https://doi.org/10.1016/j.cmpb.2020.105466

Title: Deep Stacked Support Matrix Machine Based Representation Learning for Motor Imagery EEG Classification

Authors: Wenlong Hang[1,2], Wei Feng[1], Shuang Liang[3*], Qiong Wang[2], Xuejun Liu[1], Kup-Sze Choi[4]

[1]*School of Computer Science and Technology, Nanjing Tech University, Nanjing 211816, China*

[2]*CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Shenzhen 518055, China*

[3]*Smart Health Big Data Analysis and Location Services Engineering Lab of Jiangsu Province, Nanjing University of Posts and Telecommunications, Nanjing 210093, China*

[4]*School of Nursing, Hong Kong Polytechnic University, Hung Hom, Hong Kong*

*Corresponding Author:

Shuang Liang, Ph.D.

Nanjing University of Posts and Telecommunications

Email: shuang.liang@njupt.edu.cn

Abstract

Background and Objective: Electroencephalograph (EEG) classification is an important technology that can establish a relationship between EEG features and cognitive tasks. Emerging matrix classifiers have been successfully applied to single-trial EEG classification, but they belong to shallow classifiers, making potentially powerful stacked generalization principle unexploited for learning deep predictive representations of EEG features. To learn the high-level representation and abstraction, we proposed a novel deep stacked support matrix machine (DSSMM) to improve the performance of existing shallow matrix classifiers in EEG classification.

Methods: The main idea of our framework is based on stacked generalization principle, in which support matrix machine (SMM) is introduced as the base building block of deep stacked network. The weak predictions of all previous layers obtained via SMM are randomly projected to help move apart the manifold of the original input EEG feature, and then the newly generated features are fed into the next layer of DSSMM. The framework only involves an efficient feed-forward rather than parameter fine-tuning with backpropagation, each layer of which is a convex optimization problem, thus simplifying the objective function solving process.

Results: We conduct extensive experiments on three benchmark EEG datasets and a self-collected EEG dataset. Experimental results demonstrate that the proposed DSSMM is competitive when compared with the available state-of-the-art methods.

Conclusion: Our method inherits the characteristic of matrix classifiers that can learn the structural information of data as well as the powerful capability of deep representation learning, which makes it adapted to classify complex matrix-form EEG data.

Keywords: Electroencephalograph; brain-computer interface; support matrix machine; stacked generalization; deep architecture

1. Introduction

Brain-computer interfaces (BCIs) are capable of establishing a direct communication pathway between a brain and various external device. Motor imagery (MI) is one of the most typical types of BCI paradigm, which is promising to apply MI-based BCIs in various areas (*e.g.*, entertainment and convenient life for healthy people [1, 2, 3]). More importantly, MI-based BCIs can be developed for assisting, augmenting or repairing sensory-motor functions of disabled patients [4, 5, 6, 7], especially those with neuromuscular disorders. Electroencephalograph (EEG) is commonly used to record brain electrical activity from the scalp in MI-based BCIs, due to its non-invasiveness, simplicity, and high temporal resolution. Considering that EEG recognition method can establish the mapping relationship between EEG signals and complex brain activities, accurate recognition of movement intentions from EEG signals is essential for achieving MI-based BCIs [8, 9].

With the rapid development of pattern recognition algorithms, various classification methods have been widely used for EEG recognition, such as support vector machine (SVM) [10, 11] linear discriminant analysis (LDA) [12, 13], k-nearest neighbor (KNN) [14], Bayesian classifier [15], extreme learning machine (ELM) [16, 17], and so on. These classic classifiers are constructed on the basis that the input features are in vector form. In practical applications, single-trial EEG signal records the voltage fluctuations of multiple channels over a period of time, making it more naturally represented in the form of two-dimensional matrix [18]. To accommodate format requirements of traditional classifiers for input data, it is often necessary to reshape EEG feature matrices into vectors or extract features in vector form [19]. However, because of the high correlation in multi-channel EEG signal [20, 24], the structural information between the columns or rows of the EEG feature matrix will inevitably be destroyed after vectorization.

To overcome the aforementioned problem, researchers proposed a series of matrix classification methods that can directly deal with the data in matrix form. The cornerstone of most matrix classification methods is the application of low-rank constraints to the regression matrix. For example, rank-k SVM [21] method is a representative matrix classifier that regularizes the regression matrix into the sum of k rank-one orthogonal matrices. Pirsiavash *et al*. [22] proposed a bilinear SVM (BSVM) classifier that decomposes the regression matrix into the product of two low-rank matrices. Although the above classification methods can preserve the structural information between columns or rows within the feature matrix, they all need to determine the rank of the

regression matrix in advance, resulting in a problem of difficulty in adjusting parameters. Inspired by the idea that the nuclear norm is the better convex approximation of the matrix rank on the unit ball of the matrices, Kobayashi *et al*. [23] proposed a method similar to BSVM classifier, which can automatically determine the rank of the regression matrix by using nuclear norm regularization. In addition, by introducing the squared Frobenius matrix norm and kernel norm, Luo *et al*. [24] proposed a novel support matrix machine (SMM), which can grasp the structural information of the feature matrix. The optimization problem of SMM can be easily solved through the alternating direction method of multipliers (ADMM). This method has proven to be more suitable for image classification and EEG classification than other matrix classification methods. Furthermore, researchers proposed many variants of SMM. Zhu *et al*. [28] proposed an entropy-based support matrix machine used to classify unbalanced data in matrix form. Zheng *et al*. [25, 29] extended the SMM to anti-noise SMM version and multiclass SMM version for EEG classification.

Despite the success of the above matrix classification methods, they belong to shallow classifiers, which leave the potentially powerful stacked generalization principle unexploited for learning deep representations of data, especially for the complex matrix-form EEG features. Deep architecture is built to achieve a more complex function approximation, which allows it to capture higher-level representations and abstractions of EEG signals. Most of the current deep architectures need to solve a difficult and non-convex optimization problem. Following the philosophy of stacked generalization [26, 27], recently researchers proposed several layer-by-layer models that can learn deep representations via a convex stacking architecture, where the module is simply classifier. For example, Cohen *et al*. [30] proposed a stacking architecture, which using Conditional Random Field as its basic block. Vinyals *et al*. [31] proposed a random recursive linear support vector machine (R^2SVM), which uses linear SVM as the base building module and modifies the original feature via a random shift. In a similar way, Yu *et al*. [32] proposed another deep architecture, called DrELM, with ELM as the basic module. By further introducing transfer learning mechanism, Wang *et al*. [33] proposed a deep transfer additive kernel least square support vector machine (DTA-LS-SVM), which stacks multiple AK-LS-SVM classifiers. In particular, the newly generated feature for the module in the higher layer comes from the concatenation of the original feature and the predicted output from the adjacent front layer.

