CenEEGs: Valid EEG Selection for Classification

CHENGLONG DAI and DECHANG PI, Nanjing University of Aeronautics and Astronautics
STEFANIE I. BECKER, The University of Queensland
JIA WU, Macquarie University
LIN CUI, Suzhou University
BLAKE JOHNSON, Macquarie University

This article explores valid brain electroencephalography (EEG) selection for EEG classification with different classifiers, which has been rarely addressed in previous studies and is mostly ignored by existing EEG processing methods and applications. Importantly, traditional selection methods are not able to select valid EEG signals for different classifiers. This article focuses on a source control-based valid EEG selection to reduce the impact of invalid EEG signals and aims to improve EEG-based classification performance for different classifiers. We propose a novel centroid-based EEG selection approach named CenEEGs, which uses a scale-and-shift-invariance similarity metric to measure similarities of EEG signals and then applies a globally optimal centroid strategy to select valid EEG signals with respect to a similarity threshold. A detailed comparison with several state-of-the-art time series selection methods by using standard criteria on 8 EEG datasets demonstrates the efficacy and superiority of CenEEGs for different classifiers.

CCS Concepts: • **Computing methodologies** → **Instance-based learning**; **Feature selection**; **Classification and regression trees**; • **Applied computing** → *Bioinformatics*;

Additional Key Words and Phrases: Electroencephalography (EEG) selection, classification, EEG similarity, centroid searching

ACM Reference format:
Chenglong Dai, Dechang Pi, Stefanie I. Becker, Jia Wu, Lin Cui, and Blake Johnson. 2020. CenEEGs: Valid EEG Selection for Classification. *ACM Trans. Knowl. Discov. Data* 14, 2, Article 18 (February 2020), 25 pages.
https://doi.org/10.1145/3371153

1 INTRODUCTION

Electroencephalography (EEG) relies on a weak, complex, non-stationary, high-dimensional, and low signal-to-noise ratio bioelectrical potentials generated by numbers of neurons on cerebral cortex [21]. Since these potentials reflect brain functions, EEG is widely applied in two fields. (1) Cerebral disorder or disease diagnosis through classification techniques, such as diagnosing

This work was partially supported by the National Natural Science Foundation of China (Grant nos. U1433116 and 61702355), Fundamental Research Funds for the Central Universities (Grant no. NP2017208), and Funding of Jiangsu Innovation Program for Graduate Education (Grant no. KYZZ16_0171).

Authors' addresses: C. Dai and D. Pi, Nanjing University of Aeronautics and Astronautics, 29 Jiangjun Avenue, Nanjing, Jiangsu Province 211106, China; emails: {chenglongdai, dc.pi}@nuaa.edu.cn; S. I. Becker, The University of Queensland, Brisbane, Queensland, Australia; email: s.becker@psy.uq.edu.au; J. Wu, Macquarie University, Sydney, New South Wales, Australia; email: jia.wu@mq.edu.au; L. Cui, Suzhou University, Suzhou, China; email: jsjxcuilin@nuaa.edu.cn; B. Johnson, Macquarie University, Sydney, New South Wales, Australia; email: blake.johnson@mq.edu.au.

1556-4681/2020/02-ART18 $15.00
https://doi.org/10.1145/3371153

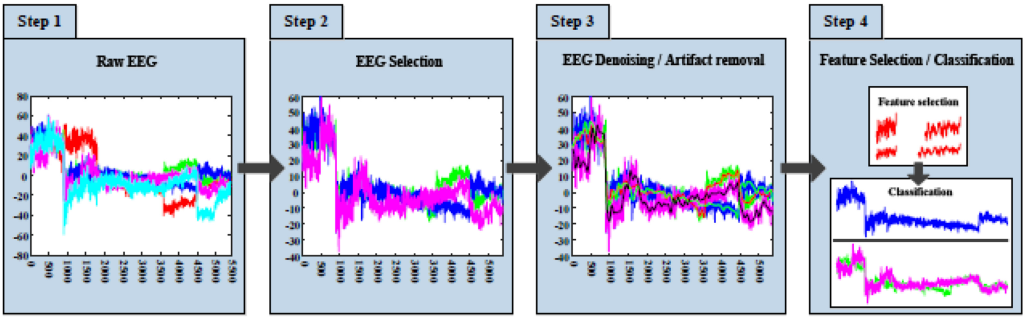


Fig. 1. The process of EEG analysis (Step 1: The original EEG signals that contain invalid EEG recordings, e.g., the light blue and the red recordings. They are directly used by researchers to study EEG-related problems and seldom being pre-selected before analyzing. Step 2: The original EEG signals are pre-processed through selecting relatively valid ones for subsequent studies as Step 3 and Step 4 illustrate. Step 3 and Step 4: With such pre-selected EEG signals, such following-up procedures are performed with less invalid EEG signals, enhancing and improving the performances).

Alzheimer's disease (AD) [46], epileptic seizures [24], strokes [1], amyotrophic lateral sclerosis (ALS) [19], and so on [5, 11]; (2) Brain-Computer Interface (BCI) for rehabilitation [3]. In these applications, EEG artifact removal or denoising, feature extraction, and EEG classification [9, 10] are widely studied, see Step 3 and Step 4 in Figure 1. In practice, EEG artifact removal or denoising aims to remove interference EEG segments such as eye movements or eye blinks embedded in original EEG signals, to provide clean EEG for EEG feature extraction and selection. Further, different with EEG signal selection that is from the view of source control of all EEG signals, this step mainly focuses on EEG segment processing that is from the view of an individual EEG signal. Then, EEG features are extracted out by feature extraction algorithms to represent original EEG signals, with which the dimension of EEG signals can be significantly reduced and the original EEG signals can be represented by several distinct features. Namely, the best features are selected from such extracted features to finally represent original EEG signals. With such selected EEG features, classifiers can learn a model to classify EEG signals into different classes according to their features, and according to the classification results, the cerebral disorders could be diagnosed, and real-world tools such as wheelchair and robotic arms in BCI can be controlled by disabled people according to the predictive EEG classes. Clearly, the subsequent processes for EEG signals are determined by the prior processing and it obviously indicates the importance of EEG selection on such following processes in Figure 1. In other words, analyzing performances for EEG artifact removal, feature extraction/selection and classification rely on the valid EEG signals, but in existing studies, they directly applied to raw EEG data, ignoring the impact of invalid EEG signals on the applications. In other words, valid EEG selection was ignored by most EEG researchers, who skipped this important step and immediately proceeded to subsequent steps such as feature selection [42, 51] and classification [41]) (corresponding to Step 3 and 4; see Figure 1 for an illustration). However, this ignorance of valid EEG selection likely limits and degrades the performance of follow-up analyses [16]. Invalid EEG[1] significantly degrades the diagnosis and detection accuracy due to the features of invalid

[1] Invalid EEG singles, strictly speaking, are those bioelectrical potentials stimulated by non-target cerebral activities or contaminated by noises. For example, when recording the motor imagery EEG of hands and legs (the target activities), the subject imagines the tongue rolls, then the generated potentials are invalid EEG signals. Further, the interferences from environmental noises, such as noise of EEG recording equipment, and so on, can cause invalid EEG even though the subjects are simulating target cerebral activities. Correspondingly, the valid EEG signals are strictly those potentials from target cerebral activities without noise interferences.

EEG mixed in valid ones. Especially, more invalid EEG signals exist in those EEG signals recorded from patients suffering from cerebral diseases that are due to the uncontrolled cerebral activities in such patients' brain. Hence, it is a difficult task for neurologists to accurately analyze and diagnose the diseases with such raw EEG signals. A necessary step is to pre-process raw EEG signals (from the view of source control) to reduce the degradation of invalid EEG on classification performance in EEG applications. Therefore, we in this article explored valid EEG selection from the view of source control[2] to enhance or improve EEG classifications and also propose a novel method for it.

1.1 Motivation

In practice, many invalid EEG signals exist with valid ones, degrading classification accuracy for EEG applications. Invalid EEG mainly comes from (1) the environmental noises and (2) the non-target bioelectrical potentials activated by other uncontrolled cerebral activities (especially from cerebral disorder patients). Valid EEG selection, from the source controlling view, is an important and necessary step in EEG analysis, which provides more target EEG for follow-up processes such as artifact removal, feature selection and classification, and so on. Unfortunately, to the best of our knowledge, EEG selection is rarely addressed by EEG researches in previous studies, and most researchers skipped this source controlling step and directly proceeded to the follow-up analyses, as illustrated in Figure 1. Consequently, these direct analyses performed over raw EEG signals seem to result in relatively poor classification and limit their applications in real world such as in BCI-based applications.

EEG, as one type of potentials with non-stationary, non-linear, high-dimensional features, contains not only frequency information but also spatial information (i.e., correlations among multiple EEG channels, which means that several EEG channels together rather than a single one can reflect the cerebral activity. In other words, for some specific cerebral activities, multiple channels are required to record EEG signals from different cortex area. So these EEG channels have correlations with each other.) [51], which make it a challenging task to select valid EEG. Furthermore, the phase shift and amplitude scale of EEG signals also make it another challenge to select valid EEG from originally recorded EEG. Fortunately, time series selection techniques may be considered as one way to solve such problems, since EEG signal is regarded as one special time series with specific characteristics. In recent decades, several time series selection methods have been proposed for time series classification, but these are only suitable for specified classifiers. For example, the condensation methods are suitable to reduce training set for SVM and edition methods are for Nearest Neighbor (NN) classifiers [34], and the hybrid methods are highly adoptable to kNN classifiers as well [20].

In order to reduce the impact of invalid EEG signals on the follow-up EEG analyses and enhance the classification performance for different classifiers, this article proposes a novel centroid-inspired approach to select valid EEG signals from the original EEG signals (which is regarded as the source control for EEG analyses), which is based on a scaling and shifting invariant similarity metric [13].

1.2 Contributions and Outline

This article explores valid EEG selection for different classifiers and proposes a novel centroid-based method for valid EEG selection. In detail, this article made such contributions that are highlighted as follows.

[2] Source control in this article is a signal pre-processing methodology that is to prescreen valid EEG signals from "the original raw EEG signals," that means original raw EEG recordings are pre-processed before being applied in the subsequent analyses such as denoising, artifact removal, feature extraction, classification, and so on. In other words, source control aims to enhance and improve the performance through reducing invalid EEG signals and selecting more valid ones from the original raw EEG signals.

- Aiming at reducing the degradation of invalid EEG signals on EEG classification, a source control-based selection is explored in this article. To the best of our knowledge, it is seldom addressed in previous researches.
- A centroid-based EEG selection approach is proposed in this article, which we call CenEEGs, which is suitable for different classifiers.
- Besides, a scale-and-shift-invariance distance metric is utilized to measure the similarities among EEG signals. Then, an EEG centroid extraction function is proposed to search the most representative EEG centroid for valid EEG selection based on their similarities, which is also globally optimal.
- The efficacy of CenEEGs to improve the classification performance for several classifiers is demonstrated through a detailed experimentation by comparing with several state-of-the-art time series selection methods on 8 EEG datasets. The results clearly show that CenEEGs not only improves the classification accuracy compared with raw EEG signals, but also outperforms those state-of-the-art time series selection methods.

The reminder of this article is organized as follows. In Section 2, the related works on EEG time series selection are presented. After that, the proposed method is introduced in detail in Section 3, including the introduction of similarity metric and the centroid-based CenEEGs approach. Subsequently, EEG datasets, criteria and baselines to evaluate the efficacy of CenEEGs are outlined in Section 4 respectively. Finally, conclusions and some directions for the future work are summarized in Section 5.

