Feature Selection method based on Menger Curvature and LDA Theory for a P300 Brain-computer Interface

Shurui Li[1], Jing Jin[1], Ian Daly[2], Chang Liu[1], and Andrzej Cichocki[3,4]

[1] Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, People's Republic of China
[2] Brain-Computer Interfacing and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Colchester, Essex, CO4 3SQ, UK
[3] Skolkovo Institute of Science and Technology (Skoltech), 121205 Moscow, Russia
[4] Department of Applied Computer Science, Nicolaus Copernicus University (UMK), 87-100 Torun, Poland

Corresponding author: Jing Jin, E-mail: jinjingat@gmail.com

Abstract

Brain–computer interface (BCI) systems decode electroencephalogram signals to establish a channel for direct interaction between the human brain and the external world without the need for muscle or nerve control. The P300 speller, one of the most widely used BCI applications, presents a selection of characters to the user and performs character recognition by identifying P300 event-related potentials from the EEG. Such P300-based BCI systems can reach good levels of accuracy but are difficult to use in day-to-day life due to redundancy and noisy signal. A room for improvement should be considered. We propose a novel hybrid feature selection method for the P300-based BCI system to address the problem of feature redundancy, which combines the Menger curvature and linear discriminant analysis. First, selected strategies are applied separately to a given dataset to estimate the gain for application to each feature. Then, each generated value set is ranked in descending order and judged by a predefined criterion to be suitable in classification models. The intersection of the two approaches is then evaluated to identify an optimal feature subset. The proposed method is evaluated using three public datasets, i.e., BCI Competition III dataset II, BNCI Horizon dataset, and EPFL dataset. Experimental results indicate that compared with other typical feature selection and classification methods, our proposed method has better or comparable performance. Additionally, our proposed method can achieve the best classification accuracy after all epochs in three datasets. In summary, our proposed method provides a new way to enhance the performance of the P300-based BCI speller.

Keywords: brain–computer interface, P300 speller, feature selection, Menger curvature, linear discriminant analysis

1. Introduction

A brain–computer interface (BCI) system provides a direct communication pathway between the brain and external devices. This system can offer a form of assistive technology for individuals with motor disabilities, such as spinal cord injury or Parkinson's disease [1, 2], and identify the user's intentions from their brain activity to issue control commands. This technology can help severely disabled individuals to have effective interactions with the environment or aid in their recovery. Given its easy attachment and no need for surgery [3], the noninvasive electroencephalogram (EEG) is the most commonly used signal acquisition method for BCI systems.

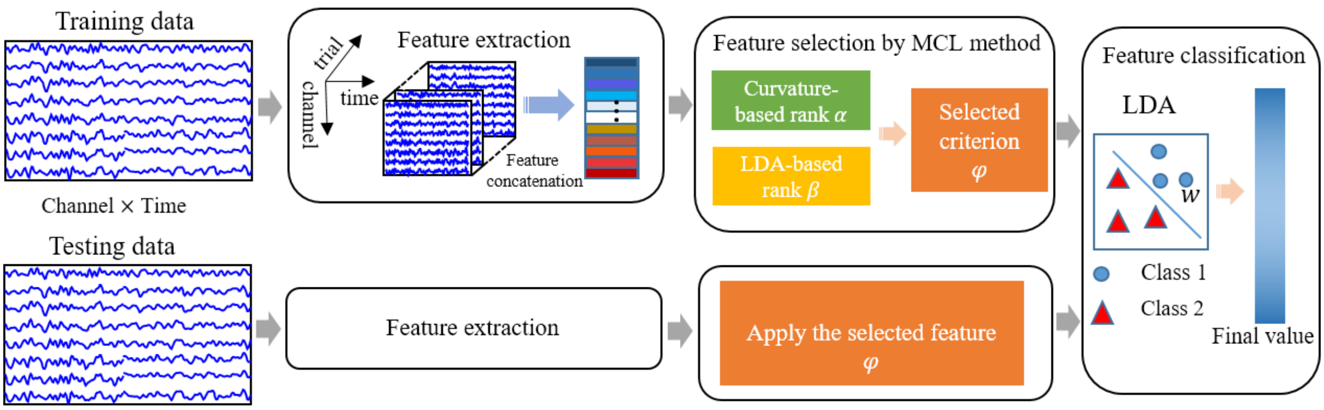

Fig. 1 Overall framework of our proposed MCL feature selection method for P300-based BCI systems.

The EEG measures electrophysiological activity originating from cortical neurons at the surface of the scalp.

Several key types of neural response apparent in the EEG are commonly used in BCI control. These types include the event-related potential (ERP) [4, 5], steady-state visual-evoked potentials (SSVEP) [6], and motor imagery (MI) [7, 8]. In this work, we consider the P300 ERP. The P300 ERP is widely used in BCI systems due to its high temporal predictability and signal-to-noise ratio (SNR), thereby enabling reliable detection without the need for extended user training. The P300 signal is a positive deflection in the EEG recorded over parietal and occipital EEG electrodes, which usually occur approximately 300 ms after the onset of an unexpected (out-of-sequence) stimulus [9]. The commonly used P300 speller BCI, based on the oddball paradigm, is initially designed by Farewell and Donchin in 1988 [10]. The amplitude of the P300 signal is small, and the EEG collected during the use of the P300 speller has low SNR. Consequently, the performance of the P300 speller is susceptible to disturbance from artifacts, such as eye blink, muscular movements, and environmental noise, which can lead to low recognition performance. Consequently, considerable research effort has focused on developing filter, feature extraction, feature selection, and classification methods to improve the P300 speller system's performance.

In general, P300-based systems usually require considerable amounts of training data to build accurate classification models [11]. EEG is typically obtained with high sample rates by using dense placements of large numbers of EEG. For this reason, a large number of possible sample features may result in overfitting and data redundancy when training classification models. Therefore, an effective feature selection method should be developed to reduce dimensionality. Feature selection may be achieved using machine learning methods, which can be trained to discard irrelevant data for improved classification model accuracy. To date, some studies described various feature selection methods for P300-based BCI systems. For example, Citi et al. [12] applied a wrapper-based feature selection approach by using a genetic algorithm to identify optimal feature sets. Long et al. [13] came up with an iterative semisupervised support vector machine for joint spatiotemporal feature selection and classification by using labeled training data and unlabeled test data. Typically, many P300 spellers use some fixed EEG channels, a choice that is based on the assumption that the P300 may be reliably spatially localized within the same specific brain region overall users. However, selecting fixed EEG channels is not guaranteed for all individuals and mental tasks [14]. The channel selection may be used to remove task-irrelevant and redundant channels to overcome this spatial variability [15]. For example, Colwell et al. [16] proposed a new channel selection method called jumpwise regression. Experimental results indicate that this active channel selection method can enhance P300 classification performance compared with the use of a standard channel set. In another example, Yu et al. [17] proposed a group-sparse Bayesian model to identify channels automatically. The classification accuracy of their method is higher or comparable with those of traditional fixed-channel methods. Thus, developing an efficient feature/channel selection strategy offers considerable benefits to P300 classification.