Despite encouraging experimental performance of above-mentioned deep convex stacking

architectures, they are predominantly built on the assumption that the input features are vectors, which sometimes cannot directly process the input data represented as two-dimensional matrices. Motivated by structural information extraction and stacked generalization, we propose a novel deep architecture that uses SMM as a base building block and random shift as the stacking element. Specifically, the proposed deep stacked support matrix machine (DSSMM) is built in a layer-by-layer fashion. Each layer of the DSSMM contains an SMM module that can retain structural information between columns or rows within the EEG feature matrix. Furthermore, the random projections of weak predictions of each previous layer obtained by SMM are used to modify the original feature, and then the newly generated data is fed into the next layer of DSSMM. The random shift can help to move apart the manifolds in the original feature space in a stacked fashion, so that the EEG features are more linearly separable. In this way, DSSMM combines the virtue of SMM with the powerful feature representation derived from the deep architecture, making it suitable for EEG classification. Besides, DSSMM involves an efficient feed-forward instead of parameter fine-tuning using backpropagation, where each layer is a convex optimization problem. The experimental results show that DSSMM gives superior classification performance on three public EEG datasets and a self-collected EEG dataset, compared to the state-of-the-art matrix classifiers.

The remainder of the paper is organized as follows. In Section 2, we first give some notations and preliminaries involved in this study, and then introduce SMM. The proposed DSSMM and its learning procedure are presented in Section 3. In section 4, we conduct extensive experiments to evaluate the performance of the proposed method. Section 5 gives a conclusion of our work.

2. Brief Review of Support Matrix Machine

2.1 Notations and preliminaries

First, we briefly introduce the mathematical notations used in the following sections. Given a matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ of rank *r*, the singular value decomposition (SVD) of $\mathbf{X}$ can be represented as $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$ satisfy $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, and $\mathbf{\Sigma} = diag(\sigma_1, \sigma_2, \ldots, \sigma_r)$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$. $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} X_{ij}^2}$ and $\|\mathbf{X}\|_* = \sum_{i=1}^{r} \sigma_i$ are denoted as the Frobenius norm and the nuclear norm of matrix $\mathbf{X}$ respectively. For any $\tau \geq 0$, the singular value thresholding (SVT) operater [36,37] can be formulated as $\mathcal{D}_\tau[\mathbf{X}] = \mathbf{U}\mathcal{S}_\tau[\mathbf{\Sigma}]\mathbf{V}^T$, where

$\mathcal{S}_\tau[\boldsymbol{\Sigma}] = diag\left(\{\sigma_i - \tau\}_+\right)$ and $\{u\}_+ = \max(0,u)$.

2.2 Support matrix machine

Given a set of training data $\{\mathbf{X}_i, y_i\}_{i=1}^n$, $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$ represents the *i*-th trial of input EEG matrix and its ground truth label is denoted as $y_i \in \{1,-1\}$, where d_1 represents the number of EEG recording channels and d_2 is the number of sampling points. In particular, SMM is implemented by the hinge loss function plus the spectral elastic net penalty as follow:

$$\begin{aligned} \underset{\mathbf{W},b}{\arg\min} \quad & \frac{1}{2}\mathrm{tr}\left(\mathbf{W}^T\mathbf{W}\right) + \tau\|\mathbf{W}\|_* + C\sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i\left[\mathrm{tr}\left(\mathbf{W}^T\mathbf{X}_i\right) + b\right] \geq 1 - \xi_i, \forall i = 1,2,\ldots,n \end{aligned}, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ is the regression matrix, and b is the bias term. The spectral elastic net is composed of the squared Frobenius norm and nuclear norm. The squared Frobenius norm $\|\mathbf{W}\|_F^2 = \mathrm{tr}\left(\mathbf{W}^T\mathbf{W}\right)$ is taken to control model complexity and prevent overfitting. $\|\mathbf{W}\|_*$ is the nuclear norm which can be approximately alternative to the rank of the regression matrix. τ is the positive scalar used to penalize the nuclear norm term, and C is the trade-off parameter. Due to the property of grouping effect of the spectral elastic net, SMM can strongly capture the intrinsic structural information within the input EEG matrix.

Since the nuclear norm is the best convex approximation of matrix rank, it makes SMM can be easily optimized using the alternating direction method of multipliers (ADMM) method [34, 38]. In particular, after introducing an auxiliary variable $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$, the optimization problem in Eq. (1) can be reformulated as:

$$\begin{aligned} \underset{\mathbf{W},b,\mathbf{Z}}{\arg\min} \quad & P(\mathbf{W},b) + Q(\mathbf{Z}) \\ \text{s.t.} \quad & \mathbf{Z} - \mathbf{W} = 0 \end{aligned}, \tag{2}$$

where

$$\begin{aligned} P(\mathbf{W},b) &= \frac{1}{2}\mathrm{tr}\left(\mathbf{W}^T\mathbf{W}\right) + C\sum_{i=1}^n \left\{1 - y_i\left[\mathrm{tr}\left(\mathbf{W}^T\mathbf{X}_i\right) + b\right]\right\}_+ . \\ Q(\mathbf{Z}) &= \tau\|\mathbf{Z}\|_* \end{aligned} \tag{3}$$

Eq. (2) can be optimized using Augmented Lagrangian Multiplier (ALM) method:

$$L(\mathbf{Z},\mathbf{W},b,\mathbf{M}) = P(\mathbf{W},b) + Q(\mathbf{Z}) + \mathrm{tr}\left(\mathbf{M}^T(\mathbf{Z} - \mathbf{W})\right) + \frac{\beta}{2}\|(\mathbf{Z} - \mathbf{W})\|_F^2 , \tag{4}$$

where $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ is the Lagrangian multiplier, and β is a positive hyperparameter. The update

equation of $\mathbf{Z}$, $\mathbf{W}$, b, and Lagrangian multiplier $\mathbf{M}$ can be represented as:

$$\mathbf{Z}^{t+1} = \arg\min_{\mathbf{Z}} L\left(\mathbf{Z}, \mathbf{W}^t, b^t, \mathbf{M}^t\right), \tag{5}$$

$$\left(\mathbf{W}^{t+1}, b^{t+1}\right) = \arg\min_{(\mathbf{W},b)} L\left(\mathbf{Z}^{t+1}, \mathbf{W}, b, \mathbf{M}^t\right), \tag{6}$$

$$\mathbf{M}^{t+1} = \mathbf{M}^t - \beta\left(\mathbf{W}^{t+1} - \mathbf{Z}^{t+1}\right), \tag{7}$$

where t is the index of iteration. According to [24], the optimal solutions of Eq. (5)-(7) can be written as follow:

$$\begin{aligned} \hat{\mathbf{W}} &= \frac{1}{\beta+1}\left(\mathbf{M} + \beta\mathbf{Z} + \sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{X}_i\right), \\ \hat{b} &= \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \mathrm{tr}\left(\hat{\mathbf{W}}^T \mathbf{X}_i\right)\right\} \end{aligned} \tag{8}$$

$$\hat{\mathbf{Z}} = \frac{1}{\beta}\mathcal{D}_\tau\left(\beta\mathbf{W} - \mathbf{M}\right). \tag{9}$$

Finally, the decision function of SMM can be formulated as:

$$f(\mathbf{X}) = \mathrm{sgn}\left(\mathrm{tr}\left(\mathbf{W}^T\mathbf{X}\right) + b\right). \tag{10}$$

3. The Proposed DSSMM Method

To further improve the classification performance of SMM on EEG data in matrix form, we propose a novel deep stacked architecture, *i.e.*, DSSMM, which combines the matrix classification capability of SMM with the merit derived from deep architectures.

