Shapelet-transformed Multi-channel EEG Channel Selection

CHENGLONG DAI, Nanjing University of Aeronautics and Astronautics, China
DECHANG PI, Nanjing University of Aeronautics and Astronautics, China
STEFANIE I. BECKER, The University of Queensland, Australia

This article proposes an approach to select EEG channels based on EEG shapelet transformation, aiming to reduce the setup time and inconvenience for subjects, and to improve the applicable performance of Brain-Computer Interfaces (BCIs). In detail, the method selects top-k EEG channels by solving a logistic loss-embedded minimization problem with respect to EEG shapelet learning, hyperplane learning, and EEG channel weight learning simultaneously. Especially, to learn distinguished EEG shapelets for weighting contributions of each EEG channel to the logistic loss, EEG shapelet similarity is also minimized during the procedure. Furthermore, the gradient descent strategy is adopted in the article to solve the non-convex optimization problem, which finally leads to the algorithm termed StEEGCS. In a result, classification accuracy, with those EEG channels selected by StEEGCS, is improved compared to that with all EEG channels, and classification time consumption is reduced as well. Additionally, the comparisons with several state-of-the-art EEG channel selection methods on several real-world EEG datasets also demonstrates the efficacy and superiority of StEEGCS.

CCS Concepts: • **Mathematics of computing** → **Time series analysis**; **Dimensionality reduction**; • **Computing methodologies** → **Supervised learning by classification**; • **Applied computing** → *Bioinformatics*;

Additional Key Words and Phrases: EEG channel selection, EEG shapelets, channel contribution, shapelet similarity minimization

ACM Reference Format:
Chenglong Dai, Dechang Pi, and Stefanie I. Becker. 2020. Shapelet-transformed Multi-channel EEG Channel Selection. *ACM Trans. Intell. Syst. Technol.* 1, 1, Article 1 (January 2020), 29 pages.
https://doi.org/10.1145/3397850

1 INTRODUCTION

Electroencephalogram (EEG) signal is widely used to diagnose neuropsychiatric disorders such as Alzheimers Disease (AD) [6, 24], epileptic seizure [49, 50], stroke [45], and so on [52], and it also practically applied in Brain-Computer Interfaces (BCIs or Human-Machine Interfaces (HMIs)) [3, 14, 48], since it can reflect the states and functions of human brain and even the whole body [16]. For specific functions, they are activated in accordingly specific positions of brain. For instance, the motor tasks including motor imagery are related to

Authors' addresses: Chenglong Dai, Nanjing University of Aeronautics and Astronautics, 29 Jiangjun Avenue, Nanjing, Jiangsu Province, 211106, China, chenglongdai@nuaa.edu.cn; Dechang Pi, Nanjing University of Aeronautics and Astronautics, Nanjing, China, dc.pi@nuaa.edu.cn; Stefanie I. Becker, The University of Queensland, Brisbane, Queensland, Australia, s.becker@psy.uq.edu.au.

2157-6904/2020/1-ART1 $15.00
https://doi.org/10.1145/3397850

motor cortex; the visual-based tasks are located in the primary visual cortex area, and the frontoparietal regions correspond to decision making [19, 47]. As one type of biological potentials that can be recorded in a non-invasive way, many channels or electrodes of EEG acquisition equipment are commonly used in wide applications. Although more channels theoretically provide more information of brain functions, it correspondingly causes high dimensional and redundant EEG data, since (1) as introduced above, specific functions are activated in specific cortex of brain, channels attached on non-specific areas are useless or redundant for analysis; (2) channels that have small contributions for EEG analysis result in extra time or space to process and analyze, without significantly improving the performance of analytical methods such as classification; (3) it increases the inconvenience for subjects in BCI-based applications when using more EEG channels. Consequently, EEG channel selection is a necessary process for its follow-up analysis, especially for BCI-based applications in our daily life, because it can significantly reduce the impact of noisy/redundant channels and promote the contributions of informative channels for EEG analysis. Particularly, less but more informative EEG channels are also beneficial to BCI applications. Besides, channel selection methods can identify informative and suitable recording sites from a large amount of sites without any prior knowledge of specific cerebral tasks, and reduce the redundancy of EEG electrodes for EEG signal classification without losing its performance [30]. Figure 1 briefly presents the disadvantages (or problems) of existing researches and applications using a large number of EEG channels. Besides, it also shows the reasons why we are going to solve the problem by listing the advantages of EEG channel reduction for its applications. However, how to perform optimal EEG channel selection is not a trivial task, since manually selecting EEG channels based on experts' knowledge does not guarantee to achieve better results compared to that with all channels [7].

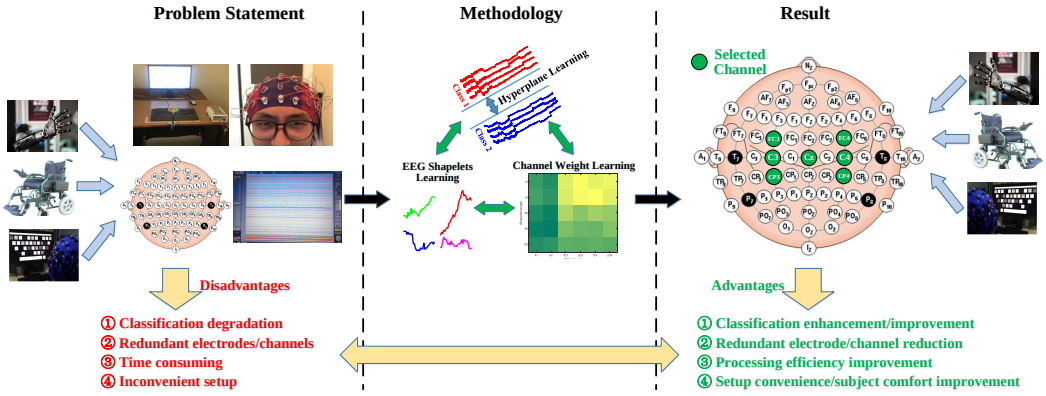


Fig. 1. The aim of the proposed method. Problem statement presents that most researches used too many EEG channels/electrodes to record and analyze EEG signals for BCI-based applications (*i.e.,* wheelchair navigation, robotic arm control, and EEG-based spelling, and so on), which involves in some shortcomings; Methodology briefly explains the EEG shapelet-transformed channel selection proposed in the article, which mainly aims to solve a logistic loss-embedded minimization function to learn distinct EEG shapelets, hyperplane, and channel weights/contributions simultaneously. In the end, EEG channels are selected based on the channel weights; Result indicates the selected EEG channels based on their channel contributions to classification performance, with which the efficacy of BCI-based applications can be significantly enhanced or improved. The advantages of EEG channel selection are also summarized in the figure.

In this article, we handled the challenging task of EEG channel selection for its classification. In detail, it utilizes EEG shapelets to represent original multi-channel EEG signals and then weights EEG channel contributions according to their logistic loss in shapelet-transformed space. Finally, the top-k EEG channels that contribute more to the logistic loss are selected out. The procedure of the proposed method is also roughly introduced in Fig. 1. Further, the contributions of this article are highlighted below.

- Shapelet-transformed EEG channel selection is mapped to an minimization objective function of logistic loss that simultaneously considers distinct EEG shapelets, optimal EEG channel contributions to classification performance, and good hyperplane for classifier.
- EEG shapelet similarity minimization is also considered to learn distinct EEG shapelets that highly represent original multi-channel EEG signals.
- A novel approach of EEG channel selection is proposed that we call StEEGCS by using gradient descent to learn EEG shapelets, channel weights, and hyperplane of classifier.
- Experimental results compared to five classic and state-of-the-art methods on 10 real-world EEG datasets demonstrate the efficacy and superiority of StEEGCS for multi-channel EEG channel selection.

The remainder of this article is organized as follows: The related work on EEG channel selection is reviewed in Section 2. The proposed method is introduced in Section 3, followed by the selection algorithm StEEGCS in Section 4. Then a detailed experiment is carried out in Section 5. Finally, a summary of the article is presented in Section 6.

2 RELATED WORK

Brain-Computer Interface (BCI) provides a good way for rehabilitation or a tool to improve living quality of the disabled by using brain to control wheelchair or robotic arms. In BCI-based applications, EEG classification plays an important role. Netzer et al [38] proposed an approach for EEG classification in BCI applications by using core sets that link to BCI with data summarization. Jafarifarmand [23] proposed a new framework that firstly applied an artifact rejected common spatial pattern to reduce artifacts, and then utilized a strategy named self-regulated supervised Gaussian fuzzy adaptive system Art to classify motor imagery EEG signals. He et al [20] classified motor imagery EEG signals by using a common Bayesian network constructed by related EEG channels connected by common and varying edges. Additionally, many state-of-the-art methods for EEG classification are emerged out recently, such as [1, 17, 27]. But the classification accuracy and efficiency influence the performance of BCI applications in daily life. To improve the efficiency and accuracy of EEG classification, EEG channel selection is subsequently adopted. In recent decades, many studies on EEG channel selection or reduction for BCI are merged out and have achieved a lot in this field [2]. In this article, we generally categorized these methods into three classes: common spatial pattern-based methods, entropy/mutual information-based methods, and classifier-embedded methods.

Common spatial pattern (CSP)-based methods [36, 51] select EEG channels based on their CSP coefficients without deriving the features corresponding to each EEG channel. CSP-based channel selection methods consider all the EEG channels and they are highly sensitive to EEG artifacts, such as electrooculography and electromyography. Furthermore, as is known to all, CSP-based methods are also suffering from over-fitting. To deal with this problem, a regularized CSP and a sparse CSP are proposed in [12, 28, 35] and [4], respectively, both of which aim to remove irrelevant EEG channels and obtain sparse spatial

filters for EEG classification. They all achieved better performance than the classic CSP, but they still more or less suffer from overfitting and artifact sensitivity inherited from CSP.

Entropy [40, 41, 48, 53] or mutual information [29, 31] -based methods select EEG channels by ranking them with entropy or mutual information between EEG channels as well as the EEG classes. In detail, the most relevant channels according to their entropies or mutual information to corresponding classes are selected out. These methods are independent from specific classifiers, but they suffer from low accuracy due to their ignorance of correlations among different EEG channels [2].

Classifier-embedded EEG channel selection methods such as introduced in [15, 30], which are mainly based on the classification performance of a classifier to select EEG channels in a way that the channels with lower contribution to classification are eliminated. Actually, these classifier-embedded methods probably achieve higher classification accuracy than CSP or mutual information-based methods, since classifier-embedded channel selection methods directly consider the classification accuracy into the objective function that reflects the goal of channel selection for EEG classification. However, they rely on a specific classifier such as support vector machine (SVM) [4].

Despite those various methods, selecting optimal EEG channels for BCI-based applications remains a challenging task. Therefore, this article proposed an EEG shapelet-transformed approach to select EEG channels. In detail, the original EEG data are firstly mapped into EEG shapelet representation space, and then based on EEG shapelets, those EEG channels contribute more to the logistic loss are selected. Namely, EEG channels with higher weights or contributions to the classification are probably selected. Particularly, EEG shapelets significantly influence the procedure. To improve the efficiency and accuracy for EEG time series classification, many approaches for shapelet transform, learning, and selection, especially for time series [11, 21, 22] are proposed in existing studies. Ji et al [26] proposed an efficient shapelet discovery approach for time series classification based on important data points. Then, in their work [25], they proposed another algorithm for time series shapelet selection, which firstly sampled time series and then selected shapelets efficiently based on the local farthest deviation points (LFDPs) from sampled time series, which reduced time consumption a lot. Li et al [32] also proposed a shapelet discovery approach for time series classification that was named Pruning Shapelets with Key Points (PSKP) in a way that they applied standard deviation to search the key points of time series and then extract time series shapelets based on such key points. In their work [42], Rakthanmannon et al presented an efficient scalable algorithm of shapelets discovery for time series classification, which used SAX [39] strategy to extract shapelets of time series. Grabocka et al [18] exploited an objective function embedded classification to learn top-k shapelets for time series. Similar to [18], our method also utilized the classification-embedded strategy to select EEG channels. It exploited a logistic loss minimization function to simultaneously learn EEG shapelets, hyperplane, and EEG channel weights, which built a direct correlation between relevant EEG channels and classification performance. In other words, the proposed method can help classifier (*e.g.,* SVM) achieve the highest classification accuracy with selected EEG channels.

3 THE METHOD

In this section, we introduce the proposed method that is transformed to a minimization objective function with respect to EEG shapelets, channel contributions, and hyperplane weights. Additionally, the brief procedure of the method is also illustrated in Fig. 1. In detail, the proposed method selects EEG channels based on EEG channel weights/contributions to its classification performance. That is to say, to get EEG channel weights is the essential

goal of the method. To the end, the solution to achieve EEG channel weights is mainly transformed to EEG shapelet learning in a shapelet representation space. Firstly, a similarity minimization evaluation is brought in, which aims to learn the most informative and distinct EEG shapelets to represent original EEG. Then, a hyperplane is required to learn as well, which aims to achieve a better classifier for EEG classification along with selected EEG channels. In the method, a logistic loss function simultaneously integrated with distinct EEG shapelet learning, hyperplane learning, and EEG channel contribution learning is eventually proposed, and its aim is to minimize the logistic loss (namely, higher classification accuracy). As the objective function is non-convex and not differentiable, the gradient descent strategy is adopted to solve it. During the procedure, EEG shapelets, hyperplane, and EEG channel weights are learnt iteratively with gradient descent strategy until their integrated logistic loss is locally minimized. Besides, the detailed descriptions for gradient descent-based EEG shapelet learning, hyperplane learning, and EEG channel weight learning are respectively presented in detail as follows. Finally, the methodology leads to the algorithm of StEEGCS for EEG channel selection.

3.1 EEG Shapelet

Shapelet, widely applied in time series data mining [33, 54, 55], is a subsequence from original time series [37]. Similarly, EEG shapelet is a continuous EEG subsequence that inherits structures from the original EEG, and it is usually much shorter than the original EEG. Particularly, EEG shapelet is a small subsequence, as a pattern of original EEG, that can represent the original EEG data and also able to separate EEG into different groups based on their distance to EEG shapelets.

Definition 3.1. Given an EEG $\boldsymbol{e}$ *of length* m*, a shapelet* $\boldsymbol{s}_{i,l}$ *of* $\boldsymbol{e}$ *is a continuous subsequence with length* $l \leq m$*, that starts at position* i*. That is,* $\boldsymbol{s}_{i,l} = t_i, \cdots, t_{i+l-1}$*, where* $1 \leq i \leq m - l + 1$.