Classical feature selection can be categorized into three types [18], i.e., filter, wrapper, and embedded methods. Filter methods [19] evaluate the correlation of feature subsets as scores, which are independent of learning methods. Features are then ranked in accordance with the score. Finally, the optimal feature set based on selection criterion is retained. Wrapper methods [20] consider the specific classification model, that is, the prediction performance of the specific model is assessed as the selection criterion. The optimal feature subset is identified by obtaining the highest performance. Embedded methods provide a tradeoff between filter and wrapper methods [21] by automatically embedding the feature selection into the training process of the learner. In recent years, hybrid feature selection methods are extensively used to process EEG signals. Specifically, filter methods integrate two or more methods and take advantage of those methods to achieve improved performance. By contrast, filter methods are computationally efficient and learning invariant and can be combined with any machine learning model. This finding can allow considerable reduction to the run time of machine learning algorithms. Sahu et.al [22] applied frequent pattern mining (FPM) for feature selection, and association rule mining was created and a ranking method was used, which reduced feature space and work extracted features using wavelet and power spectral density on eight channels of P300 speller. Another research has used the Genetic Algorithm (GA) and High-Performance Computing. The developed feature selection model saved 89% of the original time consumed [23]. Based on P300 based BCI dataset, GA as a search evolutionary algorithm and Linear Discriminant Analysis (LDA) as a supervised learning classifier formed the feature selection model, which saved 98.1% of the original consumed time and maintained an accuracy rate of 72.5% selecting 46.2% of the original features only [24]. Alharbi et.al [25] presented an improved Branch and bound (BB) algorithm for feature selection using an approximate monotonic criteria function, which enhanced the algorithm noticeably in terms of the number of computations and tree size.

In recent years, many processing algorithms have achieved great success in image processing problems and then used in the BCI research, such as median filtering and facet method[26], ICA[27]. According to the study of Ganguly[28], it has been proven that curvature points preserve more detailed information about the face surface when human face recognition is experienced. During the process of creating edge curvature images, the data points below a threshold are discarded to retain the most informative data from a combination of images, which can be regarded as the feature selection strategy. In addition, there is previous work showed that curvature as wave measure has been employed for feature extraction[29], which is suitable to process P300 signal. Therefore, we give the hypothesis that whether the curvature value of EEG signal can be used to select features and yield superior performance. For this reason, we put forward the curvature ranking as the criterion to select the feature subset. Based on the experimental results evaluated on three public datasets, we can make the conclusion that the curvature method make a positive effect on feature selection and gain the best performance by combining it with LDA theory. In this work, we incorporate two kinds of filter methods into our feature selection method: Menger curvature and linear discriminant analysis (MCL). Our method is described in the following procedures. First, considering a given set of candidate features (a given EEG dataset), our proposed selected strategies are applied to estimate gain values for each feature separately. Each generated set of gains is then ranked in descending order and judged by a predefined criterion. As a result, the intersection of the two approaches is selected as the optimal feature subset. Fig.1 describes the overall framework of our proposed MCL feature selection method for P300-based BCI systems. The whole process can be divided into training and testing phases. The optimal feature subset of training data is selected and then the corresponding feature of testing data is classified to recognize characters.

The remainder of this work is organized as follows. Section 2 describes our methods, the experimental datasets used in this work, and the corresponding data processing methods. Section 3 describes the result. Section 4 gives a discussion. Section 5 presents our conclusions.

2. Method and Materials

In this section, we present our proposed method for selecting optimal features. At first, we describe four traditional filter-based feature selection methods, i.e., mutual information (MI)-, F-test-, T-test-, and Kolmogorov–Smirnov (KS) test-based methods. Then, our proposed feature selection method that combines two kinds of feature selection methods is introduced to solve the problem of high data dimensionality. In our selection strategy, the number of features selected from the training data for each participant is inconsistent.

2.1 Feature selection methods

2.1.1 Mutual Information

Mutual information (MI) [30] is derived from information theory and can be used to measure statistical relationships between features. This method uses the MI between features and labels to determine the optimal feature order. Given a set of discretized feature values X and class labels Y, MI $I(X;Y)$ can be defined as A mutual information [31]:

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)log(\frac{p(x,y)}{p(x)p(y)}) \qquad (1)$$

where $p(x,y)$ is the joint probability distribution function, and $p(x)$ and $p(y)$ are the marginal probability distribution functions for X and Y, respectively. This can be also expressed as follows:

$$I(X;Y)=H(X)-H(X|Y) \qquad (2)$$

where $H(X)$ is the marginal entropy, and $H(X|Y)$ is the conditional entropy. The MI method is used to compute the information gain among features and between features and labels. It can result in a ranking of features, and a chosen number of features with the highest values can be selected.

2.1.2 F-test

The F-test [32] gives an F-score by computing the ratio of variances between features. The F-test assumes the data following a Gaussian distribution and determines if two samples have the same variance. Thus, this method focuses on selecting features capable of separating classes by differentiating variances. The F-score can be given by:

$$F = \frac{V_b}{V_w} \tag{3}$$

where V_b is the variance between groups indicated by the target feature, and V_w is the sum of variances within each group. High F-scores indicate that the distances within groups are minimal, and distances between groups are increased. In this feature selection method, features are ranked and selected on the basis of their F-scores.

2.1.3 T-test

The T-test [33] is often used to determine whether a significant difference is observed between the means of two groups of features by calculating the ratio between the difference of two class means and the variability of two classes. The T-test assumes that two groups have equal standard deviations and are normally distributed. This method selects features in accordance with their capability to differentiate different classes from their mean values. The p-value estimate of the significance of the T-test reflects the probability of a finding occurring by chance alone and is used to identify a significant difference between two groups. In this paper, we explore the use of two-sample T-tests to select features.

2.1.4 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (KS) test [34] is a nonparametric test for measuring the equality of continuous variables that do not make any assumption about how the data are distributed. The KS test measures the maximum difference between the cumulative distribution of two random variables. The empirical distribution function (F_n) for n independent and identical observations X_i is defined as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x} \tag{4}$$

where $I_{X_i \le x}$ represents the indicator function. If $X_i \le x$, it is one; otherwise, it is zero. In this case, the KS test statistics can be denoted by:

$$D = \max_x |F_n(x) - F(x)| \tag{5}$$

A high D value results in high difference between the two classes, indicating a strong ability to distinguish between positive and negative samples. In this paper, we use a two-sample KS test to select features.

2.2 Proposed MCL feature selection method

In this section, we first introduce two filter methods for feature selection, i.e., MC- and LDA-based feature selection methods. We then combine these feature selection methods to identify relevant information from EEG data, called MCL method in this paper.

The high-dimensional input data set to our feature selection method is $X = \{x_1, \dots, x_{N_t}\}$, where N_t is the number of samples, and each sample is represented by an N_s-dimensional vector (i.e., $x_i \in \mathbb{R}^{N_s}$). N_s features are present for all sample data in X. The corresponding single-output feature is denoted by: $y_i \in \mathbb{R}^1, i = 1, \dots, N_t$. At first, the input data set X is decomposed into N_s 2-dimensional (2-D) planes, which is performed by combining all input features $F'_j (1 \le j \le N_s)$ and output feature y, denoted as $P_{(F'_j, y)}$. As a result, the high-dimensional data can be broken down into a set of corresponding decomposed 2-D planes. Then, for each 2-D plan P_i^j, the MC method [35] is applied to calculate the average curvature of the feature F'_j.

The MC measures the curvature of a triplet of points in N-D Euclidean space E^N, which is the inverse of the radius of the circumcircle that passes through three points, i.e., A, B, and C (Fig. 2). The MC on point B can be denoted as:

$$MC(A, B, C) = \frac{1}{R} = \frac{2\sin(\theta)}{\|A, C\|} \tag{6}$$

where R is the radius of a circle, including points A, B, and C; where the triplet points A, B, and C are the sample features which are in the same feature dimension and different samples. Therefore, the MC values of all samples in the same dimension are obtained when each calculation is carried out. $||A, C||$ represents the Euclidean distance between points A and C; and θ is the angle of point B that exists in triangle spanned by points A, B, and C. According to the law of cosines, this results in the following formula:

$$cos(\theta) = \frac{\|A, B\|^2 + \|B, C\|^2 - \|A, C\|^2}{2 \cdot \|A, B\|^2 \cdot \|B, C\|^2} \tag{7}$$

Notably, two boundary points are not calculated. In Fig. 2, points A and C are boundary points.

