Neural Computing and Applications (2023) 35:12001–12027
https://doi.org/10.1007/s00521-023-08336-z

ORIGINAL ARTICLE

Multi-classification for EEG motor imagery signals using data evaluation-based auto-selected regularized FBCSP and convolutional neural network

Yang An[1] · Hak Keung Lam[2] · Sai Ho Ling[3]

Received: 3 June 2022 / Accepted: 25 January 2023 / Published online: 19 February 2023

Abstract
In recent years, there has been a renewal of interest in brain–computer interface (BCI). One of the BCI tasks is to classify the EEG motor imagery (MI). A great deal of effort has been made on MI classification. What seems to be lacking, however, is multiple MI classification. This paper develops a single-channel-based convolutional neural network to tackle multi-classification motor imagery tasks. For multi-classification, a single-channel learning strategy can extract effective information from each independent channel, making the information between adjacent channels not affect each other. A data evaluation method and a mutual information-based regularization parameters auto-selection algorithm are also proposed to generate effective spatial filters. The proposed method can be used to tackle the problem of an inaccurate mixed covariance matrix caused by fixed regularization parameters and invalid training data. To illustrate the merits of the proposed methods, we used the tenfold cross-validation accuracy and kappa as the evaluation measures to test two data sets. BCI4-2a and BCI3a data sets have four mental classes. For the BCI4-2a data set, the average accuracy is 79.01%, and the kappa is 0.7202 using data evaluation-based auto-selected filter bank regularized common spatial pattern voting (D-ACSP-V) and single-channel series convolutional neural network (SCS-CNN). Compared to traditional FBRCSP, the proposed method improved accuracy by 7.14% for the BCI4-2a data set. By using the BCI3a data set, the proposed method improved accuracy by 9.54% compared with traditional FBRCSP, the average accuracy of the proposed method is 83.70%, and the kappa is 0.7827.

Keywords Brain–computer interface · Motor imagery · Electroencephalogram · Convolutional neural network · Common spatial pattern

1 Introduction

Brain–computer interface has many applications in various fields. Motor imagery classification is one of the BCI applications. People hope to control the machine in an imaginary way only with brain imaging. In biomedical engineering, a brain-controlled wheelchair is one of the applications that can help the disabled use the brain's imagination to complete the activities of moving the wheelchair [1]. So far, many researchers have made certain progress in the research of imaginary movement. However, most of them mainly study the binary classification problem, especially the classification of left-hand and right-hand motor imagery [2]. Multi-classification is essential. Take the brain-controlled wheelchair as an example, the two-class method can only classify the forward and stop commands, so it is insufficient for practical applications [3]. There are two essential steps for the motor imagery multi-classification task: feature extraction and classification.

Feature extraction can obtain useful information from EEG signals. Common spatial pattern (CSP) is an excellent

✉ Sai Ho Ling
Steve.Ling@uts.edu.au

[1] College of Artificial Intelligence and Big Data for Medical Science, Shandong First Medical University & Shandong Academy of Medical Sciences, Jinan, Shandong 250117, China

[2] Department of Engineering, King's College London, London, UK

[3] School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, NSW, Sydney, Australia

method to extract EEG features in the space domain [4]. It extracts the two-class EEG signal with the most significant difference in information. CSP cannot be used for multi-classification problems directly because CSP is a binary feature extraction method. If CSP is used for multi-classification tasks, some strategies must be used. The existing common CSP strategies are the one-vs-one (OVO) strategy and one-vs-remain (OVR) strategy [5]. Besides, some novel strategies can also be used to extend CSP into a multi-class CSP method. For example, [5] proposed two new extension strategies, namely divide-and-conquer (DC) and pair-wise (PW) strategies. Compared to OVR and OVO strategies, the proposed techniques can retain more effective information on EEG features. In addition to the extension strategies, some advanced algorithms can also be used for multiple feature extraction tasks. [6] applied joint approximate diagonalization (JAD) algorithm to the CSP method, which can tackle the multi-classification tasks. The advantage of this method is that it can effectively against the effects of the artifacts. In addition, they also use self-regulated supervised Gaussian fuzzy adaptive system art (SRSG-FasArt) as the classifier, which decreases neuron proliferation and over-training probability. Additionally, [7] proposed local temporal common spatial patterns (LTCSP), which tackle multi-classification tasks by maximizing the harmonic mean of the KL divergences. Compared to the traditional CSP method, LTCSP incorporates the EEG information before calculating the spatial filter, which can better improve discriminant ability.

Although CSP can extract the EEG features from the spatial domain, EEG features also exist in the time–frequency domain. Some papers extracted features from the time and frequency domain and combined the spatial features generated by CSP as imaginary motion classification features. In [8], wavelet packet transform was used to divide EEG signals into equal frequency bands, and then, they found the sub-signals that were easy to distinguish on the frequency bands. CSP is used to continue to obtain the spatial features of the extracted frequency band signals. This method can adaptively select the effective frequency band according to different individuals. [9] proposed the sparse time–frequency segment common spatial pattern (STFSCSP) algorithm for feature extraction. The advantage of this algorithm is that they use sparse regions based on time–frequency characteristics to select important spatial features so that features have obvious spatial and time–frequency differentiation. [10] combined multivariate empirical mode decomposition (MEMD) with CSP and first decomposed the EEG signal using MEMD and then extracted the effective components from the sub-signal using CSP. They identified the subject's specific MIMFs based on the mean frequency. According to these MIMFs, they can select the specific frequency range of the subject, which contributes to the motor-related rhythms. This is helpful in extracting more effective frequency-domain features.

The performance of CSP relies on the estimation of sample covariance. If the number of sample data is small, the performance of the spatial filter may not be good. The small sample data may be prone to overfitting. [11] used regularized common spatial pattern (RCSP) to resolve this problem. RCSP is an improved method of CSP. It adds data from other subjects when calculating the covariance matrix. It introduces two regularization parameters. Two regularization parameters are used to control the weight of the covariance, thereby reducing the estimation error of the covariance matrix. It solves the problem of insufficient data information for small sample data. The original and other subjects' data are jointly used to calculate the spatial filter. [12] also used RCSP, in which the authors counted the results of using different regularization parameters and proved the effectiveness of RCSP. In addition, [13] combined FBCSP with RCSP. They used multiple frequency filters to get multiple sub-signals and then applied RCSP on each sub-signal. In this way, different frequency data of other subjects were added to each frequency band. [14] also used multiple band-pass filters to filter the signals and then applied RCSP on each band signal, making the final characteristics more diverse.

In addition to CSP, there are also some other spatial transform-related algorithms that can be used to extract EEG features. [15] summarized the existing EEG signal recognition methods. Some traditional feature extraction methods include CSP, independent component analysis (ICA) and principle component analysis (PCA), and common classifiers include support vector machine (SVM), linear discriminant analysis (LDA) and k-nearest neighbors (KNN). By comparing the experimental results, the best classification effect can be obtained by CSP and SVM. Although these signal decomposition-related algorithms are not as effective as CSP, they have the potential to reduce the size of the features. PCA is an extension method of the singular value decomposition (SVD) algorithm. They can decompose the main components of the target signal. Some papers proposed improved SVD algorithms to extract EEG features. [16] used linear prediction singular value decomposition (LP-SVD) to decompose EEG signals. Its purpose is to reduce the dimension of data. The author adjusted LP coefficients, error variance and transform coefficients and finally corrected the outputs using auto-regression (AR) model. In this paper, the authors compared the proposed method to discrete cosine transform (DCT), which is a widely used unsupervised signal independent linear feature extraction method. The proposed method can improve the classification accuracy by 25%. Similarly, [17] proposed a feature extraction

algorithm of linear prediction in conjunction with QR decomposition (LPQR), which is an improved method based on the LP-SVD algorithm. Although they both compress the common information of multi-channel EEG signals through matrix decomposition, the spatial features extracted by the LPQR algorithm are more informative, while LP-SVD can achieve better classification accuracy.

Manifold learning is a nonlinear dimensionality reduction technique. Compared to PCA or ICA algorithm, this method can better compress the complicated EEG signal. [18] tried to preserve the useful information of the EEG data by distance preservation to local means (DPLM). By using this method, some nonlinear EEG features can be converted into a new dimension, in which EEG features have better discrimination. However, this algorithm has two disadvantages. The first drawback is that the calculation speed may decrease when we use more samples. Another one is that some outliers may be caused if the data is too complicated. [19] proposed two manifold learning methods, namely minimum distance to sub-manifold mean (MDSM) and tangent space of sub-manifold (TSSM). The main idea of this paper was to treat the signal as a spatial figure using spatial geometric transformation. It expanded the spatial graphics and kept the distance between adjacent points unchanged. It can also be understood as keeping the signal information unchanged and compressing effective information. This method belongs to manifold learning, but the main idea is to obtain EEG features in the space domain.

A neural network is not only a suitable feature extraction method but also a classifier. [20] proposed a modular network to classify the brain signals. The modular network was composed of four expert CNNs, each expert CNN performs binary classification, and a fully connected network was used to integrate their outputs. Besides, they also used the Bayesian optimization algorithm to optimize training hyperparameters, which is helpful in avoiding the overfitting problem. [21] first used short-time Fourier transform (STFT) to convert EEG signals into two-dimensional images and then proposed a capsule network to classify time–frequency-domain feature maps. In the capital network, they introduced activity vectors that represent variant properties of the features, such as position, size and rotation. The activity vector can be regarded as a capsule, which replaces the pooling layer because some original spatial information of EEG signals may be lost when applying pooling layers. Thus, compared to the traditional network, the proposed method can reduce feature information loss.

Some papers proposed advanced network structures to resolve EEG classification tasks. [22] applied a convolutional neural network to EEG motor imagery recognition tasks. The author proposed a new shadow network and compared three different convolutional network structures, including the traditional network, shadow network and residual network. A shallow network has a larger kennel size when convolving the features. In addition, the shadow network involves the voluntary computation and filter bank common spatial pattern (FBCSP) algorithm in a single network, and thus, all steps can be optimized jointly. [23] used 3DCNN, which composed 22 channels into a two-dimensional image and then obtained features on each channel as the height of the three-dimensional input. The network learned the feature from 3 dimensions together. By applying this network, the kappa was improved by 0.073 compared to the FBCSP method. [24] compared three different convolution methods of convolution networks, in which the convolution strategy of channel-wise convolution with channel mixing (C2CM) is to learn more complex features from time and space domains. This method can increase the flexibility of the network but with the cost of increasing the number of parameters due to the introduction of a new computational layer. [25] proposed a convolutional recurrent attention model (CRAM) where a convolutional neural network was used to encode the EEG signals, and a recurrent attention mechanism was applied to explore the temporal dynamics of the EEG signals. This method effectively takes advantage of both CNN and recurrent neural network (RNN), which can better learn the EEG information along the time sequence.

There are also some other effective techniques to improve the feature extraction methods or the classifiers, which are used to classify the EEG motor imagery tasks. [26] proposed an attractor metagene bat algorithm SVM. Attractor metagene is an unsupervised learning method used to filter the features, and the bat algorithm was used to optimize the parameters in SVM. Eventually, they combined these two methods with SVM, which improved the kappa value by 0.14 compared to the traditional SVM. In [27], the current source density (CSD) method is used to preprocess the signal before using CSP to extract features. CSD is a Laplacian method that standardizes the EEG signal according to its energy distribution. This method can increase spatial resolution. Compared to the traditional common average reference (CAR) processing method, it is easier to extract more distinct spatial features by using the data processed by CSD. [28] proposed a multi-class F-score-based time–frequency selection method, which uses Fisher discriminant analysis (FDA) to select the effective frequency bands and time periods of EEG signals. By selecting the effective EEG information, this method can effectively improve the inter-subject robustness.

From the literature review mentioned above, there are three limitations we need to overcome.

- *Limitation 1* RCSP can be used to tackle the problem of inaccurate covariance matrix caused by using small sample data. Although many researchers have used the regularization parameters to control the ratio of the covariance matrix calculated by using other subjects' data, little attention has been paid to dynamic regularization parameters. The selection of the parameters significantly affects the performance of the spatial filters. However, selecting the appropriate regularization parameters is difficult to fuse with other subjects' covariance matrixes.
- *Limitation 2* Although we can select suitable parameters based on mutual information, there is no guarantee that all data used to calculate the mixed covariance matrix have motor imagery characteristics. If the selected data is invalid or the motor imagery features are not significant, the obtained spatial filter may still not be accurate. The quality of motor imagery data can also affect the performance of the spatial filter.
- *Limitation 3* It is difficult for traditional methods to classify motor imagery tasks. Little research has been devoted to multiple classification tasks. Although some methods can be used on multi-classification tasks, they cannot achieve good performance.

In this paper, three approaches are proposed to overcome these three limitations:

Approach 1 We propose an auto-selected filter bank regularized common spatial pattern (ACSP) algorithm, which can automatically select the regularization parameters. Two regularization parameters are used to control the proportion of additional data from other subjects and correct the error of the mixed covariance matrix. We use mutual information to evaluate the degree of the difference among the generated features by using multiple groups of regularization parameters. The mutual information estimation matrix can be used to determine whether the selected regularization parameters are appropriate to the mixed covariance matrix. The distribution of the filtered data selected by this method is close to the distribution of the target classification features. Thus, the spatial features obtained by the selected regularization parameters are more suitable for the network to do classification. In addition, it can automatically adjust the parameters based on the extracted features from different subjects for different motor imagery tasks. This method significantly improves the accuracy compared to traditional FBRCSP using the fixed regularization parameters. This method is introduced in the Methodology Section C Method 1, used to resolve limitation 1.

Approach 2 We propose a motor imagery data evaluation algorithm which can be used to check whether the data has the motor imagery characteristics. The activation channels can be found when performing the motor imagery tasks. Two indicators are proposed to evaluate the degree of energy changes in these activation channels' frequency and time–frequency domains. These indicators can also check the ERD or ERS in the motor-related frequency bands. Thus, this model can more comprehensively measure the quality of motor imagery data. In addition, the fuzzy model is proposed to integrate different levels of evaluation indicators and fuzzify the indicators in desired ranges. This model is more flexible because the fuzzy rules can be designed by experience. The threshold limits the data quality levels based on the experience or the subjects' mental states. This data evaluation model considers all the information from the spatial, frequency and time–frequency domains to check motor imagery data quality. The high-quality motor imagery data is selected to calculate further the RCSP mixed covariance matrix, which can improve the performance of spatial filters. This method is introduced in the Methodology Section D Method 2, which is used to resolve limitation 2.