3.2 Shapelet-transformed Representation

Given an EEG dataset $\boldsymbol{E} = \{\boldsymbol{e}_i, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n\}_{N \times C \times M}$, where $\boldsymbol{e}_i$ denotes an EEG signal of C channels (*i.e.,* $\boldsymbol{e}_i = [\boldsymbol{e}_{i,1}, \boldsymbol{e}_{i,2}, \cdots, \boldsymbol{e}_{i,c}]_{C \times M}$) with M samples for each channel, and a set of channel shapelets $\boldsymbol{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_k\}_{K \times L}$, where $\boldsymbol{s}_k$ denotes the k^{th} shapelet with length of L, the distance of an EEG $\boldsymbol{e}_i$ to shapelet s_k is represented as $D_{i,k} \in \boldsymbol{D} \in \mathbb{R}^{L \times K}$ and $\boldsymbol{D}$ denotes their distance matrix or representation. Formally, we define $\boldsymbol{e}_{i,c,j}$ as the EEG segment of channel c and j is the starting point of segment. Obviously, there are $J = M - L + 1$ EEG segments for each EEG channel. The distance $D_{i,k}$ between channel shapelet $\boldsymbol{s}_k$ and EEG $\boldsymbol{e}_i$ can be calculated with Eq.1.

$$D_{i,k} = \min_{j=1,\ldots,J} \frac{1}{C \times L} \sum_{c=1}^{C} \left(\pi_{k,c} \cdot \sum_{l=1}^{L} \left(\boldsymbol{e}_{i,c,j+l-1} - \boldsymbol{s}_{k,c,l} \right)^2 \right) \tag{1}$$

where $\pi_{k,c} \in [0, 1)$ is the weight of each shapelet for a particular EEG channel, *i.e.,* the channel contribution with respect to a particular shapelet.

The shapelets learned from original EEG data can be transformed into a new representation $\boldsymbol{D} \in \mathbb{R}^{N \times C^* \times K}$ of $\boldsymbol{E} \in \mathbb{R}^{N \times C \times M}$, where C^* denotes the number of selected EEG channels. Obviously, this transformation reduces the dimension of original EEG, since $C^* < C$ and $K < C$.

Since Eq.1 is not a differentiable equation [43], the soft minimum function [18] is adopted to approximately compute $D_{i,k}$. In detail,

$$D_{i,k} \approx \frac{1}{C}\sum_{c=1}^{C}\left(\frac{\sum_{j=1}^{J} Q_{i,k,j}e^{\alpha Q_{i,k,j}}}{\sum_{j=1}^{J} e^{\alpha Q_{i,k,j}}}\pi_{k,c}\right) \tag{2}$$

and

$$Q_{i,k,j} = \frac{1}{L}\sum_{l=1}^{L}\left(\boldsymbol{e}_{i,c,j+l-1} - \boldsymbol{s}_{k,c,l}\right)^2 \tag{3}$$

where α is a precision control parameter of the approximation function. In the case of $\alpha \to -\infty$, Eq.2 approximates to Eq.1. In this article, we set $\alpha = -100$ according to [18].

3.3 Similarity Minimization for EEG Channel Shapelets

To learn distinct shapelets for each EEG class, we consider to minimize similarities for shapelets within and between EEG classes, as well introduced in [56]. For k shapelets, their similarity matrix is defined as $\boldsymbol{A} \in \mathbb{R}^{k\times k}$. Furthermore, let $A_{i,j} \in \boldsymbol{A}$ be the similarity between two shapelets $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, and $A_{i,j}$ can be computed as

$$A_{i,j} = e^{-\frac{\|Q_{i,j}\|^2}{\sigma^2}} \tag{4}$$

where $Q_{i,j}$ can be computed as similarly as Eq.3.

3.4 Learning Shapelets

The shapelet-transformed distance matrix is regarded as EEG features, and we utilize a linear classifier to predict the approximate target variable $\hat{\boldsymbol{Y}} \in \mathbb{R}^{N\times C\times K}$ with $\boldsymbol{D}$ and linear classification weights $\boldsymbol{W} \in \mathbb{R}^{K\times V}$. In detail,

$$Y_{i,v} = W_{0,v} + \sum_{k=1}^{K} D_{i,k}W_{k,v}, \quad \forall i \in \{1,2,\ldots,N\} \tag{5}$$

where $W_{0,v} \in \mathbb{R}$ denotes the bias for the v^{th} class. And we also use a logistic function to transform Eq.5. Namely,

$$\hat{Z}_{i,v} = \frac{e^{Y_{i,v}}}{1+e^{Y_{i,v}}}, \quad \forall i \in \{1,2,\ldots,N\} \tag{6}$$

Given that, EEG data commonly contain V classes, we can also transform the learning model into a one-to-all binary problem. In detail,

$$Z_{i,v} = \begin{cases} 1, & Z_i = v \\ 0, & Z_i \neq v \end{cases} \quad \forall i \in \{1,2,\ldots,N\}, \quad \forall v \in \{1,2,\ldots,V\} \tag{7}$$

Finally, we learn the model by minimizing the logistic loss between the true target $\boldsymbol{Z}$ and the estimated one $\hat{\boldsymbol{Z}}$,

$$\mathcal{L}(\boldsymbol{Z},\hat{\boldsymbol{Z}}) = -\boldsymbol{Z}\ln\hat{\boldsymbol{Z}} - (1-\boldsymbol{Z})\ln(1-\hat{\boldsymbol{Z}}) \tag{8}$$

The aim of the article is to select more important EEG channels based on EEG shapelet learning. Hence, we aim to minimize the logistic loss by jointly learning optimal EEG shapelets $\boldsymbol{S}$, channel contributions $\boldsymbol{\pi}$, and the optimal hyperplane $\boldsymbol{W}$, simultaneously.

$$\min_{\boldsymbol{S},\boldsymbol{\pi},\boldsymbol{W}} \mathcal{F} = \sum_{i=1}^{N}\sum_{v=1}^{V}\mathcal{L}\left(Z_i,\hat{Z}_i\right) + \frac{\lambda_W}{2}\|\boldsymbol{W}\|^2 + \frac{\lambda_S}{2}\|\boldsymbol{A}\|^2 \tag{9}$$

At right of Eq.9, the first term is the logistic loss embedded with channel contribution, which enhances the approximate target variable to approach the real one. The second term is the hyperplane that learns the linear classifier, and the last term diversifies the EEG channel shapelets through minimizing their similarities.

3.5 Shapelet-transformed EEG Channel Selection

In our method, we apply gradient descent technique to solve the non-convex optimization objective function. As stated in Eq.9, it mainly contains three variables such as EEG shapelet $\boldsymbol{S}$, channel contribution $\boldsymbol{\pi}$, and the optimal hyperplane $\boldsymbol{W}$. We update each of them with gradient descent technique by respectively fixing the other two variables.

3.5.1 Shapelet Gradient. To analyze the gradients of the objective function with respect to $\boldsymbol{S}$, we firstly fix $\boldsymbol{\pi}$ and $\boldsymbol{W}$, and then the objective function Eq.9 degenerates to Eq.10.

$$\min_{\boldsymbol{s},\boldsymbol{\pi},\boldsymbol{W}} \mathcal{F} = \sum_{i=1}^{N}\sum_{v=1}^{V} \mathcal{L}\left(Z_i, \hat{Z}_i\right) + \frac{\lambda_S}{2}\|\boldsymbol{A}\|^2 \tag{10}$$

The derivative of Eq.10 with respect to $\boldsymbol{S}$ is defined as Eq.11, which accordingly contains two terms: the derivative of logistic loss $\mathcal{L}$ and the derivative of the similarity between shapelets $\boldsymbol{A}$ with respect to EEG channel shapelet point $\boldsymbol{S}$.

$$\frac{\partial \mathcal{F}_{i,v}}{\partial S_{k,c,l}} = \frac{\partial \mathcal{L}\left(Z_{i,v}, \hat{Z}_{i,v}\right)}{\partial \hat{Z}_{i,v}} \frac{\partial \hat{Z}_{i,v}}{\partial D_{i,k}} \frac{\partial D_{i,k}}{\partial Q_{i,k,j}} \frac{\partial Q_{i,k,j}}{\partial S_{k,c,l}} + \lambda_S A_k \frac{\partial A_k}{\partial S_{k,c,l}} \tag{11}$$

Subsequently, the derivative of the logistic loss $\mathcal{L}$ with respect to the estimated target $\hat{\boldsymbol{Z}}$ is defined as Eq.12, and the derivative of the estimated target $\hat{\boldsymbol{Z}}$ with respect to EEG-vs-shapelet distance $\boldsymbol{D}$ is also shown in Eq.13.

$$\frac{\partial \mathcal{L}\left(Z_{i,v}, \hat{Z}_{i,v}\right)}{\partial \hat{Z}_{i,v}} = \hat{Z}_{i,v} - Z_{i,v} \tag{12}$$

$$\frac{\partial \hat{Z}_{i,v}}{\partial D_{i,k}} = W_{k,v} \tag{13}$$

Furthermore, the derivative of $\boldsymbol{D}$ with respect to a EEG channel segment distance $\boldsymbol{Q}$ is defined in Eq.14, and subsequently, its derivative of the channel segment distance $\boldsymbol{Q}$ with respect to EEG channel shapelet point $\boldsymbol{S}$ is shown as Eq.15.

$$\frac{\partial D_{i,k}}{\partial Q_{i,k,j}} = \frac{1}{C}\sum_{c=1}^{C}\left(\frac{\pi_{k,c}}{E_1^2}\sum_{j=1}^{J=M-L+1}\left(e^{\alpha Q_{i,k,j}}\left(\left(1+\alpha Q_{i,k,j}\right)E_1 - \alpha E_2\right)\right)\right) \tag{14}$$

where $E_1 = \sum_{j=1}^{J=M-L+1} e^{\alpha Q_{i,k,j}}$ and $E_2 = \sum_{j=1}^{J=M-L+1} Q_{i,k,j} e^{\alpha Q_{i,k,j}}$.

$$\frac{\partial Q_{i,k,j}}{\partial S_{k,c,l}} = \frac{2}{L}\left(S_{k,c,l} - e_{i,c,j+l-1}\right) \tag{15}$$

As stated in Eq.11, the second term for the derivative of shapelet similarity $\boldsymbol{A}$ with respect to EEG channel shapelet point $\boldsymbol{S}$ is then defined in Eq.16.

$$\frac{\partial A_k}{\partial S_{k,c,l}} = \frac{\partial A_k}{\partial Q_{i,k,j}} \frac{\partial Q_{i,k,j}}{\partial S_{k,c,l}} = -\frac{2}{\sigma^2} Q_{i,k,j} e^{-\frac{Q_{i,k,j}^2}{\sigma^2}} \frac{\partial Q_{i,k,j}}{\partial S_{k,c,l}} \tag{16}$$

3.5.2 Channel Contribution Gradient. Since the EEG channel contribution $\boldsymbol{\pi}$ is embedded in logistic loss $\mathcal{L}$, it can be learnt by degenerating Eq.9 to Eq.17 while fixing $\boldsymbol{W}$ and $\boldsymbol{S}$.

$$\min_{\boldsymbol{s},\boldsymbol{\pi},\boldsymbol{W}} \mathcal{F} = \sum_{i=1}^{N}\sum_{v=1}^{V} \mathcal{L}\left(Z_i, \hat{Z}_i\right) \tag{17}$$

Again, with the gradient descent technique, the derivative of Eq.17 with respect to $\boldsymbol{\pi}$ is defined as

$$\frac{\partial \mathcal{F}_{i,v}}{\partial \pi_{k,c}} = \frac{\partial \mathcal{L}\left(Z_{i,v}, \hat{Z}_{i,v}\right)}{\partial \hat{Z}_{i,v}} \frac{\partial \hat{Z}_{i,v}}{\partial D_{i,k}} \frac{\partial D_{i,k}}{\partial \pi_{k,c}} \tag{18}$$

In addition, $\frac{\partial \mathcal{L}\left(Z_{i,v}, \hat{Z}_{i,v}\right)}{\partial \hat{Z}_{i,v}}$ and $\frac{\partial \hat{Z}_{i,v}}{\partial D_{i,k}}$ are introduced in Eq.12 and Eq.13, respectively. So here we only introduce the derivative of $D_{i,k}$ with respect to EEG channel contribution $\boldsymbol{\pi}$, which is defined in Eq.19.

$$\frac{\partial D_{i,k}}{\partial \pi_{k,c}} = \frac{1}{C}\sum_{c=1}^{C} \frac{\sum_{j=1}^{J} Q_{i,k,j} e^{\alpha Q_{i,k,j}}}{\sum_{j=1}^{J} e^{\alpha Q_{i,k,j}}} \tag{19}$$

To the end, the derivative of Eq.17 is achieved as

$$\frac{\partial \mathcal{F}_{i,v}}{\partial \pi_{k,c}} = \frac{\left(\hat{Z}_{i,v} - Z_{i,v}\right) W_{k,v}}{C} \sum_{c=1}^{C} \frac{\sum_{j=1}^{J} Q_{i,k,j} e^{\alpha Q_{i,k,j}}}{\sum_{j=1}^{J} e^{\alpha Q_{i,k,j}}} \tag{20}$$

where $Q_{i,k,j}$ is introduced in Eq.3.

3.5.3 Hyperplane Weight Gradient. For achieving the hyperplane weight $\boldsymbol{W}$ of classifier to minimize the objective function, we also use gradient descent to update it by fixing the EEG shapelets $\boldsymbol{S}$ and the EEG channel contributions $\boldsymbol{\pi}$. Then Eq.9 degenerates to Eq.21.

$$\min_{\boldsymbol{S},\boldsymbol{\pi},\boldsymbol{W}} \mathcal{F} = \sum_{i=1}^{N}\sum_{v=1}^{V} \mathcal{L}\left(Z_i, \hat{Z}_i\right) + \frac{\lambda_W}{2}\|\boldsymbol{W}\|^2 \tag{21}$$

Along with $\frac{\partial \mathcal{L}\left(Z_{i,v}, \hat{Z}_{i,v}\right)}{\partial \hat{Z}_{i,v}}$ is given in Eq.12, the derivative of Eq.21 with respect to $\boldsymbol{W}$ is then defined in Eq.22.

$$\frac{\partial \mathcal{F}_{i,v}}{\partial W_{k,v}} = \frac{\partial \mathcal{L}\left(Z_{i,v}, \hat{Z}_{i,v}\right)}{\partial \hat{Z}_{i,v}} \frac{\partial \hat{Z}_{i,v}}{\partial W_{k,v}} + \lambda_W W_{k,v} = \left(\hat{Z}_{i,v} - Z_{i,v}\right) D_{i,k} + \lambda_W W_{k,v} \tag{22}$$

And especially,

$$\frac{\partial \mathcal{F}_{i,v}}{\partial W_{0,v}} = \hat{Z}_{i,v} - Z_{i,v} \tag{23}$$

4 THE ALGORITHM

We firstly introduce the proposed algorithm StEEGCS for EEG channel selection in this section, followed by its convergence analysis, model initialization, and computational complexity analysis.

