Bayesian-based Classification Confidence Estimation for Enhancing SSVEP Detection

Yue Zhang, Sheng Quan Xie, *Senior Member, IEEE,*, He Wang, *Member, IEEE*, Chaoyang Shi, *Member, IEEE*, and Zhi-Qiang Zhang, *Member, IEEE*

***Abstract*—The brain-computer interface (BCI) enables paralyzed people to directly communicate with and operate peripheral equipment. The steady-state visual evoked potential (SSVEP)-based BCI system has been extensively investigated in recent years due to its fast communication rate and high signal-to-noise ratio. Many present SSVEP recognition methods determine the target class via finding the largest correlation coefficient. However, the classification performance usually degrades when the largest coefficient is not significantly different from the rest of the values. This study proposed a Bayesian-based classification confidence estimation method to enhance the target recognition performance of SSVEP-based BCI systems. In our method, the differences between the largest and the other values generated by a basic target identification method are used to define a feature vector during the training process. The Gaussian mixture model (GMM) is then employed to estimate the probability density functions of feature vectors for both correct and wrong classifications. Subsequently, the posterior probabilities of being an accurate and false classification are calculated via Bayesian inference in the test procedure. A classification confidence value (CCValue) is presented based on two posterior probabilities to estimate the classification confidence. Finally, the decision-making rule can determine whether the present classification result should be accepted or rejected. Extensive evaluation studies were performed on an open-access benchmark dataset and a self-collected dataset. The experimental results demonstrated the effectiveness and feasibility of the proposed method for improving the reliability of SSVEP-based BCI systems.**

***Index Terms*—Brain-computer interface (BCI), electroencephalography (EEG), steady-state visual evoked potential (SSVEP), classification confidence estimation, Bayesian inference**

I. Introduction

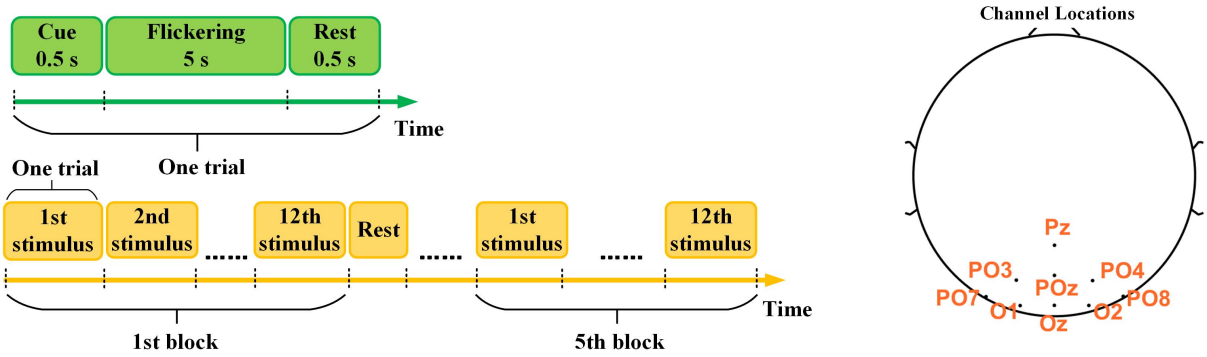
BRAIN-computer interface (BCI) systems can detect brain activity and then translate neural signals directly into commands to operate external devices without relying on peripheral nerves and muscles [1]–[3]. The electroencephalogram (EEG)-based BCI is a popular non-invasive technique due to portability, low cost, and high temporal resolution [4]–[6]. Three paradigms in the EEG signal are most widely explored, namely, the steady-state visual evoked potential (SSVEP), the P300 event-related potential (ERP), and the event-related desynchronization (ERD) [7]. These paradigms have come to light in several practical applications, including assistance robots [8], augmented reality (AR) glasses [9], [10], and entertainment [11], [12]. Among these paradigms, SSVEP-based BCI systems have attracted extensive research attention because of their advantages of high information transfer rate (ITR) and signal-to-noise ratio (SNR) [13]–[16].