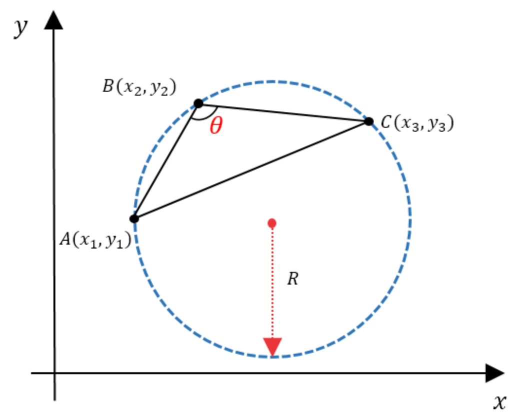

Fig. 2 Menger curvature of a triplet of data points on a single 2-D space.

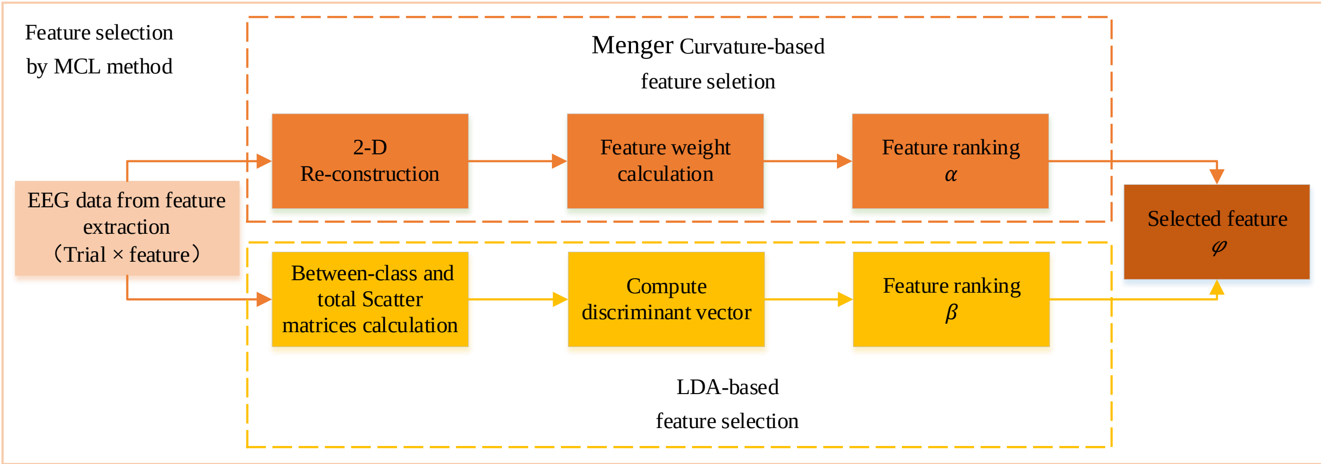

Fig. 3 Flowchart of the proposed MCL feature selection method.

Therefore, the mean MC for feature F_j' can be calculated as:

$$\widehat{\mathrm{MC}}_{F_j'} = \frac{1}{N_t - 2} \sum_{i=1}^{N_t - 1} MC_{N_{ti}}^{j} \tag{8}$$

where $MC_{N_{ti}}^{j}$ is the curvature of the N_{ti}^{th} data point in feature F_j'. $\widehat{\mathrm{MC}}_{F_j'}$ represents the corresponding weight of the feature. A high value indicates a high importance of the corresponding feature for input data set X. In the end, we adopt the traditional descending method to rank features on the basis of the mean of the MC value and select the first d largest values. During the process of Menger Curvature-based feature selection in our proposed method, it can be divided into three steps, i.e., 2-D data reconstruction, feature weight calculation by MC, and feature ranking. The EEG data after feature extraction is broken into data of 2-D planes, and then the feature weight is calculated and ranked, finally according to the feature selection strategy achieve the optimal feature subset.

The input data set $X \in \mathbb{R}^{N_t \times N_s}$, where N_t and N_s are the number of samples and features, and the corresponding label $y_i \in \{-1, +1\}, i = 1, \dots, N_t$. We denote N_1 as the number of samples for $y = 1$, and N_2 as the number of samples for $y = 2$. The objective function to compute the discriminant vector $w \in \mathbb{R}^D$ is described as:

$$\mathrm{J}(w) = \frac{(\mu_1 - \mu_2)}{\sigma_1^2 + \sigma_2^2} \tag{9}$$

where

$$\mu_k = \frac{1}{N_k} \sum_{i \in class\ k} w^T x_i$$

$$\sigma_k^2 = \sum_{i \in class\ 1} (w^T x_i - \mu_k)^2 \quad k = 1,2$$

The between-class S_B and total S_W scatter matrices of LDA are computed as follows:

$$S_B = (m_1 - m_2)(m_1 - m_2)^T \tag{10}$$

$$S_w = \sum_{k=1}^{2} \sum_{i \in class\ k} (x_i - m_k)(x_i - m_k)^T \tag{11}$$

where

$$m_k = \frac{1}{N_k} \sum_{i \in class\ k} x_i$$

The eigen-equation of LDA is defined as follows:

$$S_B w = \lambda S_w w \tag{12}$$

The first d largest eigenvalues are selected and denoted by $\{\lambda_1, \dots, \lambda_d\}$. A feature selection criterion is required to measure the relevance of each feature after ranking and remove an irrelevant feature. Certain evaluation criteria where we retain 95% of all number of features are used and are similar to the principal component analysis (PCA). As a result, the feature set of the MC-based feature selection method is $\alpha = \{x_1', \dots, x_d'\}$. For consistency with the curvature method, we retain 95% of all features as the selected criteria. Therefore, the corresponding selected features set are $\beta = \{x_1'', \dots, x_d''\}$. In the end, the selected feature set φ that results from combining the MC- and LDA-based feature selection methods can be described as $\varphi = \{\alpha \cap \beta\}$. Notably, different input data result in a different numbers of selected features. The overall feature selection process is shown in Fig. 3.

2.3 Dataset description

In this study, we evaluate our feature selection method with three public datasets, i.e., the BCI Competition III dataset II, BNCI Horizon dataset, and EPFL dataset.

The BCI Competition III dataset II, called Dataset I in this study, consists of two EEG data from two participants, who are denoted as participants A and B. The data from each participant contain training and testing subsets of data. Each

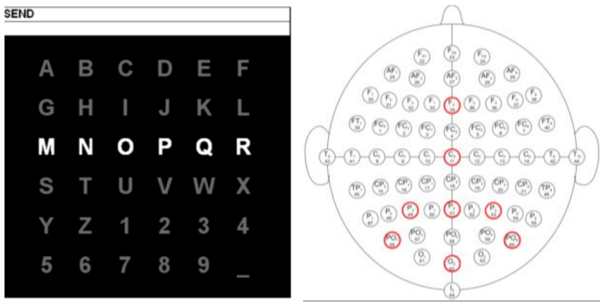

Fig. 4 Interface used in a P300 speller BCI (left). The electrode positions recorded the EEG data (right).

subset is recorded participants view characters. A total of 85 and 100 characters are in the training and testing subsets, respectively. In the experiment, participants are presented with a grid of characters. All rows and columns of this letter grid are flashed at a rate of 5.7 Hz, and the brightness of the grid is intensified for 100 ms followed by a blank presentation for 75 ms. An epoch in the experiment contains 12 trials/sample, in which every row and every column are flashed once in random order, and two of the flashes contain the target character in that epoch. The sets of 12 intensifications are repeated 15 times for each target character, resulting in 180 total intensifications for each character epoch. A set of 64 EEG channels is placed on different positions of the participant's scalp for recording EEG signals. The collected EEG signals are filtered using a bandpass filter with a cutoff frequency of 0.1–60 Hz and digitized at 240 Hz. Details on Dataset I can be found at http://www.bbci.de/competition/iii/desc_II.pdf.