Approach 3 We propose a single-channel serial convolutional neural network (SCS-CNN) within a voting strategy to resolve the multi-class motor imagery classification task using the BCI4-2a and BCI3a data sets. The classification accuracy using this network is better than using some traditional classifiers. The single-channel learning strategy used in the network can extract the EEG information from each independent channel. It is more suitable for spatially transformed data because the whitening matrix in the CSP filters removes the correlation among all the channels. In addition, this network can get a broader learning horizon in the learning process so that the learned information can express more data features. The network also contains the residual net structure, which is used to tackle the problem of gradient disappearance. Combining the network with a voting strategy further increases the diversity of the extracted features. Using more features makes the model learn more EEG information. Integrating the information from multiple features to classify makes the system more stable. This method is introduced in the Methodology Section D Method 3, which is used to resolve limitation 3.

2 Methodology

2.1 Overview of the system

The block diagram of the entire system is presented in Fig. 1. Firstly, the raw data is preprocessed, including re-reference and average removal. Then, the motor imagery data evaluation model evaluates other subjects' data, and the high-quality data is selected. After that, the

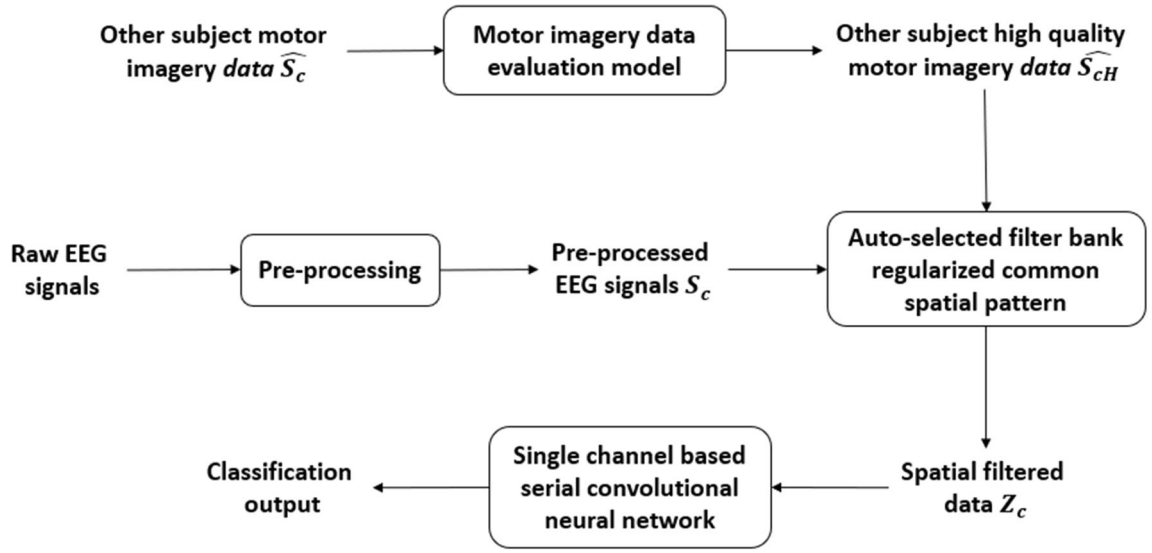

Fig. 1 Block diagram of the entire system

preprocessed EEG signal S_c and other subjects' high-quality data $\widehat{S_{cH}}$ are input into the auto-selected filter bank regularized common spatial pattern model to calculate the spatial filtered data Z_c which can be used as the final EEG features. Eventually, the extracted features are input into the single-channel-based serial convolutional neural network to perform classification. Important abbreviations and variable names are listed in Table 1.

2.2 Method 1: auto-selected filter bank regularized common spatial pattern

Auto-selected filter bank regularized common spatial pattern algorithm is proposed to automatically select the regularization parameters and construct spatial filters. The two regularization parameters selected by this algorithm can better fuse the motor imagery data from the target subject and other subjects to make the extracted spatial features more accurate and improve the classification performance. The block diagram of this method is shown in Fig. 2. Firstly, multiple band-pass filters are used to filter the preprocessed signal S_c. The filtered data and other subjects' EEG data $\widehat{S_c}$ within multiple regularization parameters β_c, γ_c are used to calculate RCSP filters. Then, the multiple groups of spatial filters are applied to the preprocessed signal S_c to obtain the feature matrix X_c. Use mutual information to evaluate these features, select suitable regularization parameters, and use the parameters to rebuild the final RCSP filter W_c, which can be applied to the preprocessed data to extract the EEG spatial filtered data Z_c.

2.2.1 Regularized common spatial pattern

Common spatial pattern (CSP) is an effective method for extracting the features of the EEG signal. Its main principle is to perform matrix decomposition on two EEG signals to extract effective components to maximize the variance difference between two signals. However, the covariance matrix calculated by the CSP under a small sample is not accurate. There is a way to resolve a small sample's inaccurate estimation problem, which is the regularization method. When estimating the covariance matrix and using the original data, the data from other subjects of the same task is also used. It indirectly increases the number of samples and improves the accuracy of the estimation. Considering S_c is the preprocessed target EEG data and $\widehat{S_c}$ is the EEG data of other subjects, we can calculate R_c that is the covariance of S_c and $\widehat{R_c}$ that is the covariance of $\widehat{S_c}$:

$$R_c = \sum_{n=1}^{N_t} \frac{S_{cn} S_{cn}^T}{\text{trace}\left(S_{cn} S_{cn}^T\right)} \tag{1}$$

$$\widehat{R_c} = \sum_{\hat{n}=1}^{\widehat{N_t}} \frac{\widehat{S_{c\hat{n}}}\,\widehat{S_{c\hat{n}}}^T}{\text{trace}\left(\widehat{S_{c\hat{n}}}\,\widehat{S_{c\hat{n}}}^T\right)} \tag{2}$$

where trace(S_c) is the sum of elements on the diagonal of the matrix S_c; N_t is the number of trials of S_c; and $\widehat{N_t}$ is the number of trials of $\widehat{S_c}$. Then we can obtain J_c which is the regularized covariance matrix and $\Sigma(\beta_c, \gamma_c)$ which is the mixed covariance matrix by using the regularization parameters β_c and γ_c. We have:

Table 1 Important abbreviations and variable names

Abbreviation	Full name
ACSP	Auto-selected filter bank regularized common spatial pattern
BCI	Brain–computer interface
BSML	Bilinear sub-manifold learning
C2-CNN	Channel mixing convolutional neural network
CAR	Common average reference
CNN	Conventional neural network
CR	Common reference
CSD	Current source density
CSP	Common spatial pattern
CWCNN	Channel-wise convolutional neural network
CWT	Continuous wavelet transform
D-ACSP	Data evaluation-based auto-selected filter bank regularized common spatial pattern
DFFN	Densely feature fusion deep learning network
EEG	Electroencephalogram
EMD	Empirical mode decomposition
ERD	Event-related desynchronization
ERS	Event-related synchronization
FBCSP	Filter bank common spatial pattern
FBRCSP	Filter bank regularized common spatial pattern
FD	Frequency distance
FFT	Fast Fourier transform
ICA	Independent component analysis
IMFs	Intrinsic mode functions
JAD	Joint approximate diagonalization
KNN	K-nearest neighbors
LDA	Linear discriminant analysis
LSTM	Long short-term memory
MDSM	Minimum distance to sub-manifold mean
MEMD	Modified empirical mode decomposition
MI	Motor imagery
OVO	One-vs-one
OVR	One-vs-remain
PCA	Principal component analysis
RCSP	Regularized common spatial pattern
SCS-CNN	Single-channel-based series convolutional neural network
SVM	Support vector machine
TFD	Time–frequency distance
TSSM	Tangent space of sub-manifold
WT	Wavelet transform

$$J_c(\beta_c) = \frac{(1-\beta_c)\cdot R_c + \beta_c \cdot \widehat{R_c}}{(1-\beta_c)\cdot N_t + \beta_c \cdot \widehat{N_t}} \tag{3}$$

$$\Sigma_c(\beta_c, \gamma_c) = (1-\gamma_c)\cdot J_c(\beta_c) + \frac{\gamma_c}{N_c}\text{trace}[J_c(\beta_c)]\cdot I \tag{4}$$

where N_c is the total channel numbers; β_c controls the variance of the estimated covariance; and γ_c is the second regularized parameter, which can reduce large eigenvalues and increase small eigenvalues. Then, decompose the mixed covariance matrix and obtain eigenvalue λ_c and eigenvector U_c. Sort eigenvalue U_c in descending order and obtain the whiting matrix P_w.

$$U_c \lambda_c U_c^T = \Sigma_c = \Sigma_{c1} + \Sigma_{c2} \tag{5}$$

$$P_w = \sqrt{\lambda_c^{-1}} U_c^T \tag{6}$$

where Σ_{c1} is the mixed covariance matrix of the first-class data and Σ_{c2} is the mixed covariance matrix of the second-

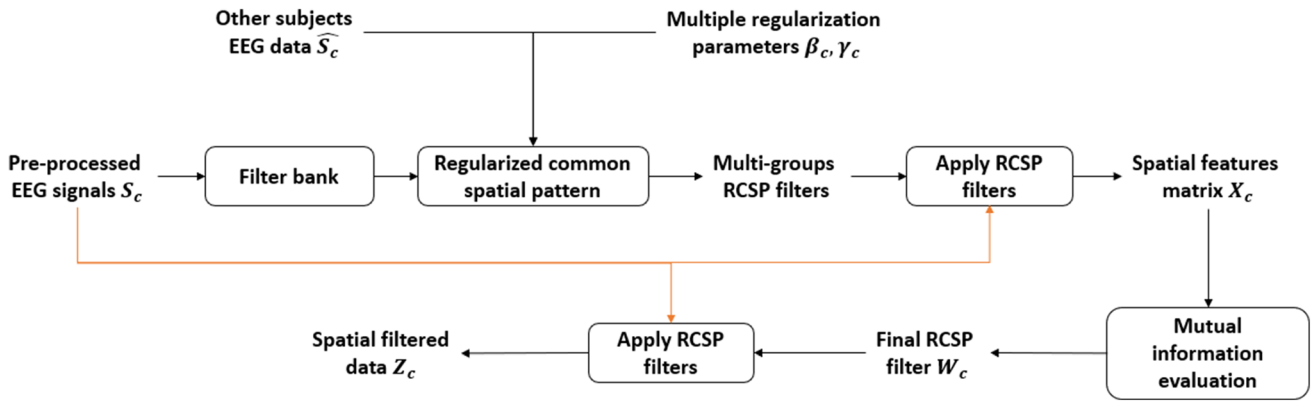

Fig. 2 Block diagram of the auto-selected filter bank regularized standard spatial pattern algorithm

class data. Apply P_w to the two classes mixed matrix to obtain the whiten matrix of the first-class data S_{w1} and the whiten matrix of the second-class data S_{w2}. After that, continue to decompose one of the class matrixes S_{w1} to obtain the eigenvalues λ_B and eigenvectors U_b.

$$S_{w1} = P_w \Sigma_{c1} P_w^T \tag{7}$$

$$S_{w2} = P_w \Sigma_{c2} P_w^T \tag{8}$$

$$U_b \lambda_B U_b^T = S_{w1} \tag{9}$$

Eventually, we can obtain the spatial filter W_c. We apply the filter to the preprocessed signal to obtain the feature matrix X_c.

$$W_c = U_b^T P_w \tag{10}$$

$$X_c = \mathrm{var}(W_c * S_c) \tag{11}$$

where var() is the function of calculating variance.

2.2.2 Auto-selected RCSP weight selection

After using the RCSP spatial filter, the variance feature can be obtained. The corresponding labels are defined for the two types of variance features. The feature vector is X_c, and the label vector is Y_c. Their information entropy $H_I(X_c)$ and $H_I(Y_c)$ can be calculated. Then, use their joint probability density function to calculate their mutual information $M_I(X_c, Y_c)$.

$$H_I(X_c) = -\sum_{x \in X_c} P(x) \log_2 P(x) \tag{12}$$

$$H_I(Y_c) = -\sum_{y \in Y_c} P(y) \log_2 P(y) \tag{13}$$

$$M_I(X_c, Y_c) = \frac{2 \sum_{y \in Y_c} \sum_{x \in X_c} P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right)}{H_I(X_c) + H_I(Y_c)} \tag{14}$$

where $p(x)$ is the probability of x; $p(y)$ is the probability of y; and $p(x, y)$ is the joint probability of x and y. We assume that the X_c and Y_c are relatively independent, then $p(x, y) = p(x)p(y)$. The final calculated mutual information $M_I(X_c, Y_c) = 0$. If the value of M_I is larger, the feature is closely related to the label. In this way, M_I can be used to evaluate whether the obtained EEG signal characteristics are suitable for distinguishing the two categories. The two parameters of RCSP can take values within a certain range. Then calculate the mutual information matrix of the feature labels of each set of RCSP parameters. In this matrix, the γ_c and β_c corresponding to the maximum value are used as the final RCSP parameters. Eventually, we apply the spatial filter corresponding to these two parameters to the preprocessed signal and obtain the spatial filtered data Z_c.

$$Z_c = W_c * S_c \tag{15}$$

2.3 Method 2: motor imagery data evaluation model

A motor imagery data evaluation algorithm is proposed to check whether the data has motor imagery characteristics. This method can determine how much the subject is doing motor imagery tasks by calculating the energy changes of the signal at a specific brain location in a specific frequency domain. Thus, this algorithm can select the high-quality motor imagery data, which can improve the accuracy of the mix covariance matrix when calculating the spatial filter. The block diagram of the data evaluation model is shown in Fig. 3. When EEG data is coming, it is re-referenced, filtered in motor-related frequency bands and processed by removing the mean value. Then the preprocessed data can be divided into baseline data and motor data. The variance of each data is calculated, and the energy distribution of both data can be obtained. Based on the distribution difference, it can obtain effective channel data. The selected channel data is analyzed in the frequency and time–frequency domains using continuous wavelet transform (CWT) and Welch method. Different motor-related