In the past decades, many target recognition methods have been proposed to analyze the SSVEP features and detect the subject's intent to operate the peripheral device [17]. In particular, canonical correlation analysis (CCA) is the most popular target detection method because of its simplicity of use and robustness [18], [19]. However, the performance of CCA is still influenced by the interference from spontaneous EEG signals [20]. In recent years, many improved approaches have been proposed for SSVEP detection. Generally, the literature presents three major optimization directions, i.e., individual templates [21], [22], time filters [23], and spatial filters [20], [24]. Among many methods, sum of squared correlations (SSCOR) [24] and task-related component analysis (TRCA) [20] have attained nice performance in SSVEP detection. In the recognition stage of the aforementioned methods, the target class is identified by the largest correlation coefficient. It may lead to misclassification when the maximum coefficient has a low confidence level. The detection performance may deteriorate if the maximum value is not remarkably different from the other values. Therefore, evaluating the reliability of classification results is another crucial direction for enhancing the performance and applicability of SSVEP-based BCI systems.

The classification confidence analysis process could facilitate detection methods to reject results with a low level of confidence [25]. In recent research, many confidence evaluation methods for the SSVEP-based BCI system have been introduced. For instance, Zhao *et. al* [26] designed a decision-making selector to select a more reliable result from a pair of CCA-based methods. The overall recognition performance was enhanced by the fusion strategy, but the average detection time increased accordingly. Currently, many researchers have focused on confidence estimation based on a single decision. Chen *et.al* [27] created a hypothesis testing model

This work was supported in part by Engineering and Physical Sciences Research Council (EPSRC) (Grant No. EP/S019219/1), in part by Royal Society (Grant No. IEC\NSF\211360) in part by China Scholarship Council (CSC) (Grant No. 201906460007). (Corresponding authors: Zhi-Qiang Zhang and Chaoyang Shi)

Yue Zhang, Sheng Quan Xie, and Zhi-Qiang Zhang are with the Institute of Robotics, Autonomous System and Sensing, School of Electrical and Electronic Engineering, University of Leeds, Leeds LS2 9JT, U.K. (e-mail: elyzh@leeds.ac.uk; s.q.xie@leeds.ac.uk; z.zhang3@leeds.ac.uk). He Wang is with School of Computing, University of Leeds, LS2 9JT, UK (H.E.Wang@leeds.ac.uk). Chaoyang Shi is with School of Mechanical Engineering, Tianjin University, 300072, Tianjin, China (chaoyang.shi@tju.edu.cn)



(a) Experimental paradigm (b) Channel location representation

Fig. 1. (a) experimental paradigm and (b) channel location of SSVEP recording.

for evaluating the credibility of results using the coefficients of filter-bank CCA. Cecotti [28] investigated the impact of different dynamic time segment selections on the confidence of CCA's outputs. Similarly, Jiang *et.al* [29] estimated the classification confidence based on the largest two coefficients and then determined the optimal data length. According to several previous studies, the difference between the first and the second-largest feature values provides useful information for the classification estimation [30]. In general, the probability of correct recognition is higher as this difference is larger [27]. However, these methods simply exploit the first two coefficients or their difference, which is insufficient to construct informative features for enhancing SSVEP detection.

In this paper, Bayesian-based classification confidence estimation method was proposed for improving the recognition reliability of SSVEP-based BCI systems, which is crucial for SSVEP-based human-robot interaction [31], [32]. Wrong classifications can cause the external device to carry out the wrong actions, perhaps resulting in adverse incidents and serious physical harm to humans. In the practical usage scene, it is essential to enhance subjects' safety and security, particularly in rehabilitation and assistive technology. The main contributions of this work include: 1) In the training step, the feature vector involving the differences between the largest correlation coefficient and the other values was constructed. Gaussian mixture model (GMM) was used to estimate the conditional probability density functions of feature vectors given correct and wrong results. 2) In the test step, Bayesian inference was used to calculate the posterior probabilities of being a correct and wrong classification using the newly obtained feature vector. A classification confidence value (CCValue) was then presented to estimate the classification confidence. 3) The decision-making rule decides whether the present classification result should be accepted or rejected.

For this study, SSCOR and TRCA were selected as the basic target recognition methods. The proposed methods that incorporate CCValue estimation based on SSCOR/TRCA are named SSCOR+CCValue and TRCA+CCValue, respectively. The performance was assessed on a 40-class publicly available benchmark dataset [33] and a 12-class self-collected dataset. Extensive comparisons were performed among the four methods. The effectiveness and reliability of SSCOR+CCValue and TRCA+CCValue were demonstrated via experimental evaluation studies on two datasets.