The BNCI Horizon dataset, called Dataset II in this study, consists of eight EEG data from participants (S1–S8) with amyotrophic lateral sclerosis [36] . A total of 35 predefined characters are given for each participant to spell using a P300-based BCI speller. We assign the EEG signals recorded from spelling the first 20 characters to training data and the EEG data from the rest of the character spelling attempts to testing data. Eight channels, placed in accordance with 10–10 standard are used for collecting the EEG data. Specifically, channels Fz, Cz, Pz, Oz, P3, P4, PO7, and PO8 are used. The collected EEG signals are filtered using a bandpass filter with a cutoff frequency of 0.1–30 Hz and digitized at 256 Hz. In the experiment, all rows and columns of the letter grid are intensified for 125 ms with an interstimulus interval of 125 ms. For each character, all rows and columns are intensified 10 times. Details about Dataset II can be found at https://lampx.tugraz.at/~bci/database/008-2014/description.pdf. In addition, Datasets I and II share the same interface as the P300 speller, as shown in Fig. 4. Fig. 4 also illustrates the configuration of the 64 electrode positions based on the international 10–20 system as used in Dataset I, where the eight channels of Dataset II are indicated with red circles.

The EPFL dataset, called Dataset III in this study, covers EEG data of 4 disabled and 4 healthy participants (P1–P8). A 6-choice P300 paradigm is applied on the experiment, and participants are required to focus on desired target images. Images are shown in random order, and each image flashing lasts for 100 ms and followed by no flash for 300 ms. Four sessions are completed for each participant, and each session consists of six runs. An average of 22.5 epochs of six flashes each run is observed. In this study, we randomly choose two sessions as training data, and the rest of the sessions as testing data. The sample rate of EEG data is 2048 Hz recorded from 32 electrodes (i.e., Fp1, AF3, F7, F3, FC1, FC5, T7, C3, CP1, CP5, P7, P3, Pz, PO3, O1, Oz, O2, PO4, P4, P8, CP6, CP2, C4, T8, FC6, FC2, F4, F8, AF4, Fp2, Fz, and Cz) according to the international 10–20 system. Details on Dataset III can be found at https://www.epfl.ch/labs/mmspg/research/page-58317-en-html/bci-2/bci_datasets/.

2.4 Experiment setup

Considering that raw EEG signals contain various kinds of noise that are difficult to distinguish from neural signals, EEG data should be preprocessed to improve SNR. The typical processing pipeline for EEG data consists of signal preprocessing, downsampling, feature extraction, and feature selection. The feature extraction of training and testing parts keep consistent. As Fig.1 shows, the temporal features are extracted. The type of data changes from channel × time to channel × trial × time, and temporal features are concatenated together across channels in a single feature vector for each trial.

For Dataset I, the 64-channel EEG data is bandpass filtered from 0.5 Hz to 20 Hz, and the Parks–McClellan optimal equiripple finite impulse response is used for bandpass

Table 1
Number of original and selected features for all participants in the three datasets.

Participants	Dataset I		Dataset II								Dataset III							
	A	B	S1	S2	S3	S4	S5	S6	S7	S8	P1	P2	P3	P4	P5	P6	P7	P8
Original number	1152		160								640							
Selected number	1055	1049	146	147	148	149	146	146	146	148	582	583	581	581	582	585	586	581

Table 2
Comparison of character recognition accuracy (%) for Dataset I with different classification methods.

Method	Epochs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	p
xDAWN	A	12	30	44	49	56	69	74	78	80	86	90	91	92	94	**95**	**0.002**
	B	41	65	74	76	81	88	89	93	93	94	95	96	93	95	**95**	
	AVG	26.5	47.5	59	62.5	68.5	78.5	81.5	85.5	86.5	90	92.5	**93.5**	**92.5**	**94.5**	**95**	
STDA	A	17	31	41	50	57	60	64	68	78	84	85	88	87	91	92	0.7
	B	40	57	66	69	77	81	86	87	91	94	95	97	99	96	97	
	AVG	28.5	44	53.5	59.5	67	70.5	75	77.5	84.5	89	90	92.5	93	93.5	94.5	
FC+LDA	A	11	17	35	44	48	52	55	66	69	76	79	86	83	86	86	**9.97×10^{-6}**
	B	34	48	58	65	73	81	82	86	90	92	91	94	93	92	94	
	AVG	22.5	32.5	46.5	54.5	60.5	66.5	68.5	76	79.5	84	85	90	88	89	90	
SWLDA	A	16	24	41	46	52	59	65	69	75	80	81	85	87	88	88	**0.004**
	B	39	60	61	69	69	81	82	83	88	87	90	93	91	92	93	
	AVG	27.5	42	51	57.5	60.5	70	73.5	76	81.5	83.5	85.5	89	89	90	90.5	
Proposed method	A	10	27	35	48	47	60	64	71	78	84	85	90	89	92	96	/
	B	32	63	68	70	79	82	86	91	91	92	96	97	97	98	99	
	AVG	21	45	51.5	59	63	71	75	81	84.5	88	90.5	93.5	93	95	97.5	

"AVG" means the averaged classification accuracy of all participants.

Table 3
Average character recognition accuracy (%) for Dataset I with different feature selection methods when using LDA as classification method.

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	p
MI	29	47.5	55	59	64	73.5	75.5	82.5	84.5	**87**	**86.5**	**90**	**92**	**92.5**	**94.5**	0.71
F-test	27	45	53.5	57	67.5	71	74.5	79	82.5	**86**	**87**	**89.5**	**94**	**94.5**	**94.5**	0.59
T-test	25.5	45.5	54	56	64.5	71.5	73.5	79.5	84	**86.5**	**86**	**88.5**	**92**	**92**	**94**	0.15
KS-test	27	46.5	56.5	57	64.5	72	77	81.5	83	**85.5**	**86.5**	**89.5**	**92.5**	**93.5**	**94**	0.86
KS-T test	25.5	47	54	58	65	72	76	79	84	**85.5**	**87**	**88**	**91.5**	**92**	**94.5**	0.38
KS-F test	27	45	54.5	58	64	71.5	75	79.5	82.5	**85.5**	**86**	**90**	**92.5**	**92.5**	**93.5**	0.30
T-F test	26	43.5	54	56.5	65	71.5	75	79.5	82.5	**86**	**86**	**90.5**	**92.5**	**93**	**94**	0.20
MC	26	48	53	57	60	71	73	81	80.5	**87**	**86.5**	**90.5**	**93**	**94**	**95**	**3.6×10^{-9}**
MCL	21	45	51.5	59	63	71	75	81	84.5	**88**	**90.5**	**93.5**	**93**	**95**	**97.5**	-

filtering. Based on the knowledge of the P300 signal, a positive deflection at approximately 300 ms after stimulus onset is observed. We extract EEG at a range of 0–600 ms from the stimulus presentation from each EEG channel in each trial. The filtered data are then decimated by a factor of 8 to remove high frequencies and reduce the dimensionality of feature vectors, that is, we downsample from 240 Hz to 30 Hz. Therefore, we obtain an input feature vector with size of 64 × 18, where 64 is the number of channels, and 18 is the number of samples per channel. The signal is then normalized to 0 mean and unit variance. The numbers of training and testing trials are 15 300 and 18 000, respectively.

For Dataset II, the 8-channel EEG data are filtered with a second-order butterworth bandpass filter from 0.1 Hz to 30 Hz. A Notch filter of 50 Hz is applied to reduce electrical noise. An EEG epoch from 0 ms to 600 ms after the stimulus presentation from each channel is extracted and downsampled from 256 Hz to 32 Hz. The resulting size of the feature vector for each stimulus is 8 × 20, where 8 is the number of channels, and 20 is the number of samples per channel. The numbers of training and testing trials are 7020 and 2400, respectively.