4.1 StEEGCS

The algorithm, we call StEEGCS, is a gradient descent algorithm that iteratively learns the distinct EEG shapelets $\boldsymbol{S}$, channel contributions $\boldsymbol{\pi}$, and linear hyperplane weights $\boldsymbol{W}$ based on the gradient descent technique. Finally, according to channel contributions $\boldsymbol{\pi}$, $\boldsymbol{c}_{sel} \in \mathbb{R}^{sel}$ with top-sel $\boldsymbol{\pi}_c$ contribution channels for K shapelets can be selected out. Algorithm 1 shows the details, in which $\boldsymbol{S}$, $\boldsymbol{\pi}$, and $\boldsymbol{W}$ are updated iteratively under the learning rate η. Finally, sel channels $\boldsymbol{c}_{sel}$ are selected according to the top-sel channel contributions $\boldsymbol{\pi}$.

Algorithm 1: StEEGCS for EEG channel selection

Input: EEG data $\boldsymbol{E} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n\}_{N \times C \times M}$; Labels of EEG $\boldsymbol{Y} \in \mathbb{R}^{N \times V}$; Number of shapelets K; Length of shapelet $l_{min} < l < L$; Weight parameters λ_W and λ_S; Learning rate η; Precision control parameter α; Kernel parameter σ and maximum iteration I_{iter}; Number of selected channel sel.

Output: Best shapelets $\boldsymbol{S} \in \mathbb{R}^{K \times V \times L}$; Channel contributions $\boldsymbol{\pi} \in \mathbb{R}^{K \times C}$; Hyperplane weights $\boldsymbol{W} \in \mathbb{R}^{K \times V}$; bias $\boldsymbol{W}_0 \in \mathbb{R}^{V}$; Selected EEG channel $\boldsymbol{c}_{sel} \in \mathbb{R}^{K \times sel}$.

Initialize $\boldsymbol{S}_0$, $\boldsymbol{\pi}_0$, $\boldsymbol{W}_0$;
for $t = 1$ *to* I_{iter} **do**
  for $i = 1$ *to* N **do**
    for $k = 1$ *to* K **do**
      Calculate $D_{i,k}$ with Eq.2 and Eq.3;
    end
    for $v = 1$ *to* V **do**
      Calculate $Y_{i,v}$ and $\hat{Z}_{i,v}$ with Eq.5 and Eq.6 respectively;
      for $k = 1$ *to* K **do**
        for $c = 1$ *to* C **do**
          for $l = 1$ *to* L **do**
            $S_{k,c,l} \leftarrow S_{k,c,l} - \eta \frac{\partial \mathcal{F}_{i,v}}{\partial S_{k,c,l}}$ with Eqs.11–16;
            $\pi_{k,c} \leftarrow \pi_{k,c} - \eta \frac{\partial \mathcal{F}_{i,v}}{\partial \pi_{k,c}}$ with Eq.20;
          end
        end
        $W_{k,v} \leftarrow W_{k,v} - \eta \frac{\partial \mathcal{F}_{i,v}}{\partial W_{k,v}}$ with Eq.22;
      end
      $W_{0,v} \leftarrow W_{0,v} - \eta \frac{\partial \mathcal{F}_{i,v}}{\partial W_{0,v}}$ with Eq.23;
      $\boldsymbol{c}_{sel} \leftarrow$ channel indexes of top-sel $\boldsymbol{\pi}$ for K shapelets;
    end
  end
end
return $\boldsymbol{S}$, $\boldsymbol{\pi}$, $\boldsymbol{W}$, $\boldsymbol{W}_0$, $\boldsymbol{c}_{sel}$;

4.2 Convergence Analysis

Algorithm 1 selects sel EEG channels with top-sel contributions for K shapelets by simultaneously learning shapelets $\boldsymbol{S}$, channel contributions $\boldsymbol{\pi}$, and linear hyperplane $\boldsymbol{W}$ based on the gradient descent strategy. As the objective function is non-convex, it, in the gradient descent

strategy, just can converge into a local optima under two parameters that interferes with each other, such as the learning rate η and the maximum iteration I_{iter}. A proper setting of η and I_{iter} can obtain a good convergence in a relatively short time. In particular, a larger η can help algorithm operate less iterations to minimize the objective function (see Eq.9), but it likely deteriorates the convergence of the algorithm. On the contrary, if the algorithm aims to converge with a smaller learning rate η, it needs more iterations. Consequently, it is also a trade-off between learning rate and maximum iteration to set for the algorithm. But considering to the acceptable time cost to converge, we recommend to set $\eta = 0.01$ and $I_{iter} \leq 100$ in the article.

4.3 Model Initialization

The objective function (*i.e.,* Eq.9) is non-convex, so the gradient descent-based algorithm in which the variables of shapelets $\boldsymbol{S}$, channel contributions $\boldsymbol{\pi}$, and linear hyperplane $\boldsymbol{W}$ are required to learn simultaneously just converge to the local optima. The gradient descent strategy to solely learn each of them cannot guarantee the global optima of objective function, but it's also widely applied to solve non-convex problems as a trade-off technique. In practice, the performance of gradient descent-based algorithms are significantly influenced by initializations of its parameters that we are addressing in the section. As the patterns of every EEG class can be represented by their centroid, for $\boldsymbol{S}_0$, it's initialized by the k-means centroid that contains same length of segments from EEG data. Then, according to the initialization of shapelets $\boldsymbol{S}$, the original EEG data can be represented by the shapelet-transformed matrix $\boldsymbol{D}$. To initialize channel contribution $\boldsymbol{\pi}_0$, we transformed the average distances between shapelets and EEG channel segment to its initialization contributions with sigmoid function, *i.e.,* $\boldsymbol{\pi}_0(i) = \text{sigmoid}\left(\frac{1}{K}\sum_{k=1}^{K} D_{i,k}\right)$ (where i denotes the i^{th} EEG channel, and $D_{i,k}$ refers to Eq.2). And $\boldsymbol{W}_0$ is simply randomly initialized close to 0 based on a normal distribution.

4.4 Computational Complexity

As shown in Algorithm 1, StEEGCS solves the problem for n EEG trials in maximum iterations I_{iter}. In each iteration, it mainly takes $O(nckl^2)$ for calculating $\boldsymbol{D}$; $O(nvk)$ for $\boldsymbol{Y}$, $\hat{\boldsymbol{Z}}$, and $\boldsymbol{W}$, respectively; $O(nvkc^2l^3+nvck^3l^3+nvckl^2)$ for $\boldsymbol{S}$; $O(nvkc^2l^2)$ for $\boldsymbol{\pi}$, and $O(nv)$ for $\boldsymbol{W}_0$, respectively, where k denotes the number of shapelets that needs to learn; v denotes the number of EEG classes; c denotes the number of EEG channels; l is the maximum length of shapelets. In sum, the total time complexity of StEEGCS is $O(I_{iter}(\max\{nckl^2, nvk, nvkc^2l^3+nvck^3l^3+nvkcl^2, nvkc^2l^2, nv\}))$. Since commonly $k, c \ll l$, the computational complexity of StEEGCS is $O(I_{iter}(nvkc^2l^3+nvck^3l^3))$.

5 EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we firstly introduce the details of EEG datasets, evaluation methodology, and baseline methods. Then, we carry out a detailed experimentation to compare the proposed StEEGCS with state-of-the-art EEG channel selection approaches on several real-world EEG datasets.

5.1 EEG Datasets

10 EEG datasets are used to evaluate the efficacy of the proposed method, which includes the slow cortical potentials (SCPs), motor imagery EEG, and wrist movement EEG data. All the EEG datasets and their detailed descriptions are publicly available as online archives at

http://www.bbci.de/competition/ii/ (Dataset:II) and http://www.bbci.de/competition/iv/ (Ddataset:IV), respectively. To evaluate EEG channel selection methods with respect to classification accuracy, we randomly divide the original EEG dataset into two parts: training dataset and testing one. All the selection methods are operated on the training data and evaluated with the testing one. The detailed descriptions of each EEG dataset are shown in Table 1.

Table 1. EEG Datasets

Dataset	Description	Number×Channel×Sample (training data:testing data)	Classes
Ia	SCPs from one healthy subject	268×6×896 (200:68)	2
Ib	SCPs from one ALS patient	200×7×1152 (120:80)	2
IV_1_calib_1a IV_1_calib_1b IV_1_calib_1f	Motor imagery of 2-class of left hand, right hand, or foot from 3 healthy subjects	200×59×800 (120:80)	2
IV_2a_s1 IV_2a_s2 IV_2a_s3	Motor imagery of left hand right hand, both feet, and tongue from 3 healthy subjects	288×22×313 (200:88)	4
IV_3_s1 IV_3_s2	Wrist movement to left, right, forward, backward from 2 healthy subjects	160×10×400 (120:40)	4

5.2 Baselines

To further establish the superiority of StEEGCS for EEG channel selection, we compare it to several classic and state-of-the-art approaches, such as CSP [51], RCSP [12], SCSP [4], IMOCS [19], and CCSE [53].

CSP: Common spatial pattern operates on a covariance matrix between EEG channels. In detail, it is effective in discriminating two classes of EEG data by maximizing the variance of one class while minimizing the variance of the other class. CSP-based channel selection method selects EEG channels based on CSP coefficients, *i.e.,* channels corresponding to maximal CSP vector coefficients are selected as the optimal EEG channels.

RCSP: Regularized common spatial pattern-based algorithm selects EEG channels by inducing the sparsity in the spatial filters, which actually uses 1-norm regularization. The solution of RCSP is sparser than conventional CSP.

SCSP: Sparse common spatial pattern-based algorithm selects EEG channels by scattering the common spatial filters within a constraint of classification accuracy. In detail, SCSP scatters the CSP spatial filters to emphasize on a limited number of EEG channels with high variances between classes, and to discard the rest of the channels with low variances.

IMOCS: Iterative multiobjective optimization for channel selection selects EEG channels by firstly initializing a reference candidate solution and subsequently finding a set of most relevant channels in an iterative manner.

CCSE: This method selects EEG channels via using correlation coefficient of spectral entropy (CCSE). EEG channels are selected based on the ranking of correlation coefficient, in a way that the spectral entropy of each channel across all frequencies is considered by taking sum of the squared correlation coefficient.

Honestly, there are many classifiers for EEG time series classification, including Deep Neural Networks (DNNs)-based classifier [13], COTE [5], HIVE-COTE [34], St-TSC [33], RPCD [46], and SAX-SEQL [39], but we just apply SVM classifier in the article, since (1) SVM is the most widely used and promising classifier for EEG time series classification; (2) as introduced in [9, 10] that SVM also performs as well as such classifiers as COTE, St-TSC, RPCD, and SAX-SEQL on EEG classification; (3) besides, SVM also achieves the highest accuracy for EEG classification compared with Fisher linear discriminant analysis (FLDA), Generalized Andersons Task linear classifier (GAT), Linear Discriminant Analysis (LDA) [44]. Therefore, SVM classifier with LIBSVM toolbox [8] was used in EEG classification step to assist to analyze the efficacy of channel selection approaches. Additionally, SVM is operated 10 times on testing dataset and their average is reported as the final classification accuracy.

For baseline methods, the parameters are tuned according to their original articles. For StEEGCS learning shapelets, we set a minimum length of shapelets $l_{min} = 10$ to learn, and other shapelets are expanded to different lengths by a scaler $r \in \mathbb{Z}^+$, *i.e.,* $\{l_{min}, \cdots, rl_{min}\}$. Besides, the weight parameters λ_W and λ_S are searched in $\{10^{-4}, 10^{-2}, 10^0, 10^2, 10^4\}$, and the learning rate η is set as $\eta = 0.01$.

5.3 Sensitivity Analysis

The StEEGCS is a shapelet-transformed EEG channel selection and the selected channel seems to be influenced by channel contribution (*i.e.,* channel weights), shapelet length, and shapelet number. Hence, we respectively discuss their impacts in this section.

5.3.1 Impact of Shapelets on Channel Weights. We analyze the impact of shapelet length and shapelet number on StEEGCS. We firstly show shapelets learned from each class of EEG datasets, see Fig.2. Actually, we just show one EEG shapelet of each class in Fig.2, and it seems to indicate that shapelet length of 30 contains better distinguishing patterns compared to other shapelet lengths such as 10, 40, 60, and so on. In detail, a shorter shapelet (*e.g.,* 10 or 20) cannot clearly distinguish the patterns of different EEG channels, so it probably does not select the most representative channels for EEG classification. Meanwhile, a longer shapelet probably contains patterns that can be presented by a relatively shorter shapelet, such as 120 to 60, 60 to 30, so it likely contains redundant patterns of EEG data. Besides, as introduced above, we only analyze shapelet lengths of 10 at least while 50 or 120 at most for corresponding EEG datasets in the article.

Moreover, shapelet numbers are affected by shapelet length as well as the sample length of each EEG channel. For example, learning too many shapelets (more than 5) with length of 120, it may require more than 600 samples for each EEG channel, otherwise the learned shapelets are not distinct with each other, since they probably overlap with each other or contain too many redundant/common shapelet segments; If learning too many shapelets with length of 10, it can be transformed to learn a smaller amount of shapelets with longer shapelets such as 20, 30, or 60, *etc*, since a longer shapelet probably contains several shorter shapelets. In other words, learning too many shapelets no matter longer or shorter may be not good for learning EEG channel weights and maybe finally degrade EEG classification. A larger number of shapelets learned to select EEG channels seem to provide more redundant EEG patterns for classifier, and it also costs more time. Consequently, we, according to the length of EEG shapelets and the corresponding channel sample, we just discuss the impact of shapelet numbers from 1 to 5 on EEG channel selection and classification.