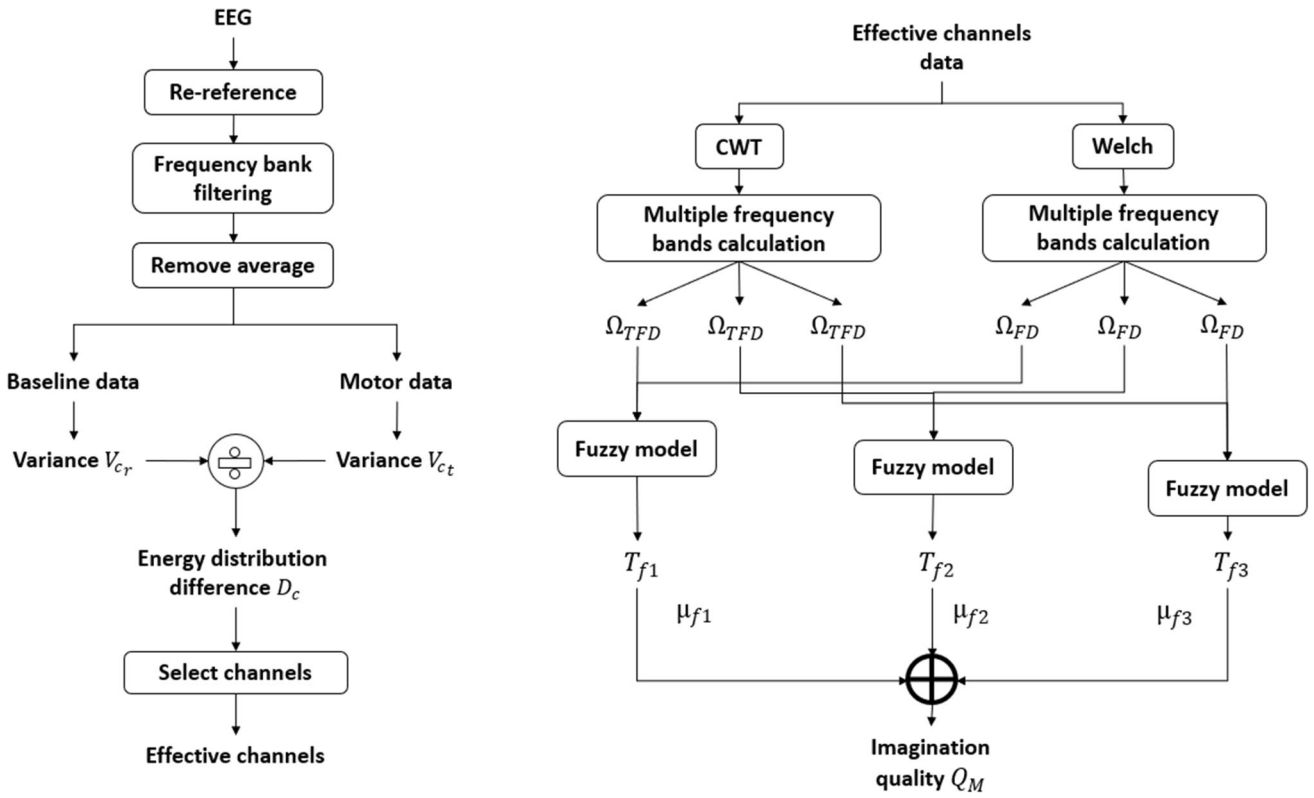

Fig. 3 Block diagram of the data evaluation model

frequency bands can get various groups of time–frequency distance (TFD) and frequency distance (FD). Each group of motor imagery feature evaluation indicators is integrated by fuzzy models. The outputs of all fuzzy models are combined by the weights to obtain the final imagination level. This final indicator can be used to evaluate the motor imagery data quality.

2.3.1 Energy distribution calculation

Re-reference and filter the original data. Then, calculate the variance V_c of each channel of motor state data and the rest state data.

$$V_c = \frac{1}{N_p}\sum_{i=1}^{N_p}(S_{ci} - \overline{S_c})^2 \tag{16}$$

where S_{ci} is the i_{th} sample point; $\overline{S_c}$ is the mean value; and N_p is the total number of sample points. Using Eq. (16), we can obtain the variance of motor state data V_{ct} and the variance of rest state data V_{cr}. We can regard the rest state variance as the baseline variance. Calculate the change in energy distribution D_c compared to the baseline state when the subject is doing an imaginary task:

$$D_c = \begin{cases} -\dfrac{V_{cr}}{V_{ct}}, & V_{ct} < V_{cr} \\ \dfrac{V_{ct}}{V_{cr}}, & V_{ct} \geq V_{cr} \end{cases} \tag{17}$$

The obtained variance of all channels can describe the distribution of brain energy changes in a specific frequency range. Generally, for left- and right-hand motor imagery tasks, the EEG signal energy will decrease in the motor-related frequency bands and event-related desynchronization. However, for the foot or tongue motor imagery task, the energy may not decrease. Sometimes, it could even increase. Also, the area of spatial activation is slightly different, and the state of each experiment can also affect the change of energy amplitude. Therefore, the brain area energy changes obtained from all experiments should be superimposed and averaged to reduce the error. The superimposed and averaged energy change E_c can be calculated:

$$E_c = \frac{1}{N_t}\sum_{i=1}^{N_t} D_{ci} \tag{18}$$

where D_{ci} is the brain energy change of the i_{th} trial and N_t is the total number of trials. E_c can roughly present the

relevant activation area of the specific motor imagery task. Select the top N_{cs} channels with the largest absolute value of E_c, and perform frequency-domain analysis and time–frequency analysis on the selected channels, respectively.

2.3.2 Frequency-domain analysis

In frequency-domain analysis, we use Welch method. Firstly, apply the window function to the EEG signals and taper the signals. Then, apply fast Fourier transform (FFT) to each window signal and calculate the average of squared absolute values of FFT outputs. Finally, integrate all the outputs of different frequency ranges. The frequency spectrum O_c can be obtained:

$$O_c = \frac{1}{N_w}\sum_{k=1}^{N_w}|\mathrm{FFT}_k(f)|^2 \tag{19}$$

where N_w is the number of segmented signals and $\mathrm{FFT}(f)$ is the FFT output on frequency f. Apply the Welch method to the motor state signal and the rest state signal to obtain the frequency spectrum O_{ct} and O_{cr} of the two signals. Eventually, the frequency spectrums got from each window are normalized and averaged. The normalized frequency spectrums of motor state signal O_{ctn} and the normalized frequency spectrums of rest state signal O_{crn} can be calculated:

$$O_{ctn} = \frac{O_{ct} - \mathrm{Min}(O_{ct}, O_{cr})}{\mathrm{Max}(O_{ct}, O_{cr}) - \mathrm{Min}(O_{ct}, O_{cr})} \tag{20}$$

$$O_{crn} = \frac{O_{cr} - \mathrm{Min}(O_{ct}, O_{cr})}{\mathrm{Max}(O_{ct}, O_{cr}) - \mathrm{Min}(O_{ct}, O_{cr})} \tag{21}$$

where N_p is the total number of the sample; frequency spectrums are used to calculate the frequency distance Ω_{FD} which is used to describe the distance of spectrums in the frequency domain.

$$\Omega_{\mathrm{FD}} = \frac{O_{crn} - O_{ctn}}{N_p} \tag{22}$$

2.3.3 Time–frequency-domain analysis

The imaginary movement may be intermittent for a while. Thus, it is not only related to frequency, but it is also related to time. Therefore, we also use continuous wavelet transform (CWT) to analyze the time–frequency features of EEG signals. The idea of CWT is to use a wavelet to apply convolution calculation to the original signal. By stretching and transforming the wavelet, we can obtain time and frequency information with different precision from the original data. Morlet wavelet is suitable for EEG analysis. The reason of using this wavelet is that Morlet wavelet is non-orthogonal, so we can obtain continuous wavelet amplitudes when analyzing EEG signals. Moreover, the Morlet wavelet is more similar to an EEG signal, which is helpful for signal compression. In addition, in order to obtain the amplitude and phase information of the time series, it is necessary to select the complex wavelet because the complex wavelet has an imaginary part, which can express the phase well. Morlet wavelet is not only non-orthogonal but also complex exponential wavelet regulated by Gaussian. Morlet wavelet has a good balance between time and frequency information. Therefore, the Morlet wavelet is better for time–frequency analysis of EEG signals. First, calculate the width of Gauss window s_w and the amplitude A_w:

$$s_w = \frac{l_w}{2\pi f} \tag{23}$$

$$A_w = \frac{1}{(s_w\sqrt{\pi})^{\frac{1}{2}}} \tag{24}$$

where l_w is the wavelet cycle and f is frequency. Then combine the Gauss window and complex trigonometric function to get the complex Morlet wavelet G_w:

$$G_w = A_w e^{\frac{-t^2}{2s_w^2}} e^{i2\pi ft} \tag{25}$$

where t is signal time. The obtained Morlet wavelet and the original signal S_c are convolved to obtain the time–frequency energy spectrum M_c:

$$M_c = \mathrm{Convolute}(G_w, S_c) \tag{26}$$

Similar to the frequency-domain analysis, we not only perform CWT on the motor state signal but also the rest state signal. We want to compare the time–frequency energy change when performing motor imagery tasks. Therefore, CWT is performed on the rest state signal and the motor state signal to obtain the rest state time–frequency energy spectrum M_{cr} and the motor state time–frequency energy spectrum M_{ct}. After that, the time–frequency energy spectrum of a specific frequency band is normalized. The normalized time–frequency spectrum of motor state signal M_{ctn} and the normalized time–frequency spectrum of rest state signal M_{crn} are calculated:

$$M_{ctn} = \frac{M_{ct} - \mathrm{Min}(M_{ct}, M_{cr})}{\mathrm{Max}(M_{ct}, M_{cr}) - \mathrm{Min}(M_{ct}, M_{cr})} \tag{27}$$

$$M_{crn} = \frac{M_{cr} - \mathrm{Min}(M_{ct}, M_{cr})}{\mathrm{Max}(M_{ct}, M_{cr}) - \mathrm{Min}(M_{ct}, M_{cr})} \tag{28}$$

The obtained spectrum energy range is between [0,1]. Segment the energy by multiple frequency ranges. The energy of each frequency band at each time point is counted. Then the energy distribution P_N of the target signal is obtained. Compare the time–frequency energy

distributions of the rest state signal and the motor state signal. If the two distributions are almost overlapping or similar, the subject does not perform imagery tasks in this segment of the signal. If the two distributions are different, the subject may be doing imagery tasks. The way to judge the similarity of two distributions is to calculate the center of gravity of the distribution K_c:

$$K_c = \frac{\sum_{i=1}^{N_p} P_{Ni}}{N_p} \tag{29}$$

where N_p is the total number of the points and P_{Ni} is the i_{th} counting number of the distribution P_N. Then the difference between the center of gravity is used to measure the distance of the energy distribution of the two signals in time-frequency domain. The larger the difference, the farther the distance is, indicating that the energy of the motor state signal has a significant change compared to the baseline. Positive or negative represents a decrease or increase in energy. Ω_{TFD} is the time–frequency distance used to describe the nonlinear distance of spectrums in the time–frequency domain.

$$\Omega_{\mathrm{TFD}} = K_{cr} - K_{ct} \tag{30}$$

where K_{cr} is the center of gravity of the rest state time–frequency energy distribution and K_{ct} is the center of gravity of the motor state time–frequency energy distribution.

2.3.4 Calculate data quality level

These two evaluation indicators can determine the degree of the motor imagery signal energy changes in different rhythms. These indicators should jointly determine whether the data is a high-quality motor imagery signal. Therefore, after getting the evaluation indicators of the selected channels, a fuzzy logic model should be built to integrate these indicators and output a final evaluation indicator. The fuzzy logic is used because it has great freedom in designing the fuzzy model. Thus, we can find the most suitable fuzzy logic model according to the actual situation. The fuzzy model is shown in Fig. 4.

There are two inputs, namely time–frequency distance and frequency distance. According to experience, seven degrees of energy changes are set, which are negative big (NB), negative medium (NM), negative small (NS), zero (ZE), positive small (PS), positive medium (PM) and positive big (PB). The time–frequency distance and frequency distance in the same channel are fuzzified. Thus, 49 rules should be developed. Finally, the output is de-fuzzified following the rules, and the de-fuzzification method calculates the centroid of the area under the fuzzy output set. The output variable T_f is used to describe the energy

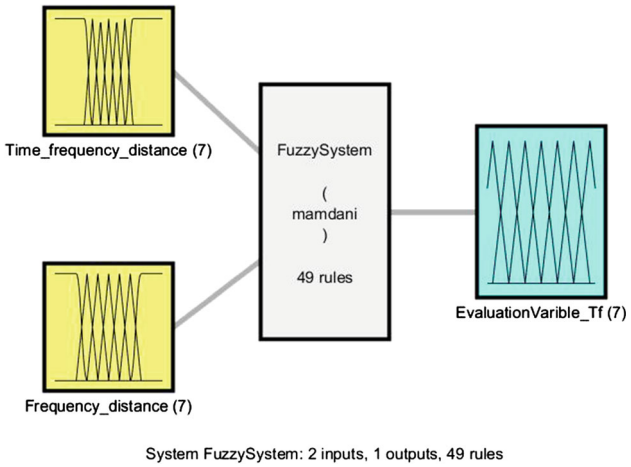

Fig. 4 Structure of the fuzzy model with 2 inputs and 1 output

changes of the selected channel under a specific frequency range. The sign of the variable obtained can be defined by the D_c obtained in (31). Negative indicates an increase, and positive indicates a decrease.

$$T_f = \begin{cases} T_f, & D_c < 0 \\ -T_f, & D_c \geq 0 \end{cases} \tag{31}$$

It is to calculate the evaluation indicator for one channel. If there are multiple channels, the confidence of each channel should gradually decrease. Thus, we set a decreasing variable μ_f. The final quality evaluation level Q_M of the motor imagery signal can be obtained:

$$Q_M = \frac{1}{N_{cs}} \sum_{i=1}^{N_{cs}} \mu_f^{(i-1)} T_f \tag{32}$$

where N_{cs} is the number of selected channels. The energy change of each frequency band should be independent. We set a threshold ξ_d. If the Q_M of at least one frequency band is higher than the threshold, then the signal can be seen as that it has motor imagery characteristics. This data can be regarded as high-quality motor imagery data. When the other subjects' EEG data $\widehat{S_c}$ is input into this model, the high-quality motor imagery data $\widehat{S_{cH}}$ can be selected.

2.4 Method 3: single-channel-based serial convolutional neural network within a voting strategy

We propose the single-channel-based serial convolutional neural network (SCS-CNN) as the classifier to classify the motor imagery tasks. It only extracts the features between a fixed size of points each time. The serial network structure contains a residential network. It uses a single-channel-based learning strategy. A single-channel learning strategy means the CNN does not learn the relation between every

two channels. Each time, it will generate feature maps or do a pooling operation based on only one channel. The size of the first dimension of the image should also be unchanged. Compared to the traditional network structure, the proposed network can reduce the convolutional feature complexity, which may avoid overfitting issues. In addition, this network is specially designed for the classification features extracted by CSP-related method. Because whitening is one of the steps in CSP filtering, the information of each channel should have no relation. If we use a traditional network, the information from different channels will be mixed up. Thus, this proposed network can extract useful information from each independent channel better.