For Dataset III, the EEG signal is filtered by a sixth-order Butterworth bandpass filter, and the cutoff frequency is from 1 Hz to 12 Hz. Then, the EEG signal is downsampled from 2048 Hz to 32 Hz and extracted with the first 600 ms after stimuli as temporal features. Therefore, the resulting size of the feature vector for each stimulus is 32 × 20, where 32 is the number of channels, and 20 is the number of samples per channel. For each participant, we only use 20 epochs each run for training and 10 blocks each run for testing to validate the performance. The numbers of training and testing trials are 1440 and 720, respectively.

The proposed feature selection algorithm (section 2.2.5) is applied to the feature sets extracted from each dataset and remove minimally informative features. After completing data processing, training data are used to train the classifier. Given its simple implementation, relatively quick training, and nonadjustment of hyperparameters, the LDA classification model is successfully applied to classify the target and nontarget data of P300 signal. Specifically, in our study, we use the LDA classifier to perform character recognition based on the P300 speller. According to the objective function in formula (4), the discriminant vector w satisfies the following equation by computing the derivative of J and setting it to zero.

$$\mathrm{w} \propto S_w^{-1}(m_1 - m_2) \tag{13}$$

The output of the LDA model for a given testing vector $\hat{x}$ is simply the product $w^T \hat{x}$. In the P300-based BCI system, the target trial can be recognized by maximizing the summed output values for each epoch.

3. Result

In this section, we illustrate the performance of our proposed method. We use character recognition accuracy to evaluate the performance of our proposed method. Statistical performance comparison based on paired t-tests [37] between our proposed method and earlier reported methods is adopted to present the significance.

3.1 Performance analysis on Dataset I

In Dataset I, temporal features after feature extraction are concatenated together, and the number of input features to our feature selection method is 1152 (64 channels × 18 sample points) for participants A and B. Two ranked feature sets obtained from the MC- and LDA-based feature selection methods are intersected to obtain the final feature set. The numbers of input features after feature selection are 1055 and 1049 for participants A and B, respectively, which is shown in Table 1.

In order to further visualize our feature selection process, we take participant A as an example and select the first 3, middle 3, and last 3 features to compare the difference of features. The average curvature for nine ranked features (1st, 2nd, 3rd, 575th, 576th, 577th, 1150th, 1151st, and 1152nd) is plotted (Fig. 5 top). Notably, in conjunction with Table 2, different curvature values are present for different features, and increased curvature corresponds to relevant feature information. In this study, we select the ranked features corresponding to the highest values.

The character recognition performance using the proposed feature selection method with an LDA classifier is compared with traditional classification methods, such as xDAWN[38], STDA [39], FC+LDA [40], and SWLDA [41]. These methods do not use a feature selection process. These results are shown in Table 2 and indicate that character recognition performance improves as the number of epochs increases. The proposed method achieves average accuracy values of 63%, 88%, and 97.5% after 5, 10, and 15 epochs, respectively. As shown in Table 2, our proposed method obtains 99% accuracy for participant B after 15 epochs. Results reveal that our proposed method yields improved performance, with a significant increase between our proposed method and the xDAWN, FC+LDA, and SWLDA techniques ($p < 0.05$). Although no significant difference is present between our proposed method and STDA, our proposed method can achieve superior accuracy after 15 epochs.

The character recognition performance by using our proposed method is compared with traditional filter-based feature selection methods. To ensure consistency between our proposed method and other methods, we select the top 95th percentile of the ranked features via traditional filter-based feature selection methods as the final feature set for comparison. As shown in Table 3, our proposed method achieves higher mean accuracy than the other methods (i.e., MI-, KS test-, T-test-, and F-test-based feature selection methods) after 10 and 15 epochs. It can be found that although the proposed method obtains the best classification performance after 15 epochs compared with other methods, there are no significant differences. And we make further analysis across the final 5 epochs, which can be concluded that there existed significant differences ($p < 0.05$) between the proposed method and different feature selection methods.

Table 4
Comparison of character recognition accuracy (%) with conventional methods in existing works for Dataset I after 15 epochs.

Existed work	Method	Accuracy after 15epochs
Our work	**MCL+LDA**	**97.5**
Rakotomamonjy et al.[48]	Ensemble SVM	96.5
Cecotti et al. [49]	MCNN-1	95.5
Idaji et al. [50]	HOSRDA+LDA	96.5
Kong et al. [51]	PCA+WELM	97
Yu et al. [17]	gsLDA	97

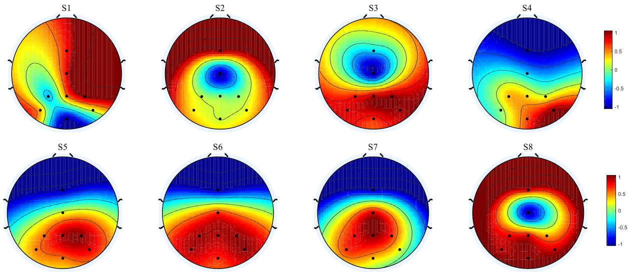

Fig. 6 2-D topographic maps of normalized curvature value across eight channels from eight participants.

In this work, we incorporate two kinds of filter methods into our feature selection method to obtain effective features. In order to validate the superiority of the proposed MCL algorithm, we also combine the KS-test and T-test, called KS-T test; the KS-test and F-test, called KS-F test; the T-test and F-test, called T-F test, and then evaluate them in Dataset I. The character recognition performances are shown in Table 3. It is seen that our feature selection algorithm MCL method still presents the best performance after 15 epochs.

3.2 Performance analysis on Dataset II

In Dataset II, temporal features after feature extraction are concentrated together, and the number of input features to which we apply feature selection is 160 (8 channels × 20 sample points) for each participant. Two ranked feature sets from the MC- and LDA-based feature selection are intersected to obtain the final feature set. The number of features selected for use in classification is shown in Table 1.

Fig. 6 presents the topographic maps of the normalized curvature values integrated over eight channels for eight participants. This finding reveals how feature distributions over channels differ among participants. Combined with Table 5, participant S1 displays low normalized curvature in occipital areas (visual response) and obtains poor performance, e.g., consistent with previous work that has used the same dataset [42]. To further understand the results, we take participant S1 as an example and select the first 2, and last 2 features to compare the difference of features. First, we analyze the number of trials across different ranges of curvature values in participant S1. Statistical results are shown

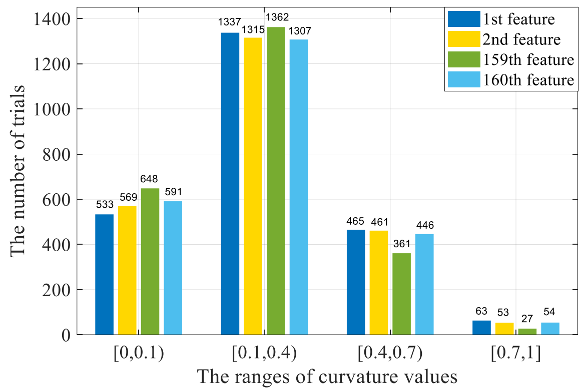

Fig. 7 Number of trials in different ranges of curvature values for participant S1 from Dataset II.

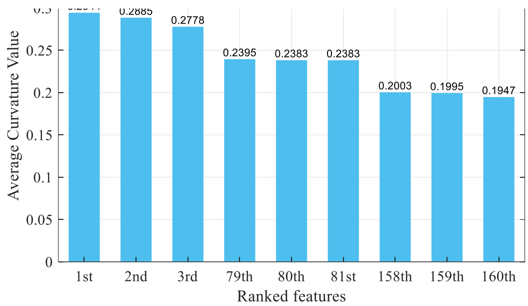

Fig. 5 Average curvature of nine ranked features in Participant A of Dataset I (top) and S1 of Dataset II (bottom).