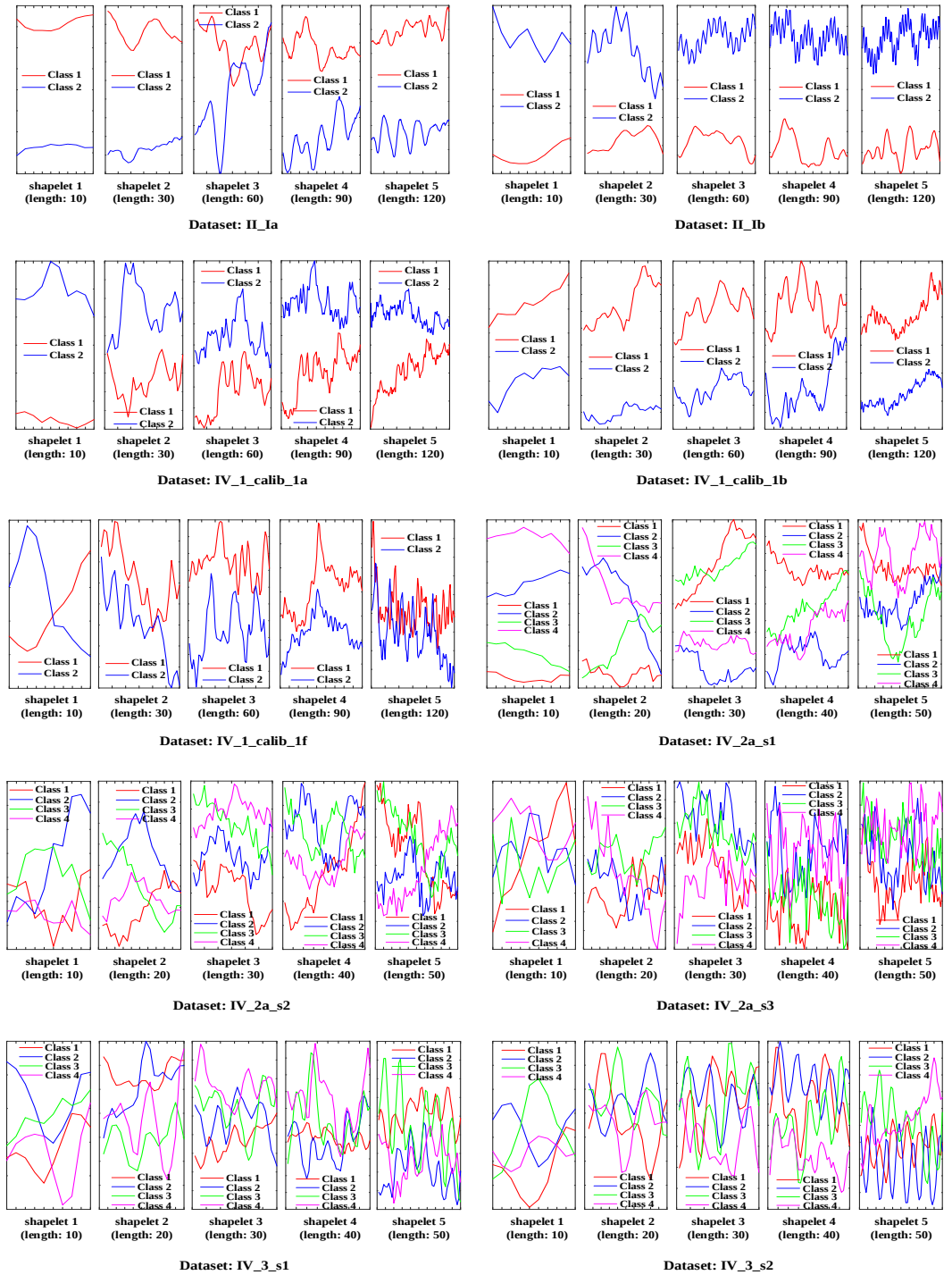


Fig. 2. Shapelets with different lengths learned from every EEG class

StEEGCS selects EEG channel based on channel contributions (*i.e.,* channel weights). In other words, EEG channel weights are actually determined by EEG shapelets. Therefore, we analyze the impact of shapelets on EEG channel weights, including shapelet length and shapelet number. The results on 10 EEG datasets are shown in Figs.3–12, respectively. Generally, using 3 shapelets, the discriminations among EEG channel weights are relatively more obvious than other numbers of shapelets. Correspondingly, Fig.13 briefly illustrates the relationship between EEG channel weights and shapelet lengths when fixing the number of learned shapelets as 3, and it shows that shapelet length of 30 reflects relatively more distinguishing EEG channel weights for StEEGCS. As we stated before, a large number of shapelets seem to contain redundant or less discriminative patterns in EEG channel data while small numbers of shapelets may contain incomplete patterns of EEG channel signals. Hence, both of the two situations result in relatively lower discrimination among EEG channel weights, which probably influences the performance of channel selection and classification. Similarly, short or long shapelets probably result in relatively lower discrimination of EEG channel weights as well, since short shapelets (*e.g.,* 10) may contain incomplete patterns of EEG channel data while long shapelets (*e.g.,* 90, 120), on the contrary, may contain redundant or less discriminative EEG patterns. Consequently, according to the results in Figs.3–13, setting EEG shapelet number as 3 and shapelet length as 30 for StEEGCS can get relatively the highest discrimination EEG channel weights, which is beneficial to achieve best EEG classification.

5.3.2 Impact of Shapelets on Classification Performance. We discuss the impact of shapelet number and shapelet length on EEG classification performance, along with the number of selected EEG channels. The impact of shapelet number (fixing shapelet length as 30

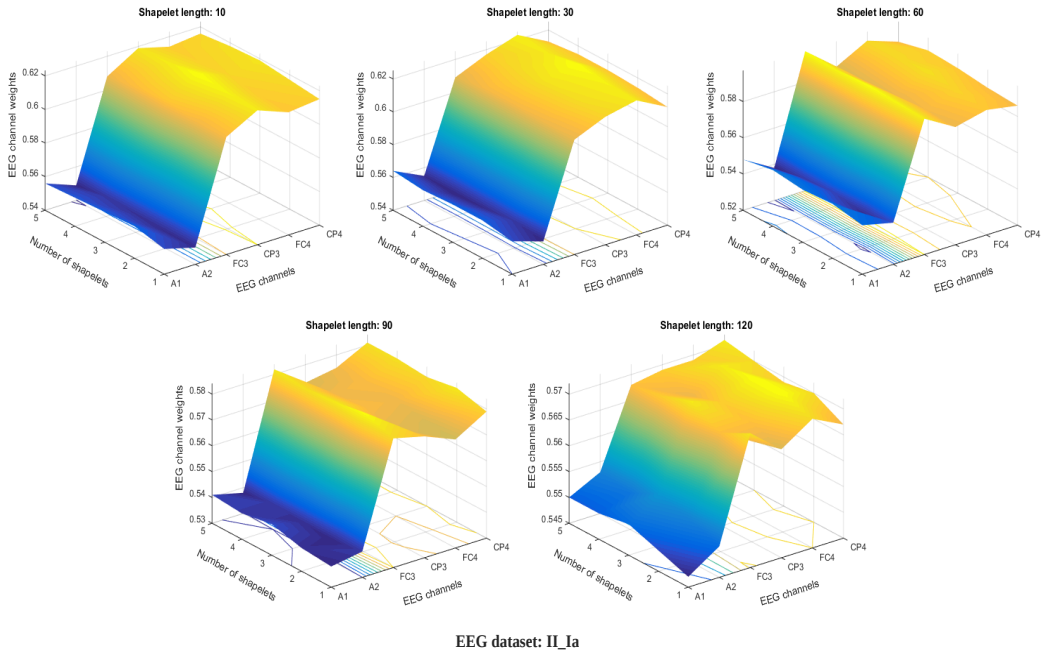
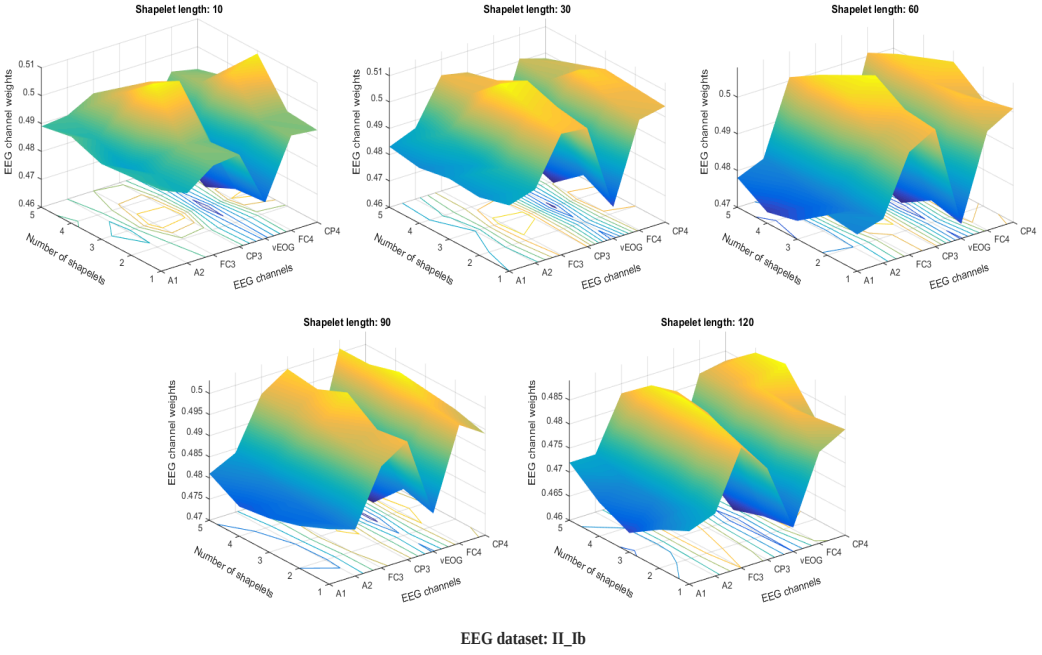
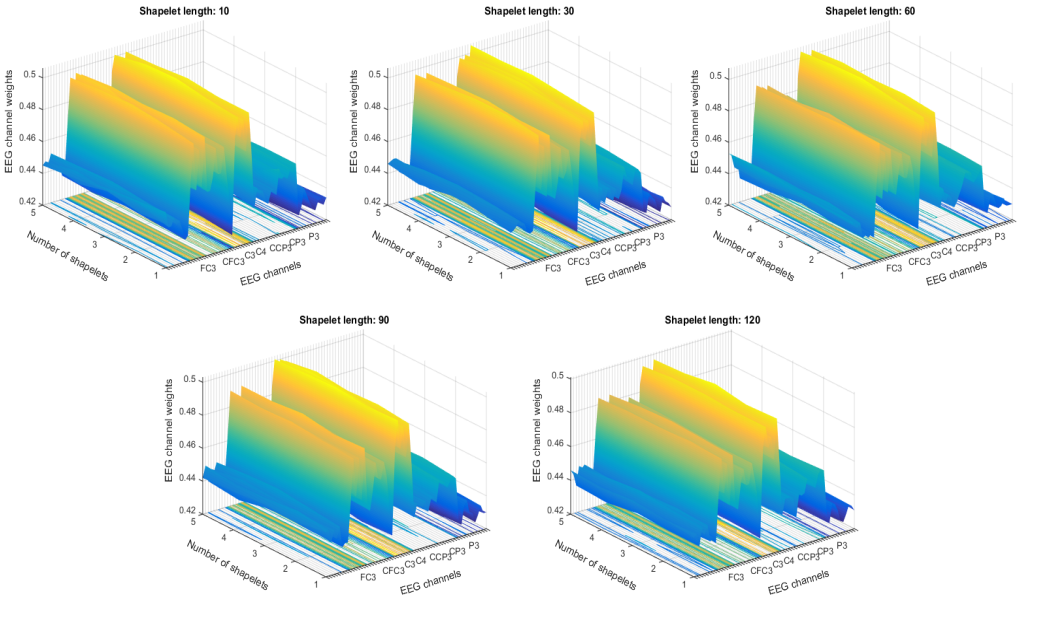