2 RELATED WORKS

EEG selection aims to select a subset of relevant EEG signals from the original EEG for classifiers and remove noisy/invalid EEG signals, though without creating new artificial data [25]. Unfortunately, only very few studies have explored possibilities for valid EEG selection. As EEG data are a specific kind of time series data, EEG selection may be performed by time series selection approaches. In practice, time series selection is broadly used in time series prediction [44], regression [4, 43] and classification [52]. Traditionally, the most widely used approaches to select time series can be categorized into condensation, edition, and hybrid methods that include both condensation and edition [7, 20, 29, 47, 53].

In detail, condensation algorithms condense (or remove) internal time series data while maintaining border time series data that are close to the threshold, as these are believed to have stronger impact on classification performance than internal time series data. The edition algorithms use the opposite strategy to condensation ones. In edition methods, the border time series data are regarded as noises and are removed to produce a smoother boundary, while internal time series data are maintained, also with the aim to improve the generalization accuracy for testing time series. The hybrid algorithms contain both strategies of condensation and edition methods. In detail, hybrid methods aim to find the smallest subset that satisfies the internal and border time series removal so as to achieve generalization accuracy in testing time series [20, 29, 34, 53]. Although they all have their own strengths and have been widely researched and applied, such instance selection methods seem more suitable for specific classifiers, as indicated in [34] that "condensation methods are probably suitable for reduction for training SVM, while edition strategies are more suitable for the scene of using k-Nearest Neighbors (kNN) classifier." Furthermore, García et al. [20] also introduced that kNN classifier is highly adaptable to hybrid methods. In other words, these methods may limit the classification-based EEG applications in real world (i.e., BCI or cerebral disease diagnosis), since they rely on a specific classifier (i.e., kNN or SVM) and may not perform well when other classifiers (i.e., st-TSC [32], RPCD [40], COTE [6], or SAX-SEQL [37]) are adopted to classify EEG signals.

Additionally, some newly state-of-the-art algorithms are presented to select time series or instances such as MLIS [36] and NNGIR [50], but they are designed for kNN classifier, not for different classifiers. Moreover, some spectral methods are merged to represent signals and then classify, such as [2, 14, 35], but their performance is probably limited because they belong to frequency domain analysis but EEG classification also implicitly embeds spatial information (i.e., correlations among multiple EEG channels). Besides, the spectral methods' performance seems to be influenced by the length of wavelet time window. In order to broaden the applications for different classifiers without losing temporal (e.g., time, frequency) or spatial information (e.g., channel correlation) of EEG recordings, weighted Naïve Bayes-based selection methods [26, 48] have recently been proposed, and have shown to yield excellent performance for different classifiers. However, the Naïve Bayes-based methods split the original data into two parts, using one part to build weighted Bayes network and the other part to test the network; hence, it may contain invalid time series in training part (i.e., building weighted Bayes network) to degrade the Bayes network and affect the time series selection for testing data. Moreover, the selection results are also sensitive to weighted Naïve Bayes constructed by the training time series.

In the present article, we also focus on this kind of strategy for EEG selection for different classifiers. Thereafter, we proposed a novel centroid-based approach to select valid EEG signals with a scaling and shifting invariant distance metric [13], which is suitable for different classifiers such SVM, shapelet-based classifier (i.e., st-TSC [32]), distance-based classifier (i.e., RPCD [40]), ensemble-based classifier (i.e., COTE [6]), and subsequence-based classifier (i.e., SAX-SEQL [37]).

3 CENTROID-BASED VALID EEG SELECTION

Invalid EEG signals degrade the EEG analysis performance, especially at the stage of classifying EEG signals. Moreover, most EEG studies analyze original EEG data without considering the impact of invalid EEG signals, thereby skipping the important source control pre-processing of EEG selection. Consequently, this article proposes an approach to select valid EEG signals so as to reduce the impact of invalid ones on EEG classification.

3.1 EEG Similarity Metric

Recently, various similarity measures for time series have been investigated and widely applied. However, as a specific time series, conventional similarity measures such as Euclidean Distance (ED) [18], Dynamic Time Warping (DTW) [27, 39], and Hausdorff Distance (HD)[45] are not suitable for EEG similarity [16]. As introduced above, EEG is weak, complex, non-stationary, high-dimensional, and high-vibration, and it probably has different distributions from different subjects during several recording sessions. Two similar EEG signals with scaled amplitude and shifted phase should be regarded as similar, whereas ED, DTW, and HD all evaluate them to be dissimilar. ED cannot capture flexible similarity of EEG signals, since it requires the same length of EEG and is sensitive to noises and outliers. DTW concentrates too much on minimizing the accumulation of all local distances between adjacent points of vibrating EEG, and DTW does not work well when signals are sampled less frequently [17]. HD is sensitive to outliers [8, 45] and it transforms EEG signals to arbitrary point sets, and hence it does not consider the point orders of EEG signals. Namely, HD is solely based on NN distances between points in EEG trials. It is possible for two dissimilar EEG trials to have small HD. In this article, we utilized a scale-and-shift-invariance metric [13] to measure the similarities among EEG signals. This similarity metric contains scale and shift parameters, so it can better measure similarities of EEG signals with different amplitudes and phases shifts. Then, this metric contributes to EEG similarities for the proposed centroid-based EEG selection method. Besides, the comparison of such distance measures is also illustrated in Figure 2, which indicates the superiority of the proposed EEG similarity measure over the widely

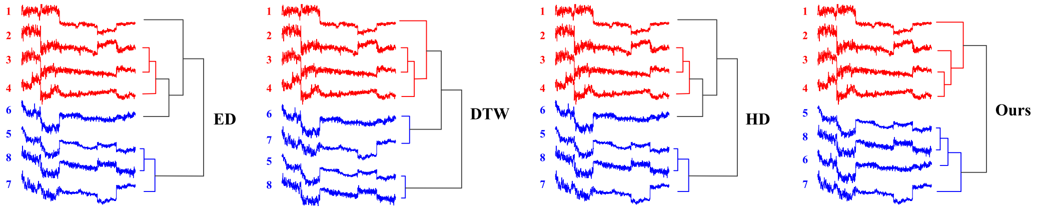


Fig. 2. An illustration of distance measures for EEG similarities. The red and blue respectively denote two-class EEG signals activated by two corresponding cerebral stimulations: moving a cursor up and moving a cursor down. Every class contains 4 trials with length of 5,376.

used ED, DTW, and HD, since the proposed measure succeeds measuring the EEG similarities and classifying them into right classes based on their similarities.

Mathematically, given two EEG signals e_x and e_y, the similarity $d(e_x, e_y)$ between them is defined by the following equation:

$$d(e_x, e_y) = \min_{\alpha, s} \frac{\| e_x - \alpha e_{y(s)} \|}{\| e_x \|}, \tag{1}$$

where α and $e_{y(s)}$ denote the scaling coefficient and the shifted EEG of e_y with s points, respectively. This similarity metric aims to find the optimal scaling coefficient α and the shifting length s for two EEG signals.

To obtain the optimally shifted EEG signal $e_y(s)$ by s points. We applied cross correlation [15] to compute it. Give two EEG signals $e_x = (e_{x_1}, e_{x_2}, \ldots, e_{x_m})$ and $e_y = (e_{y_1}, e_{y_2}, \ldots, e_{y_m})$, their cross correlation is computed by doing the point-to-point inner product in a way that keeps e_y static and slides e_x along with e_y, see the following equation:

$$CC_\tau(e_x, e_y) = R_\tau(e_x, e_y) = \sum_{k=1}^{m-|\tau|} e_{x_k} \cdot e_{y_k}, \tag{2}$$

where $\tau = (-m, \ldots, 0, 1, \ldots, m)$ denotes the shift.

According to all the potential shifting sequences, the cross-correlation sequence $CC_s(e_x, e_y) = (c_1, \ldots, c_s)$ with length of $2m - 1$ is defined as

$$CC_s(e_x, e_y) = R_{s-m}(e_x, e_y) = R_\tau(e_x, e_y), \tag{3}$$

where $s \in \{1, 2, \ldots, 2m - 1\}$, and $R_{s-m}(e_x, e_y)$ is defined by the following equation:

$$R_\tau(e_x, e_y) = \begin{cases} \sum_{k=1}^{m-\tau} e_{k+\tau} \cdot e_k, & \tau \geq 0, \\ R_{-\tau}(e_y, e_x), & \tau < 0, \end{cases} \tag{4}$$

where $\tau = s - m$. Cross correlation aims to search the optimal location of s that maximizes $CC_s(e_x, e_y)$ between e_x and e_y. Moreover, we also exploited z-normalization to normalize $CC_s(e_x, e_y)$, see the following equation:

$$zCC_s(e_x, e_y) = \frac{CC_s(e_x, e_y)}{\sqrt{R_0(e_x, e_x) \cdot R_0(e_y, e_y)}}. \tag{5}$$

Correspondingly, the optimal shift s is then given as follows:

$$s = \arg\max_s \frac{CC_s(e_x, e_y)}{\sqrt{R_0(e_x, e_x) \cdot R_0(e_y, e_y)}}. \tag{6}$$

In the meantime, Algorithm 1 indicates the computation of shifted EEG.

ALGORITHM 1: $e_{y(s)} = Shift(e_x, e_y)$

```
Input: $e_x$, $e_y$, two EEG signals.
Output: $e_{y(s)}$, the shifted EEG signal.
Compute $CC_s(e_x, e_y) = IFFT(FFT(e_x) * FFT(e_y))$[15];
$z$-normalize $CC_s(e_x, e_y)$: $zCC_s(e_x, e_y) = \frac{CC_s(e_x, e_y)}{||e_x||||e_y||}$;
$s = \arg\max_s zCC_s$;
shift $\tau = s - m$;
if $\tau \geq 0$ then
    $e_{y(s)} = [zeros(1, \tau), e_y(1 : end - \tau)]$;
end
else
    $e_{y(s)} = [e_j(1 - \tau : end), zeros(1, -\tau)]$;
end
return $e_{y(s)}$;
```

When an EEG e_y is shifted by s, optimal α can be computed with a closed-form expression $\frac{\|e_x - \alpha e_{y(s)}\|_2}{\|e_x\|_2}$, which is a convex function of α. Therefore, the optimal α in Equation (1) can be computed by setting $\frac{\partial d(e_x, e_y)}{\partial \alpha} = 0$, see the following equation in detail:

$$\alpha = \frac{e_x^T e_{y(s)}}{\| e_{y(s)} \|^2}. \tag{7}$$

3.2 EEG Centroid Extraction

In order to select valid EEG signals, we proposed a novel approach to extract the reference sequence as the centroid, with which the raw EEG signals can be pre-processed. However, it is a challenging task to extract an optimal centroid.

