kernel learning method. *J. Neurosci. Methods* 261, 19–28. doi: 10.1016/j.jneumeth.2015.11.028

Shri, T. K. P., and Sriraam, N. (2016). Spectral entropy feature subset selection using sepcor to detect alcoholic impact on gamma sub band visual event related potentials of multichannel electroencephalograms (EEG). *Appl. Soft Comput.* 46, 441–451. doi: 10.1016/j.asoc.2016.04.041

Silva, C., Maltez, J., Trindade, E., Arriaga, A., and Duclasoares, E. (2004). Evaluation of l1 and l2 minimum norm performances on EEG localizations. *Clin. Neurophysiol.* 115, 1657–1668. doi: 10.1016/j.clinph.2004.02.009

Student, T. K. P. S. R., and Sriraam, N. (2017). Comparison of t-test ranking with pca and sepcor feature selection for wake and stage 1 sleep pattern recognition in multichannel electroencephalograms. *Biomed. Signal Process. Control* 31, 499–512. doi: 10.1016/j.bspc.2016.09.016

Suk, H. I., and Lee, S. W. (2011). "Data-driven frequency bands selection in EEG-based brain-computer interface," in *International Workshop on Pattern Recognition in Neuroimaging* (Seoul), 25–28.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* 73, 273–282. doi: 10.1111/j.1467-9868.2011.00771.x

Tomida, N., Tanaka, T., Ono, S., Yamagishi, M., and Higashi, H. (2015). Active data selection for motor imagery EEG classification. *IEEE Trans. Biomed. Eng.* 62, 458–467. doi: 10.1109/TBME.2014.2358536

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

Wu, W., Gao, X., Hong, B., and Gao, S. (2008). Classifying single-trial EEG during motor imagery by iterative spatio-spectral patterns learning (ISSPL). *IEEE Trans. Biomed. Eng.* 55, 1733–1743. doi: 10.1109/TBME.2008.919125

Wu, Y., and Ge, Y. (2013). A novel method for motor imagery EEG adaptive classification based biomimetic pattern recognition. *Neurocomputing* 116, 280–290. doi: 10.1016/j.neucom.2012.03.030

Xu, H., Song, W., Hu, Z., Chen, C., Zhao, X., and Zhang, J. (2010). "A speedup SVM decision method for online EEG processing in motor imagery BCI," in *International Conference on Intelligent Systems Design and Applications* (Cairo). 149–153.

Yang, Y., Bloch, I., Chevallier, S., and Wiart, J. (2016). Subject-specific channel selection using time information for motor imagery brain-computer interfaces. *Cogn. Comput.* 8:505. doi: 10.1007/s12559-015-9379-z

Yao, D. (2001). A method to standardize a reference of scalp EEG recordings to a point at infinity. *Physiol. Meas.* 22, 693–711. doi: 10.1088/0967-3334/22/4/305

Yao, D. (2017). Is the surface potential integral of a dipole in a volume conductor always zero? A cloud over the average reference of EEG and ERP. *Brain Topogr.* 30, 1–11. doi: 10.1007/s10548-016-0543-x

Yu, K., Wang, Y., Shen, K., and Li, X. (2013). The synergy between complex channel-specific fir filter and spatial filter for single-trial EEG classification. *PLoS ONE* 8:e76923. doi: 10.1371/journal.pone.0076923

Yu, T., Yu, Z., Gu, Z., and Li, Y. (2015). Grouped automatic relevance determination and its application in channel selection for p300 bcis. *IEEE Trans. Neural Syst. Rehabil. Eng.* 23, 1068–1077. doi: 10.1109/TNSRE.2015.2413943

Zhang, J., Chen, M., Zhao, S., Hu, S., Shi, Z., and Cao, Y. (2016). Relieff-based EEG sensor selection methods for emotion recognition. *Sensors* 16, 1–15. doi: 10.3390/s16101558

Zheng, W. L., Guo, H. T., and Lu, B. L. (2015). "Revealing critical channels and frequency bands for emotion recognition from EEG with deep belief network," in *International IEEE/EMBS Conference on Neural Engineering* (Montpellier), 154–157.

Conflict of Interest Statement: 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.

APPENDIX

For any vector $\mathbf{x} \in \mathbb{R}^{n\times 1}$, the l_p-norm is defined as

$$||\mathbf{x}||_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}, \tag{A1}$$

where x_i is the i-th element of $\mathbf{x}$. For any matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$, the m_r-norm is defined as

$$||\mathbf{X}||_r = \left(\sum_{i=1}^{n}\sum_{j=1}^{m} |\mathbf{X}_{ij}|^r\right)^{\frac{1}{r}}. \tag{A2}$$