Fig. 3. Channel weights with shapelet lengths and numbers on II_Ia



EEG dataset: II_Ib

Fig. 4. Channel weights with shapelet lengths and numbers on II_Ib



EEG dataset: IV_1_calib_1a

Fig. 5. Channel weights with shapelet lengths and numbers on IV_1_calib_1a

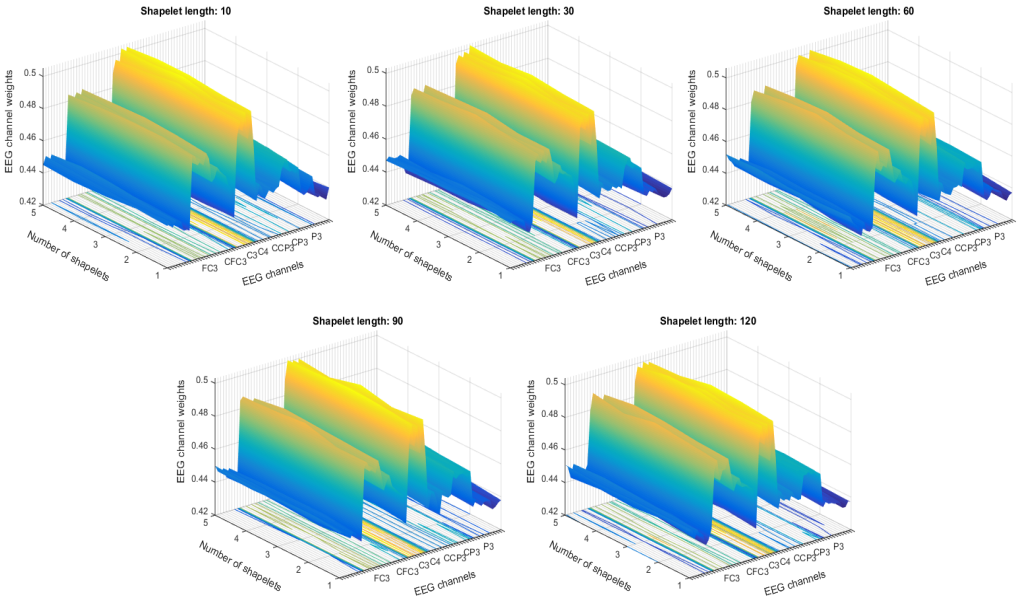


Fig. 6. Channel weights with shapelet lengths and numbers on IV_1_calib_1b

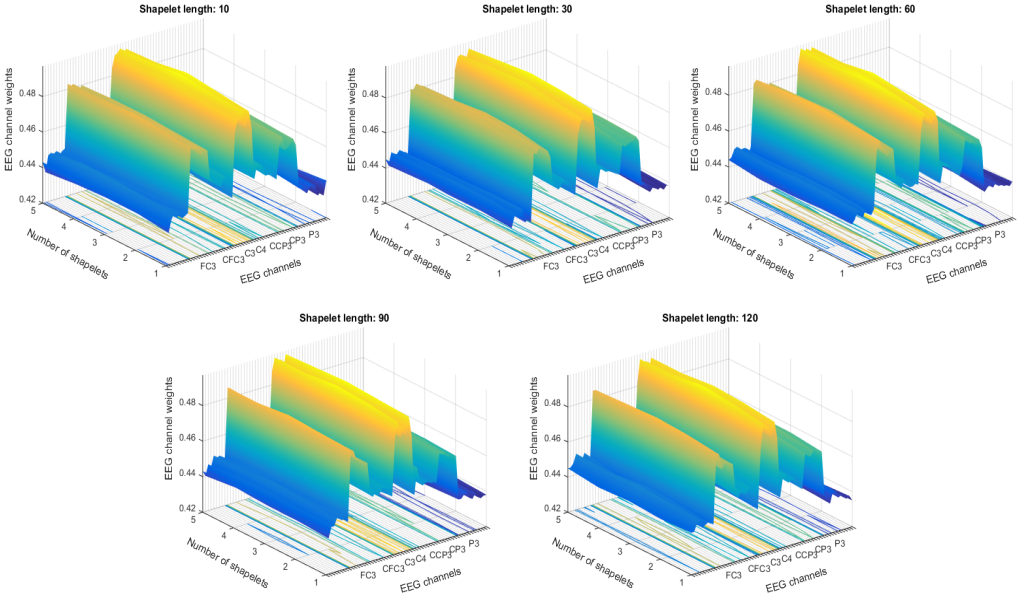


Fig. 7. Channel weights with shapelet lengths and numbers on IV_1_calib_1f

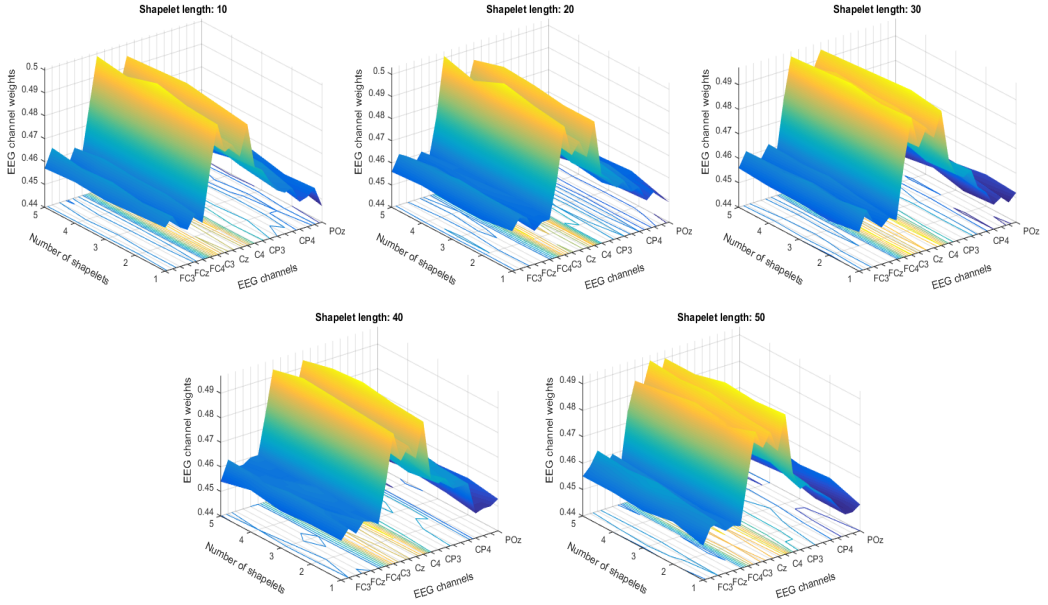


Fig. 8. Channel weights with shapelet lengths and numbers on IV_2a_s1

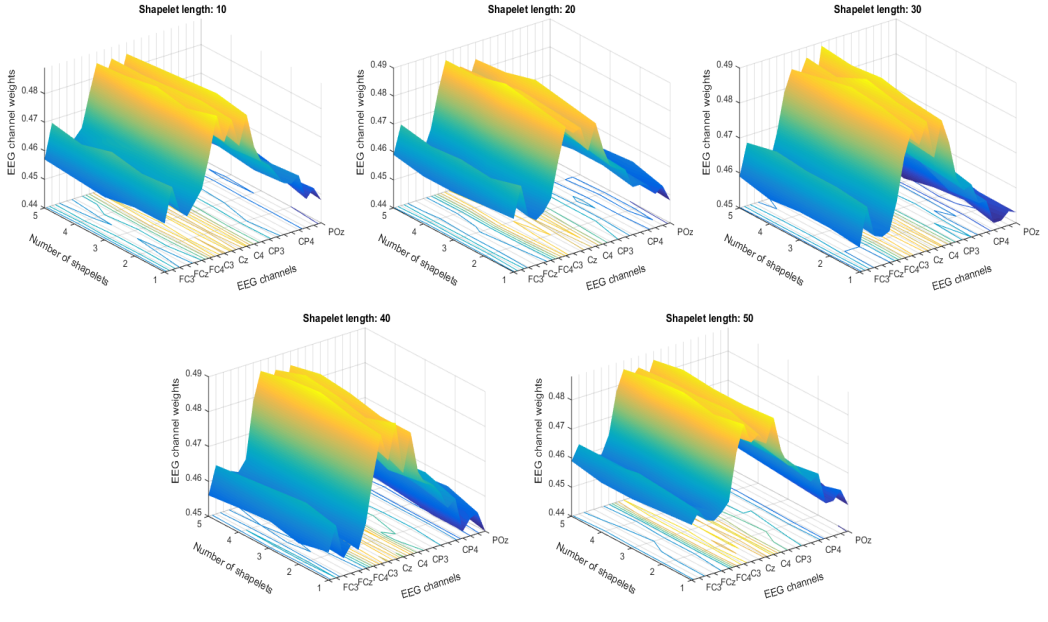


Fig. 9. Channel weights with shapelet lengths and numbers on IV_2a_s2

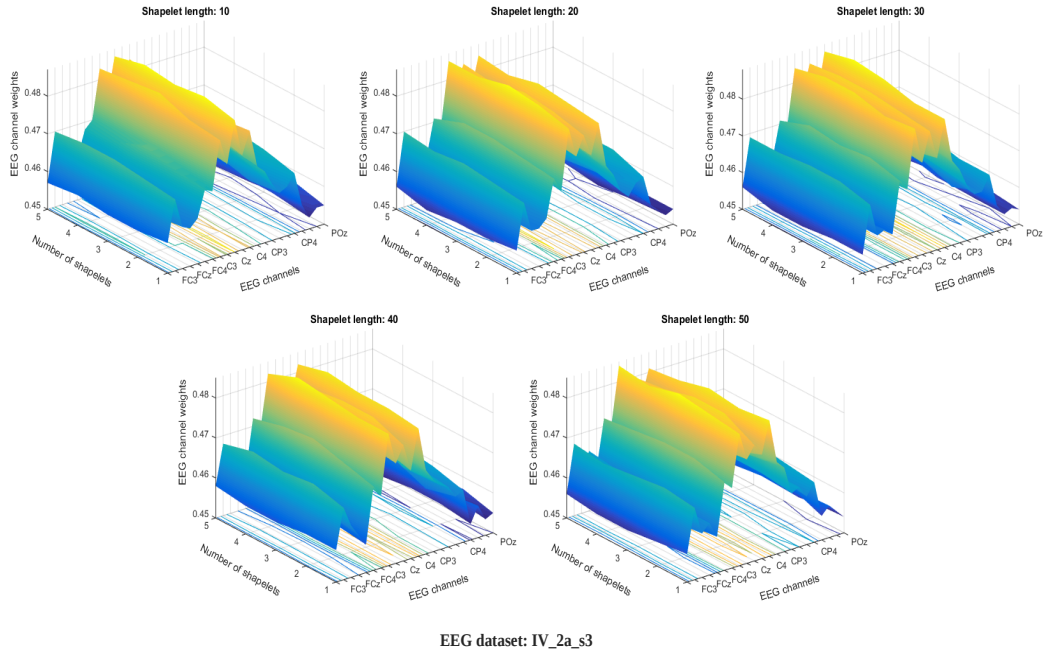


Fig. 10. Channel weights with shapelet lengths and numbers on IV_2a_s3

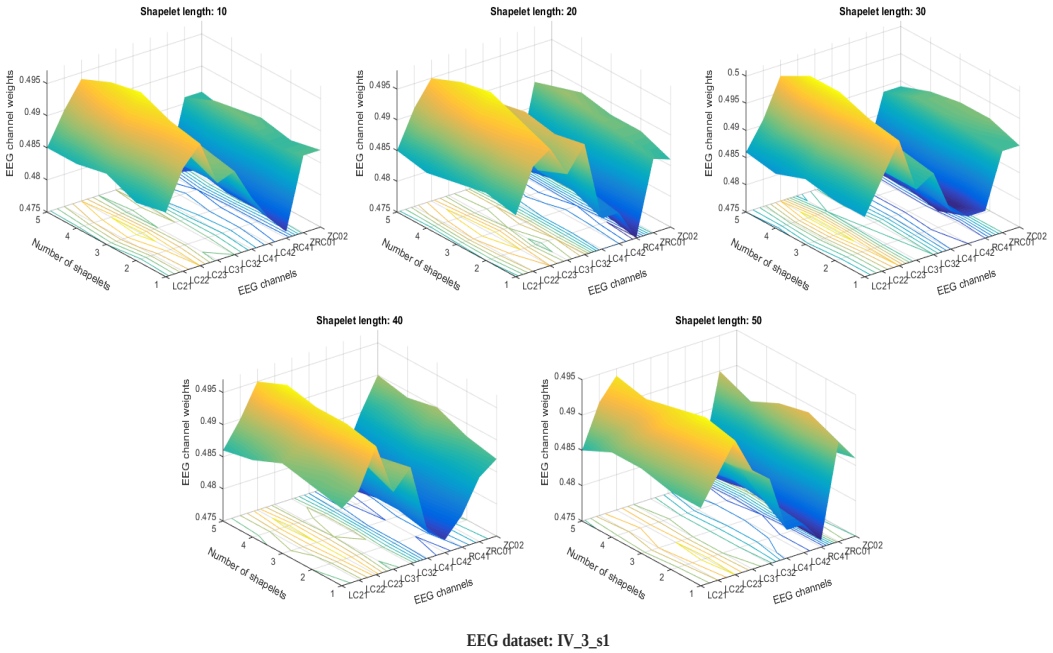


Fig. 11. Channel weights with shapelet lengths and numbers on IV_3_s1

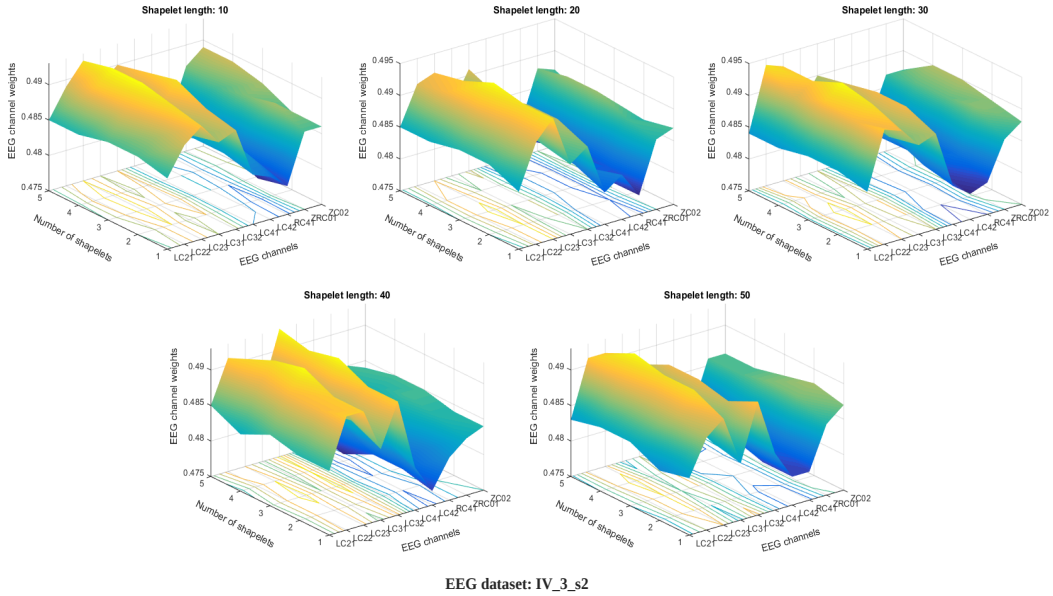


Fig. 12. Channel weights with shapelet lengths and numbers on IV_3_s2

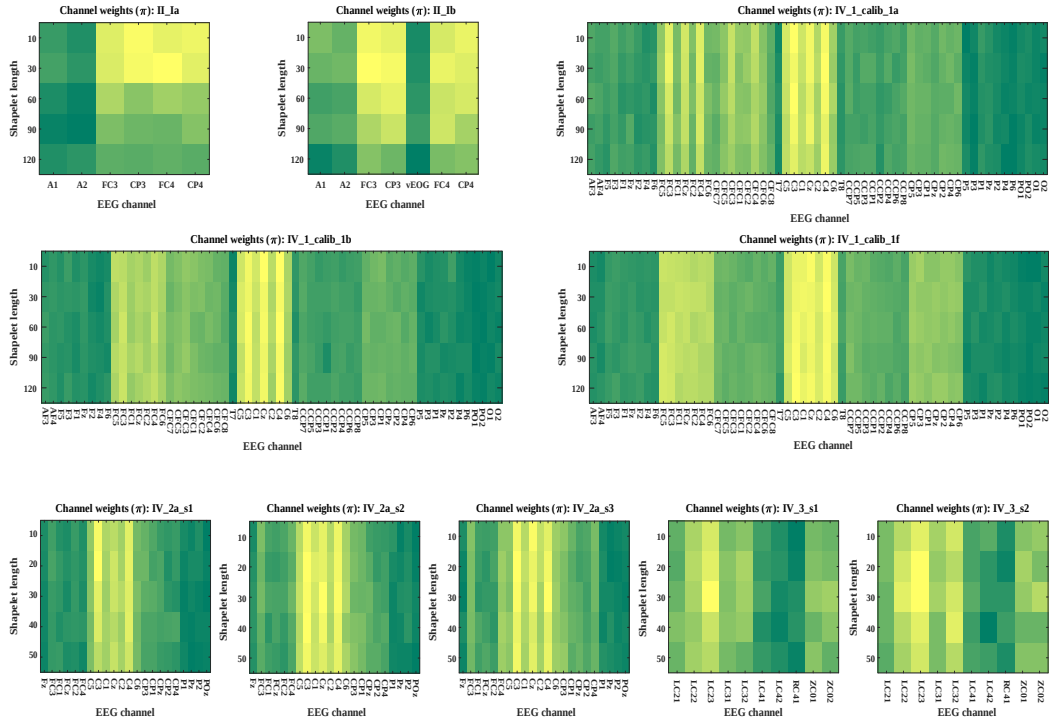


Fig. 13. EEG channel weights with respect to different shapelet lengths (shapelet number: 3)

according the discussion in subsection 5.3.1) and shapelet length (fixing shapelet number as 3 according the discussion in subsection 5.3.1) is displayed in Fig.14 and Fig.15, respectively, both of which indicate that classification performance is improved with a fewer EEG channels selected by StEEGCS, especially when the number of selected EEG channels is 2, 3 or 4.