To the best of our knowledge, the easiest way to find a centroid sequence to represent the set of candidate EEG signals is k-means, which computes the mean coordinates of all corresponding coordinates in all candidate EEGs. However, this kind of methods cannot obtain the most meaningful information of EEG signals, especially when scale and shift exist among EEG signals. Namely, such methods as k-means search EEG centroid just considering the corresponding coordinate of EEG signals rather than the flexible similarity of EEG. As we introduced before, a wide range of similarity-based methods are applied to measure EEG time series and assist to search the centroid sequence. However, the characteristics of the problem addressed in the article make the setting somewhat different with those based on common metrics, such as Euclidean, DTW, or HD. Again, this kind of similarity metrics are inappropriate in our case: (1) If two EEG signals have similar shape but different amplitudes, they should be regarded as similar. Thus, scaling EEG signals on the y-axis should not change the similarity; (2) meanwhile, if the shapes of two EEG signals are similar, even shifted on x-axis, it should not change their similarity [13]. To avoid the shortcoming of such similarity-based methods for EEG centroid extraction, we proposed a method that transforms the centroid computation to an optimization problem that aims to find a minimizer of the sum of squared distances to all EEG signals. Given n EEG signals $E = (e_1, \ldots, e_n)^T$, CenEEGs tries to find an optimal centroid c for EEG selection that minimizes the function O defined by the following equation:

$$O = \sum_{e_i \in S} d(e_i, c)^2, \tag{8}$$

where S denotes the set of selected EEG signals based on the centroid. Therefore, the centroid c of selected EEG signals is the minimizer of the sum of $d_{\leq\delta}(e_i, c)^2$ for all $e_i \in S$ such that $d \leq \delta$, where δ is the similarity threshold and those EEG signals satisfy the threshold can be considered as the valid ones and then added into S. Mathematically,

$$c^* = \arg\min_c \sum_{e_i \in S} d_{\leq\delta}(e_i, c)^2. \tag{9}$$

With Equation (9), CenEEGs requires many iterative runs until it converges. In order to find the optimal centroid efficiently, we solve the minimization problem based on its closed form.

Proposition 3.1. The minimizer of centroid function $c^* = \arg\min_c \sum_{e_i \in S} d(e_i, c)^2$ can be rewritten equivalently as $c^* = \arg\min_c \frac{c^T U c}{\|c\|^2}$.

Proof. Recall Equations (1) and (9), then combining them,

$$c^* = \arg\min_c \sum_{e_i \in S} \min_{\alpha, s} \frac{\| c - \alpha e_{i(s)} \|^2}{\| c \|^2}. \tag{10}$$

Based on the optimal scaling coefficient $\alpha = \frac{e_x^T e_{y(s)}}{\|e_{y(s)}\|^2}$ as well as the optimal shift s, the Equation (10) can be written as $c^* = \arg\min_c \frac{1}{\|c\|^2} \sum_{e_i \in S} \| c - \frac{c^T e_i}{\|e_i\|^2} e_i \|^2$. Furthermore, to simplify the solution function, we transform $c^T e_i e_i$ to $e_i e_i^T c$ (see footnote 3 below[3]), then c^* is equivalently rewritten as

$$\begin{aligned} c^* &= \arg\min_c \frac{1}{\| c \|^2} \sum_{e_i \in S} \left\| c - \frac{e_i e_i^T c}{\|e_i\|^2} \right\|^2 \\ &= \arg\min_c \frac{1}{\|c\|^2} \sum_{e_i \in S} \left\| \left(I - \frac{e_i e_i^T}{\|e_i\|^2} \right) c \right\|^2 \\ &= \arg\min_c \frac{c^T}{\|c\|^2} \sum_{e_i \in S} \left(I - \frac{e_i e_i^T}{\|e_i\|^2} \right)^2 c. \end{aligned}$$

Then, set $U = \sum_{e_i \in S} (I - \frac{e_i e_i^T}{\|e_i\|^2})^2$ and the original minimizer of centroid function is transformed to

$$c^* = \arg\min_c \frac{c^T U c}{\| c \|^2}. \tag{11}$$

□

Moreover, based on [22], the solution of Equation (11) is the eigenvector that corresponds to the smallest eigenvalue of U. In other words, when c is transformed through multiplying eigenvectors of matrix U, the $c^T U c$ equals the weighted sum of eigenvalues of matrix U. Consequently, the solution of Equation (11) is the smallest eigenvalue of U. At last, it is easy to obtain the new

[3] Obviously, $c^T e_i e_i = (c^T e_i) e_i$ and $e_i e_i^T c = e_i (e_i^T c)$. Since c^T and e_i are respectively row matrix and column matrix with same length, the values of $c^T e_i$ and $e_i^T c$ are equal to the same real number, namely, $c^T e_i = e_i^T c = a$ (a is real number). Consequently, $c^T e_i e_i = a e_i = e_i a = e_i e_i^T c$.

ALGORITHM 2: $c^* = Centroidextraction(E, c)$

```
Input: $E = (e_i, \dots, e_n)^T_{n \times m}$, $n \times m$ EEG signals; $\delta$, the similarity threshold.
Output: $c^*$, the centroid.
$E' \leftarrow [\ ]$;
for $i = 1$ to $n$ do
    $e_i \leftarrow Shift(c, e_i)$;
    $d \leftarrow d(e_i, c)$;
    if $d \leq \delta$ then
        $E' \leftarrow [E'; e_i]$;
    end
end
$U \leftarrow \sum_{e_i \in E}(I - \frac{e_i e_i^T}{\|e_i\|^2})$;
$c^* \leftarrow Eigenvector(U, thesmallestone)$;
```

centroid c^* by finding the smallest eigenvector of U. Meanwhile, Algorithm 2 shows how to extract the centroid for EEG selection.

3.3 The CenEEGs

The CenEEGs aims to minimize the sum of squared similarities with an iterative refinement procedure. CenEEGs computes EEG centroid effectively with both the scaling and shifting invariances, and contains the following two main phases: (1) the assignment phase and (2) the refinement phase. For each iteration, CenEEGs performs these two phases to select EEG signals and update the centroid. In the assignment phase, CenEEGs updates the selected EEG signals through comparing every EEG trace to the centroid with respect to the similarity threshold. In this phase, CenEEGs depends on the similarity measure introduced in Section 3.2. In the refinement phase, the centroid is updated along with the changes of selected EEG candidates by finding the smallest eigenvector of U, as described in Section 3.3. Finally, CenEEGs repeats these two phases until the optimal centroid c is searched out and all the EEG signals are evaluated. In detail, Algorithm 3 shows the process of CenEEGs for EEG selection.

Particularly, since CenEEGs originally begins with an initial EEG signal and then proceeds based on similarities, iteratively optimizes its objective function to search the optimal centroid sequence. So its convergence, intuitively, is sensitive to the initialization of centroid [31, 49]. If the centroid is initialized poorly, the CenEEGs may run very slow. Therefore, in order to decrease the impact of randomly initializing centroids on EEG selection, we simply use a standardization sequence of all EEG signals as the initial centroid. Although the centroid sequence is different to the standardized one, it is intuitively closer to the standardized sequence than a randomly selected one. Meanwhile, the standardization can successfully reduce outliers in EEG signals. Mathematically, for n-trial EEG dataset $E_{n \times m}$ with m samples, its standardization is defined as

$$E'^T = \frac{E^T - \mu(E^T)}{\sigma(E^T)}, \tag{12}$$

$$c_0 = \mu(E'^T), \tag{13}$$

where μ is the average of E^T, i.e., $\mu(E^T) = \frac{1}{n}\sum_{i=1}^{n} e_i$ and σ is its standard deviation, i.e., $\sigma(E^T) = (\frac{1}{n}\sum_{i=1}^{n}(e_i - \mu)^2)^{\frac{1}{2}}$.

ALGORITHM 3: CenEEGs

```
   **Input**: $E = (e_i, \ldots, e_n)^T_{n\times m}$, $n \times m$ EEG signals; $\delta$, the similarity threshold.
   **Output**: $S$, the selected valid EEG signals.
1  Initialize centroid $E'^T = \frac{E^T - \mu(E^T)}{\sigma(E^T)}$ and $c = \mu(E'^T)$;
2  $S \leftarrow c$;
3  for $i = 1$ to $n$ do
4      $e_i \leftarrow Shift(c, e_i)$;
5      $d \leftarrow d(e_i, c)$;
6      if $d \leq \delta$ then
7          $S \leftarrow [S; e_i]$;
8      end
9  end
10 repeat
11     $S' \leftarrow S$;
12     $c^* \leftarrow centroidextraction(S', c)$ based on Algorithm 2;
13     for $i = 1$ to $n$ do
14         $d^* \leftarrow d(e_i, c^*)$;
15         if $d^* \leq \delta$ then
16             $S \leftarrow [S; e_i]$;
17         end
18     end
19 until $S' \equiv S$;
```

To reiterate, CenEEGs begins with the initialization of centroid based on a standardizing strategy of Equations (12) and (13) (see Line 1). The standardization can reflect the membership of all EEG signals and reduce the impact of outliers, so it is better than a randomly selected EEG as the initial centroid. Then the EEG signals based on the initial centroid are shifted (with Algorithm 1; Line 4) and selected with respect to similarity threshold δ, as described in Lines 3–9. Afterwards, CenEEGs computes the new centroid based on Algorithm 2 (see Lines 11–12). Once a new centroid is computed, CenEEGs refines the selected EEG candidates by using the proposed similarity metric (Equation (1)) as well as δ, as Lines 13–18 show. Finally, CenEEGs iteratively repeats the procedure (from Line 10 to Line 19) until the selected EEG signals are constant across iterative runs (i.e., cease changing with iterative runs; see Line 19). In other words, CenEEGs stops until the centroid of selected EEG signals is fixed or constant.

3.4 Time Complexity of CenEEGs

According to Algorithm 3, CenEEGs begins with the initialization of centroid and then shifts EEG signals and originally selects EEG signals. In the process (Lines 1–9), it mainly costs $O(nm \log m)$ time for n EEG signals with length of m, since shifting an EEG signal may cost $O(m \log m)$ [38]. Afterwards, CenEEGs spends time calculating matrix U and searching for the eigenvector of it. In detail, CenEEGs uses $O(m^2)$ to compute matrix U for one e_i (Line 9 in Algorithm 2) and finally spends $O(m^3)$ time searching its eigenvectors [38] (Line 10 in Algorithm 2). Consequently, in the refinement phase (Lines 11–12 in Algorithm 3), its time complexity is $O(max\{nm^2, m^3\})$. Besides, CenEEGs selects EEG signals based on their similarities to the centroid as well as the similarity threshold δ, which takes $O(nm)$ time in assignment phase (Lines 13–18). Overall, the exact time complexity for one iteration of CenEEGs is $O(max\{nm, nm^2, m^3\})$ (Lines 10–19). In practice, the

Table 1. EEG Datasets

Dataset	Description	# of EEG	Length	# of class	Training data: testing data
#1	Ia (Traindata_0 + Traindata_1)	268	5,377	2	179:89
#2	Ib (Traindata_0 + Traindata_1)	200	8,065	2	134:66
#3	Traindata_0, Ia + Traindata_0, Ib	235	5,377/8,065	2	157:78
#4	Traindata_1, Ia + Traindata_1, Ib	233	5,377/8,065	2	156:77
#5	Ia + Ib	468	5,377/8,065	4	313:155
#6	III_V_s1	3,488	97	3	2,325:1,163
#7	IV_2a_s1	288	6,887	4	192:96
#8	IV_3_s1	160	4,001	4	107:53

number n of EEG signals is commonly smaller than its length m, i.e., $n \ll m$. Consequently, for each iteration of CenEEGs, its time complexity is $O(m^3)$.

4 EXPERIMENTAL RESULTS

In the section, we will report the results of some experiments designed to evaluate the efficacy of CenEEGs on different classifiers, which involved comparing it with several time series selection approaches.