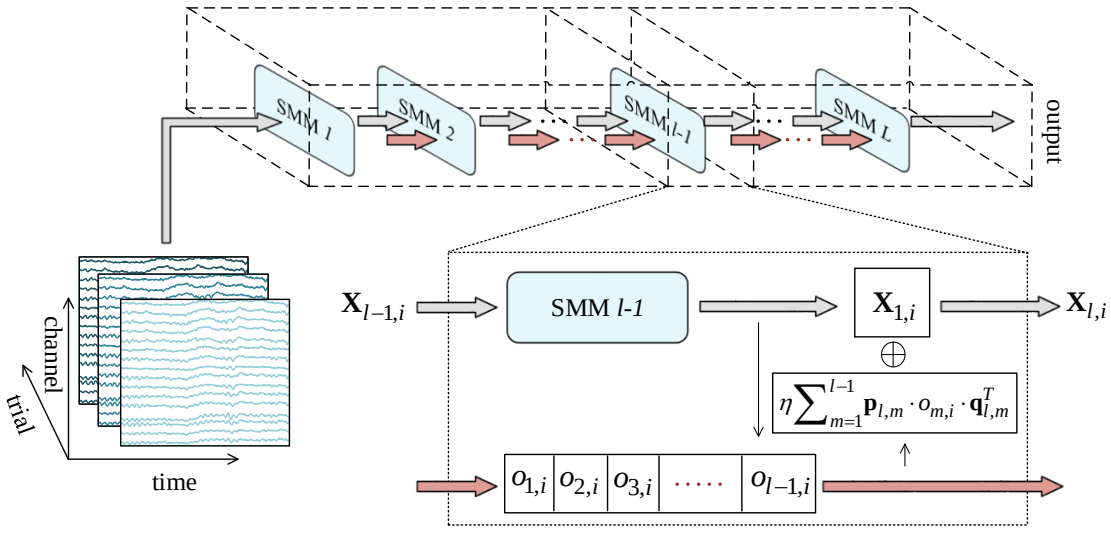


Fig.1. Overall framework of the proposed DSSMM for EEG classification. In this convex stacking architecture, the model uses simple SMMs as its base building block. The input feature $\mathbf{X}_{l,i}$ of the current layer is constructed by combining the original feature $\mathbf{X}_{1,i}$ and random projections of all previous layers.

3.1 Deep Stacked SMM

The proposed DSSMM is built in a layer-by-layer fashion using SMM as its basic module, in which the random projections of weak predictions from the previous layers SMM are combined with the original feature matrices to transform the manifold of single-trial EEG data. Fig.1 demonstrates the overall framework of the proposed DSSMM.

For the first layer of the DSSMM framework, the original EEG feature matrices are used as the input features, *i.e.*, $\mathbf{X}_{1,i}=\mathbf{X}_i, i=1,2,\ldots,n$, where $\mathbf{X}_{1,i}$ is the *i*-th input matrix data sample. The weak predictive output $o_{1,i}$ of $\mathbf{X}_{1,i}$ can be subsequently obtained via SMM, *i.e.*, $o_{1,i}=\mathrm{tr}\left(\mathbf{W}_1^T\mathbf{X}_{1,i}\right)+b$. For the second layer, we map the vector $o_{1,i}$ into a space consistent with the original feature dimension using two random projection vectors, and superimpose the obtained results on the original EEG feature matrix, so that the *i*-th input EEG matrix data sample of the second layer can be denoted as $\mathbf{X}_{2,i}$:

$$\mathbf{X}_{2,i}=\mathbf{X}_i+\eta\cdot\mathbf{p}_{2,1}\cdot o_{1,i}\cdot\mathbf{q}_{2,1}^T, \tag{11}$$

where η can regulate the degree of random shift relative to the original EEG feature matrix $\mathbf{X}_i$. Both $\mathbf{p}_{2,1}\in\mathbb{R}^{d_1}$ and $\mathbf{q}_{2,1}\in\mathbb{R}^{d_2}$ are random projection vectors whose elements are sampled from $N(0,1)$.

Without loss of generality, we can train the *l*-th layer of DSSMM model. Assume that *i*-th input sample of the *l*-th layer, *i.e.*, $\mathbf{X}_{l,i}$, is as follows:

$$\mathbf{X}_{l,i}=\mathbf{X}_i+\eta\sum_{m=1}^{l-1}\mathbf{p}_{l,m}\cdot o_{m,i}\cdot\mathbf{q}_{l,m}^T. \tag{12}$$

The random projection $\mathbf{p}_{l,m}$ and $\mathbf{q}_{l,m}$ are used to project the prediction output of the *m*-th layer simultaneously. Finally, we add random projection results of all previous layers to the original EEG feature matrix, mainly to move apart its manifold in a stacked fashion.

Due to the utilization of hinge loss function, the effectiveness of the proposed DSSMM can be guaranteed according to the theoretical analysis in [31]. Specifically, with the monotonically decreasing nature of the hinge loss function, it is possible to find an offset added to the original matrix data, making the proposed DSSMM pull these data belonging to different classes towards

their respective directions. This is the reason that can make the newly generated matrix data is more linearly separable and thus achieve better EEG classification performance. The pseudocode of DSSMM algorithm is given in *Algorithm 1*.

3.2 Computational complexity

We further analyze the time complexity of Algorithm 1. Given n EEG training samples with $d_1 \times d_2$ dimensions of each feature matrix, step 7 in *Algorithm 1* takes $O\left(n^2 d_1 d_2\right)$ to update $\mathbf{W}$ and b of SMM via solving the quadratic programming problems. Step 8 calculates the singular value decomposition for $\mathbf{Z}$, which takes $O\left(\min\left(d_1^2 d_2, d_1 d_2^2\right)\right)$. In general, both the number of EEG

Algorithm 1 Learning Algorithm for the Proposed DSSMM

Input: Training set $\left\{\mathbf{X}_i, y_i\right\}_{i=1}^n$, parameter η, number of layers L, $\beta > 0$, $\left\{C_l\right\}_{l=1}^L$, $\left\{\tau_l\right\}_{l=1}^L$.

Output: The stacked structure of DSSMM with tuned parameter values.