This paper is organized as follows: Section II introduces the SSVEP datasets and the proposed Bayesian-based classification confidence estimation method for SSVEP-based BCI systems. The experimental results are shown in Section III. Section IV discusses some issues with our method. Section V presents the conclusion.

II. METHOD AND MATERIALS

A. EEG Signals

In this paper, an open-access dataset [33] and a self-collected SSVEP dataset (referred to as Dataset I and Dataset II, respectively) were utilized to evaluate the performance of the proposed method. The benchmark dataset was recorded from thirty-five participants with forty visual stimuli. The sampling rate is 250 Hz. The frequencies range from 8 Hz to 15.8 Hz, with an interval of 0.2 Hz. The phase difference between two neighboring stimuli is 0.5π. For each participant, the data contains six blocks of forty trials associated with forty stimuli. In each trial, the subject was asked to faze at the stimulus for 5 s. More information about this publicly available dataset can be found in [33]. A detailed description of the self-collected dataset is provided below.

1) Subjects: In Dataset II, eleven subjects (five females and six males, mean age: twenty-five years) joined the SSVEP experiment. All the people were healthy and had a normal or corrected-to-normal vision. The experiment has been approved by the Research Ethics Committee of the University of Leeds. The subjects were all asked to read and sign the participant consent form.

2) Stimulus Design: In Dataset II, the visual stimulation was coded by the joint frequency and phase modulation (JFPM) method. There was a 4×3 matrix on a 23.6-inch LCD monitor, which has a resolution of 1920×1080 pixels and a refresh rate of 60 Hz, respectively. The stimulation frequencies differed from 9.25 Hz to 14.75 Hz with an interval of 0.5 Hz. The phase began from 0π to 1.5π in steps of 0.5π. The reason for frequency band selection is to collect relatively strong SSVEP signals. The experiment included five blocks for each participant. In each block, there are twelve trials corresponding

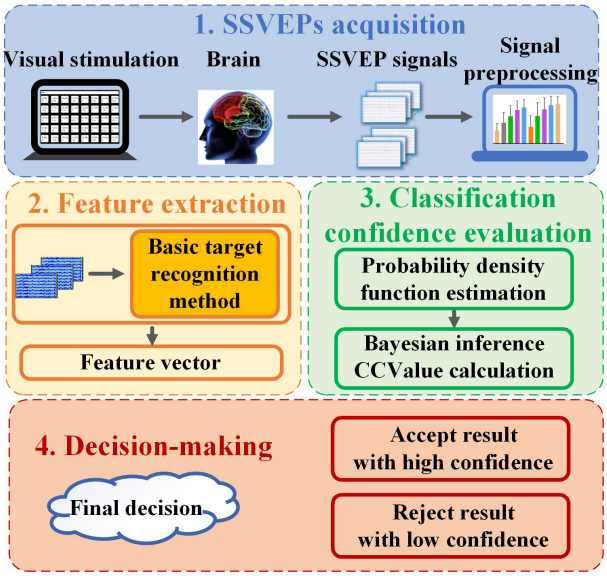


Fig. 2. Diagram of the Bayesian-based SSVEP classification confidence estimation method.

to twelve targets. Each trial started with a 0.5 s red dot cue, indicating the target stimulus. Then, twelve flickers flashed at the same time for 5 s, during which the subject was required to stare at the target flicker without moving his or her eyes. The screen was blank for 0.5 s after that. There was a short break of one minute between two adjacent blocks. During the experiment, the subject was asked to sit in a comfortable chair in a dimly lit and quiet environment. The viewing distance to the computer screen was 70 cm. To decrease body noise, each subject was requested not to talk, cough, or cry during the data collection. The experimental paradigm is shown in Fig. 1(a).