4.1 EEG Datasets

The EEG datasets we tested were slow cortical potentials (SCPs)(i.e., #1–#5), motor imagery EEG signals (MIs) (i.e., #6–#7), and hand movement EEG signals (HMs) (i.e., #8). The SCPs were recorded from two subjects, one of which was healthy (Ia) and the other being an ALS patient (Ib). Meanwhile, MIs were recorded from two healthy subjects, and HMs were also recorded from a healthy subject. Moreover, the SCP datasets and their detailed descriptions are publicly available at http://www.bbci.de/competition/ii/, and MIs' at http://www.bbci.de/competition/iii/, and HMs' at http://www.bbci.de/competition/iv/. Table 1 shows the 8 EEG datasets. Furthermore, we applied a Hold-out strategy [28] to set up our experiments in this article. In detail, the original data are divided into two parts, i.e., training data and testing data with the proportion of 2:1. All the selection algorithms are applied to training datasets, to select valid EEG trials. Then the selected EEG trials are used to train classifier. Finally, the testing EEG datasets without selecting through such selection algorithms are applied to verify the classification performance of classifiers trained by the selected valid EEG trials.

4.2 Evaluation Methodology

Three criteria, such as *rand index* (RI), *F-score*, and *Fleiss' kappa* (κ), are used to evaluate the methods.

(1) *Rand Index*: RI, also called accuracy, estimates the accuracy of classification with respect to the correct classes of EEG signals. It reflects the percentage of correct selections of classifer.

$$RI = \frac{TP + TN}{TP + TN + FP + FN}, \tag{14}$$

where TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives respectively.

(2) *F-score*: F-score unequally weighs FP and FN in RI with a scale parameter $\beta \geq 0$ on *recall*, commonly $\beta = 1$. Mathematically,

$$F - score = \frac{(1+\beta^2)pr}{\beta^2 p + r} \tag{15}$$

where *precision* $p = \frac{TP}{TP+FP}$ and *recall* $r = \frac{TP}{TP+FN}$.

(3) *Fleiss' kappa*: Fleiss' kappa (κ) measures the coherence of decision ratings among classes.

$$\kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}}, \tag{16}$$

where $\overline{P} = \frac{1}{Nn(n-1)}(\sum_{i=1}^{N}\sum_{i=1}^{n} n_{ij}^2 - Nn)$, $\overline{P_e} = \sum_{j=1}^{k}(\frac{1}{Nn}\sum_{i=1}^{N} n_{ij})^2$. $\overline{P} - \overline{P_e}$ reflects the agreement degree of actually achieved over chance; $1 - \overline{P_e}$ indicates the agreement degree of attainable above chance. N is the number of subjects; n, ratings per subject; k, number of classes.

To the end, a higher RI or F-score, or κ demonstrates a better classification performance.

4.3 Baseline Methods

We compared CenEEGs with seven promising state-of-the-art EEG time series selection methods for different classifiers, which are briefly introduced as follows.

LRIWNBs [26]: Local recursion instance weighted Naïve Bayes time series selection builds weighted Naïve Bayes with training time series data of same label, and then recursively modifies the weight with testing time series data of same label to finally produce m time series based on such modified weighted Bayes. It concentrates on local correlations of training time series to the testing ones with same label.

GRIWNBs [26]: Global recursion instance weighted Naïve Bayes time series selection constructs weighted Naïve Bayes with training time series data without considering their labels, then it recursively modifies its weights with all testing time series data. Finally, it selects m nearest time series. It mainly focuses on global similarities between the whole training time series and the entire testing ones.

LDWNBs [48]: Local dual weighted Naïve Bayes time series selection firstly weights every time series based on their similarities and then builds an attributed weighted Naïve Bayes with new training data with same label. Finally, it selects out m most similar time series from each class based on the weighted attribute for different classifiers.

GDWNBs [48]: Similarly, global dual weighted Naïve Bayes time series selection also weights time series based on the similarities and then builds attributed weighted Naïve Bayes without considering the labels, and finally selects m most similar time series from all classes for different classifiers.

UU [23]: This instance selection method considers instances uncertainty and utility simultaneously. Obviously, it is more difficult for a learning model (i.e., classifier) to identify the class of an instance with a higher uncertainty to be selected. Subsequently, the utility of an instance provides updating information for the learning model.

MLIS [36]: Metric learning-based instance selection is just designed for kNN classifier. It selects instances by using metric learning to transform an input space in a way that the points in same class are close to each other while points in different classes are separated as far as possible.

NNGIR [50]: Natural neighborhood graph-based instance reduction applies natural neighborhood graph to split original time series into three parts (i.e., noisy instances, border, and internal instances), and then achieves a subset of time series through removing the noisy data, which is

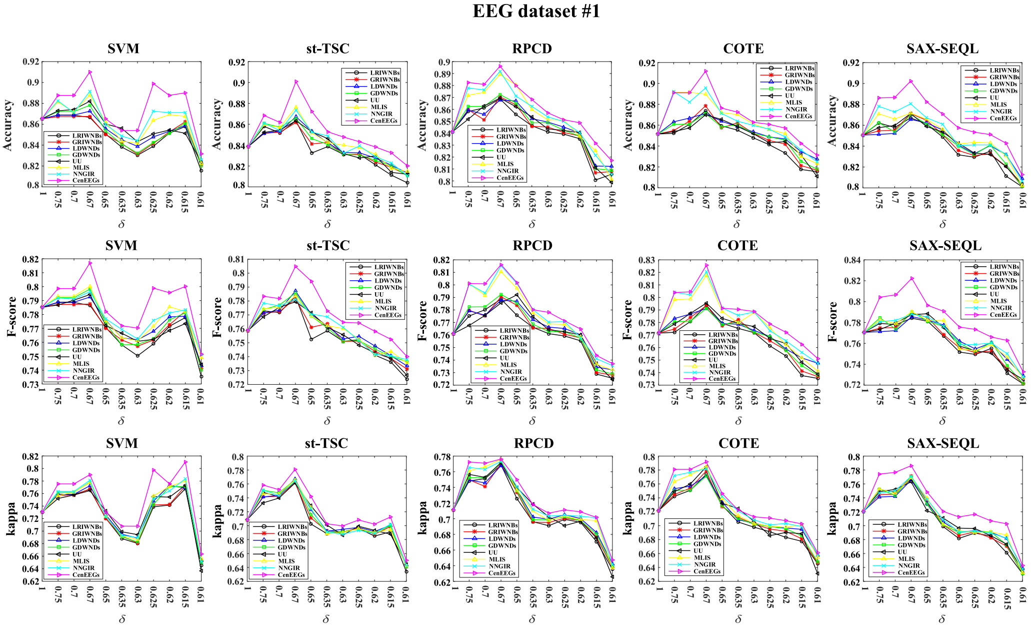


Fig. 3. Classification with different classifiers on EEG dataset #1.

reported to increase the reduction rate of original data while enhancing or improving the classification accuracy of kNN.

4.4 Classification Performance Analysis

4.4.1 Compared with Selection Methods. To demonstrate the selection capacity of CenEEGs for different classifiers, we compared it with seven selection methods such as LRIWNBs, GRIWNBs, LDWNDs, GDWNBs, UU, MLIS, and NNGIR with five classifiers: SVM [12], shapelet/distance-based (i.e., st-TSC [32][4] and RPCD [40][5]), ensemble-based (i.e., COTE [6, 33][6]), and structure-based classifier (i.e., SAX-SEQL [30, 37][7]). The classification comparisons on 8 EEG datasets are shown in Figures 3–10, respectively. Classification with selected EEG trials from training datasets achieved by selection approaches were conducted 20 times on testing datasets (without selection) for each classifier, respectively, running by Matlab R2014b, on a Windows 7 machine with 3.20 GHz CPU and 4 GB memory, and the final results are the average of these 20 operations. Besides, all the default parameter settings are as same as in corresponding references. The results with five different classifiers (SVM, st-TSC, RPCD, COTE, and SAX-SEQL) on different similarity thresholds clearly demonstrate that CenEEGs outperforms the other 7 time series selection methods on all 8

[4] st-TSC: Shapelet-transformed time series classifier firstly extracts distinct time series subsequences that can represent different classes of time series, and then utilizes an optimization function to search such subsequences of fixed length that can best identify target variables using computing distances of time series to the extracted shapelets.

[5] RPCD: Recurrence pattern compression distance classifier utilizes recurrence plots as the representation domain to classify time series based on their similarities measured by Campana-Keogh distance.

[6] COTE: A classifier with a collective of transform-based ensembles that fuses several different classifiers as one, which in detail contains shapelet-based classifiers, and spectral-based classifiers. It is an ensemble-based classifier that classifies time series via using a heterogeneous ensemble on the transformed time series representations.

[7] SAX-SEQL: An efficient linear classifier that extracts distinct discrete time series subsequences in an all-subsequences space formed with symbolic aggregate approximation (SAX) [30]. This process smoothes and compresses original time series to discrete subsequences.

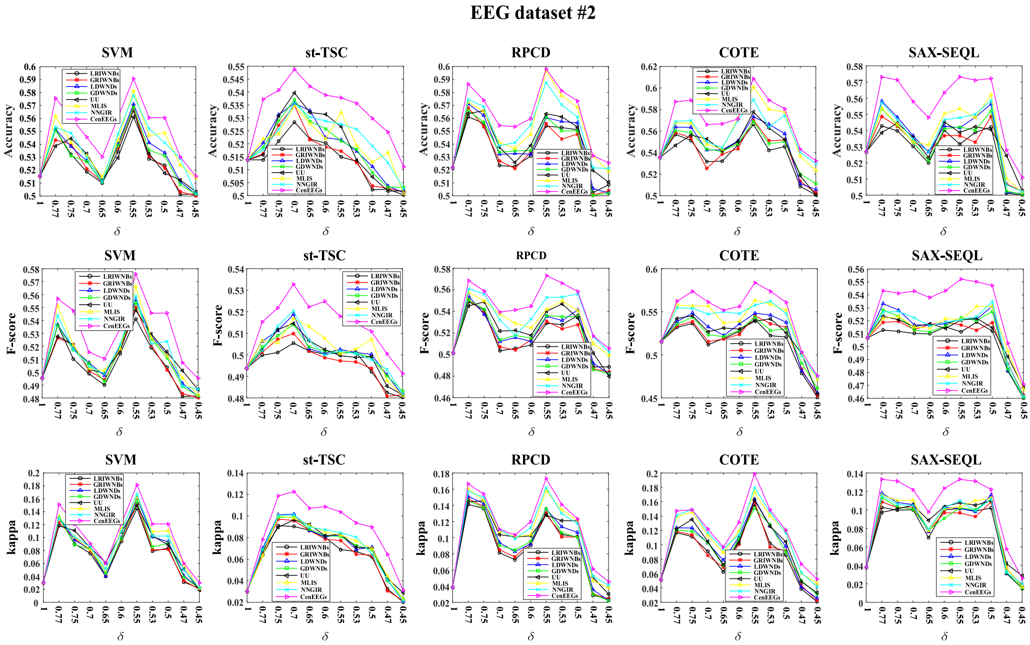


Fig. 4. Classification with different classifiers on EEG dataset #2.