1 Initialize: l=1, $t = 1$, $\left\{\mathbf{W}_l^0\right\}_{l=1}^L = \mathbf{0}$, $\left\{\mathbf{Z}_l^0\right\}_{l=0}^L = \mathbf{0}$, $\left\{\mathbf{M}_l^0\right\}_{l=0}^L = \mathbf{0}$;

2 Construct the l-th module using Eq. (1), obtain parameters $\mathbf{W}_1$, b_1 and weak prediction $\mathbf{o}_{1,i}$, $\forall i, i = 1, 2, \ldots, n$;

3 **While** $l \le L$ do

4 $l = l + 1$;

5 Generate the random projection vectors $\mathbf{p}_{l,m}$, $\mathbf{q}_{l,m}$, $m = 1, \ldots, l-1$ whose values are sampled from $N(0,1)$;

6 Compute the $\mathbf{X}_{l,i}$, $\forall i, i = 1, 2, \ldots, n$ by Eq. (14);

Repeat

7 Update $\mathbf{W}_l^t$ and b_l^t with Eq. (8);

8 Update $\mathbf{Z}_l^t$ with Eq. (9);

9 Update $\mathbf{M}_l^t$ with Eq. (7);

10 $t = t + 1$;

Until convergence

11 Obtain parameters $\mathbf{W}_l$, b_l and weak prediction $\mathbf{o}_{l,i}$, $\forall i, i = 1, 2, \ldots, n$;

End

---

channels and the number of sampling points are not too high, so that the time complexity is mainly dominated via the quadratic programming in step 7. In this regard, the computational complexity of DSSMM is $O\left(L \cdot K \cdot n^2 d_1 d_2\right)$, where K is the number of iterations and L is the number of layers.

4 Experimental Evaluation

In this section, extensive experiments are conducted to evaluate the effectiveness of the proposed DSSMM on three public EEG datasets and one self-collected EEG dataset. Based on the introduction of the experimental datasets, we describe the comparison algorithm and its parameter settings, and finally provide the performance evaluation metrics.

4.1 EEG Data Description

1) *BCI competition III Dataset IVa* (*Exp.1*): This dataset contains EEG signals recorded from five subjects (denoted as "*aa*", "*al*", "*av*", "*aw*" and "*ay*") at 118 electrodes with a sampling frequency of 100 Hz. In the experiment, each subject was asked to perform a sequential repetition of 280 trials based on a visual cue. In each trial, an arrow cue was presented to guide each subject to imagine either right-hand or foot movement for 3.5 s. The number of training (labelled) trials are 168, 224, 84, 56 and 28 for subject "*aa*", "*al*", "*av*", "*aw*", and "*ay*" respectively, and the remaining trials are used as test (unlabeled) dataset. More detail about the dataset can be found on website http://www.bbci.de/competition/iii/.

2) *BCI Competition IV Dataset IIb* (*Exp.2*): In this dataset, EEG signals are recorded from nine subjects (denoted as "*B01*", "*B02*", "*B03*", "*B04*", "*B05*", "*B06*", "*B07*", "*B08*", and "*B09*") using electrodes C3, Cz, and C4 with a sampling frequency of 250 Hz. For each trial, subjects are instructed to imagine either left hand or right hand movement for 4.5 s according to a visual cue. See website http://www.bbci.de/competition/iv/ for more detail about this BCI public dataset. In the experiment, EEG signals from session 1, 2, and 3 are used to train the classifiers, while the remaining two sessions serve as the test dataset to evaluate the performance of EEG classification.

3) *BCI Competition IV Dataset IIa* (*Exp.3*): This dataset consists of EEG signals from nine subjects (denoted as "*S1*", "*S2*", "*S3*", "*S4*", "*S5*", "*S6*", "*S7*", "*S8*", and "*S9*") acquired using 22 electrodes with a sampling frequency of 250 Hz. Each subject performs a total of 576 trials, including four-class motor imagery tasks related to left hand, right hand, foot, and tongue. More details about this dataset can refer to the website http://www.bbci.de/competition/iv/. In the

Table I. Summary of four EEG datasets

Datasets	#subset	Dimension	#pos	#neg
BCI competition III Dataset IVa (*Exp.1*)	5	250×6	140	140
BCI Competition IV Dataset IIb (*Exp.2*)	54	625×6	72	72
BCI Competition IV Dataset IIa (*Exp.3*)	9	625×2	360 ± 20	360 ± 20
Lower Limb MI-BCI Dataset (*Exp.4*)	10	640×6	100	100

("#pos" is abbreviation to "the number of positive class", and the same applies to "#subset" and "#neg")

experiment, we decompose the four-class data into $C_4^2=6$ binary subsets. In view of this, training and test datasets contain 72 trials per motor imagery tasks for each subject.

4) *Lower Limb MI-BCI Dataset* (*Exp.4*): This dataset records self-collected 32-channels EEG signals from ten subjects with a sampling rate of 256Hz. During the EEG data collection, subjects are instructed to imagine lower limb movement with the help of visual guidance provided by a virtual reality system. For each subject, a total of 200 trials are available involved two control groups ("*idle*" state or "*walking imagery*" state). In the experiment, we randomly selected 60 samples to construct the training dataset, and the remaining 140 samples are used as test data.

Herein, for both public EEG datasets and self-collected EEG dataset, we adopt the time interval of [0.5, 3] s after visual cue in each trial. EEG signals are firstly filtered with a fifth-order Butterworth band-pass filter in the frequency range of 8-30 Hz. We then use spatial filters to further detect event-related desynchronization/synchronization (ERD/ERS) patterns associated with movement imagination tasks. To extract EEG feature in matrix form, we perform the most commonly used band-power estimates method. The main information of aforementioned datasets are described in Table I.

4.2 Experimental Setup

To verify the effectiveness of the proposed DSSMM, two vector classifiers and two state-of-the-art matrix classifiers are chosen as baseline methods in the comparative experiments, which include:

1) Support Vector Machine (SVM) [35];

2) Random Recursive Linear SVM (R^2SVM) [31];

3) Bilinear SVM (BSVM) [23];

4) Support Matrix Machine (SMM) [24];

5) The proposed DSSMM.

Details of the parameter settings for each comparison method are as follows. The optimal parameters for all algorithms are determined using ten-fold cross-validation strategy. Specifically, the regularization parameters C of BSVM, SVM and SMM, as well as the parameters in each layer of R^2SVM and DSSMM are all obtained by searching for values in set $\{5e\text{-}4,2e\text{-}4,1e\text{-}4,1e\text{-}3,2e\text{-}3,5e\text{-}3,1e\text{-}2,2e\text{-}2,5e\text{-}2,1e\text{-}1,2e\text{-}1,5e\text{-}1,1e0\}$. Referring to [25], the number of iterations t is selected from the set $\{100,300,500\}$ for BSVM. For SMM and the modules in each layer of DSSMM, we determine the parameter τ from the set $\{1e\text{-}4,1e\text{-}3,1e\text{-}2,2e\text{-}2,5e\text{-}2,1e\text{-}1,2e\text{-}1,5e\text{-}1,1e0,1e1\}$ and set the number of iterations to 1000. For R^2SVM and DSSMM, the trade-off parameter η is used to control how much the original feature matrix is randomly shifted, and the value is fixed to 0.1[31]. Besides, for small or medium EEG dataset, the number of layers L is usually not set too large, mainly due to large L that is likely to cause over-fitting problem [33]. In our experiments, the number of layers L is set to 2 to 6. To meet the format requirements of the input data for SVM and R^2SVM, we reshape the EEG feature matrix into a vector as their inputs.