3) EEG Recording: For Dataset II, SSVEP data was recorded by the experiment device from g.tec medical engineering GmbH. The g.USBamp amplifier was used to sample data at 256 Hz. SSVEP signals mainly appear over parietal and occipital regions since they are closer to the visual cortex of the human brain [34]–[36]. Some studies presented that SSVEP signals near these areas have larger amplitude and SNR [33], [37]. Therefore, nine electrodes (i.e., Pz, PO3, POz, PO4, PO7, O1, Oz, O2, and PO8) located in parietal and occipital areas were chosen. Fig. 1(b) shows the channel positions. The reference channel was at the right earlobe, and the ground electrode was placed over electrode FPz.

B. Data Preprocessing

To account for the latency delay in the human visual system, the EEG signal in [0.14 s $0.14 + d$ s] was extracted for method performance evaluation [38]. The variable d in this context refers to the length of the data that is being used for analysis. The Chebyshev Type I Infinite Impulse Response (IIR) filter was applied in this work to create band-pass filters. The data was filtered between eight Hz and eighty-eight Hz for Dataset I. The data was filtered between eight Hz and forty Hz for Dataset II. In addition, a data standardization step was performed on both datasets [20].

C. Bayesian-based SSVEP Classification Confidence Estimation Method

A Bayesian-based classification confidence estimation method was proposed for improving SSVEP recognition reliability. As shown in Fig. 2, it includes four modules: EEG signal acquisition, feature extraction, classification confidence evaluation, and decision making. The dataset descriptions have been given in the previous section. In the following subsections, the work procedures of the other three modules will be explained in detail.

1) Feature Extraction: Denote a four-way tensor $\boldsymbol{\chi} \in \mathbb{R}^{N_f \times N_c \times N_s \times N_t}$, where N_f indicates the number of stimuli, N_c represents the number of channels, N_s is the number of samples, and N_t is the number of training trials. Hereafter, i refers to the stimulus index, j refers to the channel index, m refers to the sample index, and h refers to the index of training trials. Therefore, the recorded individual calibration signal for i-th stimulus is $\boldsymbol{\chi}_i \in \mathbb{R}^{N_c \times N_s \times N_t}$. The spatial filter $\boldsymbol{w}_i \in \mathbb{R}^{N_c}$ for i-th stimulus can be constructed as $\boldsymbol{w}_i = f(\boldsymbol{\chi}_i)$ by a basic target recognition method in SSVEP-based BCIs. $f(\cdot)$ represents the spatial filtering method. In this study, TRCA and SSCOR were selected. In TRCA [20], weight coefficients are optimized to maximize inter-trial covariance of brain activities. SSCOR spatial filter learns a common SSVEP representation space through the optimization of the individual SSVEP templates. It could improve the SNR of the SSVEP components embedded in the recorded EEG data [24].

The single-trial individual template signal is denoted as $\overline{\boldsymbol{\chi}}_i = \frac{1}{N_t} \sum_{h=1}^{N_t} \boldsymbol{\chi}_{ih} \in \mathbb{R}^{N_c \times N_s}$. Once the spatial filter $\boldsymbol{w}_i$ is produced, the evaluation SSVEP data $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{N_c \times N_s}$ and individual template signal $\overline{\boldsymbol{\chi}}_i$ can both be optimised. Therefore, the SSVEP feature was further extracted from recorded EEG signals. The correlation coefficient between the two spatially filtered signals corresponding to each stimulus is shown as follows:

$$r_i = \rho(\widetilde{\boldsymbol{X}}^{\mathsf{T}} \boldsymbol{w}_i, \overline{\boldsymbol{\chi}}_i^{\mathsf{T}} \boldsymbol{w}_i), \ i = 1, 2, ..., N_f \tag{1}$$

where $\rho(\boldsymbol{a}, \boldsymbol{b})$ refers the Pearson correlation coefficient between vector $\boldsymbol{a}$ and vector $\boldsymbol{b}$. The frequency of the individual template related to the largest correlation coefficient is decided as the frequency f of the test signal:

$$f = \arg\max_{f_i} r_i, i = 1, 2, ..., N_f \tag{2}$$

Considering this type of decision-making rule may result in poor classification performance when the maximal coefficient is not much different from others, a Bayesian-based classification confidence estimation method was proposed in this study. The coefficient vector calculated by (1) is denoted as $\boldsymbol{\Phi} = [r_1, r_2, ..., r_{N_f}]$. The coefficient vector was rearranged in descending order, resulting in a new vector $\widetilde{\boldsymbol{\Phi}} = [\widetilde{r}_1, \widetilde{r}_2, ..., \widetilde{r}_{N_f}]$. It means that the largest coefficient is $\widetilde{r}_1$, and the smallest one is $\widetilde{r}_{N_f}$. Subsequently, it is possible to calculate the differences between the largest coefficient $\widetilde{r}_1$ and other values $\widetilde{r}_j, (j = 2, 3, \ldots, N_f)$, and thus yield $(N_f - 1)$ differences.