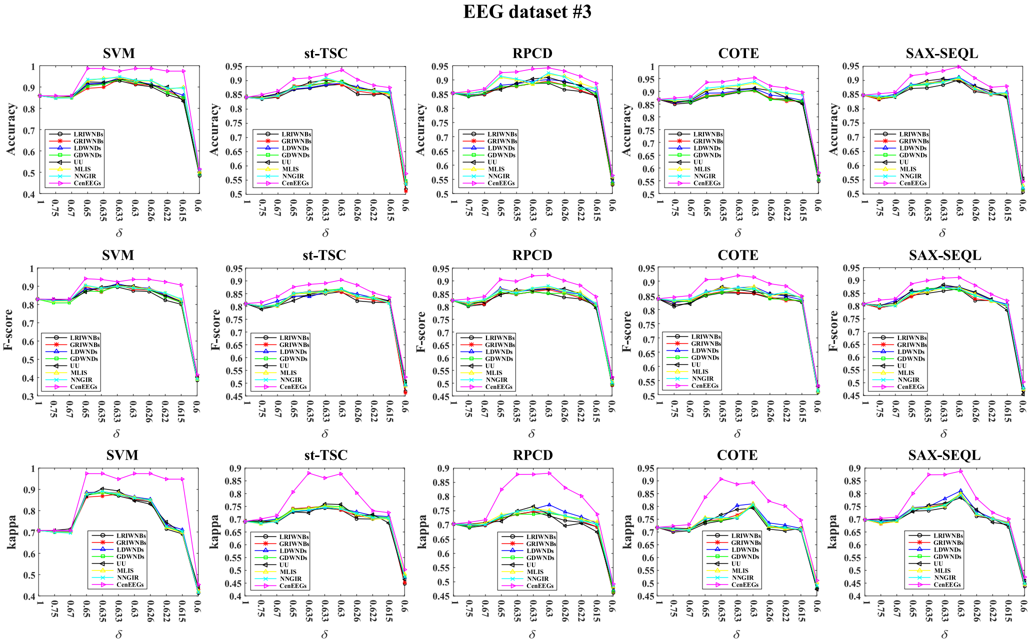


Fig. 5. Classification with different classifiers on EEG dataset #3.

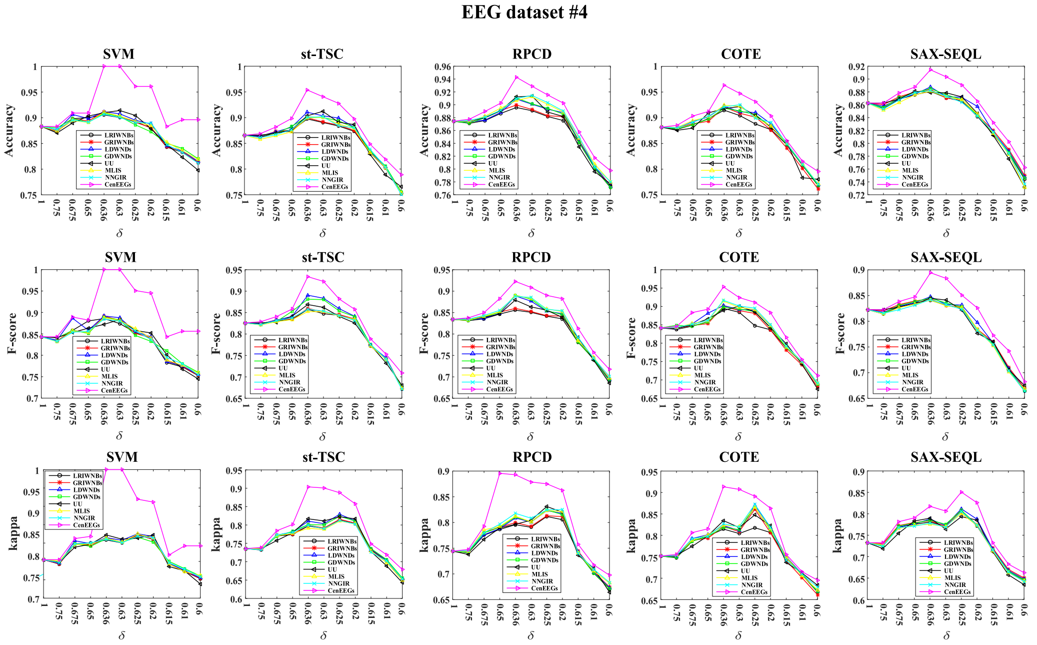


Fig. 6. Classification with different classifiers on EEG dataset #4.

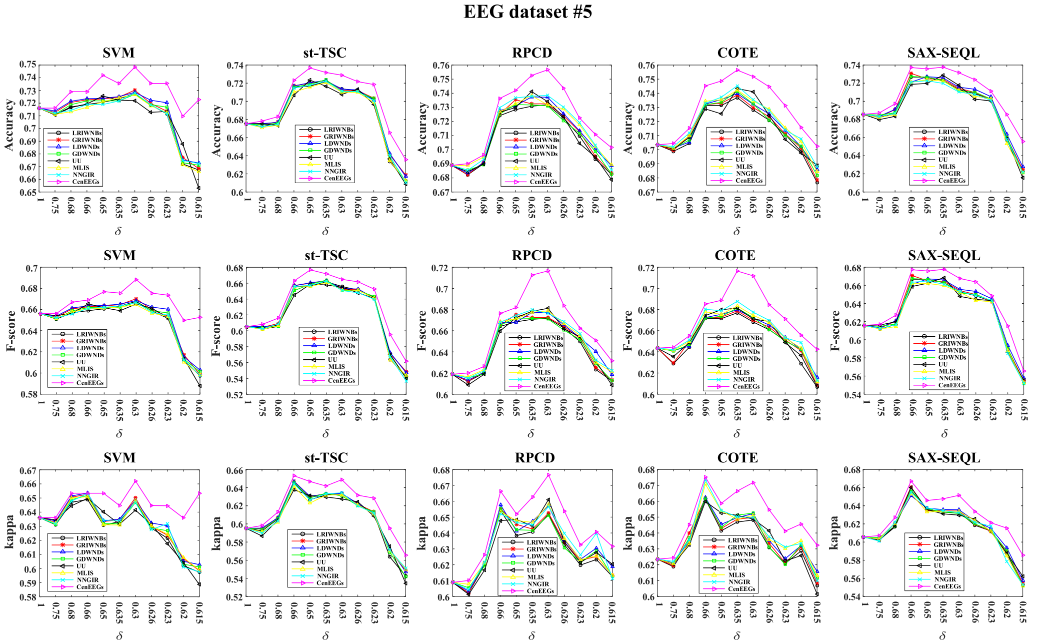


Fig. 7. Classification with different classifiers on EEG dataset #5.

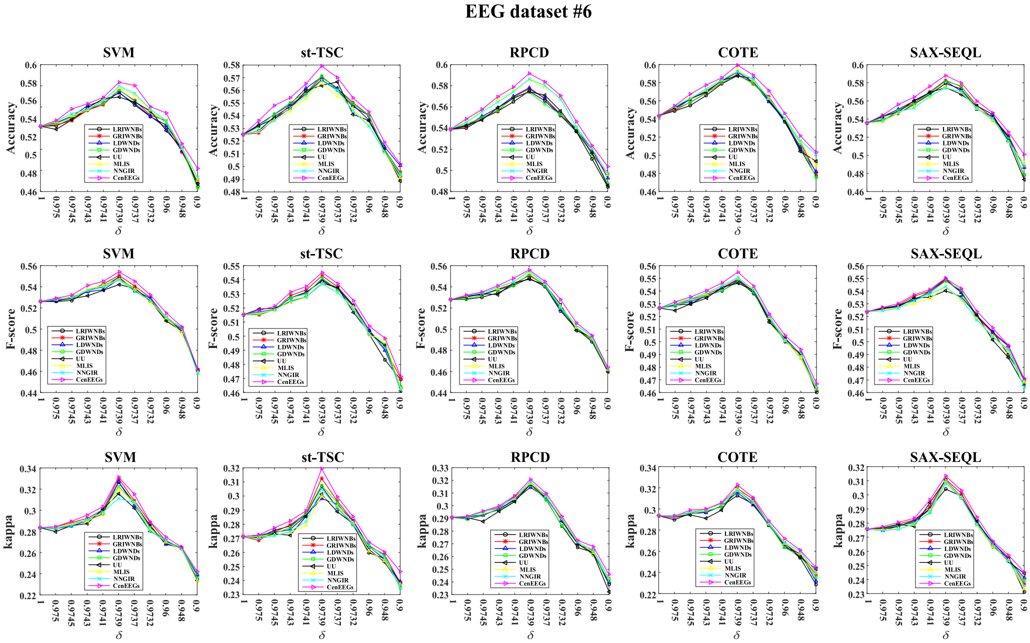


Fig. 8. Classification with different classifiers on EEG dataset #6.

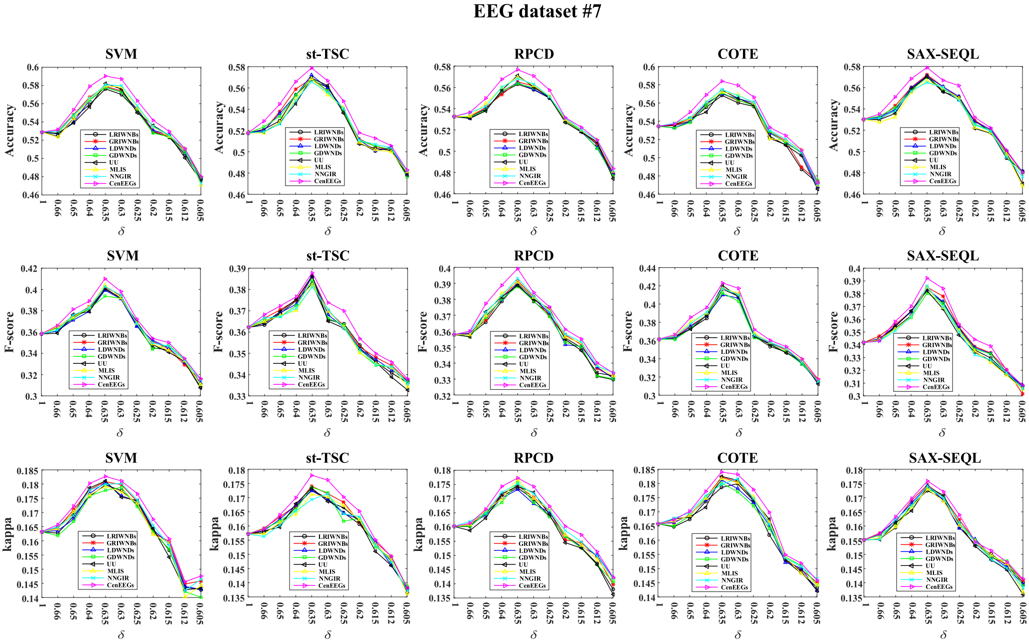


Fig. 9. Classification with different classifiers on EEG dataset #7.