To measure the classification performance of different comparison methods, the following metrics: Accuracy (ACC), F1-score (F1), and area under the receiver operating characteristics curve (AUC) are adopted to evaluate experimental results. In detail, according to the denotation in [39], ACC = (TP+TN)/(TP+FN+FP+TN) and F1 = 2×PPV×SEN/(PPV+SEN), where positive predictive value (PPV) is equal to TP/(TP+ FP) and sensitivity (SEN) is equal to TP/(TP+ FN). We used *t*-test statistical analysis to determine whether there is a significant difference between the proposed DSSMM and other comparison methods in improving EEG classification results.

5.3 Experimental Results and Analysis

In this part, we firstly give the experimental results of all comparison methods on three public EEG datasets, followed by the experimental results of the self-collected EEG dataset.

5.3.1 Results on Public EEG Datasets

To evaluate the improvement of classification performance after integrating EEG feature matrix learning into the convex stacking architecture, DSSMM is firstly compared to other competitive methods on Dataset IVa of BCI Competition III and Dataset IIb of BCI Competition

IV. Table II shows the performance comparison of DSSMM with two vector classifiers SVM and R^2SVM, and two matrix classifiers BSVM and SMM. Fig. 2 illustrates average evaluation values

Table II. Classification performance of different algorithms on Dataset IVa of BCI Competition III and Dataset IIb of BCI Competition IV

Datasets	Subjects	Metrics	Methods				
			SVM	R^2SVM	BSVM	SMM	DSSMM
Exp.1	*aa*	ACC	0.7232	0.7321	0.7143	0.7411	**0.7589**
		F1	0.6353	0.6591	0.6279	0.6742	**0.6966**
		AUC	0.7865	0.7952	0.7804	0.7955	**0.8144**
	al	ACC	**1**	**1**	**1**	**1**	**1**
		F1	**1**	**1**	**1**	**1**	**1**
		AUC	**1**	**1**	**1**	**1**	**1**
	av	ACC	0.7398	0.7500	0.7500	0.7500	**0.7653**
		F1	0.7411	0.7513	0.7200	0.7416	**0.7538**
		AUC	0.7564	0.7702	0.7854	0.7682	**0.7984**
	aw	ACC	0.8080	0.8214	0.8616	0.8929	**0.8973**
		F1	0.8072	0.8214	0.8517	0.8889	**0.8950**
		AUC	0.7953	0.8068	0.8616	0.8924	**0.8973**
	ay	ACC	0.6508	0.6627	0.7540	0.7460	**0.7619**
		F1	0.4943	0.5198	0.6869	0.6735	**0.7000**
		AUC	0.6830	0.6925	0.7941	0.7849	**0.8025**
Exp.2	*B01*	ACC	0.6750	0.6813	0.6719	0.6688	**0.6909**
		F1	0.5806	0.5526	0.5755	0.5310	**0.5959**
		AUC	0.7127	0.7220	0.7079	0.6990	**0.7289**
	B02	ACC	0.5536	0.5679	0.5286	0.5571	**0.5714**
		F1	0.5583	0.5714	0.5221	0.5540	**0.5724**
		AUC	0.5561	0.5678	0.5346	0.5619	**0.5740**
	B03	ACC	0.5344	0.5406	0.5281	0.5688	**0.5844**
		F1	0.5387	**0.5612**	0.5266	0.4964	0.5333
		AUC	0.5345	0.5420	0.5245	0.5726	**0.5868**
	B04	ACC	0.9531	0.9563	0.9594	0.9688	**0.9719**
		F1	0.9533	0.9565	0.9585	0.9686	**0.9718**
		AUC	0.9460	0.9475	0.9620	0.9669	**0.9671**
	B05	ACC	0.6656	0.6844	0.6594	0.6938	**0.7000**
		F1	0.5771	**0.6039**	0.5512	0.5984	0.6033
		AUC	0.6464	0.6679	0.6350	0.6650	**0.6714**
	B06	ACC	0.7688	0.7781	**0.7844**	0.7781	**0.7844**
		F1	0.7921	0.8033	**0.8130**	0.8076	**0.8130**
		AUC	0.8006	0.8169	0.8212	0.8171	**0.8216**
	B07	ACC	0.7250	0.7344	0.7625	0.7813	**0.7844**
		F1	0.6812	0.6863	0.7266	0.7697	**0.7723**
		AUC	0.7332	0.7421	0.7711	0.7877	**0.7894**
	B08	ACC	0.9125	**0.9156**	0.8813	0.9125	**0.9156**
		F1	0.9125	**0.9164**	0.8841	0.9125	0.9159
		AUC	0.9192	0.9240	0.8996	0.9192	**0.9256**
	B09	ACC	0.8250	0.8250	0.8375	0.8500	**0.8594**
		F1	0.8158	0.8158	0.8485	0.8491	**0.8580**
		AUC	0.8517	0.8517	0.8712	0.8803	**0.8883**

across all subjects. From these experimental results, we can make the following observations.

In most cases, as shown in Table II, it can be found that the DSSMM has higher values than other methods on all evaluation metrics. Specifically, for Dataset IVa of BCI Competition III, as shown in Fig. 2(a), the average results of DSSMM on all five subjects are 83.67%, 80.91%, and 86.25%, corresponding to ACC, F1, and AUC. The absolute accuracy increase of 1.07%, 1.35% and 1.43% against the best baseline SMM. Fig. 2(b) shows the average results of evaluation metrics on Dataset IIb of BCI Competition IV. We can find that DSSMM outperforms the best competitive classifier SMM by 0.93%, 1.65%, and 0.93%, respectively.

It can be seen that the classification performances of matrix classifiers are in most cases better than that of vector classifiers. The major limitation of SVM is that reshaping the input matrices into vectors may destroy the structural information of EEG feature matrix, resulting in the classification performance degradation. Besides, R^2SVM is superior to the benchmark SVM, and it proves the effectiveness of the deep stacked architecture.

Compared with the two matrix classifiers BSVM and SMM, our proposed DSSMM achieves better classification performance. This is because our method can exploit the potentially powerful stacked generalization principle to find the predictive deep representations of EEG feature matrix. The evaluation values of DSSMM are obviously higher than R^2SVM. This further implies that the intrinsic structural information of feature matrix indeed helps to improve the performance of EEG classification.