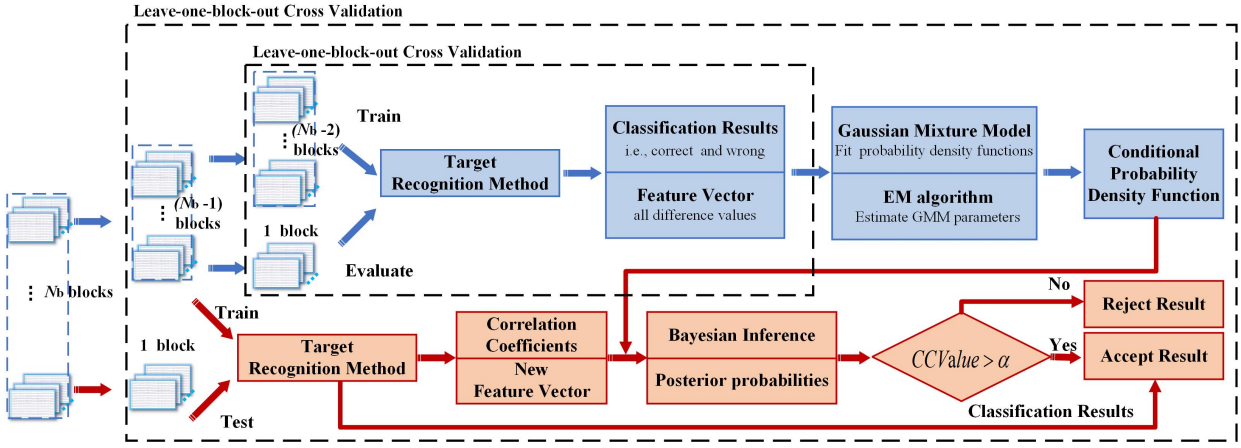


Fig. 3. The detailed framework of the Bayesian-based classification confidence estimation method for SSVEP detection. The leave-one-block-out cross-validation was performed in the experiment evaluation.

Therefore, the difference values $\Delta r_i, (i = 1, 2, \ldots, N_{f-1})$ can be expressed as:

$$\begin{aligned} \Delta r_1 &= \widetilde{r}_1 - \widetilde{r}_2 \\ \Delta r_2 &= \widetilde{r}_1 - \widetilde{r}_3 \\ &\vdots \\ \Delta r_{N_{f-1}} &= \widetilde{r}_1 - \widetilde{r}_{N_f}. \end{aligned} \tag{3}$$

The final feature vector can be expressed as $\boldsymbol{F} = [\Delta r_1, \Delta r_2, \ldots, \Delta r_{N_{f-1}}]$ by $(N_f - 1)$ differences.

2) Bayesian-based Classification Confidence Evaluation: As illustrated in Fig. 3, performance was assessed by the leave-one-block-out cross-validation. Specifically, for N_b blocks of EEG signals, $(N_b - 1)$ blocks were selected for training conditional probability density functions, and one block was used for testing. Moreover, in the training process, leave-one-block-out cross-validation was again employed to collect classification results and construct feature vectors (blue part in Fig. 3). Specifically, $(N_b - 2)$ blocks were selected to train the target recognition method, and the left-out block was used as evaluation data. The signal in each block is represented as $\boldsymbol{\chi}_h \in \mathbb{R}^{N_f \times N_c \times N_s}$. Therefore, there are total $(N_b - 1) \times N_f$ trials that can be evaluated, thus classification results and feature vectors can be collected to train GMM accordingly.