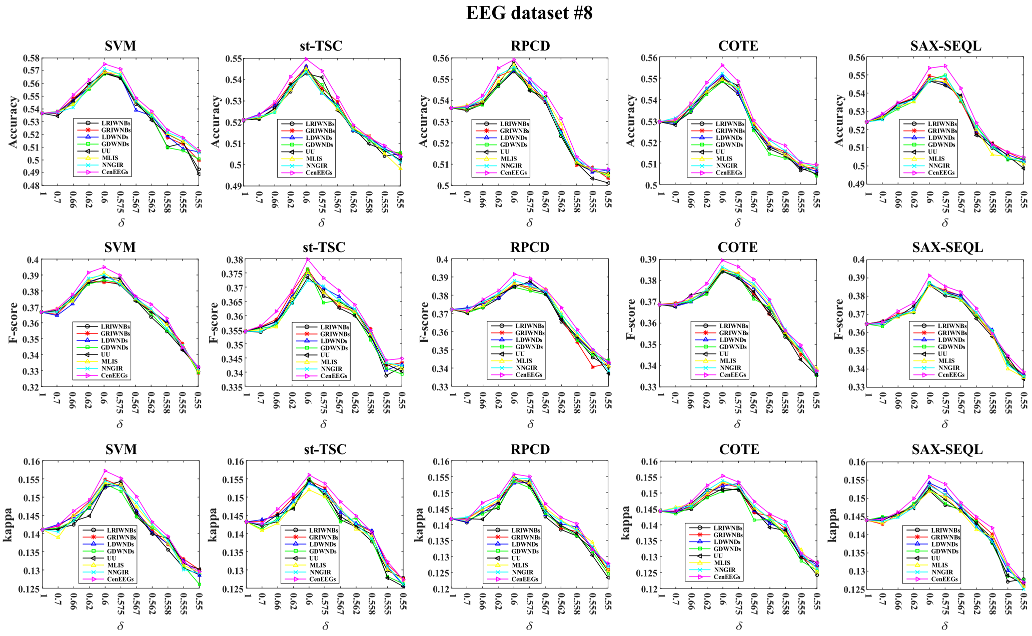


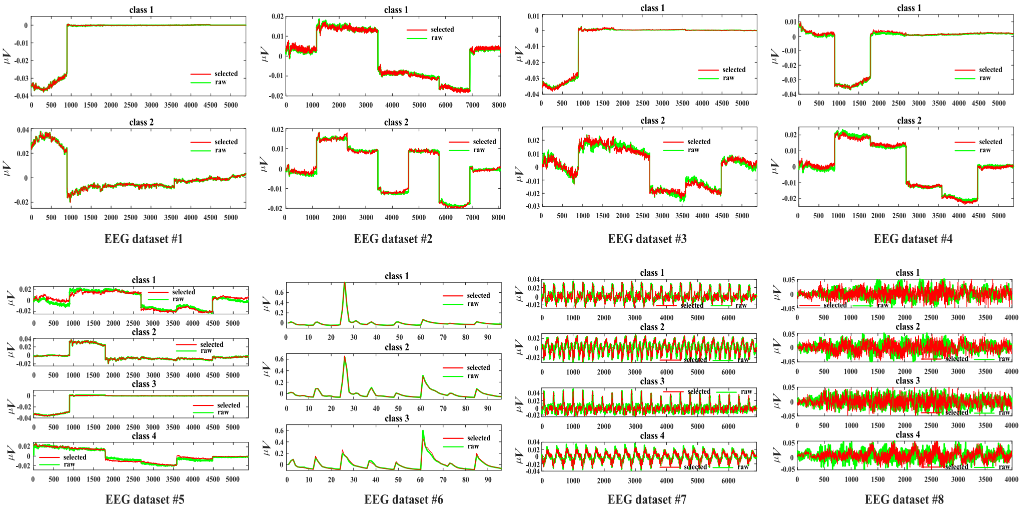
Fig. 10. Classification with different classifiers on EEG dataset #8.

EEG datasets, since the results of classification accuracy, F-score, and kappa on different similarity thresholds are all higher than those of LRIWNBs, GRIWNBs, LDWNDs, GDWNBs, UU, MLIS, and NNGIR. Namely, CenEEGs can select more valid EEG signals so as to reduce the impact of invalid ones on such classifiers for EEG classification and provide more distinguished EEG features for the classifiers.

4.4.2 Compared with Raw EEG without Selection. To firmly show the efficacy of CenEEGs for valid EEG selection, we also computed the classification accuracy improvements of five different classifiers with selected valid EEG signals by CenEEGs (under best similarity threshold that is discussed in Section 4.5) over that without selecting (i.e., raw EEG signals). Namely, the improvement is mathematically defined as $improvement = \frac{Acc_{sel} - Acc_{raw}}{Acc_{raw}} \times 100\%$ (where Acc_{sel} and Acc_{raw} denote the classification accuracy with selected EEG signals and with non-selected EEG, respectively). The improvements of classification accuracy on 8 EEG datasets are highlighted in boldface in Table 2, which demonstrates that classifiers achieve higher accuracy with selected valid EEG by CenEEGs than that without selection (raw EEG signals). Figures 3–10 actually also show the improvements in classification accuracy of five classifiers with CenEEGs selecting valid EEG signals. In the meantime, we also compared the selected EEG centroid sequence with raw EEG's and it clearly shows that the pattern (shape) of selected EEG are similar to that of raw EEG and the patterns of different classes are significantly distinct from each other. In other words, the proposed centroid extraction method can select the majority of valid EEG containing distinctly similar patterns of raw EEG (see Figure 11) (strictly, if the invalid EEG signals that have similar patterns with valid ones, they would be selected as valid EEG signals). Besides, the classification results of different classifiers in Figures 3–10 also correspondingly show the efficacy of the proposed method on centroid sequence extraction as well as on EEG classification improvement with selected valid EEG.

Table 2. EEG Classification Accuracy Improvements with CenEEGs

Classifier	Accuracy	#1	#2	#3	#4	#5	#6	#7	#8
SVM	Raw	0.8652	0.5152	0.8590	0.8831	0.7161	0.5418	0.5289	0.5366
	Selected	0.9101	0.5909	0.9872	1	0.7484	0.5833	0.5904	0.5753
	Improvement (%)	**5.19**	**14.69**	**14.92**	**13.24**	**4.51**	**7.66**	**11.63**	**7.21**
st-TSC	Raw	0.8386	0.5138	0.8415	0.8657	0.6752	0.5252	0.5176	0.5211
	Selected	0.9008	0.5488	0.9377	0.9538	0.7368	0.5792	0.5788	0.5498
	Improvement (%)	**7.42**	**6.81**	**11.43**	**10.18**	**9.12**	**10.28**	**11.82**	**5.51**
RPCD	Raw	0.8411	0.5216	0.8537	0.8743	0.6889	0.5389	0.5326	0.5365
	Selected	0.8962	0.5982	0.9425	0.9428	0.7566	0.5917	0.5766	0.5592
	Improvement (%)	**6.55**	**14.69**	**10.40**	**7.83**	**9.83**	**9.80**	**8.26**	**4.23**
COTE	Raw	0.8515	0.5355	0.8678	0.8816	0.7035	0.5435	0.5345	0.5294
	Selected	0.9117	0.6088	0.9533	0.9633	0.7564	0.5994	0.5839	0.5561
	Improvement (%)	**7.07**	**13.69**	**9.85**	**9.27**	**7.52**	**10.29**	**9.24**	**5.04**
SAX-SEQL	Raw	0.8506	0.5268	0.8485	0.8625	0.6857	0.5357	0.5302	0.5241
	Selected	0.9023	0.5732	0.9479	0.9144	0.7377	0.5881	0.5789	0.5549
	Improvement (%)	**6.08**	**8.81**	**11.71**	**6.02**	**7.58**	**9.78**	**9.19**	**5.88**
	Average improvement	**6.462**	**11.738**	**11.662**	**9.308**	**7.712**	**9.562**	**10.028**	**5.574**



EEG dataset #1

EEG dataset #2

EEG dataset #3

EEG dataset #4

EEG dataset #5

EEG dataset #6

EEG dataset #7

EEG dataset #8

Fig. 11. Centroid sequences of selected EEG and raw EEG.

4.4.3 *Compared with Non-selected EEG Centroid.* To firmly establish the efficacy of the proposed method for valid EEG selection, we also compared the centroid sequences of selected EEG signals with non-selected ones in the article. The results are illustrated in Figure 12 and they clearly show that the centroid sequences of selected EEG signals are significantly different with those of non-selected ones, which indicates that the non-selected EEG signals have different patterns (shapes) with selected EEG. Furthermore, the amount of non-selected EEG is quite smaller than selected EEG, which also results in the difference between their patterns of selected EEG and

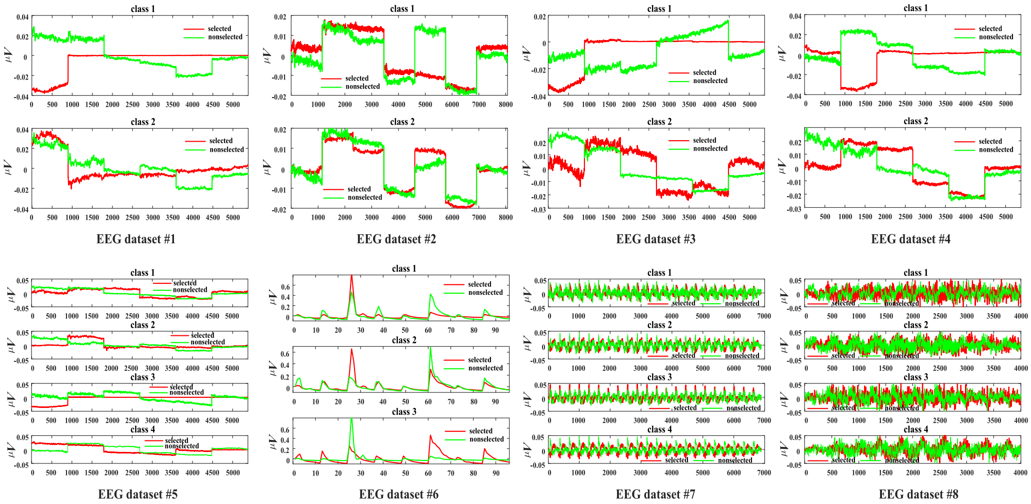


Fig. 12. Centroid sequences of selected EEG and non-selected EEG.

non-selected ones. In other words, the minority of non-selected EEG trials containing different patterns likely degrades the learning performance of classifiers from general patterns of EEG trials and they may eventually influence the classification. Additionally, this fact is also verified by the comparison between selected EEG and raw ones as shown above. But honestly, some actually valid EEG signals whose pattern is dissimilar with selected ones may be regarded as invalid EEG and left in non-selected ones. Similarly, the selected valid EEG signals may be mixed with few invalid EEG whose pattern is similar to that of selected ones.

4.5 Impact of Similarity Threshold

As we described above, the similarity threshold δ contributes to EEG selection, in that it determines the number of selected EEG signals. The larger the δ is, the smaller the similarities it requires. The smaller δ, the fewer EEG trials it selects. Here, we discuss the impact of δ on the classification results of different selection approaches.

As Figure 13 shows, the amount of selected valid EEG signals decreases with decreases of the similarity threshold, since a smaller δ requires a higher degree of similarity between two EEG signals, as it rejects those EEG signals with lower similarities to the centroid. Furthermore, with smaller δ, CenEEGs selects more specific EEG signals to be included in the centroid, thus providing classifiers with more specific features. Conversely, a larger δ provides classifiers with more general features.

Furthermore, Figures 3–10 also clearly show that a moderate δ yields the best classification results. Since a too small δ requires high similarities between the EEG signals (that not yet selected) and the centroid, it results in only a few EEG signals being selected with such δ, finally producing more specific features for classifiers to learn and a more specific model to classify EEG signals. This is also the reason why the classification decreases in most cases along with relatively smaller δ. On the contrary, a too large δ selects more EEG signals that contain more invalid EEG signals, which in turn degrade the feature extraction for classifiers so as to produce a lower classification result (as indicated by Figures 3–10).