In order to achieve this, the convolution size and stride will be set to [1, N_o] and [1, S_o], where N_o is feature size and S_o is stride. The pooling size and stride will also be set in the same way. This feature extraction method is suitable for multi-dimensional data. EEG signals are multi-dimensional signals related to time, frequency and channel position in the brain. Thus, this kind of network is more suitable for processing EEG data. It first learns the information of different brain locations separately, which is equivalent to compressing the information of different locations, and finally recognizes and classifies the hybrid features of each brain location. The structure of SCS-CNN

Table 2 Network parameters

Layer	Filter size	Stride	Filter number
A1 convolution	[1, 10]	[1]	32
A1 max pooling	[1, 4]	[1, 4]	–
B1 convolution	[1, 8]	[1]	32
B1 max pooling	[1, 3]	[1, 3]	–
B2 convolution	[1]	[1]	64
B3 convolution	[1]	[1, 3]	64
C1 convolution	[1, 5]	[1]	64
C1 max pooling	[1, 2]	[1, 2]	–
C2 convolution	[1]	[1]	128
C3 convolution	[1]	[1, 2]	128
D1 convolution	[1, 3]	[1]	128
D1 max pooling	[1, 2]	[1, 2]	–
D2 convolution	[1]	[1]	256
D3 convolution	[1]	[1, 2]	256
E1 convolution	[1, 2]	[1]	256
E1 max pooling	[1, 2]	[1, 2]	–
E2 convolution	[1]	[1]	512
E3 convolution	[1]	[1, 2]	512
F1 convolution	[1, 2]	[1]	1024
G1 average pooling	[1, 3]	[1, 3]	–

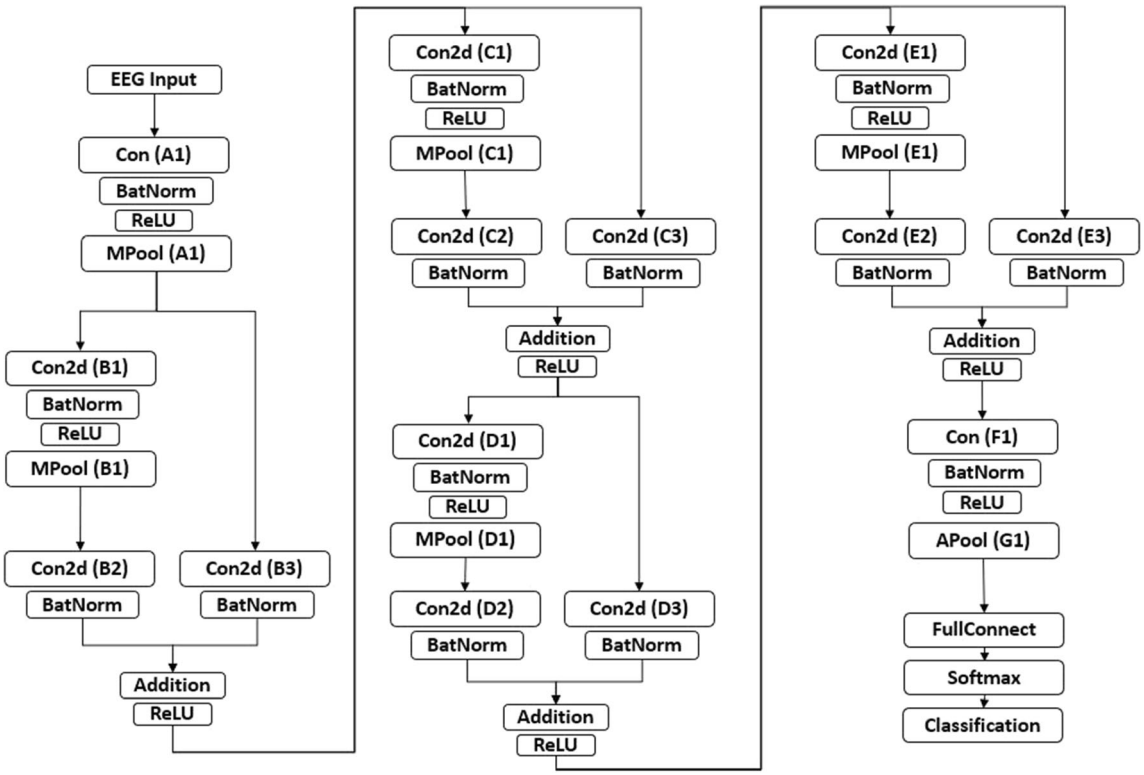

Fig. 5 Structure of single-channel-based serial convolutional neural network (SCS-CNN)

is shown in Fig. 5. The suggesting parameters are shown in Table 2.

Before classification, multiple groups of features using other subjects' data and regularization parameters can be obtained. We use the same structure network to classify each group of features. The loss function of the network is cross-entropy, and output should go through the Softmax layer. Thus, the network's output should be the categories within the highest possibility obtained by the Softmax layer. The possibility can be regarded as the confidence of this category. When we apply the voting strategy [29], we get the sum of the confidence of each category. The category within the highest sum of confidence should be the final output.

We use the soft voting method. The way is to calculate the average probability of each category obtained by all classifiers and finally select the category corresponding to the maximum probability as the output. The result of hard voting is ultimately determined by the model with a relatively low probability value, while soft voting is determined by the model with a high probability value. The soft voting method considers the additional information of prediction probability, which gives more weight to those models with high probability. Thus, its performance is better than that of hard voting. Because only CNN is used as the classifier, using multiple same type of models to soft vote the results can reduce the variance of the integrated model, thus improving the robustness and generalization ability of the model.

3 Experiments and results

3.1 Data description

There are two data sets used for EEG motor imagery classification. The first data set is BCI Competition 4-2a (BCI4-2a) [30]. This data set has 9 subjects, 22 channels and a total of 4 imaginary movement commands, which are left-hand imaginary movement, right-hand imaginary movement, foot imaginary movement and tongue imaginary movement. The data is filtered from 0.5 to 100 Hz. Each volunteer performs imaginary activities following the screen prompts. The experiment lasted for a total of 8 s. In the first two seconds of the experiment, the computer only prompts the subject to start the experiment. The screen showed specific movement prompts for 1.25 s from the third second. The participant has 2.75 s to perform the imaginary movement. The imaginary movement ends in the 6th second. There is a 1.5 s rest period. Finally, 72 trials of data are collected for each imaginary movement instruction using the same way, so 288 trials of data are collected for each subject.

The second data set is BCI Competition 3a (BCI3a) [31]. This data set has three subjects, 60 channels and a total of 4 imaginary movement commands, which are the same as the BCI competition 4-2a data set. However, the trial number for each subject is different. There are 45 trials for each imaginary movement and 180 trials for subject k3b. There are 30 trials for each imaginary movement and 120 trials for subjects k6b and l1b. The strategy of data collection is the same as the data set BCI4-2a. The specific information of these two data sets is shown in Table 3.

3.2 Experiments

3.2.1 Evaluation methods

In the experiments, we use tenfold cross-validation accuracy as the evaluation method. In addition, we also use the kappa to evaluate the performance of the proposed method. Kappa is a statistic that is used to measure inter-rater reliability for categorical items [32]. The calculating method of kappa κ and the standard error of kappa se is shown in (34) and (35).

$$p_e = \frac{\sum_i n_{ki} * n_{ik}}{N_g * N_g} \tag{33}$$

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{34}$$

$$\mathrm{se}(\kappa) = \frac{\sqrt{p_0 - p_e^2 - \sum_i [n_{ki} * n_{ik} * (n_{ki} + n_{ik})]/N_p^3}}{(1 - p_e)\sqrt{N_p}} \tag{35}$$

where n_{ki} and n_{ik} are the sums of each column and each row; p_e is a probability of agreement by chance; p_0 is the relative observed agreement; and N_p is the total number of samples.

Table 3 Description of two public data sets

Data set name	BCI4-2a	BCI3a
Trial number (total)	288 (each subject)	180/120/120
Trial number (each)	72	45/30/30
Class number	4	4
Subject number	9	3
Trial used this paper (total)	280	180/120/120
Trial used this paper (each class)	70	45/30/30

3.2.2 Experiment A: compare SCS-CNN with other classifiers

This experiment uses traditional CSP and FBCSP as feature extraction methods. We use SCS-CNN, long short-term memory network (LSTM) [33], LSTM-CNN [34], SVM [35] and random forest (RF) [36] as classifiers to classify the extracted EEG features and then compare the classification results. LSTM can extract the features along the time sequence. LSTM-CNN compresses the information of each window and extracts the features along the time sequence from multiple windows. The BCI4-2a and BCI3a data sets are used for testing. Before the experiment, the data is re-referenced and processed by removing the average. In terms of frequency filtering, for CSP, the filter band is 4–28 Hz. For FBCSP, we perform a 1–40 Hz multiple band-pass filter on the data. The filter bands are 1–4 Hz, 4–8 Hz, 8–12 Hz, 12–16 Hz, 16–20 Hz, 20–24 Hz, 24–28 Hz, 28–32 Hz, 32–36 Hz, and 36–40 Hz. All the experiments adopt tenfold cross-validation accuracy and kappa as evaluation measures.

We first divide the data into the training set and testing set. CSP is used to solve binary classification problems, but our purpose is to classify four motor imagery tasks. Thus, a multi-classification strategy is adopted. We use the one-vs-remain (OVR) strategy in this experiment. The method uses one category of data as the first class and the remaining data as the second class. The multi-classification problem is transformed into a binary classification problem.

We first calculate the CSP spatial filter. We obtain the four spatial filters because there are four motor imagery classes. Then we apply the filters to the target data to obtain the spatially filtered data. We calculate the variance of the filtered data as the CSP feature, which is input into SVM classifier. Four-channel filtered data with the top 4 largest variances difference is selected, and the data of each channel are stacked as the features to input into SCS-CNN, LSTM and LSTM-CNN. For FBCSP, we first use multiple frequency filters to filter the data from various frequency bands. Then, we calculate the CSP spatial matrix using each sub-band data. Other steps are the same as CSP.

We adopt Adam optimization algorithm as the training method. The classification results are shown in Tables 4 and 5.

From the results, using the same classifiers, the accuracy obtained by FBCSP is better than CSP. When we use CSP or FBCSP as feature extraction methods, using SCS-CNN can achieve better accuracy. The accuracy of most subjects and the mean accuracy obtained by SCS-CNN are the best. LSTM and LSTM-CNN have similar performances. Using SVM as a classifier can improve accuracy than LSTM and LSTM-CNN for the BCI3a data set but worse than that for the BCI4-2a data set. Using FBCSP as feature extraction and SCS-CNN as a classifier is more suitable for motor imagery multi-classification tasks.

3.2.3 Experiment B: auto-selected regularized FBCSP

Using FBCSP as a feature extraction method and SCS-CNN as a classifier can get the best classification results from experiment A. In this experiment, we first apply the regularization to FBCSP, which updates the method to regularized FBCSP. The main idea of regularization is to add extra data from other subjects when constructing the covariance matrix. There are two regularization parameters β_c and γ_c, which control the proportion of additional data of other subjects and the regularization ratio. In this experiment, we randomly selected the same amount of data as the training data from other subjects as the additional data. γ_c is set to 0.1, β_c is set to 0.2.

Then, we introduce the auto-selected method to adjust the two parameters in this experiment. First, filter the data in defined frequency bands. Before building the CSP matrix, calculate other subjects' data covariance and use Eq. (4) to get the mixed covariance matrix. In this step, we use multiple groups of β_c and γ_c; the range of the two parameters is between 0 and 1. γ_c takes [0,0.1,0.2,0.3], and β_c takes [0,0.1,0.2,0.3,0.4]. The reason for using these values is that the large β_c represents that most of the data used in the covariance matrix calculation are other subjects' data, which may cause much original data information loss. Thus, if the β_c exceeds 0.5, the structure of generated spatial filters could be affected significantly and the classification accuracy cannot be improved or even decreased. Therefore, we only use five β_c values, which are all less than 0.5. γ_c is used to correct the covariance error caused by the small value of the mixed matrix. The covariance matrix could also lose useful information if the correction ratio is too large. Thus, we only use four γ_c values less than 0.4.

Then, we use β_c and γ_c to calculate the CSP spatial filters. There are a total of 20 sets of spatial filters which are obtained in each frequency band. The 20 sets of filters are applied to the original data to obtain the spatially filtered data. The variance of the filtered data is calculated as the features. Use the variance features to calculate the mutual information following Eq. (14). The larger mutual information means the β_c and γ_c of this set of RCSP filters are better. Therefore, use this group of β_c and γ_c as the final regularization parameters. The testing results are shown in Tables 6 and 7.

Table 4 Accuracy (%) for BCI4-2a with different classifiers

	CSP + SVM	CSP + LSTM	CSP + LSTM-CNN	FBCSP + SVM	FBCSP + LSTM	FBCSP + LSTM-CNN	CSP + SCS-CNN	FBCSP + SCS-CNN
Sub 1	75.69	72.22	73.61	78.13	75.00	73.96	84.38	**85.42**
Sub 2	51.39	65.28	65.97	53.13	67.71	65.97	69.10	**70.14**
Sub 3	75.69	82.29	82.64	78.13	85.42	85.07	**93.06**	91.67
Sub 4	40.97	58.33	52.43	42.71	55.21	54.17	60.76	**61.46**
Sub 5	37.85	38.19	39.58	39.24	39.24	38.19	47.57	**55.21**
Sub 6	48.26	53.13	53.82	50.00	**55.56**	54.86	50.35	48.96
Sub 7	86.11	86.46	87.15	89.24	91.32	90.97	87.85	**93.06**
Sub 8	75.69	82.29	81.25	78.13	85.42	83.33	**90.28**	89.24
Sub 9	68.40	61.81	63.19	70.83	64.24	63.19	**87.50**	81.25
Mean ± std	62.23 ± 17.72	66.67 ± 15.81	66.63 ± 16.01	64.39 ± 18.24	68.94 ± 17.15	67.75 ± 17.21	74.54 ± 17.91	**75.15 ± 16.70**

Bold values represent the best ones

The data of some subjects with high classification accuracy may be significantly affected when mixed with the data from other subjects. When we use fixed regularization parameters, this may change the structure of spatial filters, which is not suitable for target classification data. Compared to regularized FBCSP with fixed regularization parameters, auto-selected regularized FBCSP can significantly improve the performance, especially for subjects 5 and 8 of BCI4-2a data set and subject l1b of BCI3-a data set. In addition, compared to FBCSP in experiment A, auto-selected regularized FBCSP can also achieve better performance for most subjects. Using suitable regularization parameters can generate better CSP filters and obtain more distinct features.

3.2.4 Experiment C: data evaluation-based auto-selected regularized FBCSP

The purpose of RCSP is to use the other subjects' data to resolve the problem of inaccurate covariance matrix due to the small amount of sample. However, if the other subjects' data quality is bad, it may also cause an inaccurate covariance matrix. Therefore, to ensure that the additional data is valid, we evaluate the data quality before adding other subjects' data.

First, perform 8–13 Hz and 16–25 Hz filtering on the data of other subjects because these two frequency bands are the two most relevant frequency bands for imaginary motion. Then we calculate the variance of each channel. We take the subject's 2 s data before the motor imagery cue as the baseline data and take the 4 s data after the cue as the motor data. Calculate the energy change distribution by using Eq. (17).