The classification results were subsequently separated into two groups. Suppose the correct classification is represented as C_1, and the corresponding feature vectors are $\boldsymbol{F}_c$. The wrong classification is denoted as C_0, and the corresponding feature vectors are $\boldsymbol{F}_w$. The probability density functions of the feature vector for correct and wrong classifications are represented as $p(\boldsymbol{F}|C_1)$ and $p(\boldsymbol{F}|C_0)$, respectively. For ease of reference, they can also be written as $p(\boldsymbol{F}_c)$ and $p(\boldsymbol{F}_w)$. In this study, GMM was applied to fit feature vectors from correct and wrong classifications. The GMM is a versatile and efficient probabilistic model, that can build any complicated probability distribution function [39]. Therefore, the two probability distribution functions can be expressed as follows:

$$\begin{aligned} p(\boldsymbol{F}|C_1) = p(\boldsymbol{F}_c) &= \sum_{k=1}^{K} \lambda_k \mathcal{N}(\boldsymbol{F}_c|\boldsymbol{\theta}_k) \\ p(\boldsymbol{F}|C_0) = p(\boldsymbol{F}_w) &= \sum_{k=1}^{K} \eta_k \mathcal{N}(\boldsymbol{F}_w|\boldsymbol{\psi}_k) \end{aligned} \tag{4}$$

where K is the number of mixed components. The $\lambda_k \in [0, 1]$ and $\eta_k \in [0, 1]$ are the mixture component weights for the k-th component, with the constraint that $\sum_{k=1}^{K} \lambda_k = 1$ and $\sum_{k=1}^{K} \eta_k = 1$. The Gaussian density functions $\mathcal{N}(\boldsymbol{F}_c)$ and $\mathcal{N}(\boldsymbol{F}_w)$ are determined by the parameter $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and $\boldsymbol{\psi}_k = (\boldsymbol{\nu}_k, \boldsymbol{\Gamma}_k)$, where $\boldsymbol{\mu}_k$ and $\boldsymbol{\nu}_k$ refer to the mean, while $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Gamma}_k$ are the covariance matrix, respectively. The GMM parameters, namely, $\lambda_k, \eta_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\nu}_k$ and $\boldsymbol{\Gamma}_k (k = 1, 2, \ldots, K)$, were estimated by the Expectation-Maximization (EM) algorithm in this study. The EM algorithm is an iterative method for estimating parameters in statistical models [40]. Each iteration of this algorithm involves two steps: the expectation (E) step and the maximization (M) step.

Consider the case of $p(\boldsymbol{F}_c)$, assuming that there are N_{co} accurate classifications and $\boldsymbol{F}_c^t, (t = 1, 2, ..., N_{co})$ is the feature vector corresponding to t-th accurate result. A latent variable $\gamma_k^t, (t = 1, 2, ..., N_{co}; k = 1, 2, ..., K)$ was defined, and its expression is:

$$\gamma_k^t = \begin{cases} 1, & \boldsymbol{F}_c^t \text{ is from } k\text{-th mixed component} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Therefore, the complete data is $(\boldsymbol{F}_c^t, \gamma_1^t, \gamma_2^t, ..., \gamma_K^t)$.

E step is to determine the Q function, which is the expectation of the log-likelihood function for complete data:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s)}) = E[\log p(\boldsymbol{F}_c, \boldsymbol{\gamma}|\boldsymbol{\theta})|\boldsymbol{F}_c, \boldsymbol{\theta}^{(s)}] \tag{6}$$

$\boldsymbol{\theta}^{(s)}$ represents the parameters obtained by the s-th iteration.

Q step is to find the model parameter corresponding to the maximum value of the Q function:

$$\boldsymbol{\theta}^{(s+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s)}) \tag{7}$$

TABLE I
A CONFUSION MATRIX OF EXPLANATION ABOUT FOUR PARAMETERS, I.E., TR, FA, FR AND TA.