Table 5
Comparison of character recognition accuracy (%) for eight participants from Dataset II with different classification methods after 10 epochs.

Participants	xDAWN	STDA	FC+LDA	SWLDA	**Proposed method**
S1	93.33	73	80	93.33	**93.33**
S2	100	86.67	80	86.67	**100**
S3	93.33	93.33	80	73.33	**100**
S4	86.67	100	93.33	100	**100**
S5	93.33	66.67	93.33	100	**100**
S6	100	100	100	100	**100**
S7	100	93.33	100	100	**100**
S8	100	100	100	100	**100**
AVG	96	89.17	90.83	94.17	**99.17**

Table 6
Comparison of character recognition accuracy (%) for Dataset II with different feature selection methods when using LDA as classification method.

Method	Epochs										p
	1	2	3	4	5	6	7	8	9	10	
MI	27.5	60.8	71.7	75.8	80.8	89.2	91.7	92.5	95.8	97.5	**2.3e-5**
F-test	30.8	56.7	71.7	75.8	79.2	89.2	90.8	94.2	95.8	96.7	**0.03**
T-test	35	60	74.2	78.3	82.5	90.8	90.8	94.2	96.7	96.7	1
KS-test	32.5	60.8	73.3	79.2	82.5	89.2	90.8	94.2	96.7	98.3	0.7
MC	29.2	60.8	71.7	77.5	82.5	90.8	94.2	94.2	98.3	99.2	0.84
MCL	**29.2**	**62.5**	**74.2**	**77.5**	**83.3**	**90**	**92.5**	**93.3**	**97.5**	**99.2**	/

in Fig. 7. The curvature value was calculated from each feature of the trials. The sum of curvature values of all trials at the same one-dimensional feature was used to rank the feature. The first feature stands for feature dimension corresponding to the highest curvature value. We employed Fig. 7 to make statistics for the number of trials in different ranges of curvature values. For the features with a higher ranking, the larger the range of curvature value is, the more the number of trials is. The average curvature varied for nine ranked features (1st, 2nd, 3rd, 79th, 80th, 81st, 158th, 159th, and 160th) is plotted. These results are shown in Fig. 5 (bottom).

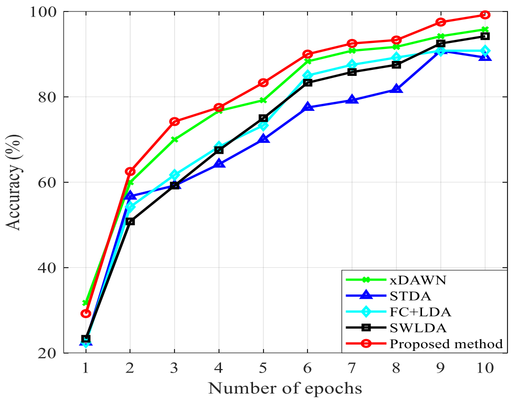

Fig. 8 Average character detection accuracy obtained by the xDAWN, STDA, FC+LDA, and SWLDA methods as well as our proposed method after 1–10 epochs.

We compare the character recognition performance by using our proposed method with traditional classification methods, which do not include a feature selection step. The character recognition accuracies after all epochs (10 epochs) for eight participants are shown in Table 5. The proposed method obtains 100% accuracy for participants S2–S8 after 10 epochs, showing the stable performance of the proposed method. In addition, the accuracy of our proposed method is superior or comparable with those of other methods. Fig. 8 shows the average character detection accuracy obtained by xDAWN, STDA, FC+LDA, and SWLDA methods along with our proposed methods after 1–10 epochs. Results indicate that character recognition performance improves as the number of epochs increases. Our proposed method achieves average accuracy values of 83.3% and 99.2% after 5 and 10 epochs, respectively. These results show that selecting high-ranking features can produce improved performance. Our proposed method yields significantly better performance than traditional classification methods, i.e., xDAWN, STDA, FC+LDA, and SWLDA techniques ($p < 0.05$).

In addition, the character recognition performance of our proposed method is compared with those of traditional filter-based feature selection methods. To ensure consistency between methods, we select the top 95th percentile of the ranked features via traditional feature selection methods and retain these features for our comparison. As shown in Table 6, our proposed MCL method achieves higher mean accuracy than other methods (i.e., MI, KS test, T-test, F-test) after 10 and 15 epochs. In addition, significant improvements are observed between our proposed method and MI ($p < 0.05$) and F-test ($p < 0.05$). Although no significant difference is observed between our proposed method and KS test- and T-test-based selection methods, our proposed method can achieve higher accuracy after 5 and 10 epochs.

Table 7
Average classification accuracy (%) obtained using xDawn, STDA, FC+LDA, LDA, MCL+LDA, SWLDA, and MCL+SWLDA for Dataset III.

Method	Epochs 1	2	3	4	5	6	7	8	9	10	p
xDAWN	62.5	70.8	84.4	91.7	90.6	91.7	93.7	93.7	94.8	95.8	0.46
STDA	60.4	62.5	80.2	81.2	83.3	85.4	87.5	90.6	91.7	92.7	**0.02**
FC+LDA	58.3	63.5	79.2	87.5	86.5	87.5	88.5	87.5	89.6	91.7	**0.01**
LDA	39.6	54.2	64.6	66.7	65.6	75	75	75	77.1	78.1	**1.03e-7**
MCL+LDA	44.8	61.5	69.8	71.9	75	79.2	79.2	82.3	84.4	88.5	**3.40e-6**
SWLDA	55.2	64.6	81.3	88.5	88.5	88.5	88.5	91.7	95.8	95.8	**0.04**
MCL+SWLDA	**52.1**	**68.8**	**81.3**	**89.6**	**91.7**	**92.7**	**95.8**	**95.8**	**95.8**	**96.9**	/

3.3 Performance analysis on Dataset III

In Dataset III, Table 1 describes the number of original and selected features. A total of 640 features is obtained by concatenating 32 channels and 20 sample points. At first, we evaluate the character recognition performance by comparing the proposed method with traditional classification method, and the average classification accuracy values obtained by xDawn, STDA, FC+LDA, LDA, MCL+LDA, MCL+SWLDA, and are shown in Table 7. Our proposed method (MCL as the feature selection method) can achieve significantly higher performance than the LDA method but cannot yield better performance. With the aid of reference [43], SWLDA is used in this study. As shown in Table 7, the MCL method combined with SWLDA has superior performance ($p < 0.05$). In addition, to reflect the superiority of feature selection method, we apply SWLDA as the same classification method and use the Pearson correlation coefficient (PCC) [44], term variance (TV) [45], and

Table 8
Average classification accuracy (%) obtained using PCC, TV, NCR, and MCL feature selection methods for Dataset III when using SWLDA as classification method.

Method	Epochs										p
	1	2	3	4	5	6	7	8	9	10	
PCC	56.3	64.6	80.2	87.5	87.5	89.6	88.5	87.5	92.7	93.8	**0.02**
TV	52.1	68.8	81.3	86.5	89.6	89.6	91.7	93.8	93.8	94.8	**0.003**
NCA	55.2	67.7	82.3	88.5	88.5	89.6	90.6	91.7	92.7	93.8	**0.03**
MC	52.1	64.6	78.1	83.3	90.6	88.5	89.6	91.7	91.7	95.8	**6×10^{-4}**
MCL	**52.1**	**68.8**	**81.3**	**89.6**	**91.7**	**92.7**	**95.8**	**95.8**	**95.8**	**96.9**	/

neighborhood component analysis (NCR) [46] as feature selection for further comparison. Results indicate that the MCL method can produce significant improvements ($p <$ 0.05) than other feature selection methods (Table 8). It should be noted that, MCL method combined the Menger curvature method and linear discriminant analysis method is proposed in this study. For better comparison, we give the recognization performance only using MC method as the feature selection method. Table 3, 6 and 8 show the experimental results based on three public datasets. From the tables, it can be obviously found that there are significant differences between the proposed MCL method and MC method in Dataset I and III. However, MC and MCL methods share similar performance in Dataset II. Combined with Table 1 to further analyze, it may be the fewer number of features that makes MCL method present no significant change in performance. On the whole, the feature selection method combined MC and LDA methods plays a rival and robust role in character recognization.