Take the N_{cs} channel data with the most significant energy change, perform CWT and Welch, and then, calculate frequency and time–frequency distances according to Eqs. (22) and (30). Then, build a fuzzy model to integrate the two indicators and calculate Q_M by using the Eq. (32). According to statistical experience, it can be concluded that Ω_{TFD} in $[-0.03,0.03]$ indicates that the time–frequency characteristics of the motor state signal are almost unchanged compared to the rest state signal. Ω_{TFD} in [0.03,0.09] indicates that the time–frequency energy of the motor state signal is reduced in different degrees compared to the rest state signal. Ω_{TFD} is higher than 0.09, indicating that the energy decreases significantly. Ω_{FD} in $[-0.2,0.2]$ indicates that the frequency-domain characteristics of the motor state signal are almost unchanged compared to the rest state signal. Ω_{FD} in [0.2, 0.6] indicates that the frequency-domain energy of the motor state signal is reduced to different degrees compared to the rest state signal. Ω_{FD} is higher than 0.6, indicating that the energy decreases significantly. We design input, output and fuzzy

Table 5 Accuracy (%) for BCI3a with different classifiers

	CSP + SVM	CSP + LSTM	CSP + LSTM-CNN	FBCSP + SVM	FBCSP + LSTM	FBCSP + LSTM-CNN	CSP + SCS-CNN	FBCSP + SCS-CNN
K3b	82.78	83.33	83.89	84.44	85.00	83.33	86.67	**87.78**
K6b	56.67	49.17	51.67	68.33	56.67	57.50	64.17	**70.00**
L1b	55.83	51.67	55.83	57.50	60.00	60.83	76.67	**80.83**
Mean ± std	65.09 ± 15.32	61.39 ± 19.04	63.80 ± 17.53	70.09 ± 13.56	67.22 ± 15.49	67.22 ± 14.05	**75.83 ± 11.27**	**79.54 ± 8.96**

Bold values represent the best ones

rules based on the above information. The two inputs are shown in Figs. 6 and 7. The output of the fuzzy model is shown in Fig. 8. The fuzzy rules are shown in Table 8.

In this experiment, C is set to 3, μ is set to 0.9. Finally, select the high-quality motor imagery data as the additional data to initialize the RCSP covariance. In this experiment, ξ is set to 0.1. The results of using DE-ARFBCSP are shown in Tables 9 and 10.

Compared to auto-selected regularized FBCSP in experiment B, the classification performance is further improved for both data sets. It proves that using high-quality data to initialize the mix covariance matrix can make the generated spatial filters more effective. Moreover, it also proves that using the proposed data evaluation method can select high-quality data with significant motor imagery features.

3.2.5 Experiment D: data evaluation-based auto-selected regularized FBCSP voting

Since the additional data used come from different subjects, the brain states of these subjects are also different. Although the validity of the data can be guaranteed, the CSP filter matrix generated each time is slightly different. The obtained EEG features are also different. In order to make the features more diversified, the voting strategy is adopted. First, randomly select different data from other subjects each time to generate multiple sets of CSP filter matrices. Then, use different sets of spatial filters each time to obtain multiple groups of filtered features. Finally, multiple SCS-CNNs with the same structure are used to classify each group of features. The multiple classified outputs are voted to get the final classification output. In this experiment, we generate five groups of spatial features. The results are shown in Tables 11 and 12.

From the results, the voting strategy can slightly improve the performance for the BCI4-2a data set and significantly improve the performance for BCI3a. The possible reason is that the number of trials of BCI3a is less than BCI4-2a. Thus, the voting strategy can increase the feature diversity for the data with a small sample size, such as BCI3a. For the BCI4-2a data set, after using the voting strategy, the performance of most subjects is improved slightly or unchanged except for subject 6. Thus, it appears that the testing performance of the model is already stable. Although the feature diversity is increased, the distribution of overall features is unchanged. Thus, the voting strategy is only used as a tip to improve the performance slightly.

Table 6 Comparison of the performance for BCI4-2a using regularized FBCSP with SCS-CNN and auto-selected regularized FBCSP with SCS-CNN

	Regularized FBCSP with SCS-CNN			Auto-selected regularized FBCSP with SCS-CNN		
	Accuracy (%)	Kappa	Se	Accuracy (%)	Kappa	Se
Sub 1	85.07	0.8009	0.0640	87.85	0.8380	0.0653
Sub 2	64.24	0.5231	0.0528	71.88	0.6250	0.0572
Sub 3	88.89	0.8519	0.0658	93.75	0.9167	0.0680
Sub 4	56.60	0.4213	0.0483	58.68	0.4491	0.0496
Sub 5	47.92	0.3056	0.0424	61.11	0.4815	0.0511
Sub 6	48.26	0.3102	0.0427	52.08	0.3611	0.0453
Sub 7	88.89	0.8519	0.0658	93.75	0.9167	0.0680
Sub 8	81.94	0.7593	0.0625	89.24	0.8565	0.0660
Sub 9	82.99	0.7731	0.0630	87.50	0.8333	0.0651
Mean ± std	71.64 ± 17.32	0.6219 ± 0.23	0.0564 ± 0.0099	77.31 ± 16.48	0.6975 ± 0.22	0.0595 ± 0.0089

Table 7 Comparison of the performance for BCI3a using regularized FBCSP with SCS-CNN and auto-selected regularized FBCSP with SCS-CNN

	Regularized FBCSP with SCS-CNN			Auto-selected regularized FBCSP with SCS-CNN		
	Accuracy (%)	Kappa	Se	Accuracy (%)	Kappa	Se
k3b	88.33	0.8444	0.0829	93.33	0.9111	0.0858
k6b	58.33	0.4444	0.0765	65.83	0.5444	0.0835
l1b	70.00	0.6000	0.0871	82.50	0.7667	0.0971
Mean ± std	72.22 ± 15.12	0.6296 ± 0.20	0.0822 ± 0.0053	80.56 ± 13.85	0.7407 ± 0.18	0.0888 ± 0.0073

4 Discussion

4.1 Auto-selected regularized FBCSP

4.1.1 The regularization parameters effects on mixed spatial filters

From the results of experiment A, using the same classifier, the classification accuracy of FBCSP is slightly higher than that of CSP. This could be because FBCSP extracts various spatial features of EEG signals in different frequency bands, while CSP only extracts a few features at a fixed frequency range. Thus, the feature diversity of FBCSP is much higher than that of CSP. However, from experiment B's results, it seems that applying regularization to FBCSP decreases the accuracy. Regularization is to use the other subjects' data and target data to calculate the covariance matrix jointly. It is possible that the selection of the regularization parameters β_c and γ_c are not suitable for the target subject's data. β_c is used to control the proportion of other subjects' data added to the calculation of the mixed covariance matrix. If β_c is large, it means that the entire model trusts the data of other subjects more. However, the states and imagination abilities of different subjects are different. Therefore, if we only use other subjects' data to calculate the covariance matrix, the spatial information distribution of other subjects may be different from the target subject, which may cause the generated spatial filter may not be suitable for the target subject's data. Eventually, it could lead to a decrease in classification accuracy.

In addition, γ_c is used to control the covariance error caused by the small value of the covariance matrix. This parameter does not provide much information on the original data covariance or newly added data covariance. It is only used to correct the deviation when adding other subjects' data. Therefore, if this parameter is significant, the model correction is strong. The correction information may be increased, but lots of information from the original covariance matrix could be lost. Therefore, the selection of these two parameters primarily affects the performance of

the final spatial filter and indirectly affects the accuracy of the extracted features.

4.1.2 Select suitable regularization parameters

We propose a parameter auto-selected method to resolve this issue. Suppose we know whether each spatial filter generated using the selected group of regularization parameters can be highly distinguished before extracting the features. In that case, we can use this group of regularization parameters to generate the best spatial filter. Mutual information can evaluate the degree of the difference between the two groups of data. Thus, we can use mutual information to evaluate whether two groups of generated features can be better distinguished. If they can be easily distinguished, the mutual information should be significant. In addition, we use the convolutional neural network as a classifier. The loss function is cross-entropy, which also belongs to another type of mutual information. Thus, when we use mutual information to evaluate the difference of the features, the features obtained are also helpful for the network to do classification.

Therefore, we can calculate the mutual information of the features obtained by using each regularization parameter group and comparing the difference. The feature mutual information generated by different parameters can be arranged into a matrix. We can draw the map of the matrix. The mutual information map of one of the subjects is generated using this subject training data. There are four motor imagery tasks. For each task, we take β_c from [0, 0.4] and γ_c from [0, 0.3]. The maps are shown in Fig. 9.

In the map, the light color represents the significant mutual information while dark color represents the low mutual information. Significant mutual information

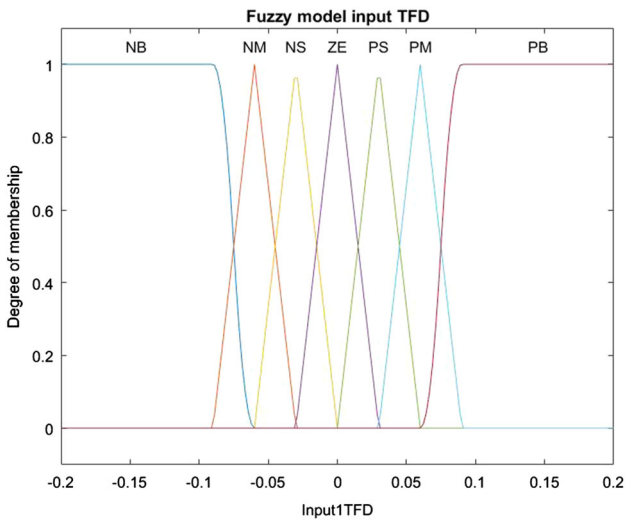

Fig. 6 Fuzzy model input 1: time–frequency distance (Ω_{TFD})

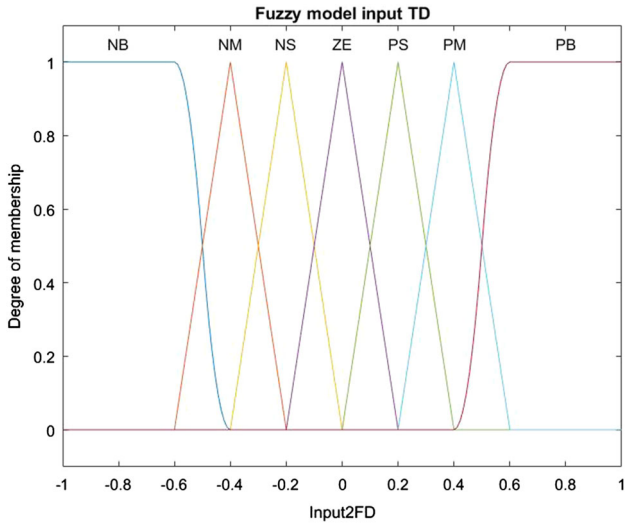

Fig. 7 Fuzzy model input 2: frequency distance (Ω_{FD})

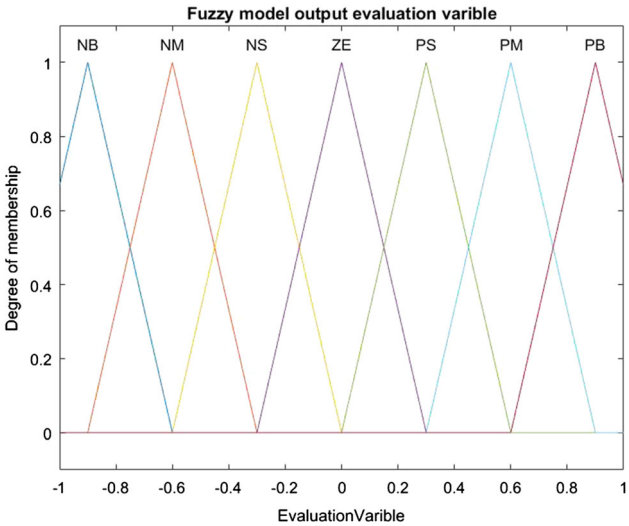

Fig. 8 Fuzzy model output: evaluation variable (T_f)

Table 8 Fuzzy rules which are used to integrate frequency distance (Ω_{FD}) and time–frequency distance (Ω_{TFD})

		Ω_{FD}						
		NB	NM	NS	ZE	PS	PM	PB
Ω_{TFD}	NB	NB	NB	NB	NB	NB	NM	NS
	NM	NB	NM	NM	NM	NM	NS	ZE
	NS	NB	NM	NS	ZE	ZE	ZE	PS
	ZE	NM	NS	ZE	ZE	ZE	PS	PM
	PS	NS	ZE	ZE	ZE	PS	PM	PB
	PM	ZE	PS	PM	PM	PM	PM	PB
	PB	PS	PM	PB	PB	PB	PB	PB

Table 9 Performance for BCI4-2a using data evaluation-based auto-selected regularized FBCSP with SCS-CNN

	Accuracy (%)	Kappa	Se
Sub 1	89.58	0.8611	0.0661
Sub 2	71.88	0.6250	0.0573
Sub 3	94.10	0.9213	0.0682
Sub 4	59.38	0.4583	0.0501
Sub 5	64.58	0.5278	0.0532
Sub 6	55.56	0.4074	0.0477
Sub 7	93.40	0.9120	0.0679
Sub 8	90.28	0.8704	0.0664
Sub 9	90.28	0.8704	0.0664
Mean ± std	78.78 ± 15.79	0.7171 ± 0.21	0.0604 ± 0.0083

Table 10 Performance for BCI3a using data evaluation-based auto-selected regularized FBCSP with SCS-CNN

	Accuracy (%)	Kappa	Se
k3b	96.11	0.9481	0.0874
k6b	67.50	0.5667	0.0849
l1b	81.67	0.7556	0.0965
Mean ± std	81.76 ± 14.31	0.7568 ± 0.19	0.0896 ± 0.0061

represents a higher degree of discrimination of the extracted features. From the map, the regularization parameter β_c of most of the better features produced is between [0, 0.2], γ_c is between [0, 0.1]. However, for this subject's right motor imagery task on the right-top in Fig. 8, the best β_c is 0.4 and γ_c is 0.1. It also proves that it should use different regularization parameters for different

Table 11 Performance for BCI4-2a using data evaluation-based auto-selected regularized FBCSP with voting strategy and SCS-CNN

	Accuracy (%)	Kappa	Se
Sub 1	89.93	0.8657	0.0663
Sub 2	72.92	0.6389	0.0578
Sub 3	95.14	0.9352	0.0687
Sub 4	59.03	0.4537	0.0499
Sub 5	65.97	0.5463	0.0540
Sub 6	54.17	0.3889	0.0468
Sub 7	94.10	0.9213	0.0682
Sub 8	90.63	0.8750	0.0666
Sub 9	89.24	0.8565	0.0660
Mean ± std	79.01 ± 16.09	0.7202 ± 0.21	0.0605 ± 0.0085

Table 12 Performance for BCI3a using data evaluation-based auto-selected regularized FBCSP with voting strategy and SCS-CNN

	Accuracy (%)	Kappa	Se
k3b	96.11	0.9481	0.0874
k6b	70.00	0.6000	0.0871
l1b	85.00	0.8000	0.0991
Mean ± std	83.70 ± 13.10	0.7827 ± 0.17	0.0912 ± 0.0068

mental tasks. Thus, automatically selecting the regularization parameters according to the mutual information map can generate better spatial filters. It could more effectively apply the regularization function to the target data. From the results of experiment B, we can also see that the accuracy of using the ACSP with auto-selected spatial filters is about 8% higher than that of the FBRCSP with the fixed parameters. It also proves that dynamically selecting the regularization parameter can generate more suitable spatial filters for the target data.