As illustrated in Figs.14–15, classification accuracies achieved with 3 shapelets of 30 are relatively the highest in all conditions, in accord with the discussion in subsection 5.3.1. In other words, StEEGCS with a relatively small number of shapelets (*i.e.,* 3) in a moderate length (*i.e.,* 30) can select the most relevant EEG channels for SVM classifier to yield the highest classification accuracy.

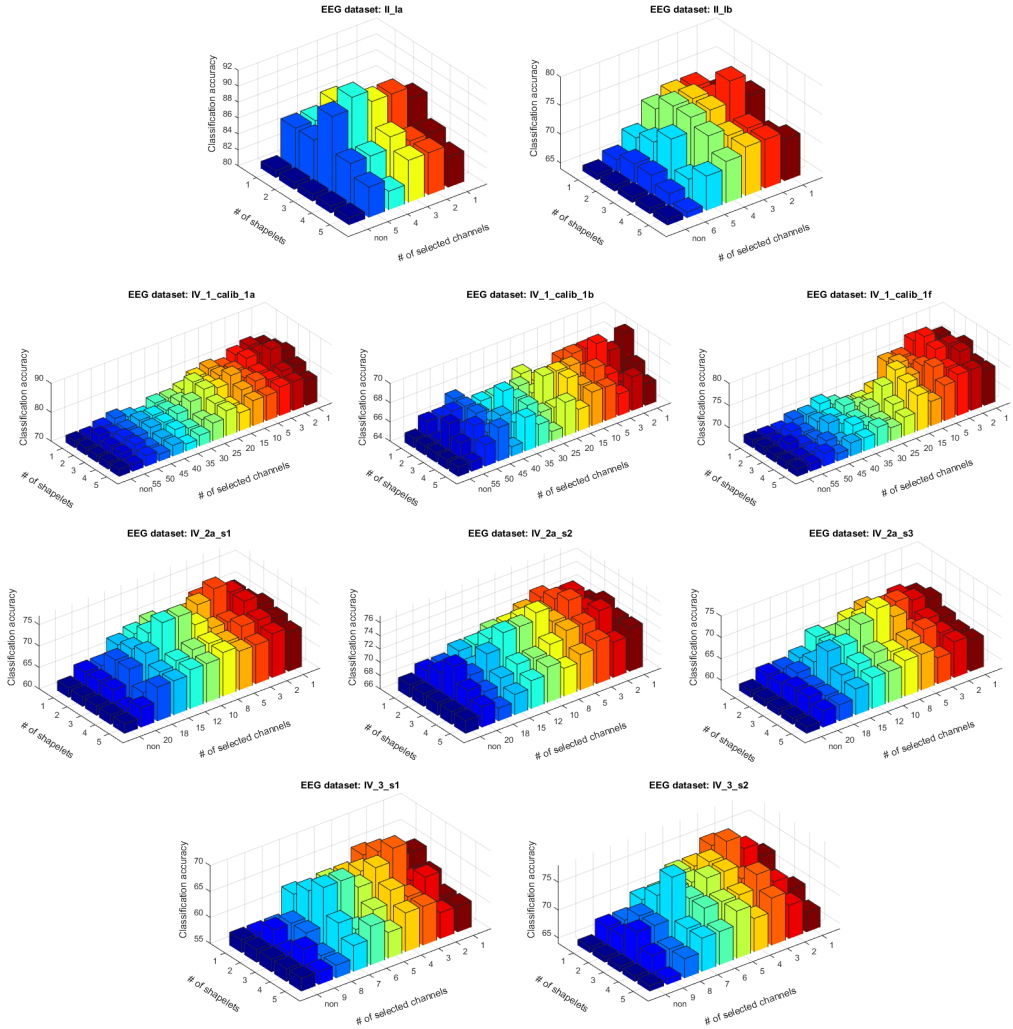


Fig. 14. Classification with respect to selected EEG channels and shapelet numbers (shapelet length: 30)

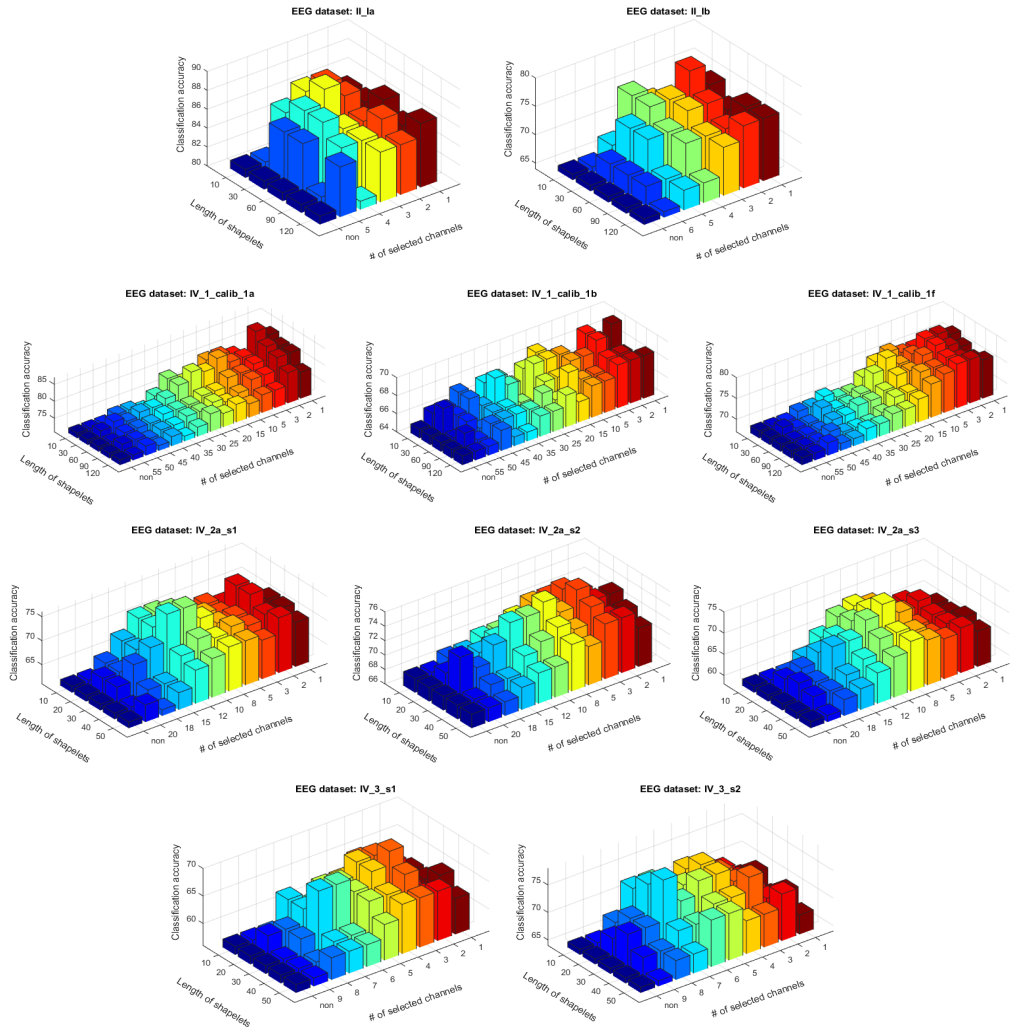


Fig. 15. Classification with respect to selected EEG channels and shapelet lengths (shapelet number: 3)

5.4 Performance Comparison with Baselines

We, in this section, analyze the efficacy of our method StEEGCS with respect to classification accuracy by comparing it to baselines on 10 EEG datasets. As we concluded in Section 5.3 that 3 shapelets with length of 30 can select the most relevant EEG channels and yield classification performance with the selected EEG channels, we set shapelet length: 30 and shapelet number: 3 for StEEGCS.

5.4.1 Comparison with Non-selected EEG Channels. With SVM, we firstly analyze the impact of StEEGCS selecting EEG channels on EEG classification accuracy. The results, with 3 shapelets of length 30, are shown in Fig.16, which indicates that the SVM-based classification accuracies on all EEG datasets increase generally along with the number of selected EEG channels decreases. Besides, Table 2 also shows that EEG classification accuracy with

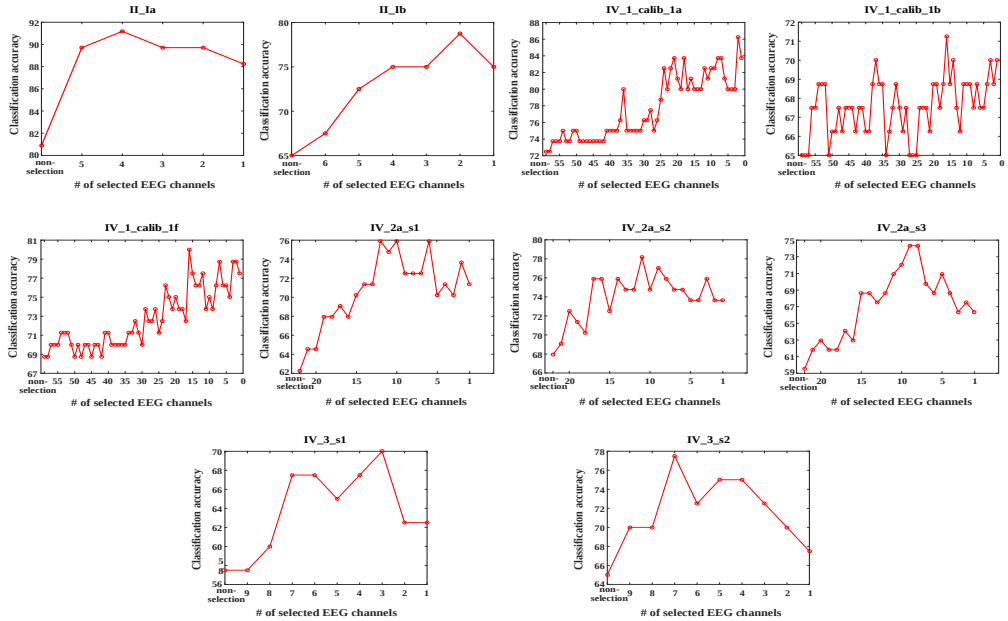


Fig. 16. Classification accuracy with StEEGCS selected EEG channels (shapelet number: 3; shapelet length: 30)

Table 2. Classification accuracy improvement (Best classification with selected EEG channels vs non-selection)

EEG dataset	Non-selection (# of all channels)	Selected (# of selected channels)	Improvement (%)
II_Ia	80.88 (6)	91.18 (4)	**12.73**
II_Ib	65 (7)	78.75 (2)	**21.15**
IV_1_calib_1a	72.5 (59)	86.25 (2)	**18.97**
IV_1_calib_1b	65 (59)	71.25 (16)	**9.62**
IV_1_calib_1f	68.75 (59)	80 (16)	**16.36**
IV_2a_s1	62.27 (22)	75.91 (6, 10, 12)	**21.90**
IV_2a_s2	67.95 (22)	78.18 (11)	**15.06**
IV_2a_s3	59.55 (22)	74.32 (8, 9)	**24.80**
IV_3_s1	57.5 (10)	70 (3)	**21.74**
IV_3_s2	65 (10)	77.5 (7)	**19.23**

StEEGCS selected channels is greatly improved by 9.5% at least for all EEG datasets, compared to those with non-selection channels (*i.e.,* all channels). StEEGCS aims to search distinct EEG shapelets that represent the original EEG data, to provide more informative EEG patterns for SVM classifier. In other words, as embedded with logistic loss, shapelet-transformed StEEGCS not only reduces redundancy of EEG data, but also strengthens important patterns for classifier modeling. Meanwhile, as the number of selected EEG channels decreases, the efficiency of SVM classifier is correspondingly significantly improved, compared to non-selection EEG channels, see Fig.17.