4. Discussion

In this study, the curvature calculation approach and LDA are used to address the feature selection problem in high-dimensional datasets. Furthermore, we provide two ways to confirm the efficacy of proposed method by using three datasets. First, we compare the proposed method with different classification methods without the feature selection process. Then, the proposed method is compared with different feature selection methods when using the same classifer. Previous study has focused on minizing curvature of the error surface in CNN method based on P300 detection [47], and we explore a novel research approach in current work. In addition, we choose previous studies for comparison and show the classification performance after 5 and 15 epochs of our method and previous work reported in the literature. As the commonly used dataset, Dataset I are discussed and the results are shown in Table 4. Rakotomamonjy and colleagues [48] proposed the channel selection method and ensemble support vector machine classifiers for P300 classification. This approach achieves accuracy values of 96.5% after 15 epochs. Idaji and Graser [49] reported seven classifiers based on the convolutional neural network and achieved the best result by using a multiclassifier solution. The recognition rate reaches 95.5% after 15 epochs. In work by Idaji and colleagues [50], higher-order spectral regression discriminant analysis is used to develop a framework for addressing different regularization constraints in higher-order feature reduction problems. This approach is used to achieve average accuracies of 96.5% after 15 epochs, respectively. Kong and colleagues [51] used PCA with a weighted extreme learning machine method to improve the P300 detection accuracy. This approach yields accuracy values of 97% after 15 epochs. Finally, Yu and colleagues [17] used an embedded channel selection group-sparse Bayesian LDA. This approach achieves accuracy values of 97% after 15 epochs. We can see that our proposed approach achieves the highest accuracy performance after 15 epochs. All results indicate that the proposed MCL feature selection method obtains superior performance. We also visualize the normalized curvature in topographic maps, which are consistent with classification results and prior study [52].

Besides, we visualize our feature selection process to compare the difference of features. The average curvature for nine ranked features is designed. Notably, different curvature values are present for different features, and increased curvature corresponds to relevant feature information. Furthermore, topographic maps of the normalized curvature values in Fig. 6 are shown that how feature distributions over channels differ among participants. Combined with Table 5, S1 of Dataset II displays low normalized curvature in occipital areas (visual response) and obtains poor performance. Recently, researchers have started to develop novel visual evoked potentials to encode the character recognization, such as miniature asymmetric visual evoked potentials (aVEPs) [53] . In future work, we will evaluate our decoding algorithm using different event-related potentials. In addition, in the study of Xiao’s work, multi-window discriminative canonical pattern matching (DCPM) achieved the character recognition accuracy of 91.84% especially using a small calibration dataset [54], which can be a promising method for enhancing the practicability of P300-speller. However, the datasets applied in this study consist of an amount of training data to construct the model. On the basis of the results of this study, further work will pursue high decoding accuracies with few epochs to widen the range of practical applications of P300 BCI systems.

5. Conclusion

Feature selection, as a data preprocessing strategy, is proven to be effective and efficient in dealing with high-dimensional data. In this work, we present a novel hybrid feature selection method for P300 character recognition to reduce data redundancy and improve signal quality. We propose to use the MCL method to select an optimal feature set by combining the Menger curvature method with linear discriminant analysis for feature selection. Our proposed method is independent of the classification model. We evaluate the efficacy of our proposed method with three public datasets (i.e., BCI Competition III dataset II, BNCI Horizon dataset, and EPFL dataset). Results indicate that our proposed method can achieve the best performance and significantly outperform most traditional filter-based feature selection and classification methods.

Acknowledgments

This work was supported by the National key research and development program 2017YFB13003002. This work was also supported in part by the Grant National Natural Science Foundation of China, under Grant Nos.61573142 and 61773164, the programme of Introducing Talents of Discipline to Universities (the 111 Project) under Grant B17017, and the "ShuGuang" project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation under Grant 19SG25. This work was also supported by the Ministry of Education and Science of the Russian Federation (grant 14.756.31.0001) and Polish National Science Center (UMO-2016/20/W/NZ4/00354).

References

[1] Bulárka S and Gontean A 2016 Brain-computer interface review *2016 12th IEEE International Symposium on Electronics and Telecommunications (ISETC)* pp 219–22

[2] Wolpaw J R *et al* 2000 Brain-computer interface technology: a review of the first international meeting *IEEE Trans. Neural Syst. Rehabil. Eng.* **8** 164–73

[3] Power S D, Falk T H and Chau T 2010 Classification of prefrontal activity due to mental arithmetic and music imagery using hidden Markov models and frequency domain near-infrared spectroscopy *J. Neural Eng.* **7** 26002

[4] Coles M G H and Rugg M D 1995 *Event-related brain potentials: An introduction* (Oxford University Press)

[5] Jin J, Chen Z, Xu R, Miao Y, Wang X and Jung T-P 2020 Developing a novel tactile P300 brain-computer interface with a cheeks-stim paradigm *IEEE Trans. Biomed. Eng.* **67** 2585–93

[6] Chen X, Wang Y, Gao S, Jung T-P and Gao X 2015 Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain-computer interface *J. Neural Eng.* **12** 46008

[7] Sitaram R, Zhang H, Guan C, Thulasidas M, Hoshi Y, Ishikawa A, Shimizu K and Birbaumer N 2007 Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain-computer interface *NeuroImage* **34** 1416–27

[8] Jin J, Xiao R, Daly I, Miao Y, Wang X and Cichocki A 2020 Internal feature selection method of CSP based on L1-norm and dempster-shafer theory *IEEE Trans. Neural Netw. Learn. Syst.*

[9] Mao Y, Jin J, Xu R, Li S, Miao Y and Cichocki A 2021 The influence of visual attention on the performance of a novel tactile P300 brain-computer interface with cheeks-stim paradigm *Int J Neural Syst* **31** 2150004

[10] Farwell L A and Donchin E 1988 Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials *Electroencephalography and clinical Neurophysiology* **70** 510–23

[11] Jin J, Li S, Daly I, Miao Y, Liu C, Wang X and Cichocki A 2019 The study of generic model set for reducing calibration time in P300-based brain-computer interface *IEEE Trans. Neural Syst. Rehabil. Eng.* **28** 3–12

[12] Citi L, Poli R, Cinel C and Sepulveda F 2008 P300-based BCI mouse with genetically-optimized analogue control *IEEE Trans. Neural Syst. Rehabil. Eng.* **16** 51–61

[13] Long J, Gu Z, Li Y, Yu T, Li F and Fu M 2011 Semi-supervised joint spatio-temporal feature selection for P300-based BCI speller *Cogn. Neurodyn.* **5** 387

[14] El Dabbagh H and Fakhr W 2011 Multiple classification algorithms for the BCI P300 speller diagram using ensemble of SVMs *2011 IEEE GCC Conference and Exhibition (GCC)* pp 393–6

[15] Wang H, Shi W and Choy C-S 2017 Integrating channel selection and feature selection in a real time epileptic seizure detection system *2017 39th Annual International Conference of the IEEE*

Engineering in Medicine and Biology Society (EMBC) pp 3206–11

[16] Colwell K A, Ryan D B, Throckmorton C S, Sellers E W and Collins L M 2014 Channel selection methods for the P300 Speller *J. Neurosci. Methods* **232** 6–15