4.2 Data quality evaluation

4.2.1 Find the activation channels

Another way to enhance the mixed covariance matrix is to ensure that the added data has distinct motor imagery features. ERS and ERD are standard EEG motor imagery characteristics. They can be regarded as energy changes in motor-related frequency bands. Thus, if we can check that the target data has distinct energy changes in these frequency bands, we can say that this EEG data may have motor imagery characteristics.

The energy changes should be considered based on the baseline. Thus, if we want to check the energy changes, we should have baseline energy. People may generate different EEG signals when they are in different states. Therefore, the baseline energy should always change as the dynamic brain activities. In the two public data sets, before the subject performs the motor imagery tasks, they all have a period where they do not know the next task. The overall state of the brain may not change a lot in a short time, so we can assume that in this short time, the overall brain state of this subject is unchanged. Therefore, the data in this period can be regarded as the baseline.

Our purpose is to evaluate whether there are energy changes when performing motor tasks. However, the imagination abilities of different subjects are different. When doing different motor imagery tasks, they may activate different channels. For example, several researchers have reported that C3 and C4 are two activation channels when the subject is performing left- or right-hand

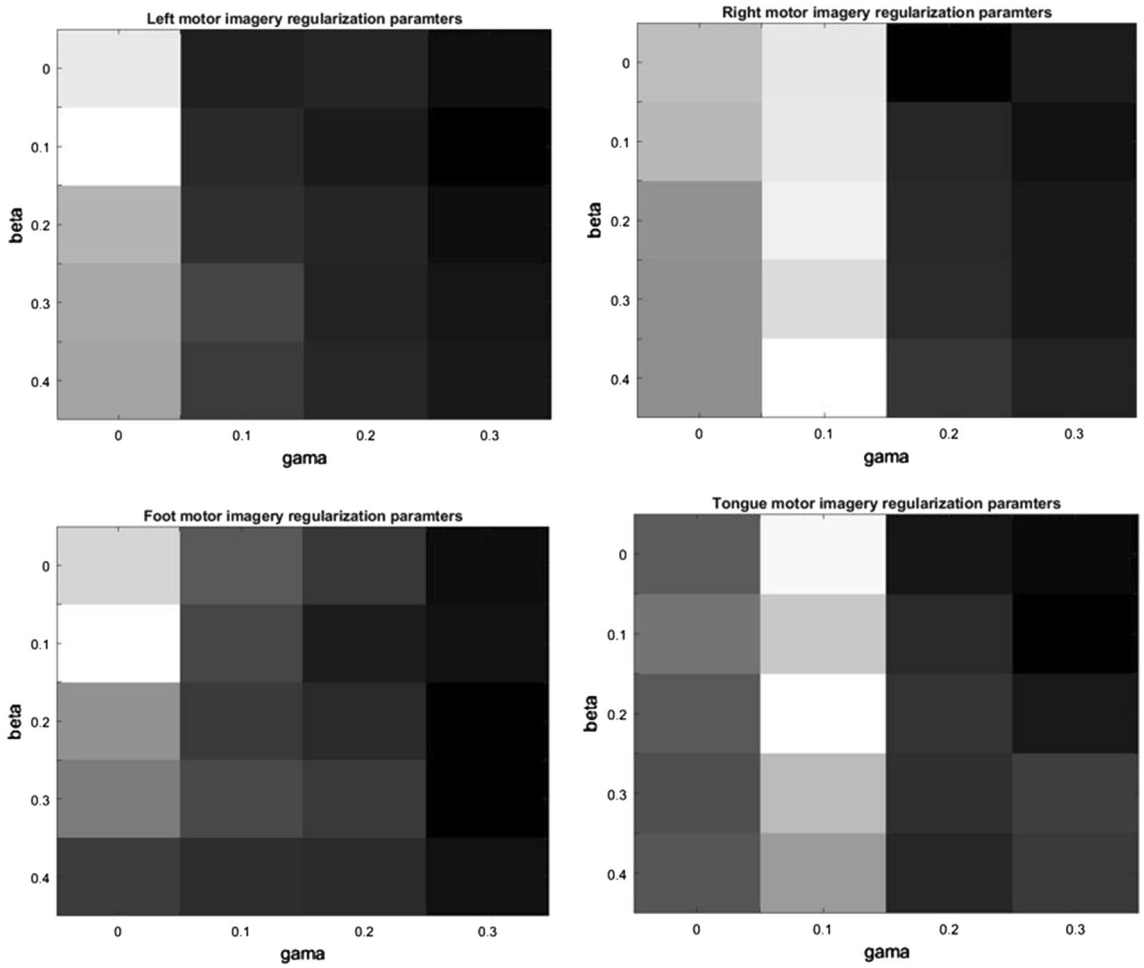

Fig. 9 Mutual information maps for different motor imagery tasks

motor imagery. However, this conclusion applies to most cases, not all situations. The actual activation channels may be close to these two channels, but not always these two. Using Eq. (17) can obtain the distribution of brain area information. We also present the energy distribution of one subject in Fig. 10, where the red color represents energy increase while blue color represents energy decrease. When the subject is doing left-hand or right-hand motor imagery, there are significant energy changes near the motion areas C3 and C4 on the top two maps in Fig. 10. When doing foot motor imagery, there are significant energy changes in the central motion area on the left bottom map in Fig. 10. When doing tongue motor imagery, there are significant energy changes in the back of the head on the right bottom map in Fig. 10. It follows the general spatial distribution when performing motor imagery tasks, but the specific activation channels have a certain difference.

However, there is no guarantee that all subjects and all data trials can follow this distribution. The distribution could be affected by the subjects' states, feelings and some external factors. This calculated distribution can only prove that most data trials follow this distribution when the subject is doing a specific motor imagery task. Therefore, this cannot prove that the data of each trial meets the motor imagery characteristics of the spatial information distribution. These findings confirm that various brain areas are activated when the subject is doing different motor imagery tasks. We should continue to extract the frequency and time–frequency features for these brain areas to prove this data contains motor imagery characteristics.

4.2.2 Evaluate the degree of energy changes

We perform frequency and time–frequency analysis on the selected channels. The purpose is to detect whether the data has motor imagery's frequency and time–frequency characteristics. We introduce two indicators to evaluate the energy changes of the data in these two domains. In the experiment, we use two frequency bands, namely 8–13 Hz and 16–25 Hz, because these two frequency bands are the most relevant frequency bands of motor imagery.

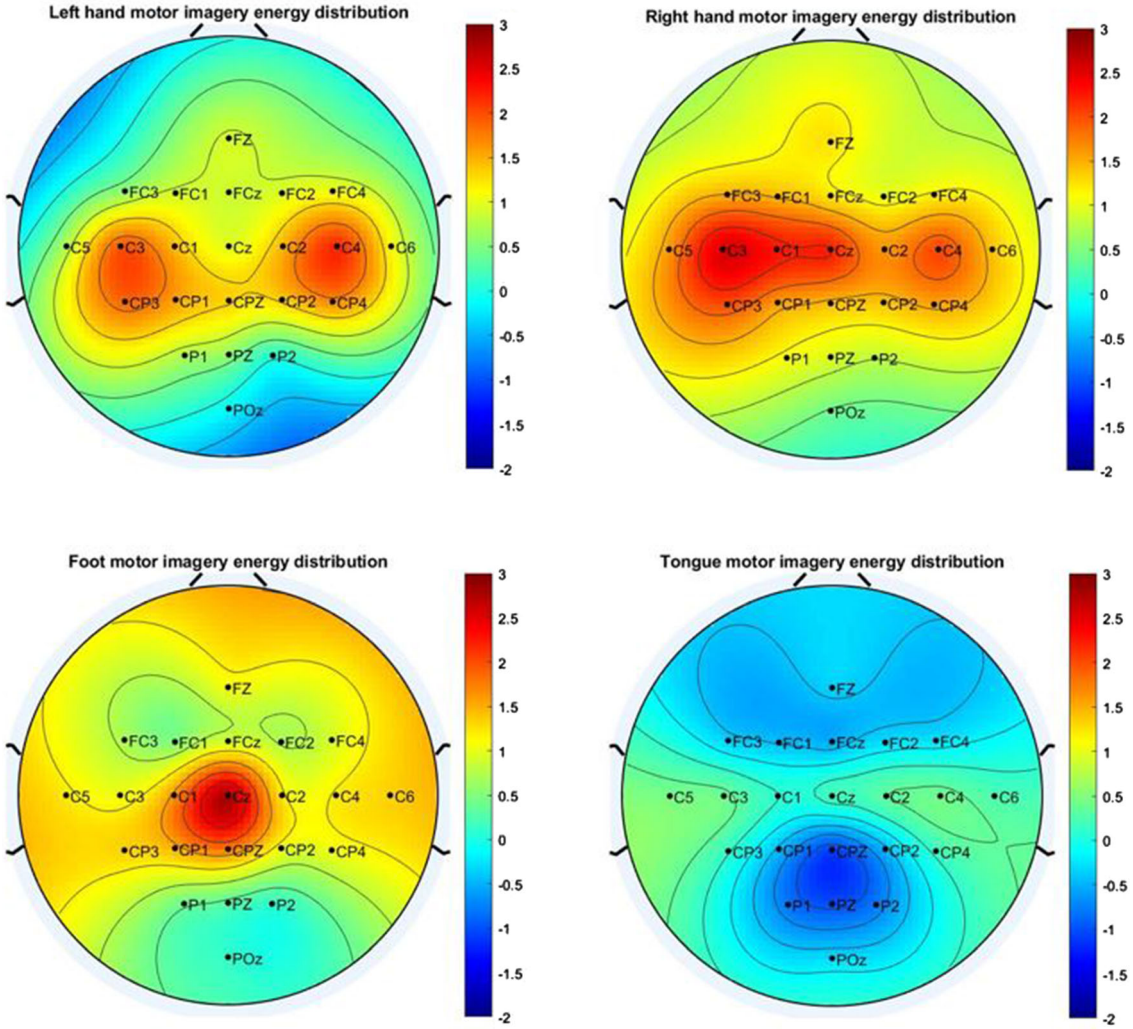

Fig. 10 Energy distribution maps for different motor imagery tasks

FD and TFD are the two indicators used to judge the degree of energy changes. After CWT transformation, we can obtain the energy feature map. From the map, we can check the energy changes. However, because the moto imagery can last for a long time, it may also occur many times in a short period. Therefore, the idea of TFD is to transform the energy feature map into distribution and use the distance between rest state distribution and motor state distribution to evaluate the degree of the energy changes. Thus, whether the motor imagery lasts for a long time or occurs many times, this indicator can still evaluate the degree of energy changes from the nonlinear term. FD is to calculate the overall difference of energy between motor state data and baseline in the motor-related frequency bands. Therefore, FD uses the linear way to evaluate the energy change. Through experiments, we can obtain the relationship between the amplitude of the indicators and the degree of energy changes. Figure 11 shows the frequency spectrum difference between motor state and baseline signals.

In Fig. 11, when the subject performs an imaginary task in the frequency domain, the motor state energy decreases to different degrees than the rest state energy. The value of FD can be used to describe the degree of energy reduction in the frequency domain. The two maps on the top in Fig. 11 show that when FD is greater than 0.5, the amplitude of energy change is significant. It means that the data has a relatively distinct frequency characteristic of motor imagery. The map on the right bottom in Fig. 11 shows that when the FD is less than 0.25, the frequency-domain energy of the rest state and the motor state are mixed. There is no significant energy change. Thus, the frequency characteristics of the motor imagery of this data are not significant.

Similarly, Fig. 12 shows the time–frequency spectrum difference between motor state and baseline signals. The first 2 s of data are baseline data, and the remaining 4 s of