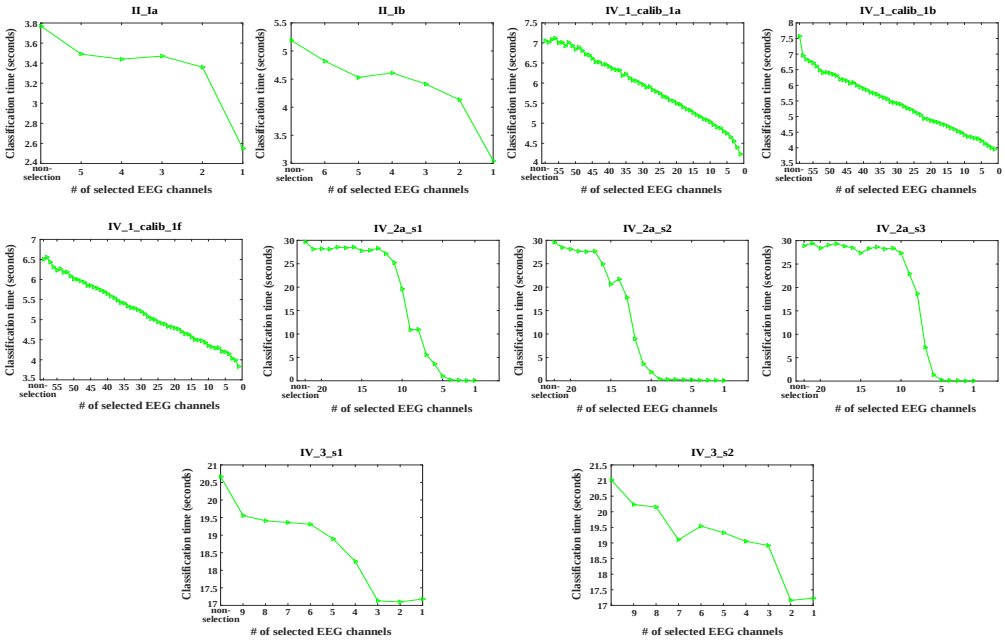


Fig. 17. Classification efficiency with EEG channels selected by StEEGCS (shapelet number: 3; shapelet length: 30)

5.4.2 Comparison with EEG Channel Selection Baselines. To further establish the efficacy of StEEGCS, we, with SVM classifier, compare it to other EEG channel selection methods such as CSP, RCSP, SCSP, IMOCS, and CCSE. As applied to 10 EEG datasets, the classification results are shown in Fig.18. The results of StEEGCS are achieved with shapelet length of 30 and shapelet number of 3 (the explanation is in detail introduced in Section 5.3). The results clearly demonstrate the superiority of StEEGCS for EEG channel selection, since the classification accuracy with StEEGCS selected channels is generally higher than those of other channel selection methods. Additionally, we also compare their averages and standard deviations of all classification accuracies achieved with different selected EEG channels on each EEG dataset, *i.e.,* the averages and standard deviations of all classification for each channel selection method, see Fig.18. The results are shown in Table 3, which demonstrates that on all EEG datasets, StEEGCS assists SVM classifier to achieve the best averaged classification accuracy. Moreover, the best and worst classification accuracies with correspondingly selected EEG channels are also displayed in Table 4, which also indicates that StEEGCS is the best method for EEG channel selection, compared to different baselines, since it yields the highest classification accuracy on all EEG datasets no matter in best or worst situation. Besides, we also analyze the significance of StEEGCS over baselines for EEG channel selection by using one-tailed t-test ($\alpha = 0.05$) on their classification accuracy, see Table 5. The p-values in Table 5 demonstrate the classification performance achieved by StEEGCS is significantly different to those baseline methods on most of EEG datasets, especially some of which are extremely significantly different. In other words, Table 5 indicates that StEEGCS outperforms baselines for EEG channel selection with respect to classification accuracy.

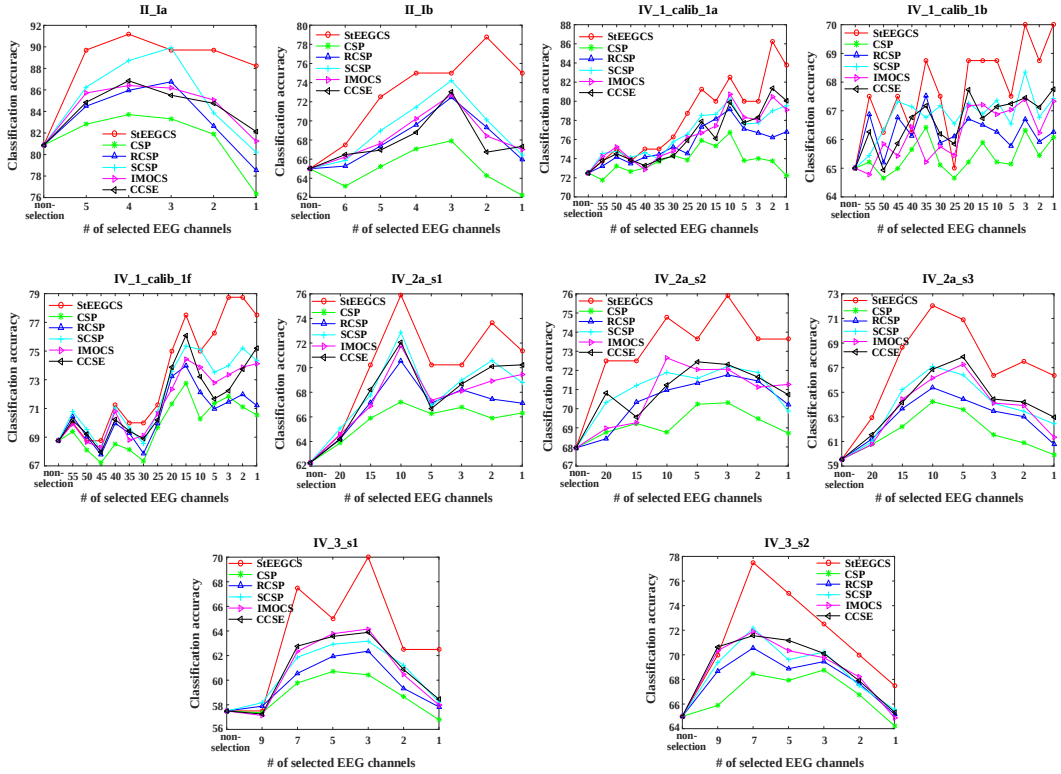


Fig. 18. Classification results of different EEG channel selection methods (shapelet number: 3; shapelet length: 30)

Table 3. Classification accuracy in all cases with different selected numbers of channels (Average ± standard deviation) (The best results are highlighted in boldface.)

EEG dataset	CSP	RCSP	SCSP	IMOCS	CCSE	StEEGCS
II_Ia	81.62 ± 3.03	83.68 ± 3.27	85.80 ± 3.89	84.92 ± 2.12	84.82 ± 1.72	**89.71 ± 1.04**
II_Ib	64.98 ± 2.22	68.33 ± 2.66	69.51 ± 3.14	68.71 ± 2.39	68.22 ± 2.50	**73.96 ± 3.74**
IV_1_calib_1a	73.88 ± 1.37	75.77 ± 1.84	76.83 ± 2.19	76.59 ± 2.51	76.49 ± 2.71	**78.66 ± 3.97**
IV_1_calib_1b	65.43 ± 0.57	66.33 ± 0.58	66.95 ± 0.66	66.30 ± 0.89	66.73 ± 0.81	**67.95 ± 1.44**
IV_1_calib_1f	69.83 ± 1.76	70.67 ± 1.84	72.11 ± 2.60	71.50 ± 2.29	71.58 ± 2.53	**73.48 ± 3.83**
IV_2a_s1	66.03 ± 1.06	67.40 ± 1.86	68.71 ± 2.53	68.15 ± 2.25	68.59 ± 2.59	**70.88 ± 3.52**
IV_2a_s2	69.36 ± 0.68	70.62 ± 1.11	71.28 ± 0.87	71.11 ± 1.43	71.25 ± 1.00	**73.80 ± 1.21**
IV_2a_s3	61.88 ± 1.57	63.14 ± 1.65	64.24 ± 2.18	64.00 ± 2.35	64.58 ± 2.17	**67.82 ± 3.06**
IV_3_s1	58.95 ± 1.63	59.99 ± 1.96	60.93 ± 2.24	60.97 ± 2.96	61.13 ± 2.78	**64.17 ± 4.38**
IV_3_s2	67.00 ± 1.73	68.40 ± 1.84	69.07 ± 2.29	69.24 ± 2.44	69.46 ± 2.39	**72.08 ± 3.68**

In addition, we also accordingly compare the execution time of EEG channel selection approaches, and the result is illustrated in Fig.19. As Fig.19 indicates, StEEGCS costs the most for EEG channel selection on most EEG datasets. As introduced in Section 4.4, the computational complexity of StEEGCS is $O(I_{iter}(nvkc^2l^3 + nvck^3l^3))$, and it shows that the time consumption of StEEGCS is mainly determined by optimal shapelet learning, which

Table 4. The best and worst classification accuracy with corresponding selected number of EEG channels (The corresponding best results are highlighted in boldface.)

EEG dataset	Classification (# of selected channels)	CSP	RCSP	SCSP	IMOCS	CCSE	StEEGCS
II_Ia	best	83.71 (4)	86.76 (3)	89.92 (3)	86.41 (4)	86.85 (4)	**91.18** (4)
	worst	76.34 (1)	78.54 (1)	80.22 (1)	81.23 (1)	82.14 (1)	**88.24** (1)
II_Ib	best	67.92 (3)	72.45 (3)	74.21 (3)	72.72 (3)	73.04 (3)	**78.75** (2)
	worst	62.22 (1)	65.27 (6)	65.87 (6)	66.22 (6)	66.48 (6)	**67.50** (6)
IV_1_calib_1a	best	76.72 (10)	79.15 (10)	80.23 (10)	80.74 (10)	81.35 (2)	**86.25** (2)
	worst	71.77 (55)	73.22 (55)	73.96 (35)	72.86 (40)	73.27 (40)	**73.75** (45,55)
IV_1_calib_1b	best	66.42 (35)	67.52 (35)	68.35 (3)	67.41 (3)	67.76 (1)	**71.25** (16)
	worst	64.65 (50)	65.19 (50)	65.45 (55)	64.78 (55)	64.94 (50)	**66.25** (40,50)
IV_1_calib_1f	best	72.76 (15)	73.97 (15)	75.34 (15)	74.45 (15)	76.09 (15)	**80** (16)
	worst	67.21 (45)	67.78 (45)	68.10 (45)	68.28 (45)	67.96 (45)	**68.75** (45,50)
IV_2a_s1	best	67.22 (10)	70.54 (10)	72.87 (10)	71.77 (10)	72.06 (10)	**75.91** (6,10,12)
	worst	63.88 (20)	64.23 (20)	65.07 (20)	64.59 (20)	64.19 (20)	**64.55** (20)
IV_2a_s2	best	70.31 (3)	71.77 (3)	72.21 (3)	72.65 (10)	72.45 (5)	**78.18** (11)
	worst	68.72 (1)	68.43 (20)	69.87 (1)	68.97 (20)	69.56 (15)	**72.5** (15,20)
IV_2a_s3	best	64.23 (10)	65.37 (10)	67.11 (10)	67.28 (5)	67.87 (5)	**74.32** (8,9)
	worst	59.92 (1)	60.78 (1)	60.98 (20)	60.77 (20)	61.54 (20)	**62.95** (20)
IV_3_s1	best	60.71 (5)	62.36 (3)	63.17 (3)	64.16 (3)	63.89 (3)	**70** (3)
	worst	56.78 (1)	57.83 (1)	58.21 (1)	57.98 (1)	58.45 (1)	**62.5** (1,2)
IV_3_s2	best	68.77 (3)	70.56 (7)	72.17 (7)	71.89 (7)	71.56 (7)	**77.5** (7)
	worst	64.23 (1)	65.18 (1)	65.56 (1)	64.89 (1)	65.34 (1)	**67.5** (1)

Table 5. Significant difference (p-value) of StEEGCS comparing to baselines with classification accuracy (Significant $*$: $p \leq 0.05$; Very significant $**$: $p \leq 0.01$; Extremely significant $***$: $p \leq 0.001$)

StEEGCS (vs.)	CSP	RCSP	SCSP	IMOCS	CCSE
II_Ia	**	**	*	**	***
II_Ib	***	**	*	**	**
IV_1_calib_1a	***	*	$-$ ($p = 0.059$)	*	*
IV_1_calib_1b	***	***	*	***	**
IV_1_calib_1f	**	*	$-$ ($p = 0.114$)	*	$-$ ($p = 0.058$)
IV_2a_s1	**	*	$-$ ($p = 0.088$)	*	$-$ ($p = 0.078$)
IV_2a_s2	***	***	***	**	***
IV_2a_s3	***	**	*	*	*
IV_3_s1	*	*	$-$ ($p = 0.075$)	$-$ ($p = 0.086$)	$-$ ($p = 0.094$)
IV_3_s2	**	*	$-$ ($p = 0.063$)	$-$ ($p = 0.074$)	$-$ ($p = 0.088$)

requires many iterations (or time) to find the optimal length and number of shapelets. But StEEGCS has competitive EEG channel selection efficiency to IMOCS and CCSE. Anyway, considering its superior classification performance, StEEGCS's execution time for EEG channel selection is acceptable.

6 CONCLUSION AND FUTURE WORKS

Multi-channel EEG is widely applied in Brain-Computer Interfaces (BCIs), but analyzing EEG signals with too many channels likely results in computational cost and inconvenience for BCI applications. EEG channel selection is a way to deal with the issue. Besides, as many

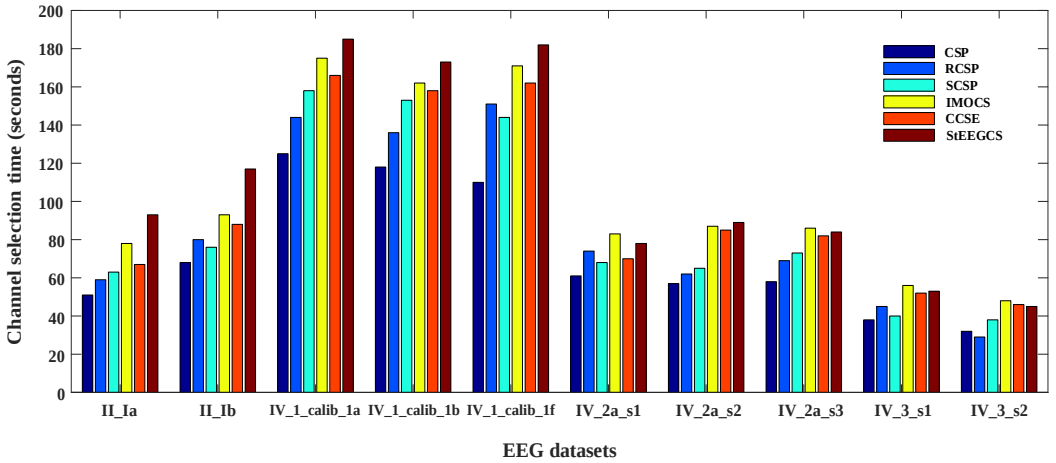


Fig. 19. Execution time on EEG channel selection (shapelet number: 3; shapelet length: 30)

studies reported before, EEG channel selection can not only improve the BCI performance by removing irrelevant or redundant EEG channels, but also enhances convenience for BCI applications with less EEG channels. Hence, we proposed an EEG shapelet-transformed channel selection approach that we call StEEGCS, which firstly used EEG shapelets to represent original EEG data, and subsequently applied gradient descent technique to learn distinct EEG shapelets, hyperplane, and EEG channel weights in a non-convex logistic loss minimization function. Finally, the most relevant EEG channels to classification performance are selected for SVM classifier. The experimental results on several real-world EEG datasets demonstrated that StEEGCS improves the classification accuracy and efficiency by selecting a small amount of EEG channels, and it also outperforms the classic and state-of-the-art EEG channel selection methods on SVM classification performance.