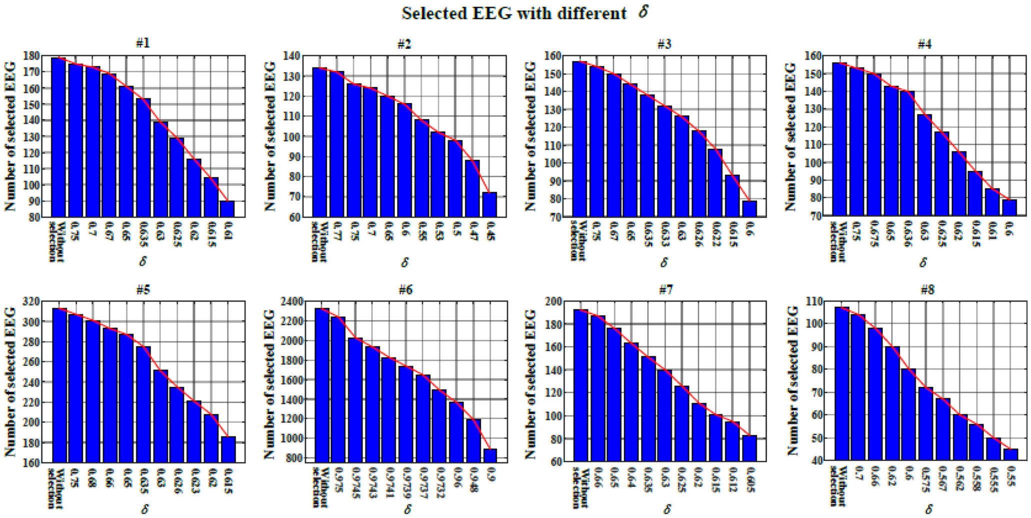


Fig. 13. Selected EEG signals with different similarity threshold δ.

4.6 Similarity Measure Analysis

CenEEGs applies a scale-and-shift-invariance metric to measure EEG similarities. To demonstrate the superiority of the similarity metric, we compared it with the classic and widely used similarity measures: ED [18], DTW [27, 39], and HD [45] for CenEEGs. Figure 14 shows the classification accuracy of CenEEGs with ED, DTW, HD, and our scaling/shifting invariant similarity metric, with optimal similarity thresholds as described in Section 4.5. The results clearly demonstrate that the scaling and shifting invariant similarity metric in CenEEGs is superior over ED, DTW, and HD.

4.7 Execution Time Analysis

To further evaluate the efficacy of CenEEGs, we analyzed execution time of CenEEGs for EEG selection by comparing CenEEGs with seven state-of-the-art time series selection methods introduced in Section 4.3. The time consumption comparison is illustrated in Figure 15. Although CenEEGs applies an optimization function to iteratively search the EEG centroid, it still has competitive efficiency compared with all the selection methods. Moreover, compared with its outstanding performance on improving classification, the runtime costs of CenEEGEs are negligible. Therefore, CenEEGs can be recommended, as it is acceptable to spend longer time selecting EEG signals for different classifiers.

5 CONCLUSIONS AND FUTURE WORK

Aiming to reduce the degradation of invalid EEG on classification for EEG-based disease diagnosis or BCI research, this article explored several EEG selection methods, and proposed a novel approach named CenEEGs for EEG selection from the view of source control, which is suitable for different classifiers. CenEEGs is a centroid-based approach for EEG selection with a scale-and-shift-invariance similarity metric, which is better than the classic and widely used ED, DTW, and HD. Subsequently, CenEEGs applies a globally optimized centroid searching strategy to find the reference sequence based on the similarity metric, and selects valid EEG signals with the centroid with respect to the similarity threshold. CenEEGs greatly improves the classification accuracy for classifiers, comparing with that of non-selected EEG signals. Besides, the results of comparing CenEEGs with several time series selection methods for five classifiers (SVM, st-TSC, RPCD, COTE,

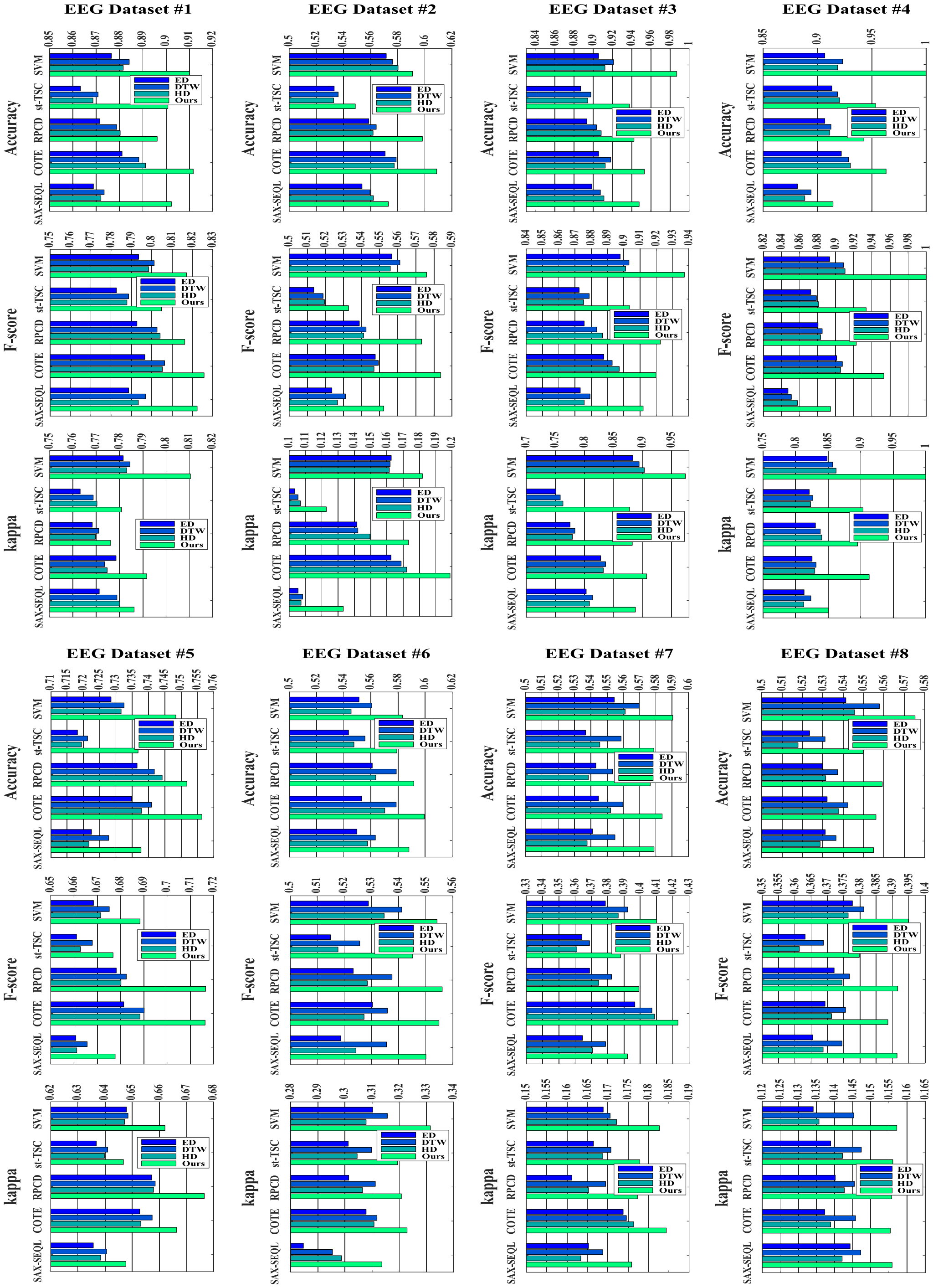


Fig. 14. Impact of different similarity measures on CenEEGs.

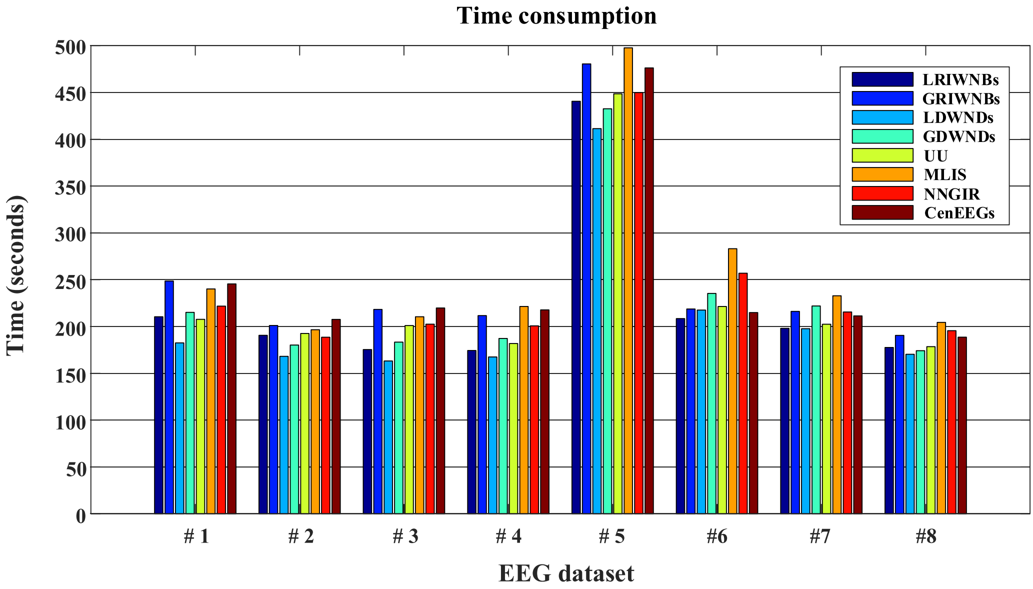


Fig. 15. Time consumption comparisons on 8 EEG datasets.

and SAX-SEQL) on 8 EEG datasets demonstrated that CenEEGs yielded the best classification performance among seven state-of-the-art time series selection methods (i.e., LRIWNBs, GRIWNBs, LDWNDs, GDWNBs, UU, MLIS, and NNGIR) with respect to classification accuracy, F-score, and kappa.

In this work, we investigated the efficacy of CenEEGs through using a particular similarity function as well as similarity threshold. In the future, it would be an interesting direction to evaluate the impact of more similarity measures on CenEEGs and to apply more heuristics to determine a suitable similarity threshold for CenEEGs on different EEG datasets. Furthermore, it is important and necessary to enhance the efficiency of CenEEGs computations as well. Apart from that, we also intend to investigate the capacity of CenEEGs to detect EEG signals represented in other domains such as wavelet-transformed EEG in the future.

REFERENCES

[1] Anna Aminov, Jeffrey M. Rogers, Stuart J. Johnstone, Sandy Middleton, and Peter H. Wilson. 2017. Acute single channel EEG predictors of cognitive function after stroke. *PLOS One* 12, 10 (2017), Article e0185841.

[2] Joakim Andén and Stéphane Mallat. 2014. Deep scattering spectrum. *IEEE Transactions on Signal Processing* 62, 16 (2014), 4114–4128.

[3] Kai Keng Ang and Cuntai Guan. 2017. EEG-based strategies to detect motor imagery for control and rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation on Engineering* 25, 4 (2017), 392–401.

[4] Álvar Arnaiz-González, José F. Díez-Pastor, Juan J. Rodríguez, and César Ignacio García-Osorio. 2016. Instance selection for regression by discretization. *Expert Systems with Applications* 54 (2016), 340–350.

[5] Claudio Babiloni, Claudio Del Percio, Roberta Lizio, Giuseppe Noce, Susanna Lopez, Andrea Soricelli, Raffaele Ferri, Flavio Nobili, Dario Arnaldi, Francesco Fama, Dag Aarsland, Francesco Orzi, Carla Buttinelli, Franco Giubilei, Marco Onofrj, Fabrizio Stocchi, Paola Stirpe, Peter Fuhr, Ute Gschwandtne, jGerhard Ransmayr, Heinrich Garn, Lucia Fraioli, Michela Pievani, Giovanni B. Frisoni, Fabrizia D'Antonio, Carlo De Lena, Bahar Guntekin, Lutfu Hanoglu, Erol Basar, Gorsev Yener, Derya Durusu Emek-Savas, Antonio Ivano Triggiani, Raffaella Franciotti, John Paul Taylor, Laura Vacca, Maria Francesca De Pandis, and Laura Bonanni. 2018. Abnormalities of resting-state functional cortical connectivity in patients with dementia due to Alzheimer's and Lewy body diseases: An EEG study. *Neurobiology of Aging* 65 (2018), 18–40.