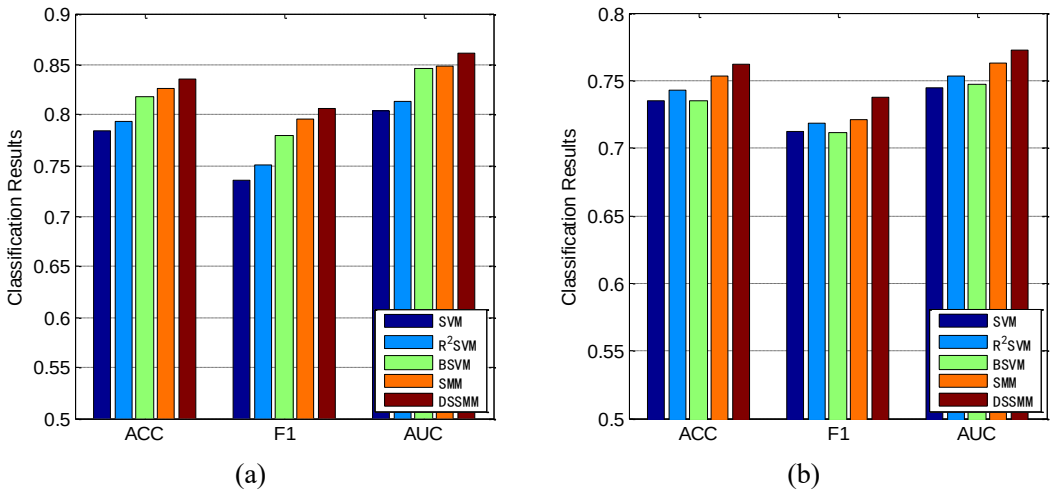
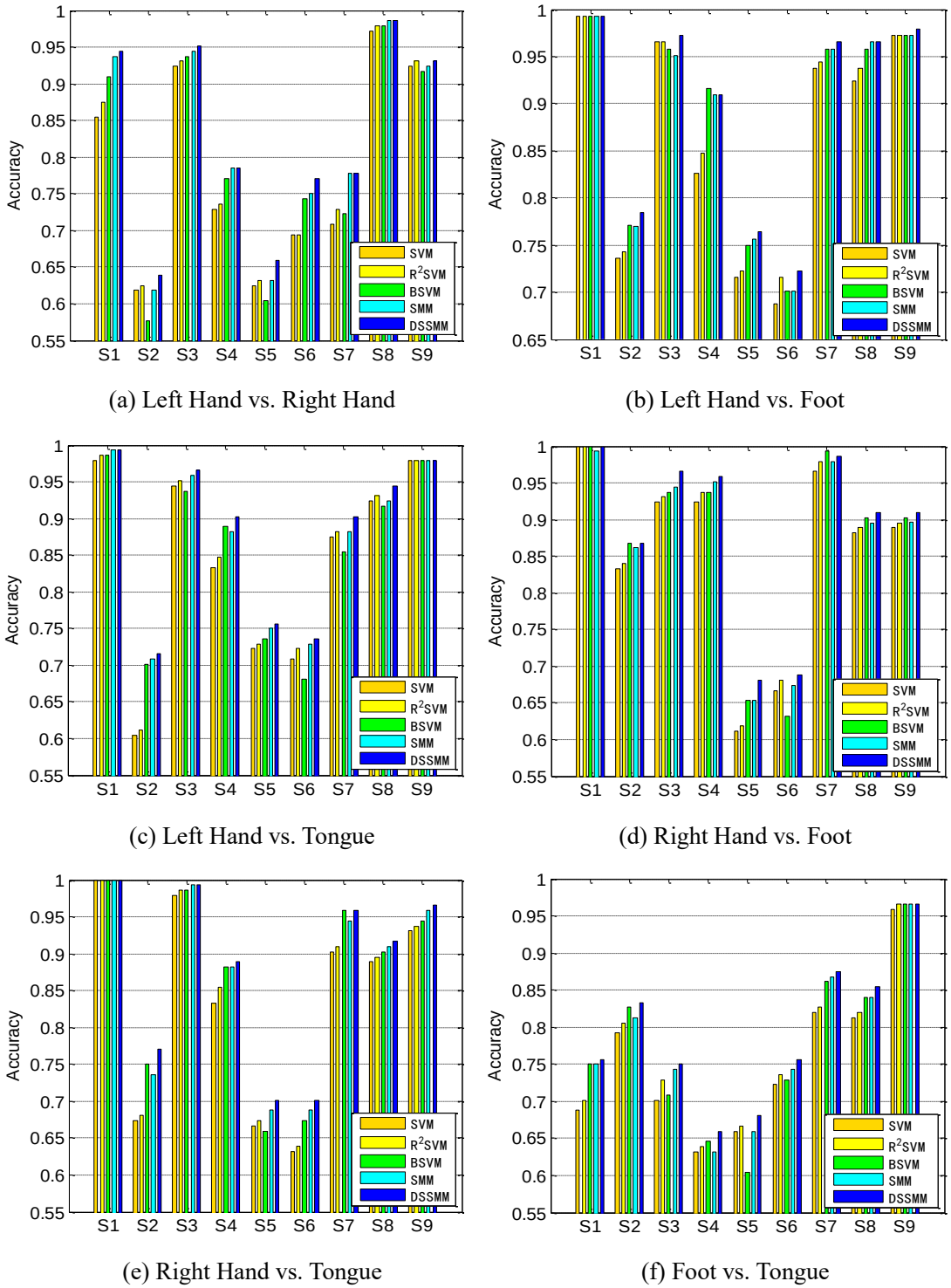


Fig. 2. Average classification performance of different algorithms on (a) Dataset IVa of BCI Competition III and (b) Dataset IIb of BCI Competition IV

We further evaluate the performance of the proposed DSSMM and other comparison methods on Dataset IIa of BCI Competition IV. Due to space constraints, only the classification accuracy results are listed, as shown in Fig. 3. Consistent with the above conclusions, it is proved again that



(a) Left Hand vs. Right Hand

(b) Left Hand vs. Foot

(c) Left Hand vs. Tongue

(d) Right Hand vs. Foot

(e) Right Hand vs. Tongue

(f) Foot vs. Tongue

Fig. 3. Classification performance of different algorithms on Dataset IIa of BCI Competition IV

the classification accuracy of the matrix classifier is better than SVM for most subjects. We can see that DSSMM achieves the best performance as expected, especially for subjects (i.e., S2, S5, and S6) whose EEG data are not easily distinguishable. The results empirically indicate the effectiveness and robustness of our approach.