		The proposed method's decision	
		Rejection	**Acceptance**
Target identification method's decision (e.g. TRCA, SSCOR)	**Wrong**	True Rejection (TR)	False Acceptance (FA)
	Correct	False Rejection (FR)	True Acceptance (TA)

The updated model parameters $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \lambda_k, (k = 1, 2, ..., K)$ are [41]:

$$\boldsymbol{\mu}_k^{(s+1)} = \frac{\sum_{t=1}^{N_{co}} \hat{\gamma}_k^t \boldsymbol{F}_c^t}{\sum_{t=1}^{N_{co}} \hat{\gamma}_k^t} \tag{8}$$

$$\boldsymbol{\Sigma}_k^{(s+1)} = \frac{\sum_{t=1}^{N_{co}} \hat{\gamma}_k^t (\boldsymbol{F}_c^t - \boldsymbol{\mu}_k^{(s+1)})(\boldsymbol{F}_c^t - \boldsymbol{\mu}_k^{(s+1)})^{\mathsf{T}}}{\sum_{t=1}^{N_{co}} \hat{\gamma}_k^t} \tag{9}$$

$$\lambda_k^{(s+1)} = \frac{1}{N_{co}} \sum_{t=1}^{N_{co}} \hat{\gamma}_k^t \tag{10}$$

where $\hat{\gamma}_k^t$ is the probability that t-th feature vector $\boldsymbol{F}_c^t$ belongs to k-th mixed component. $\hat{\gamma}_k^t, (t = 1, 2, ..., N_{co}; k = 1, 2, ..., K)$ can be calculated via the following equation:

$$\hat{\gamma}_k^t = E(\gamma_k^t | \boldsymbol{F}_c, \boldsymbol{\theta}) = \frac{\lambda_k \mathcal{N}(\boldsymbol{F}_c^t | \boldsymbol{\theta}_k)}{\sum_{k=1}^{K} \lambda_k \mathcal{N}(\boldsymbol{F}_c^t | \boldsymbol{\theta}_k)} \tag{11}$$

The iteration between the E-step and M-step continues until convergence. Finally, $p(\boldsymbol{F}_c)$, also known as $p(\boldsymbol{F}|C_1)$ can be obtained. Accordingly, the parameters of the probability density function $p(\boldsymbol{F}|C_0)$ can also be calculated by the EM iterations. The distinction is that the $\boldsymbol{F}$ here refers to the feature vector $\boldsymbol{F}_w$ associated with the wrong classifications.

The prior probabilities of the correct and wrong classifications can be formulated as follows:

$$\begin{aligned} P(C_1) &= \frac{N_{co}}{N_{co} + N_{wr}} \\ P(C_0) &= \frac{N_{wr}}{N_{co} + N_{wr}} \end{aligned} \tag{12}$$

where N_{wr} indicates the number of wrong classification results. The target recognition method is then trained using $(N_b - 1)$ blocks of SSVEP signals, and the trained model is tested using the left-out block. According to the newly obtained feature vector $\hat{\boldsymbol{F}} \in \mathbb{R}^{(N_f - 1)}$, Bayesian inference is used to calculate the posterior probabilities of being a correct classification $P(C_1|\hat{\boldsymbol{F}})$ and a wrong classification $P(C_0|\hat{\boldsymbol{F}})$:

$$\begin{aligned} P(C_1|\hat{\boldsymbol{F}}) &= \frac{p(\hat{\boldsymbol{F}}|C_1)P(C_1)}{p(\hat{\boldsymbol{F}}|C_1)P(C_1) + p(\hat{\boldsymbol{F}}|C_0)P(C_0)} \\ P(C_0|\hat{\boldsymbol{F}}) &= \frac{p(\hat{\boldsymbol{F}}|C_0)P(C_0)}{p(\hat{\boldsymbol{F}}|C_1)P(C_1) + p(\hat{\boldsymbol{F}}|C_0)P(C_0)} \end{aligned} \tag{13}$$

Based on (13), the classification confidence value (CCValue) can be defined as:

$$\text{CCValue}(\hat{\boldsymbol{F}}) = P(C_1|\hat{\boldsymbol{F}}) - P(C_0|\hat{\boldsymbol{F}}) \tag{14}$$

3) Decision-Making Rule: In the decision-making module, the CCValue needs to be compared with a threshold α. The classification result should be accepted if the CCValue is greater than α. Otherwise, this module should reject the classification result. Therefore, the decision-making rule can be written as:

$$D_{final}(\hat{\boldsymbol{F}}) = \begin{cases} \text{Accept}, & \text{if CCValue}(\hat{\boldsymbol{F}}) > \alpha \\ \text{Reject}, & \text{if CCValue}(\hat{\boldsymbol{F}}) \leq \alpha \end{cases} \tag{15}$$