[6] Anthony Bagnall, Jason Lines, Jon Hills, and Aaron Bostrom. 2015. Time-series classification with COTE: The collective of transformation-based ensembles. *IEEE Transactions on Knowledge and Data Engineering* 27, 9 (2015), 2522–2535.

[7] James C. Bezdek and Ludmila I. Kuncheva. 2001. Nearest prototype classifier designs: An experimental study. *International Journal of Intelligent Systems* 16, 12 (2001), 1445–1473.

[8] Eric Billet, Andriy Fedorov, and Nikos Chrisochoides. 2008. The use of robust local Hausdorff distances in accuracy assessment for image alignment of brain MRI. *The Insight Journal* (2008). http://hdl.handle.net/1926/1354.

[9] J. Caicedo-Acosta, D. Cárdenas-Pena, D. Collazos-Huertas, J. I. Padilla-Buritica, G. Castano-Duque, and G. Castellanos-Dominguez. 2019. Multiple-instance lasso regularization via embedded instance selection for emotion recognition. In *Proceedings of International Work-Conference on the Interplay Between Natural and Artificial Computation (IWINAC'19)*. Springer, Almería, Spain, 244–251.

[10] Julian Caicedo-Acosta, Luisa Velasquez-Martinez, David Cárdenas-Pena, and Germán Castellanos-Dominguez. 2018. Multiple instance learning selecting time-frequency features for Brain Computing Interfaces. In *Proceedings of International Workshop on Artificial Intelligence and Pattern Recognition (IWAIPR'18)*. Springer, Havana, Cuba, 326–333.

[11] James F. Cavanagh, Praveen Kumar, Andrea A. Mueller, Sarah Pirio Richardson, and Abdullah Mueen. 2018. Diminished EEG habituation to novel events effectively classifies Parkinson's patients. *Clinical Neurophysiology* 129, 2 (2018), 409–418.

[12] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 3 (2011), Article 27.

[13] Kelvin Kam Wing Chu and Man Hon Wong. 1999. Fast time-series searching with scaling and shifting. In *Proceedings of the 18th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. ACM, Philadelphia, PA, 237–248.

[14] Václav Chudáček, Joakim Andén, Stéphane Mallat, Patrice Abry, and Muriel Doret. 2014. Scattering transform for intrapartum fetal heart rate variability fractal analysis: A case-control study. *IEEE Transactions on Biomedical Engineering* 61, 4 (2014), 1100–1108.

[15] Chenglong Dai, Dechang Pi, Lin Cui, and Yanlong Zhu. 2018. MTEEGC: A novel approach for multi-trial EEG clustering. *Applied Soft Computing* 71 (2018), 255–267.

[16] Chenglong Dai, Jia Wu, Dechang Pi, and Lin Cui. 2018. Brain EEG time series selection: A novel graph-based approach for classification. In *Proceedings of SIAM International Conference on Data Ming (SDM'18)*. SIAM, San Diego, CA, 558–566.

[17] Anne Driemel, Amer Krivosǐja, and Christian Sohler. 2016. Clustering time series under the Fréchet distance. In *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete algorithms (SODA'16)*. SIAM, Arlington, Virginia, 766–785.

[18] Christos Faloutsos, M. Ranganathan, and Yannis Manolopoulos. 1994. Fast sub-sequence matching in time-series databases. In *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data (SIGMOD'94)*. ACM, Minneapolis, Minnesota, 419–429.

[19] Matteo Fraschini, Matteo Demuru, Arjan Hillebrand, Lorenza Cuccu, Silvia Porcu, Francesca Di Stefano, Monica Puligheddu, Gianluca Floris, Giuseppe Borghero, and Francesco Marrosu. 2016. EEG functional network topology is associated with disability in patients with amyotrophic lateral sclerosis. *Scientific Reports* 6 (2016), Article 38653.

[20] Salvador Garcia, Joaquin Derrac, Jose Cano, and Francisco Herrera. 2012. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 3 (2012), 417–435.

[21] Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. 2016. A multi-modal parcellation of human cerebral cortex. *Nature* 536 (2016), 171–178.

[22] Gene H. Golub and Charles F. Van Loan. 1996. *Matrix Computations* (3rd ed.). Johns Hopkins University Press, Baltimore, Maryland.

[23] Guoliang He, Yong Duan, Yifei Li, Tieyun Qian, Jinrong He, and Xiangyang Jia. 2015. Active learning for multivariate time series classification with positive unlabeled data. In *Proceedings of 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI'16)*. IEEE, Vietri sul Mare, Italy, 178–185.

[24] Ramy Hussein, Mohamed Elgendi, Z. Jane Wang, and Rabab K. Ward. 2018. Robust detection of epileptic seizures based on L1-penalized robust regression of EEG signals. *Expert Systems with Applications* 104 (2018), 153–167.

[25] Norbert Jankowski and Marek Grochowski. 2004. Comparison of instances selection Algorithms I. Algorithms survey. In *Proceedings of Artificial Intelligence and Soft Computing (ICAISC'04)*. Springer, Zakopane, Poland, 598–603.

[26] Liangxiao Jiang. 2012. Learning instance weighted naive Bayes from labeled and unlabeled data. *Journal of Intelligent Information Systems* 38 (2012), 257–268.

[27] Eamonn Keogh and Chotirat Ann Ratanamahatana. 2005. Exact indexing of dynamic time warping. *Knowledge and Information Systems* 7, 3 (2005), 358–386.

[28] Ji-Hyun Kim. 2009. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics and Data Analysis* 53, 11 (2009), 3735–3745.

[29] Wai Lam, Chi-Kin Keung, and Danyu Liu. 2002. Discovering useful concept prototypes for classification based on filtering and abstraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 8 (2002), 1075–1090.

[30] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. 2007. Experiencing sax: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery* 15, 2 (2007), 107–144.

[31] Jessica Lin, Michail Vlachos, Eamonn Keogh, and Dimitrios Gunopulos. 2004. Iterative incremental clustering of time series. In *Proceedings of International Conference on Extending Database Technology (EDBT'04)*. Springer, Heraklion, Crete, Greece, 106–122.

[32] Jason Lines, Luke M. Davis, Jon Hills, and Anthony Bagnall. 2012. A shapelet transform for time series classification. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*. ACM, Beijing, China, 289–297.

[33] Jason Lines, Sarah Taylor, and Anthony Bagnall. 2018. Time series classification with HIVE-COTE: The hierachical vote collective of transformation-based ensembles. *ACM Transactions on Knowledge Discovery from Data* 12, 5 (2018), Article 52.

[34] Chuan Liu, Wenyong Wang, Meng Wang, Fengmao Lv, and Martin Konan. 2017. An efficient instance selection algorithm to reconstruct training set for support vector machine. *Knowledge-Based Systems* 116 (2017), 58–73.

[35] Li Liu, Jiasong Wu, Dengwang Li, Lotfi Senhadji, and Huazhong Shu. 2019. Fractional wavelet scattering network and applications. *IEEE Transactions on Biomedical Engineering* 66, 2 (2019), 553–563. DOI: https://doi.org/10.1109/TBME.2018.2850356

[36] Eduardo Zárate Max, Ricardo Marcondes Marcacini, and Edson Takashi Matsubara. 2018. Improving instance selection via metric learning. In *Proceedings of 2018 International Joint Conference on Neural Networks (IJCNN'18)*. IEEE, Rio de Janeiro, Brazil, 1–6.

[37] Thach Le Nguyen, Severin Gsponer, and Georgiana Ifrim. 2017. Time series classification by sequence learning in all-subsequence space. In *Proceedings of 2017 IEEE 33rd International Conference on Data Engineering (ICDE'17)*. IEEE, San Diego, CA, 947–958.

[38] John Paparrizos and Luis Gravano. 2016. k-Shape: Efficient and accurate clustering of time series. *ACM SIGMOD Record* 45, 1 (2016), 69–76.

[39] Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. 2013. Addressing big data time series: Mining trillions of time series subsequences under dynamic time warping. *ACM Transactions on Knowledge Discovery from Data* 7, 3 (SI) (2013), Article 10.

[40] Diego F. Silva, Viníus M. A. De Souza, and Gustavo E. A. P. A. Batista. 2013. Time series classification using compression distance of recurrence plots. In *Proceedings of 2013 IEEE 13th International Conference on Data Mining (ICDM'13)*. IEEE, Dallas, TX, 687–696.

[41] Siuly and Yan Li. 2014. A novel statistical algorithm for multiclass EEG signal classification. *Engineering Applications of Artificial Intelligence* 34 (2014), 154–167.

[42] Otis Smart and Lauren Burrell. 2015. Genetic programming and frequent itemset mining to identify feature selection patterns of iEEG and fMRI epilepsy data. *Engineering Applications of Artificial Intelligence* 39 (2015), 198–214.

[43] Yunsheng Song, Jiye Liang, Jing Lu, and Xingwang Zhao. 2017. An efficient instance selection algorithm for k-nearest neighbor regression. *Neurocomputing* 251 (2017), 26–34.

[44] Milos B. Stojanović, Milos M. Bozić, Milena M. Stanković, and Zoran P. Stajić. 2014. A methodology for training set instance selection using mutual information in time series prediction. *Neurocomputing* 141, SI (2014), 236–245.

[45] Abdel Aziz Taha and Allan Hanbury. 2015. An efficient algorithm for calculating the exact Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 11 (2015), 2153–2163.

[46] Anthoula C. Tsolaki, Vasiliki Kosmidou, Ioannis (Yiannis) Kompatsiaris, Chrysa Papadaniil, Leontios Hadjileontiadis, Aikaterini Adam, and Magda Tsolaki. 2017. Brain source localization of MMN and P300 ERPs in mild cognitive impairment and Alzheimer's disease: A high-density EEG approach. *Neurobiology of Aging* 55 (2017), 190–201.

[47] D. Randall Wilson and Tony R. Martinez. 2000. Reduction techniques for instance-based learning algorithms. *Machine Learning* 38, 3 (2000), 257–286.

[48] Jia Wu, Shirui Pan, Zhihua Cai, Xingquan Zhu, and Chengqi Zhang. 2014. Dual instance and attribute weighting for naive Bayes classification. In *Proceedings of 2014 International Joint Conference on Neural Networks (IJCNN'14)*. IEEE, Beijing, China, 1675–1679.

[49] Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11)*. ACM, Hong Kong, China, 177–186.

[50] Lijun Yang, Qingsheng Zhu, Jinlong Huang, Donggong Cheng, Quanwang Wu, and Xiaolu Hong. 2018. Natural neighborhood graph-based instance reduction algorithm without parameters. *Applied Soft Computing* 70 (2018), 279–287.

[51] Zhang Yin, Yongxiong Wang, Li Liu, Wei Zhang, and Jianhua Zhang. 2017. Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination. *Frontiers in Neurorobotics* 11 (2017), Article 19. DOI: https://doi.org/10.3389/fnbot.2017.00019

[52] Tingting Zhai and Zhenfeng He. 2013. Instance selection for time series classification based on immune binary particle swarm optimization. *Knowledge-Based Systems* 49 (2013), 106–115.

[53] Shichao Zhang, Xuelong Li, Min Zong, Xiaofeng Zhu, and Debo Cheng. 2017. Learning k for kNN classification. *ACM Transactions on Intelligent Systems and Technology* 8, 3 (SI) (2017), Article 43.

Received February 2019; revised September 2019; accepted November 2019