5.3.2 Results on Self-collected EEG Dataset

To evaluate the applicability of the proposed DSSMM, we testify whether the proposed model can work well for the EEG signals recorded from real-world BCI system. Table III lists the performance comparison among SVM, R2SVM, BSVM, SMM, and DSSMM on 10 subjects using various evaluation metrics. Fig. 4 illustrates the average performance of all comparison methods. In

Table III. Classification performance of different algorithms on real-world EEG dataset

Subjects	Metrics	Methods				
		SVM	R^2SVM	BSVM	SMM	DSSMM
S01	ACC	0.7929	0.8071	0.8286	0.8429	**0.8571**
	F1	0.8027	0.8138	0.8421	0.8493	**0.8611**
	AUC	0.7861	0.8104	0.8363	0.8427	**0.8629**
S02	ACC	0.8000	0.8143	0.8214	0.8286	**0.8429**
	F1	0.7941	0.8088	0.8276	0.8235	**0.8451**
	AUC	0.8245	0.8390	0.8363	0.8424	**0.8551**
S03	ACC	0.9357	0.9429	0.9286	0.9357	**0.9571**
	F1	0.9379	0.9444	0.9324	0.9379	**0.9583**
	AUC	0.9427	0.9492	0.9396	0.9433	**0.9696**
S04	ACC	0.7929	0.8143	0.8357	0.8429	**0.8571**
	F1	0.7972	0.8219	0.8535	0.8590	**0.8701**
	AUC	0.7565	0.7731	0.8137	0.8380	**0.8427**
S05	ACC	0.9357	0.9357	0.9429	0.9500	**0.9571**
	F1	0.9388	0.9388	0.9452	0.9517	**0.9583**
	AUC	0.9604	0.9604	0.9500	0.9584	**0.9614**
S06	ACC	0.7571	0.7643	0.7357	0.7500	**0.7786**
	F1	0.7671	0.7755	0.7517	0.7586	**0.7891**
	AUC	0.8045	0.8045	0.7814	0.7884	**0.8245**
S07	ACC	0.9786	0.9786	0.9857	**0.9929**	**0.9929**
	F1	0.9787	0.9787	0.9857	**0.9929**	**0.9929**
	AUC	0.9773	0.9773	0.9859	**0.9900**	**0.9900**
S08	ACC	0.9286	0.9357	0.9429	0.9429	**0.9500**
	F1	0.9306	0.9379	0.9429	0.9437	**0.9504**
	AUC	0.9267	0.9337	0.9388	0.9333	**0.9429**
S09	ACC	0.7000	0.7143	0.6786	0.7071	**0.7429**
	F1	0.6500	0.6667	0.6400	0.6612	**0.7097**
	AUC	0.7322	0.7392	0.7010	0.7349	**0.7645**
S10	ACC	0.8857	0.8929	0.8929	**0.9071**	**0.9071**
	F1	0.8857	0.8921	0.9007	**0.9078**	**0.9078**
	AUC	0.8912	0.8953	0.8810	**0.9202**	**0.9202**

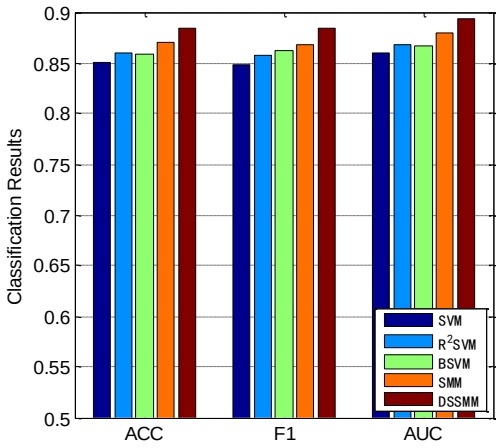


Fig. 4. Average classification performance of different algorithms on self-collected EEG dataset

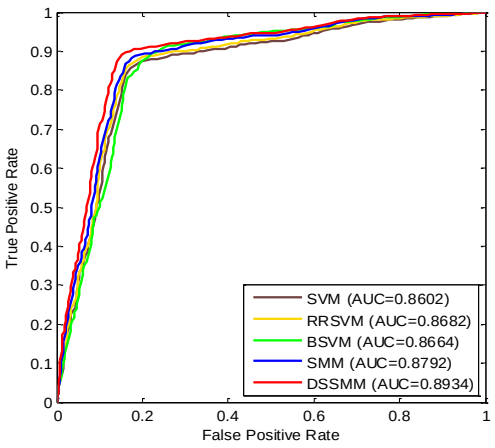


Fig. 5. ROC curves of different algorithms on self-collected EEG dataset

general, it can be seen that DSSMM still exhibits superior classification performance and substantially improves the classification accuracy on 8 out of 10 subjects. In terms of average evaluation results, ACC, F1, and AUC are 88.43%, 88.43%, and 89.34% for DSSMM, and the absolute values increase are 1.43%, 1.57%, and 1.42%, respectively, against the best competitive matrix classifier SMM.

Fig. 5 shows the average ROC curves of five comparison methods. The ROC curve of the proposed DSSMM is closer to the upper left corner of the figure than that of other methods, which also proves the superiority of our method in the classification of EEG feature matrix.

We further perform a t-test statistical analysis of comparison methods with a confidence level set to 0.05, and the results with the significant difference in statistics are marked in bold, as shown in Table IV. For the self-collected EEG data, we can see that all evaluation metrics satisfy the requirement of statistical significance (p-value ≤ 0.05), which reveals that the proposed DSSMM can significantly improve the classification performance compared to other methods. This highlights the potential value of DSSMM for practical applications.

Table IV. Statistical significance comparison between DSSMM and other classifiers

Method	DSSMM *vs*. SVM	DSSMM *vs*. R^{2}SVM	DSSMM *vs*. BSVM	DSSMM *vs*. SMM
ACC	**0.00031**	**0.00025**	**0.00151**	**0.00385**
F1	**0.00055**	**0.00040**	**0.00540**	**0.00926**
AUC	**0.00393**	**0.00391**	**0.00131**	**0.00744**

6. Conclusion

To seek predictive deep representations of extracted matrix-form EEG feature, we propose a novel deep architecture called DSSMM based on the stacked generalization principle. The proposed DSSMM uses a simple matrix classifier SMM as the basic stacking unit to deal with EEG spatial-temporal pattern, which is usually represented as a matrix with strong correlations between rows and columns. At the same time, to obtain more efficient deep representation and achieve better separability, random projections of weak prediction from all previous SMM modules are used to help open the manifold of original input EEG feature. As far as we know, this is the first attempt to incorporate a matrix classification model into the deep architecture.

We extensively evaluate our approach on three public EEG datasets and a self-collected EEG dataset. Experimental results show that DSSMM is superior to other comparison algorithms in most cases. Despite the promising performance of DSSMM, it still leaves considerable room for further improvement. For example, the development of transfer learning strategy is important to further improve the generalization capability of DSSMM in scenes with insufficient EEG data. Future work will be devoted to above issue.

Acknowledgments

This work was supported in part by the National project funding for Key R & D programs (2018YFC0808500), by the National Natural Science Foundation of China (61802177, 61902197), by the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province under Grant (18KJB520020), by the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems (2014DP173025), and by the Scientific Research Foundation for the Introduction of Talent of Nanjing Tech University (3827401749).

Conflicts of interest

None.