[17] Yu T, Yu Z, Gu Z and Li Y 2015 Grouped automatic relevance determination and its application in channel selection for P300 BCIs *IEEE Trans. Neural Syst. Rehabil. Eng.* **23** 1068–77

[18] Jin J, Liu C, Daly I, Miao Y, Li S, Wang X and Cichocki A 2020 Bispectrum-based channel selection for motor imagery based brain-computer interfacing *IEEE Trans. Neural Syst. Rehabil. Eng.* **28** 2153–63

[19] Cateni S, Colla V and Vannucci M 2017 A fuzzy system for combining filter features selection methods *International Journal of Fuzzy Systems* **19** 1168–80

[20] Gao W, Hu L and Zhang P 2018 Class-specific mutual information variation for feature selection *Pattern Recognit* **79** 328–39

[21] Li J, Cheng K, Wang S, Morstatter F, Trevino R P, Tang J and Liu H 2017 Feature selection: A data perspective *ACM Computing Surveys (CSUR)* **50** 1–45

[22] Sahu M, Verma S, Nagwani N K and Shukla S 2020 EEG signal analysis and classification on P300 speller-based BCI performance in ALS patients *International Journal of Medical Engineering and Informatics* **12** 375–400

[23] Alwadei S, Dahab M and Kamel M 2017 A Feature Selection Model based on High-Performance Computing (HPC) Techniques *International Journal of Computer Applications* **180** 11–6

[24] Alwadei S, Dahab M and Kamel M 2019 High performance GA-LDA feature selection model for Brain-Computer Interface data *International Journal of Information Technology* **3** 1–12

[25] Alharbi A N and Dahab M 2020 An improvement in branch and bound algorithm for feature selection *International Journal of Information Technology and Language Studies* **4**

[26] Mirghasemi H, Shamsollahi M B and Fazel-Rezai R 2006 Assessment of preprocessing on classifiers used in the P300 speller paradigm *2006 International Conference of the IEEE Engineering in Medicine and Biology Society* pp 1319–22

[27] Kachenoura A, Albera L, Senhadji L and Comon P 2007 ICA: a potential tool for BCI systems *IEEE Signal Process. Mag.* **25** 57–68

[28] Ganguly S, Bhattacharjee D and Nasipuri M 2017 Fuzzy matching of edge and curvature based features from range images for 3D face recognition *Intelligent Automation & Soft Computing* **23** 51–62

[29] Gevins A S, Yeager C L, Diamond S L, Spire J, Zeitlin G M and Gevins A H 1975 Automated analysis of the electrical activity of the human brain (EEG): A progress report *Proceedings of the IEEE* **63** 1382–99

[30] Liu H, Sun J, Liu L and Zhang H 2009 Feature selection with dynamic mutual information *Pattern Recognit* **42** 1330–9

[31] Hoque N, Bhattacharyya D K and Kalita J K 2014 MIFS-ND: A mutual information-based feature selection method *Expert Syst. Appl.* **41** 6371–85

[32] Dhanya R, Paul I R, Akula S S, Sivakumar M and Nair J J 2020 F-test feature selection in Stacking ensemble model for breast cancer prediction *Procedia Computer Science* **171** 1561–70

[33] Dong A and Wang B 2009 Feature selection and analysis on mammogram classification *2009 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing* pp 731–5

[34] Biesiada J and Duch W 2005 Feature selection for high-dimensional data: A kolmogorov-smirnov correlation-based filter *Computer Recognition Systems* (Springer) pp 95–103

[35] Zuo Z, Li J and Moubayed N A 2021 Curvature-based Feature Selection with Application in Classifying Electronic Health Records *arXiv preprint arXiv:2101.03581*

[36] Sahu M, Nagwani N and Verma S 2016 Applying auto regression techniques on amyotrophic lateral sclerosis patients EEG dataset with P300 speller *Indian Journal of Science and Technology* **9** 1–8

[37] Mladenovic J, Frey J, Joffily M, Maby E, Lotte F and Mattout J 2020 Active inference as a unifying, generic and adaptive framework for a P300-based BCI *J. Neural Eng.* **17** 16054

[38] Rivet B, Souloumiac A, Attina V and Gibert G 2009 xDAWN algorithm to enhance evoked potentials: application to brain-computer interface *IEEE Trans. Biomed. Eng.* **56** 2035–43

[39] Zhang Y, Zhou G, Zhao Q, Jin J, Wang X and Cichocki A 2013 Spatial-temporal discriminant analysis for ERP-based brain-computer interface *IEEE Trans. Neural Syst. Rehabil. Eng.* **21** 233–43

[40] Pires G, Nunes U and Castelo-Branco M 2011 Statistical spatial filtering for a P300-based BCI: tests in able-bodied, and patients with cerebral palsy and amyotrophic lateral sclerosis *J. Neurosci. Methods* **195** 270–81

[41] Krusienski D J, Sellers E W, Cabestaing F, Bayoudh S, McFarland D J, Vaughan T M and Wolpaw J R 2006 A comparison of classification techniques for the P300 Speller *J. Neural Eng.* **3** 299

[42] Riccio A, Simione L, Schettini F, Pizzimenti A, Inghilleri M, Olivetti Belardinelli M, Mattia D and Cincotti F 2013 Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis *Front. Hum. Neurosci.* **7** 732

[43] Krusienski D J, Sellers E W, McFarland D J, Vaughan T M and Wolpaw J R 2008 Toward enhanced P300 speller performance *J. Neurosci. Methods* **167** 15–21

[44] Zhou H, Deng Z, Xia Y and Fu M 2016 A new sampling method in particle filter based on Pearson correlation coefficient *Neurocomputing* **216** 208–15

[45] Liu L, Kang J, Yu J and Wang Z 2005 A comparative study on unsupervised feature selection methods for text clustering *2005 International Conference on Natural Language Processing and Knowledge Engineering* pp 597–601

[46] Qin C, Song S, Huang G and Zhu L 2015 Unsupervised neighborhood component analysis for clustering *Neurocomputing* **168** 609–17

[47] Shojaedini S V, Morabbi S and Keyvanpour M R 2021 A New Method to Improve the Performance of Deep Neural Networks in Detecting P300 Signals: Optimizing Curvature of Error Surface Using Genetic Algorithm *Journal of Biomedical Physics & Engineering* **11** 357

[48] Rakotomamonjy A, Guigue V, Mallet G and Alvarado V 2005 Ensemble of SVMs for improving brain computer interface P300 speller performances *International conference on artificial neural networks* pp 45–50

[49] Cecotti H and Graser A 2010 Convolutional neural networks for P300 detection with application to brain-computer interfaces *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 433–45

[50] Idaji M J, Shamsollahi M B and Sardouie S H 2017 Higher order spectral regression discriminant analysis (HOSRDA): A tensor feature reduction method for ERP detection *Pattern Recognit* **70** 152–62

[51] Kong W, Guo S, Long Y, Peng Y, Zeng H, Zhang X and Zhang J 2018 Weighted extreme learning machine for P300 detection with application to brain computer interface *Journal of Ambient Intelligence and Humanized Computing* 1–11

[52] Bell C J, Shenoy P, Chalodhorn R and Rao R P N 2008 Control of a humanoid robot by a noninvasive brain-computer interface in humans *J. Neural. Eng.* **5** 214

[53] Xiao X, Xu M, Han J, Yin E, Liu S, Zhang X, Jung T-P and Ming D 2021 Enhancement for P300-speller classification using multi-window discriminative canonical pattern matching *Journal of neural engineering* **18** 46079

[54] Xu M, Xiao X, Wang Y, Qi H, Jung T-P and Ming D 2018 A brain-computer interface based on miniature-event-related potentials induced by very small lateral visual stimuli *IEEE Transactions on Biomedical Engineering* **65** 1166–75