In the article, gradient descent is adopted in StEEGCS that probably leads to local optima, so other techniques such as heuristic approach should be considered to solve the non-convex minimization objective function of logistic loss for StEEGCS. Besides, as many promising classifiers emerged out, it would be interesting to apply new classifiers, such as DNNs, HIVE-COTE, *etc.* mentioned in Section 5.2, to analyze the universality of StEEGCS in future work. Additionally, we also intend to exploit more efficient EEG shapelet transforming/learning/selection approaches to extract distinct and informative EEG shapelets in the future, such as [11, 25, 32, 42], and so on.

ACKNOWLEDGMENTS

This work was partially supported in part by National Natural Science Foundation of China (Grant No. U1433116), Fundamental Research Funds for the Central Universities (Grant No. NP2017208).

REFERENCES

[1] F. Alimardani, R. Boostani, and B. Blankertz. 2017. Weighted spatial based geometric scheme as an efficient algorithm for analyzing single-trial EEGs to improve cue-based BCI classification. *Neural networks* 92 (2017), 69–76.

[2] T. Alotaiby, F. E. Abd El-Samie, S. A. Alshebeili, and I. Ahmad. 2015. A review of channel selection algorithms for EEG signal processing. *EURASIP journal on advances in signal processing* 2015, 66

(2015), 1–21.
[3] K. K. Ang, K. S. G. Chua, K. S. Phua, C. Wang, Z. Y. Chin, C. W. Kuah, W. Low, and C. Guan. 2015. A randomized controlled trial of EEG-based motor imagery brain-computer interface robotic rehabilitation for stroke. *Clinical EEG and neuroscience,* 46, 4 (2015), 310–320.
[4] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek. 2011. Optimizing the channel selection and classification accuracy in EEG-based BCI. *IEEE transactions on biomedical engineering* 58, 6 (2011), 1865–1873.
[5] A. Bagnall, J. Lines, J. Hills, and A. 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.
[6] X. Bi and H. Wang. 2019. Early Alzheimer's disease diagnosis based on EEG spectral images using deep learning. *Neural networks* 114 (2019), 119–135.
[7] B. Blankertz, F. Losch, M. Krauledat, G. Dornege, G. Gurio, and K. R. Müller. 2008. The Berlin brain-computer interface: Accurate performance from first-session in BCI-naïve subjects. , *IEEE transactions on biomedical engineering* 55, 10 (2008), 2452–2462.
[8] C. Chang and C. Lin. 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology* 2, 3 (2011), 1–27.
[9] C. Dai, D. Pi, Stefanie I. Becker, J. Wu, L. Cui, and B. Johnson. 2020. CenEEGs: Valid EEG selection for classification. *ACM transactions on knowledge discovery from data* 14, 2, Article 18 (2020).
[10] C. Dai, J. Wu, D. Pi, and L. Cui. 2018. Brain EEG time series selection: A novel graph-based approach for classification. In *Proceeding of SIAM international conference on data mining.* SIAM, San Diego, CA, USA, 558–566.
[11] Z. Fang, P. Wang, and W. Wang. 2018. Efficient learning interpretable shapelets for accurate time series classification. In *Proceeding of 34th IEEE international conference on data engineering.* IEEE, Paris, France, 497–508.
[12] J. Farquhar, N. J. Hill, T. N. Lal, and B. Schölkopf. 2006. Regularised CSP for sensor selection in BCI. In *Proceeding of 3rd international conference on brain-computer interface workshop training course.* Verlag der Technischen Universität Graz, Graz, Austria, 14–15.
[13] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. Muller. 2019. Deep learning for time series classification: a review. *Data mining and knowledge discovery* 33 (2019), 917–963.
[14] J. Feng, E. Yin, J. Jin, R. Saab, I. Daly, X. Wang, D. Hu, and A. Cichocki. 2018. Towards correlation-based time window selection method for motor imagery BCIs. *Neural networks* 102 (2018), 87–95.
[15] A. Ghaemi, E. Rashedi, A. M. Pourahimi, M. Kamandar, and F. Rahdari. 2017. Automatic channel selection in EEG signals for classification of left or right hand movement in Brain Computer Interfaces using improved binary gravitation search algorithm. *Biomedical signal processing and control* 3 (2017), 109–118.
[16] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C Van Essen. 2016. A multi-modal parcellation of human cerebral cortex. *Nature* 536 (2016), 171–178.
[17] C. D. V. Gonzalez, J. H. S. Azuela, J. M. Antelis, and L. E. Falcón. 2020. Spiking neural networks applied to the classification of motor tasks in EEG signals. *Neural networks* 122 (2020), 130–143.
[18] J. Grabocka, N. Schilling, M. Wistuba, and L. Schmidt-Thieme. 2014. Learning time-series shapelets. In *Proceeding of the 20th ACM SIGKDD international conference on knowledge discovery and data mining.* ACM, New York, NY, USA, 392–401.
[19] V. S. Handiru and V. A. Prasad. 2016. Optimized bi-objective EEG channel selection and cross-subject generalization with brain-computer interfaces. *IEEE transactions on human-machine systems* 46, 6 (2016), 777–786.
[20] L. He, D. Hu, M. Wan, Y. Wen, K. M. von Deneen, and M. Zhou. 2016. Common Bayesian network for classification of EEG-based multiclass motor imagery BCI. *IEEE transactions on systems, man, and cybernetics: systems* 46, 6 (2016), 843–854.
[21] J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. J. Bagnall. 2014. Classification of time series by shapelet transformation. *Data mining and knowledge discovery* 28, 4 (2014), 851–881.
[22] L. Hou, J. T. Kwok, and J. M. Zurada. 2016. Efficient learning of timeseries shapelets. In *Proceeding of the 30th AAAI conference on artificial intelligence.* ACM, Phoenix, Arizona, USA, 1209–1215.
[23] A. Jafarifarmand, M. A. Badamchizadeh, S. Khanmohammadi, M. A. Nazari, and B. M. Tazehkand. 2018. A new self-regulated neuro-fuzzy framework for classification of EEG signals in motor imagery BCI. *IEEE transactions on fuzzy systems* 26, 3 (2018), 1485–1497.

[24] J. S. Jeong. 2004. EEG dynamics in patients with Alzheimer's disease. *Clinical neurophysiology,* 115, 7 (2004), 1490–1505.
[25] C. Ji, C. Zhao, S. Liu, C. Yang, L. Pan, L. Wu, and X. Meng. 2019. A fast shapelet selection algorithm for time series classification. *Computer networks* 148 (2019), 231–240.
[26] C. Ji, C. Zhao, L. Pan, S. Liu, C. Yang, and L. Wu. 2017. A fast shapelet discovery algorithm based on important data points. *International journal of web services research* 14, 2 (2017), 67–80.
[27] Y. Jiao, Y. Zhang, X. Chen, E. Yin, J. Jin, X. Wang, and A. Cichocki. 2019. Sparse group representation model for motor imagery EEG classification. *IEEE journal of biomedical and health informatics* 23, 2 (2019), 631–641.
[28] J. Jin, Y. Miao, I. Daly, C. Zuo, D. Hu, and A. Cichocki. 2019. Correlation-based channel selection and regularized feature optimization for MI-based BCI. *Neural networks* 118 (2019), 262–270.
[29] R. Lahiri, P. Rakshit, and A. Konar. 2017. Evolutionary perspective for optimal selection of EEG electrodes and features. *Biomedical signal processing and control* 36 (2017), 113–137.
[30] T. N. Lal, M. Schröder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Schölkopf. 2004. Support vector channel selection in BCI. *IEEE transactions on biomedical engineering* 51, 6 (2004), 1003–1010.
[31] T. Lan, D. Erdogmus, A. Adami, M. Pavel, and S. Mathan. 2005. Salient EEG channel selection in brain computer interfaces by mutual information maximization. In *Proceeding of 27th Annual international conference of IEEE engineering in medicine and biology society*. IEEE, Shanghai, China, 7064–7067.
[32] G. Li, W. Yan, and Z. Wu. 2019. Discovering shapelets with key points in time series classification. *Expert systems with applications* 132 (2019), 76–86.
[33] J. Lines, L. M. Davis, J. Hills, and A. Bagnall. 2012. A shapelet transform for time series classification. In *Proceeding of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, Beijing, China, 289–297.
[34] J. Lines, S. Taylor, and A. 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, Article 52 (2018).
[35] F. Lotte and C. Guan. 2011. Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms. *IEEE transactions on biomedical engineering* 58, 2 (2011), 355–362.
[36] J. Meng, G. Liu, G. Huang, and X. Zhu. 2009. Automated selecting subset of channels based on CSP in motor imagery brain-computer interface system. In *Proceeding of 2009 IEEE international conference on robotics and biomimetics*. IEEE, Guilin, China, 2290–2294.
[37] A. Mueen, E. Keogh, and N. Young. 2011. Logical-shapelets: an expressive primitive for time series classification. In *Proceeding of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, San Diego, CA, USA, 1154–1162.
[38] E. Netzer, A. Frid, and D. Feldman. 2020. Real-time EEG classification via coresets for BCI applications. *Engineering applications of artificial intelligence* 89, Article 103455 (2020).
[39] T. Le Nguyen, S Gsponer, and G. Ifrim. 2017. Time series classification by sequence learning in all-subsequence space. In *Proceeding of 2017 IEEE 33rd international conference on data engineering*. IEEE, San Diego, CA, USA, 947–958.
[40] E. S. Pane, A. D. Wibawa, and M. H. Purnomo. 2018. Channel selection of EEG emotion recognition using stepwise discriminant analysis. In *Proceeding of 2018 international conference on computer engineering, network and intelligent multimedia*. IEEE, Surabaya, Indonesia, 14–19.
[41] H. Peng, Y. Wang, J. Chao, X. Huo, and D. Majoe. 2017. Stability study of optimal channel selection for emotion classification from EEG. In *Proceeding of 2017 IEEE international conference on bioinformatics and biomedicine*. IEEE, Kansas City, MO, USA, 2031–2036.
[42] T. Rakthanmanon and E. Keogh. 2013. Fast shapelets: A scalable algorithm for discovering time series shapelets. In *Proceeding of the 2013 SIAM international conference on data mining*. SIAM, Austin, Texas, USA, 668–676.
[43] D. S. Raychaudhuri, J. Grabocka, and L. Schmidt-Thieme. 2017. Channel masking for multivariate time series shapelets. *arXiv:1711.00812* (2017).
[44] A. E. Selim, M. A. Wahed, and Y. M. Kadah. 2008. Machine learning methodologies in brain-computer interface systems. In *Proceeding of 2008 Cairo International Biomedical Engineering Conference*. IEEE, Cairo, Egypt, 1–5.
[45] R. V. A. Sheorajpanday, G. Nagels, A. J. T. M. Weeren, and P. P. De Deyn. 2011. Quantitative EEG in ischemic stroke: Correlation with functional status after 6 months. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 122, 5 (2011), 874–883.

[46] D. F. Silva, Vinícius M. A. De Souza, and Gustavo E. A. P. A. Batista. 2013. Time series classification using compression distance of recurrence plots. In *Proceeding of 2013 IEEE 13th international conference on data mining*. IEEE, Dallas, TX, USA, 687–696.

[47] S. Sonoda, K. Nakamura, Y. Kaneda, H. Hideitsu, A. Shotaro, M. Noboru, M. Eri, and K. Masahiro. 2018. EEG dipole source localization with information criteria for multiple particle filters. *Neural networks* 108 (2018), 68–82.

[48] P. Trujillo, A. Mastropietro, A. Scano, A. Chiavenna, S. Mrakic-Sposta, M. Caimmi, F. Molteni, and G. Rizzo. 2017. Quantitative EEG for predicting upper limb motor recovery in chronic stroke robot-assisted rehabilitation. *IEEE transactions on neural systems and rehabilitation engineering* 25, 7 (2017), 1058–1067.

[49] N. D. Truong, L. Kuhlmann, M. R. Bonyadi, J. Yang, A. Faulks, and O. Kavehei. 2017. Supervised learning in automatic channel selection for epileptic seizure detection. *Expert systems with applications* 86 (2017), 199–207.

[50] N. D. Truong, A. D. Nguyen, L. Kuhlmann, M. Bonyadi, J. Yang, S. Ippolito, and O. Kavehei. 2018. Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neural networks* 105 (2018), 104–111.

[51] Y. Wang, S. Gao, and X. Gao. 2005. Common spatial pattern method for channel selection in motor imagery based brain-computer interface. In *Proceeding of 27th Annual international conference of IEEE engineering in medicine and biology society*. IEEE, Shanghai, China, 5392–5395.

[52] H. G. Wieser, K. Schindler, and D. Zumsteg. 2006. EEG in Creutzfeldt-Jakob disease. *Clinical neurophysiology* 117, 5 (2006), 935–951.

[53] T. Yang, K. K. Ang, K. S. Phua, J. Yu, V. Toh, W. H. Ng, and R. Q. So. 2018. EEG channel selection based on correlation coefficient for motor imagery classification: A study on healthy subjects and ALS patient. In *Proceeding of 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, Honolulu, HI, USA, 1996–1999.

[54] L. Ye and E. Keogh. 2009. Time series shapelets: As new primitive for data mining. In *Proceeding of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Paris, France, 947–956.

[55] J. Zakaria, A. Mueen, and E. Keogh. 2012. Clustering time series using unsupervised-shapelets. In *Proceeding of the 2012 IEEE 12th International Conference on Data Mining*. IEEE, Brussels, Belgium, 785–794.

[56] Q. Zhang, J. Wu, P. Zhang, G. Long, and C. Zhang. 2019. Salient subsequence learning for time series clustering. *IEEE transactions on pattern analysis and machine intelligence* 41, 9 (2019), 2193–2207.