References

[1] A. Vourvopoulos, S. B. i Badia, F. Liarokapis, EEG correlates of video game experience and user profile in motor-imagery-based brain–computer interaction, The Visual Computer 33 (2017) 533-546.

[2] Z. Wang, Y. Yu, M. Xu, et al., Towards a hybrid BCI gaming paradigm based on motor imagery and SSVEP, International Journal of Human–Computer Interaction 35 (2019) 197-205.

[3] S. Stober, D. J. Cameron, J. A. Grahn, Using Convolutional Neural Networks to Recognize Rhythm Stimuli from Electroencephalography Recordings, in: Advances in Neural Information Processing Systems, 2014, pp. 1449-1457.

[4] Z. T. Al-Qaysi, B. B. Zaidan, A. A. Zaidan, et al., A review of disability EEG based wheelchair control system: Coherent taxonomy, open challenges and recommendations, Computer methods and programs in biomedicine 164 (2018) 221-237.

[5] R. Sitaram, H. Zhang, C. Guan, et al., Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain–computer interface, NeuroImage 34 (2007) 1416-1427.

[6] K. K. Ang, C. Guan, K. S. G. Chua, et al., A large clinical study on the ability of stroke patients to use an EEG-based motor imagery brain-computer interface, Clinical EEG and Neuroscience 42 (2011) 253-258.

[7] N. Birbaumer, L. G. Cohen, Brain–computer interfaces: communication and restoration of movement in paralysis, The Journal of physiology 579 (2007) 621-636.

[8] M. E. M. Mashat, C. T. Lin, D. Zhang, Effects of Task Complexity on Motor Imagery Based Brain-Computer Interface, IEEE Transactions on Neural Systems and Rehabilitation Engineering 2019, to be published.

[9] G. Cheron, Motor imagery in children with cerebral palsy: one step beyond with EEG dynamics, Developmental Medicine & Child Neurology 58 (2016) 223-224.

[10] Y. Zhang, C. S. Nam, G. Zhou, et al., Temporally constrained sparse group spatial patterns for motor imagery BCI, IEEE transactions on cybernetics 49 (2018) 3322-3332.

[11] S. Siuly, Y. Li, Improving the separability of motor imagery EEG signals using a cross correlation-based least square support vector machine for brain–computer interface, IEEE Transactions on Neural Systems and Rehabilitation Engineering 20 (2012) 526-538.

[12] R. Aldea, M. Fira, Classifications of motor imagery tasks in brain computer interface using linear discriminant analysis, International Journal of Advanced Research in Artificial Intelligence 3 (2014) 5-9.

[13] A. Subasi, M. I. Gursoy, EEG signal classification using PCA, ICA, LDA and support vector machines, Expert systems with applications 37 (2010) 8659-8666.

[14] J. Kevric, A. Subasi, Comparison of signal decomposition methods in classification of EEG signals for motor-imagery BCI system, Biomedical Signal Processing and Control 31 (2017) 398-406.

[15] Y. Zhang, G. Zhou, J. Jin, et al., Sparse Bayesian classification of EEG for brain–computer interface, IEEE transactions on neural networks and learning systems 27 (2015) 2256-2267.

[16] N. Y. Liang, P. Saratchandran, G. B. Huang, et al., Classification of mental tasks from EEG signals using extreme learning machine, International journal of neural systems 16 (2006) 29-38.

[17] Q. Yuan, W. Zhou, S. Li, et al., Epileptic EEG classification based on extreme learning machine and nonlinear features, Epilepsy research 96 (2011) 29-38.

[18] Q. Zheng, F. Zhu, J. Qin, et al., Sparse support matrix machine, Pattern Recognition 76 (2018) 715-726.

[19] H. U. Amin, A. S. Malik, R. F. Ahmad, et al., Feature extraction and classification for EEG signals using wavelet transform and machine learning techniques, Australasian physical & engineering sciences in medicine 38 (2015) 139-149.

[20] H. Zhou, L. Li, Regularized matrix regression, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76 (2014) 463-483.

[21] L. Wolf, H. Jhuang, T. Hazan, Modeling appearances with low-rank SVM, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-6.

[22] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, Bilinear classifiers for visual recognition, in: Advances in Neural Information Processing Systems, 2009, pp.1482-1490.

[23] T. Kobayashi, N. Otsu, Efficient optimization for low-rank integrated bilinear classifiers, in: European Conference on Computer Vision (ECCV), 2012, pp. 474-487.

[24] L. Luo, Y. Xie, Z. Zhang, et al., Support matrix machines, in: International Conference on Machine Learning, 2015, pp. 938-947.

[25] Q. Zheng, F. Zhu, P. A. Heng, Robust Support Matrix Machine for Single Trial EEG

Classification, IEEE Transactions on Neural Systems and Rehabilitation Engineering 26 (2018) 551-562.

[26] D. H. Wolpert, Stacked generalization, Neural networks 5 (1992) 241-259.

[27] L. Breiman, Stacked regressions, Machine learning 24 (1996) 49-64.

[28] C. Zhu and Z. Wang, Entropy-based matrix learning machine for imbalanced datasets, Pattern Recognition Letters 88 (2017) 72–80.

[29] Q. Zheng, F. Zhu, J. Qin, et al., Multiclass support matrix machine for single trial EEG classification, Neurocomputing 275 (2018) 869-880.

[30] W. W. Cohen, V. R. Carvalho, Stacked sequential learning, in: Proceedings of the 19th International Joint Conference on Artificial Intelligence, 2005, pp. 671-676.

[31] O. Vinyals, Y. Jia, L. Deng, et al., Learning with recursive perceptual representations, in: Advances in Neural Information Processing Systems, 2012, pp. 2825-2833.

[32] W. Yu, F. Zhuang, Q. He, et al., Learning deep representations via extreme learning machines, Neurocomputing 149 (2015) 308-315.

[33] G. Wang, G. Zhang, K. S. Choi, et al., Deep additive least squares support vector machines for classification with model transfer, IEEE Transactions on Systems, Man, and Cybernetics: Systems 2017, to bepublished.

[34] N. Parikh, S. P. Boyd, et al., Proximal algorithms, Foundations and Trends in optimization 1 (2014) 127–239.

[35] S. Li, W. Zhou, Q. Yuan, et al., Feature extraction and recognition of ictal EEG using EMD and SVM, Computers in biology and medicine 43 (2013) 807-816.

[36] E. J. Candès, B. Recht, Exact matrix completion via convex optimization, Foundations of Computational mathematics 9 (2009) 717-772.

[37] J. F. Cai, E. J. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM Journal on optimization 20 (2010) 1956-1982.

[38] T. Goldstein, B. O'Donoghue, S. Setzer, et al., Fast alternating direction optimization methods, SIAM Journal on Imaging Sciences 7 (2014) 1588-1623.

[39] B. Lei, X. Liu, S. Liang, et al., Walking imagery evaluation in brain computer interfaces via a multi-view multi-level deep polynomial network, IEEE transactions on neural systems and rehabilitation engineering 27 (2019) 497-506.