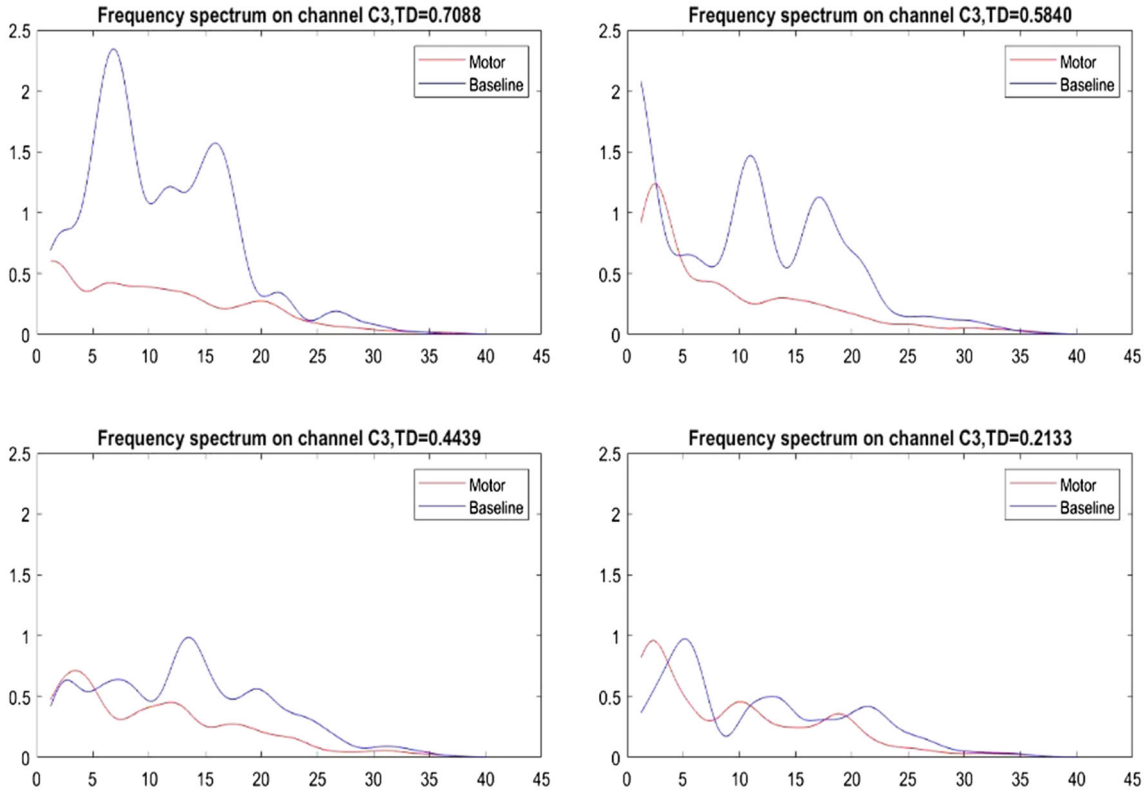

Fig. 11 Frequency spectrum for different FDs

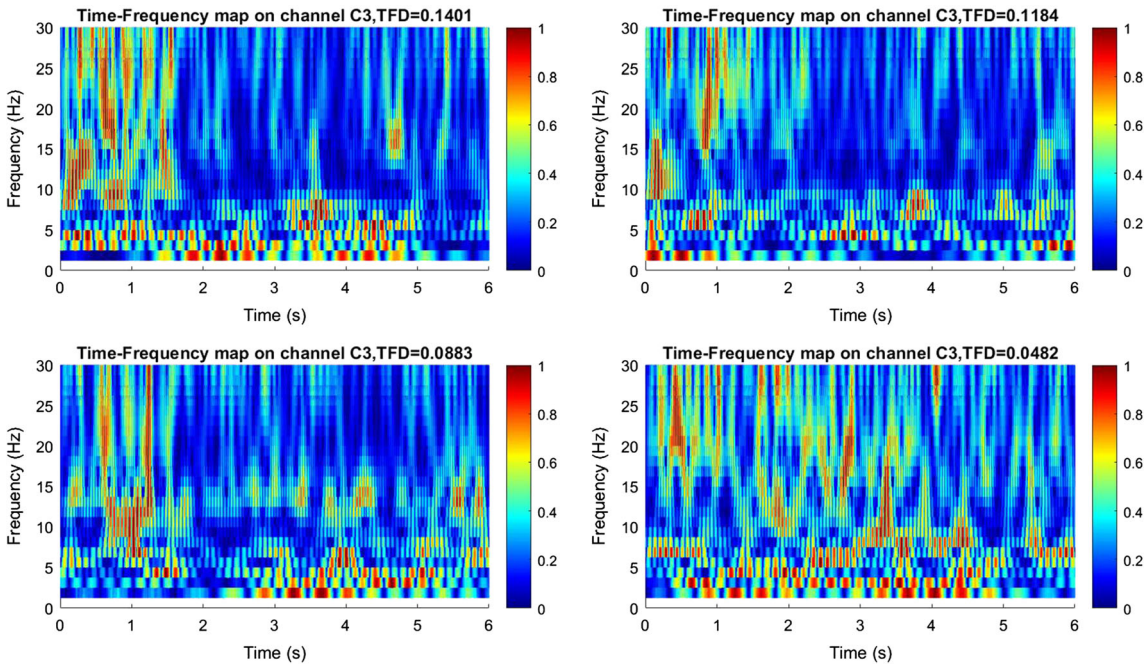

Fig. 12 Time–frequency spectrum map for different TFDs

data is motor data. Different TFDs also indicate that the data has different degrees of time–frequency characteristics. The two maps on the top in Fig. 12 show when TFD is greater than 0.1, the time–frequency energy of the motor state data in the last four seconds is significantly reduced compared to the rest state data in the first two seconds. The

map on the right bottom in Fig. 12 shows when TFD is less than 0.05, the motor state energy does not change significantly compared to the rest state time–frequency energy. Therefore, these findings confirm that the two proposed indicators can be used to describe the degree of energy changes.

4.2.3 Fuzzify energy description indicators to evaluate data quality

Consider that if we want to evaluate whether the data is good or not, we need a threshold to measure the level of the quality. However, only using the threshold method cannot describe the energy difference well in the two domains. For the same trial of data, it may have significant time–frequency characteristics, but the frequency-domain characteristics may not be significant. In this case, the threshold cannot be set to reasonably select effective data. The characteristics of these two domains should jointly determine whether the data has motor imagery features. Therefore, although we can obtain both frequency and time–frequency energy information, it cannot describe the overall degree of the energy changes. Thus, the fuzzy logic model is proposed to solve this problem. The fuzzy model can fuzzify the data to a fixed range. It transforms the real evaluation value into a variety of continuous fuzzy values. In addition, the fuzzy model is developed based on statistical experience. We can use the rules to define the degree of energy changes.

FD contains the frequency-domain energy information, representing the overall energy changes. TFD contains the time–frequency-domain energy information, representing the detailed energy changes. Generally, TFD should contain more information than FD, because TFD contains both time and nonlinear frequency information. Nevertheless, we cannot ignore FD because it provides comprehensive linear frequency information. Therefore, when we design the fuzzy rules, we can trust more on TFD. If TFD is very large and FD is small, the output should be large.

In contrast, if FD is very large and TFD is small, it is possible that the frequency energy changes a lot which a suddenly unnormal fluctuation may cause. In this case, the output should be medium because we are not sure whether the fluctuation is what we expect. Thus, FD as the overall evaluation indicator also functions to correct unnormal information.

The output is de-fuzzified according to the rules to obtain a comprehensive indicator. This indicator can better integrate the information in the frequency and the time–frequency domain. It can describe whether the data has motor imagery features on a selected activation channel. The sign represents the energy increase or decrease. The amplitude represents the degree of motor imagery.

We used three channels with the most significant energy changes in the experiments. They are independent and have different contributions to the motor imagery task. Therefore, the contribution weight is introduced to control the confidence. Confidence means how much the model should trust the information from this channel. The lower ranking indicates the smaller confidence. After applying the confidence to the fuzzy outputs, the final obtained parameter can be used to evaluate the degree of imaginary movement in different frequency ranges. This evaluating parameter considers spatial, frequency and time–frequency information. Thus, it can be used to describe the motor imagery data quality.

We can evaluate the quality of the data by using this parameter. It can be used to select the EEG data from other subjects. Selecting high-quality data to calculate the mixed covariance can better reduce the error. The mixed covariance matrix has more effective motor imagery information, which can indirectly improve the performance of the spatial filters. Compared to AFBCSP, the spatial filter obtained by using selected high-quality data can extract effective features more accurately and improve the accuracy.

4.3 SCS-CNN and voting

4.3.1 SCS-CNN extracts multiple time interval features from an independent channel

From the results of experiment A, using the same feature extraction method, SCS-CNN as a classifier can better classify the EEG motor imagery tasks. SVM and RF both use the variance of filtered data as the features, while SCS-CNN, LSTM and LSTM-CNN all use the spatially filtered data as the features. Although SVM and RF can be trained fast, this way may lose some useful information. Compared to these classifiers, a neural network can retain more original data information and extract new features from the spatially filtered features.

The ideas of both LSTM and LSTM-CNN are to extract time series information. EEG can be regarded as a time series signal, but the time feature of motor imagery is not continuous. Motor imagery may occur over some time, or it can also occur multiple times within a period. Thus, using LSTM or LSTM-CNN to extract continuous-time imaginary motion features may not get good results.

The best classification performance is obtained from the results by using SCS-CNN as the classifier. A single-channel learning strategy is adopted. CNN will not learn the relation between every two channels. Each time, it generates feature maps or does a pooling operation based on only one channel. The size of the first dimension of the image should also be unchanged. The feature extraction method is a CSP-related method. It can be regarded as

Table 13 Compare the accuracy γ and kappa κ of different models on the data set BCI Competition IV 2a

BCI4-2a	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Mean ± std
FBCSP + PW strategy [5]	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%
	κ = 0.78	κ = 0.45	κ = 0.86	κ = 0.47	κ = 0.63	κ = 0.33	κ = 0.85	κ = 0.79	κ = 0.78	κ = 0.66 ± 0.20
CSP + AM-BA-SVM [26]	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%
	κ = 0.87	κ = 0.55	κ = 0.89	κ = 0.60	κ = 0.58	κ = 0.41	κ = 0.88	κ = 0.84	κ = 0.80	κ = 0.71 ± 0.18
KL_CSP [7]	γ = 73.10%	γ = 64.20%	γ = 84.20%	γ = 73.80%	**γ = 79.40%**	γ = 57.80%	γ = 62.70%	γ = 80.90%	γ = 74.10%	γ = 72.24% ± 8.95%
	κ = –	κ = –	κ = –	κ = –	**κ = –**	κ = –	κ = –	κ = –	κ = –	κ = –
KL_LTCSP [7]	γ = 74.50%	γ = 65.10%	γ = 83.50%	γ = 75.10%	γ = 78.00%	γ = 58.30%	γ = 63.40%	γ = 81.40%	γ = 75.10%	γ = 72.71% ± 8.56%
	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –
STFS_CSP [9]	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%
	κ = 0.63	κ = 0.43	κ = 0.74	κ = 0.54	κ = 0.19	κ = 0.26	κ = 0.63	κ = 0.62	κ = 0.69	κ = 0.53 ± 0.19
MCSP [6]	γ = 77.10%	γ = 40.30%	γ = 70.50%	γ = 38.90%	γ = 35.40%	γ = 47.20%	γ = 62.50%	γ = 75.70%	γ = 75.00%	γ = 58.07% ± 17.50%
	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –
CSD-CSP [27]	γ = 67.70%	γ = 49.65%	γ = 58.33%	γ = 58.68%	γ = 48.26%	γ = 43.40%	γ = 66.32%	γ = 58.68%	γ = 72.22%	γ = 58.14% ± 9.64%
	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –
3D CNN [23]	γ = 77.40%	γ = 60.14%	γ = 82.93%	γ = 72.29%	γ = 75.84%	γ = 68.99%	γ = 76.04%	γ = 76.86%	γ = 84.67%	γ = 75.01% ± 7.35%
	κ = 0.70	κ = 0.46	κ = 0.79	κ = 0.59	κ = 0.65	κ = 0.54	κ = 0.65	κ = 0.70	κ = 0.71	κ = 0.64 ± 0.10
C2CM [24]	γ = 87.50% κ = 0.83	γ = 65.28% κ = 0.54	γ = 90.28% κ = 0.87	γ = 66.67% κ = 0.56	γ = 62.50% κ = 0.50	γ = 45.49% κ = 0.27	γ = 89.58% κ = 0.86	γ = 83.33% κ = 0.78	γ = 79.51% κ = 0.73	γ = 74.46% ± 15.33% κ = 0.66 ± 0.20
CWCNN [24]	γ = 86.11% κ = 0.82	γ = 60.76% κ = 0.48	γ = 86.81% κ = 0.82	γ = 67.36% κ = 0.57	γ = 62.50% κ = 0.50	γ = 45.14% κ = 0.27	γ = 90.63% κ = 0.88	γ = 81.25% κ = 0.75	γ = 77.08% κ = 0.69	γ = 73.07% ± 15.11% κ = 0.64 ± 0.20
Monolithic network [20]	γ = 83.13%	γ = 65.45%	γ = 80.29%	**γ = 81.60%**	γ = 76.70%	**γ = 71.12%**	γ = 84.00%	γ = 82.66%	γ = 80.74%	γ = 78.41% ± 6.27%
	κ = –	κ = –	κ = –	**κ = –**	κ = –	**κ = –**	κ = –	κ = –	κ = –	κ = –
SS-MEMDBF [10]	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%
	κ = 0.86	κ = 0.24	κ = 0.70	κ = 0.68	κ = 0.36	κ = 0.34	κ = 0.66	κ = 0.75	κ = 0.82	κ = 0.60 ± 0.23
TSSM + LDA [19]	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%
	κ = 0.77	κ = 0.33	κ = 0.77	κ = 0.51	κ = 0.35	κ = 0.36	κ = 0.71	κ = 0.72	κ = 0.83	κ = 0.59 ± 0.21
CRAM [25]	γ = 61.02%	γ = 42.35%	γ = 73.11%	γ = 50.43%	γ = 50.74%	γ = 51.48%	γ = 67.26%	γ = 69.72%	γ = 66.85%	γ = 59.10% ± 10.85%
	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –	κ = –
DPLM [18]	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%	γ = –%
	κ = 0.75	κ = 0.49	κ = 0.76	κ = 0.49	κ = 0.34	κ = 0.36	κ = 0.68	κ = 0.76	κ = 0.76	κ = 0.60 ± 0.18
Proposed D-ACSP-V + SCS-CNN	**γ = 89.93%** **κ = 0.87**	**γ = 72.92%** **κ = 0.64**	**γ = 95.14%** **κ = 0.94**	γ = 59.03% κ = 0.45	γ = 65.97% κ = 0.55	γ = 54.17% κ = 0.39	**γ = 94.10%** **κ = 0.92**	**γ = 90.63%** **κ = 0.88**	**γ = 89.24%** **κ = 0.86**	**γ = 79.01% ± 16.09%** **κ = 0.72 ± 0.22**

Bold values represent the best ones

Table 14 Compare the accuracy γ and kappa κ of different models on the data set BCI Competition IIIa

BCI3a	K3b	K6b	L1b	Mean ± std
KL-CSP [7]	γ = 83.40%	γ = 61.50%	γ = 67.10%	γ = 70.67% ± 11.38%
	κ = –	κ = –	κ = –	κ = –
KL-LTCSP [7]	γ = 84.30%	γ = 62.10%	γ = 65.60%	γ = 70.67% ± 11.94%
	κ = –	κ = –	κ = –	κ = –
LP-SVD + AR + error variance [16]	γ = 58.75%	**γ = 76.66%**	γ = 66.66%	γ = 67.35% ± 8.98%
	κ = –	**κ = –**	κ = –	κ = –
LP-SVD + logistic model tree [17]	γ = 86.38%	γ = 74.58%	γ = 77.08%	γ = 79.35% ± 6.22%
	κ = –	κ = –	κ = –	κ = –
LPQR + logistic model tree [17]	γ = 90.00%	γ = 76.25%	γ = 77.91%	γ = 81.38% ± 7.51%
	κ = –	κ = –	κ = –	κ = –
ICA + PCA + SVM [15]	γ = –%	γ = –%	γ = –%	γ = –%
	κ = 0.95	κ = 0.41	κ = 0.52	κ = 0.63 ± 0.29
CSP + SVM, LDA, KNN bagging [15]	γ = –%	γ = –%	γ = –%	γ = –%
	κ = 0.90	κ = 0.43	κ = 0.71	κ = 0.69 ± 0.24
WPT + CSP [8]	γ = 83.21%	γ = 76.17%	γ = 76.17%	γ = 78.52% ± 4.06%
	κ = –	κ = –	κ = –	κ = –
DCT [16]	γ = 43.75%	γ = 38.05%	γ = 45.83%	γ = 42.54% ± 4.03%
	κ = –	κ = –	κ = –	κ = –
ARR [39]	γ = –%	γ = –%	γ = –%	γ = –%
	κ = 0.69	κ = 0.36	κ = 0.39	κ = 0.48 ± 0.18
Proposed D-ACSP-V + SCS-CNN	**γ = 96.11%**	γ = 70.00%	**γ = 85.00%**	**γ = 83.70% ± 13.10%**
	κ = 0.95	κ = 0.60	**κ = 0.80**	**κ = 0.78 ± 0.18**

Bold values represent the best ones

Table 15 Accuracy's p-value of the purposed method to traditional FBCSP and FBRCSP method

	FBCSP	FBRCSP	Proposed ACSP	Proposed D-ACSP
p-value (FBCSP)	–	0.0045	3.6563e−4	2.1819e−4
p-value (FBRCSP)	0.0045	–	9.7594e−4	6.9928e−4
p-value (ACSP)	3.6563e−4	9.7594e−4	–	0.0173
p-value (D-ACSP)	2.1819e−4	6.9928e−4	0.0173	–

Table 16 Accuracy's p-value of the purposed method to other methods

	CRAM [25]	MCSP [6]	CSD-CSP [27]	C2CM [24]	CWCNN [24]	STFS_CSP [9]
T-test (p-value)	2.2141e−4	1.4919e−4	4.7676e−4	0.0300	0.0233	0.0016
	DPLM [18]	FBCSP + PW strategy [5]	TSSM + LDA [19]	KL-CSP [7]	DCT [16]	ARR [16]
T-test (p-value)	0.0028	0.0385	0.0097	0.0408	0.0204	0.0299

transforming the information of EEG signals into another space. Whitening is one of the steps in spatial transformation, which means that it removes the correlation among all the channels. Therefore, the information in each channel of the transformed signal should be independent. Thus, a single-channel learning strategy can extract effective information from each independent channel, making the information between adjacent channels not affect each other. It can be better to extract useful information from spatial transformed EEG signals. The purpose of using an SCS-CNN is that, in the time domain of the EEG signal, it is difficult to determine how many points are relative, so

the network is used to generate a certain number of feature maps in different time intervals, and then, after the feature maps are merged and compressed, it continues to learn features at different time intervals. Finally, all time sequences are converted into feature vectors to achieve the purpose of second stage feature extraction. In other words, this network can get a broader learning horizon so that the learned information can express more data features in the learning process.

In addition, the residual part is added to the network. The main reason is that due to the increase in the number of layers, there may be a problem of gradient disappearance. The residual network may better solve the problem of gradient disappearance [37]. The entire network update could be more stable. As the number of network layers increases, the network learning performance may be better, but it may also worsen the learning. The function of short-circuit learning is to prevent if the learning process becomes worse, and the network can go directly over the redundant network layer [38]. In our structure, the learning effect of a 52-layer network can be at least as good as a 29-layer network. The one size convolution added in the middle is to reduce the parameters and increase the number of feature maps.

4.3.2 Voting strategy increases feature diversity

Voting is a tip to increase the diversity of the extracted features. When we use RCSP to get the features, we add other subjects' data. Although we have already used the data evaluation method to select the high-quality data, we still cannot guarantee that the data from different subjects has similar spatial energy activation distribution. Therefore, developing many groups of spatial matrixes generated using different groups of other subjects' data can make the spatial energy activation distribution closer to the target distribution. In addition, although we can automatically select the RCSP regularization parameters based on mutual information, it cannot completely estimate the distance between the distributions of the two target classes. It can only be used to describe the difference in data information roughly. Thus, these findings support the hypothesis that developing various groups of spatial matrixes and selecting various groups of regularization parameters can obtain several different features. Using a voting strategy to integrate the classification results using multiple features can slightly improve the performance and make the system more stable.

4.4 Compare to other methods

We compare the results of our method with other methods proposed in recent years. We use the evaluation measures of tenfold cross-validation accuracy, kappa and standard error. The data sets are BCI4-2a and BCI3a. The comparison results are shown in Tables 13 and 14. References [6–8, 16, 17, 20, 25, 27] only provided accuracy, while references [5, 9, 10, 15, 18, 19, 26, 39] only provided kappa value. References [23, 24] shown both of them.

From the results, for the BCI4-2a data set, compared to other papers' methods, our method obtains the highest accuracy by using the data of subjects 1, 2, 3, 7, 8, 9. Most of them can reach over 0.8 kappa and over 75% accuracy except subject 2. For the BCI3a data set, our method obtains the highest accuracy by using the data of subjects k3b and l1b. It can reach over 0.8 kappa.

In Tables 13 and 14, lots of paper proposed improved CSP methods to get the feature vectors and classify the task using SVM or LDA. This classification is easy to achieve, but it may lose some useful EEG information. Thus, this way is difficult to get a good result compared to network-related methods. It also follows the conclusion we drew from experiment A. Some other methods shown in Table 13 use CSP as the feature extraction method and networks with different structures. From the results, our network structure is better than some other network structures, such as the popular networks C2CM and CWCNN.

However, compared to some unique feature extraction methods or advanced classifiers, only using an SCS-CNN is not enough. Combining D-ACSP-V and SCS-CNN can significantly improve the classification performance. From the result, for the BCI4-2a data set, expect subjects 4, 5 and 6, we can get the highest performance compared to other papers' methods. For the BCI3a data set, we can get the highest performance on subject k3b and subject l1b. The mean performance is also better than others using both two data sets. Thus, these testing results support the original hypothesis that our proposed method is more effective in classifying motor imagery multi-classification tasks.

Tables 15 and 16 show the p-value of the purposed method to traditional methods and other papers' methods. The p-values shown in the results are calculated from two-tailed paired t-test. From the results, the p-values are all less than 0.05, which means the proposed algorithm is more robust and more effective than the traditional FBCSP or FBRCSP methods and other publications' methods.

5 Conclusion

In this paper, we propose a mutual information-based regularization parameters selection method and a data quality evaluation method to improve the regularized spatial filters' performance. When mixing the covariance matrixes, the mutual information-based parameter selection

method can automatically adjust the regularization weights. The data evaluation method can analyze the motor imagery features in the spatial, frequency and time–frequency domains. This method can improve the performance of RCSP, but it can also be used to check whether the collected data has motor imagery features. In addition, a single-channel-based series convolutional neural network is introduced to classify the motor imagery multi-classification tasks. Also, a voting strategy can be used as a tip to improve the classification accuracy slightly. We use the tenfold cross-validation method and kappa to test two data sets in the experiments. For the BCI4-2a data set, the method with the best accuracy is to use D-ACSP-V and SCS-CNN. It obtains an average accuracy of 79.01% and a kappa of 0.7202. We also tested the BCI3a data set, the average accuracy is 83.70%, and the kappa is 0.7827. Compared to the methods proposed in recent papers, our method has higher accuracy. Therefore, D-ACSP-V is a suitable motor imagery feature extraction method. SCS-CNN is also a suitable classifier for motor imagery multi-classification tasks.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Code availability Code for proposed algorithm is provided as part of the replication package. It is available at https://studentutsedu-my.sharepoint.com/:u:/r/personal/yang_an-1_student_uts_edu_au/Documents/Code/AFBRCSP.zip?csf=1&web=1&e=IDUPss and https://studentutsedu-my.sharepoint.com/:u:/r/personal/yang_an-1_student_uts_edu_au/Documents/Code/DataEvaluation.zip?csf=1&web=1&e=19xAGQ for review. It will be uploaded to the NCCA once the paper has been conditionally accepted.

Declarations

Conflict of interest All authors declare that they have no conflicts of interest.

References

1. Trad D, Al-ani T, Jemni M (2015) A feature extraction technique of EEG based on EMD-BP for motor imagery classification in BCI. In: 2015 5th international conference on information & communication technology and accessibility (ICTA), pp 1–6
2. Wu Y, Huang TH, Lin CY, Tsai SJ, Wang P (2018) Classification of EEG motor imagery using support vector machine and convolutional neural network. In: 2018 international automatic control conference (CACS), pp 1–4
3. Li S, Feng H (2019) EEG signal classification method based on feature priority analysis and CNN. In: 2019 international conference on communications, information system and computer engineering (CISCE), pp 403–406
4. Li D, Wang J, Xu J, Fang X (2019) Densely feature fusion based on convolutional neural networks for motor imagery EEG classification. IEEE Access 7:132720–132730
5. Ang KK, Chin ZY, Wang C, Guan C, Zhang H (2012) Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. Front Neurosci. https://doi.org/10.3389/fnins.2012.00039
6. Jafarifarmand A, Badamchizadeh MA (2020) Real-time multiclass motor imagery brain-computer interface by modified common spatial patterns and adaptive neuro-fuzzy classifier. Biomed Signal Process Control 57:101749
7. Wang H (2012) Harmonic mean of Kullback–Leibler divergences for optimizing multi-class eeg spatio-temporal filters. Neural Process Lett 36:161–171. https://doi.org/10.1007/s11063-012-9228-y
8. Li M, Lin L, Yang J-F (2011) Adaptive feature extraction of four-class motor imagery EEG based on best basis of wavelet packet and CSP. In: 2011 international conference on electric information and control engineering, pp 3918–3921
9. Miao M, Zeng H, Wang A, Zhao C, Liu F (2017) Discriminative spatial-frequency-temporal feature extraction and classification of motor imagery EEG: an sparse regression and Weighted Naïve Bayesian Classifier-based approach. J Neurosci Methods 278:13–24
10. Gaur P, Pachori RB, Wang H, Prasad G (2018) A multi-class EEG-based BCI classification using multivariate empirical mode decomposition based filtering and Riemannian geometry. Expert Syst Appl 95:201–211
11. Lu H, Plataniotis KN, Venetsanopoulos AN (2009) Regularized common spatial patterns with generic learning for EEG signal classification. In: 2009 annual international conference of the IEEE engineering in medicine and biology society, pp 6599–6602
12. Lu H, Eng H, Guan C, Plataniotis KN, Venetsanopoulos AN (2010) Regularized common spatial pattern with aggregation for EEG classification in small-sample setting. IEEE Trans Biomed Eng 57:2936–2946
13. Park SH, Lee D, Lee SG (2018) Filter bank regularized common spatial pattern ensemble for small sample motor imagery classification. IEEE Trans Neural Syst Rehabil Eng 26:498–505
14. Park S, Lee S (2017) Small sample setting and frequency band selection problem solving using subband regularized common spatial pattern. IEEE Sens J 17:2977–2983
15. Blankertz B, Muller KR, Krusienski DJ, Schalk G, Wolpaw JR, Schlogl A, Pfurtscheller G, Millan JR, Schroder M, Birbaumer N (2006) The BCI competition III: validating alternative approaches to actual BCI problems. IEEE Trans Neural Syst Rehabil Eng 14:153–159
16. Baali H, Khorshidtalab A, Mesbah M, Salami MJE (2015) A transform-based feature extraction approach for motor imagery tasks classification. IEEE J Transl Eng Health Med 3:1–8

17. Khorshidtalab A, Salami MJE, Akmeliawati R (2017) Motor imagery task classification using transformation based features. Biomed Signal Process Control 33:213–219
18. Davoudi A, Ghidary SS, Sadatnejad K (2017) Dimensionality reduction based on distance preservation to local mean for symmetric positive definite matrices and its application in brain–computer interfaces. J Neural Eng IOP Publishing 14:036019. https://doi.org/10.1088/1741-2552/aa61bb
19. Xie X, Yu ZL, Lu H, Gu Z, Li Y (2017) Motor imagery classification based on bilinear sub-manifold learning of symmetric positive-definite matrices. IEEE Trans Neural Syst Rehabil Eng 25:504–516
20. Olivas-Padilla BE, Chacon-Murguia MI (2019) Classification of multiple motor imagery using deep convolutional neural networks and spatial filters. Appl Soft Comput 75:461–472
21. Ha K-W, Jeong J-W (2019) Motor imagery EEG classification using capsule networks. Sensors 19:2854
22. Schirrmeister RT, Springenberg JT, Fiederer LDJ, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W, Ball T (2017) Deep learning with convolutional neural networks for EEG decoding and visualization. Hum Brain Mapp 38:5391–5420. https://doi.org/10.1002/hbm.23730
23. Zhao X, Zhang H, Zhu G, You F, Kuang S, Sun L (2019) A multi-branch 3D convolutional neural network for EEG-based motor imagery classification. IEEE Trans Neural Syst Rehabil Eng 27:2164–2177
24. Sakhavi S, Guan C, Yan S (2018) Learning temporal information for brain-computer interface using convolutional neural networks. IEEE Trans Neural Netw Learn Syst 29:5619–5629
25. Zhang D, Yao L, Chen K, Monaghan J (2019) A convolutional recurrent attention model for subject-independent EEG signal analysis. IEEE Signal Process Lett 26:715–719
26. Selim S, Tantawi MM, Shedeed HA, Badr A (2018) A CSP\AM-BA-SVM approach for motor imagery BCI system. IEEE Access 6:49192–49208
27. Rathee D, Raza H, Prasad G, Cecotti H (2017) Current source density estimation enhances the performance of motor-imagery-related brain–computer interface. IEEE Trans Neural Syst Rehabil Eng 25:2461–2471
28. Yang Y, Chevallier S, Wiart J, Bloch I (2017) Subject-specific time-frequency selection for multi-class motor imagery-based BCIs using few Laplacian EEG channels. Biomed Signal Process Control 38:302–311
29. An Y, Hu T, Wang J, Lyu J, Banerjee S, Ling SH (2019) Lung nodule classification using a novel two-stage convolutional neural networks structure. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp 6259–6262
30. Brunner C, Leeb R, Müller-Putz G, Schlögl A, Pfurtscheller G (2008) BCI competition 2008–Graz data set A. In: Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology, pp 1–6
31. Blankertz B, Müller K-R, Krusienski D, Schalk G, Wolpaw JR, Schlögl A, Pfurtscheller G, Millán JdR, Schröder M, Birbaumer N (2005) Bci competition iii. Fraunhofer FIRST. IDA, http://ida.first.fraunhofer.de/projects/bci/competition_iii
32. McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med 22:276–282
33. Fischer T, Krauss C (2018) Deep learning with long short-term memory networks for financial market predictions. Eur J Oper Res 270:654–669
34. Pak U, Kim C, Ryu U, Sok K, Pak S (2018) A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction. Air Qual Atmos Health 11:883–895. https://doi.org/10.1007/s11869-018-0585-1
35. Auria L, Moro RA (2008) Support vector machines (SVM) as a technique for solvency analysis
36. Pal M (2005) Random forest classifier for remote sensing classification. Int J Remote Sens 26:217–222
37. He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. In: Computer vision—ECCV 2016, Springer International Publishing, pp 630–645
38. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
39. Mesbah M, Khorshidtalab A, Baali H, Al-Ani A (2015) Motor imagery task classification using a signal-dependent orthogonal transform based feature extraction. In: Neural Information Processing. Springer, Cham, pp 1–9