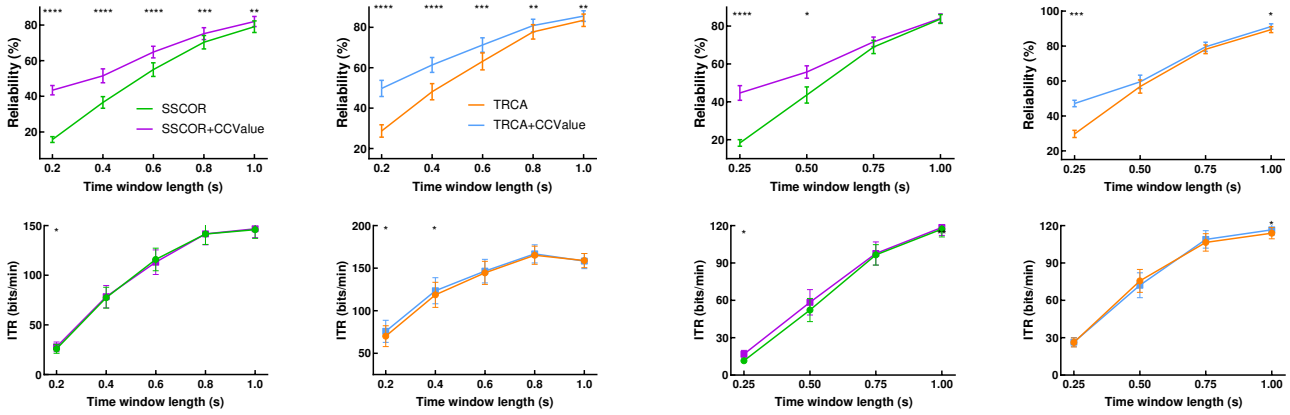
As shown in (15), $D_{final}(\hat{\boldsymbol{F}})$ works as a binary classifier. The grid-search method was used to determine α via $(N_b - 1)$ blocks training data. The range of α is specified as [-1, 1] according to (14). An exhaustive search is performed on the threshold values of the method with an interval of 0.1. In the search process, leave-one-block-out cross-validation was employed. Finally, the value that provides the highest average classification reliability across subjects was determined as α.

TRCA/SSCOR's classification result will be compared with the label of this classification. The classification results will be given a new label, i.e., "correct" or "wrong", which represents the ground truth. It is a gold standard that can be used to compare and evaluate the proposed method's results. If the proposed detection method could accept the "correct" classification or reject the "wrong" classification successfully, it means that the proposed method is effective.

The details of the proposed Bayesian-based classification confidence estimation method are shown in Fig. 3. This framework aims to reduce the number of low-confidence results and thus improve recognition reliability.

III. RESULTS

In this section, the proposed Bayesian-based classification confidence estimation method was applied to a 40-target benchmark dataset [33] as well as a 12-target self-collected dataset. TRCA+CCValue, SSCOR+CCValue, TRCA, and SSCOR are compared extensively. The number of channels and training blocks were set to nine, and five for Dataset I and nine, and four for Dataset II, respectively. The two datasets have different numbers of training blocks because of their different sizes. The selections of these hyperparameters were made to ensure that the model had access to all the available information and to facilitate the training process for each dataset. The number of Gaussian mixture components was set to two. The optimal number of components in the GMM was selected using the Akaike information criterion (AIC), which provides a trade-off between the goodness of fit of the model and its complexity. The effects of parameters, such as the number of channels, training blocks and correlation coefficients on recognition performance were further investigated.



(a) Dataset I (b) Dataset II

Fig. 4. Average recognition reliability and ITRs across subjects of various methods (i.e., SSCOR, TRCA, SSCOR+CCValue, and TRCA+CCValue) using different time windows (TWs) on (a) Dataset I and (b) Dataset II. The error bars represent standard error of mean (SEM), $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$ where $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$. x_i is the classification reliability or ITRs of i-th subject, $\overline{x}$ is the mean of samples, and n is the number of subjects. The asterisks indicate a significant difference between the two methods obtained by paired t-test analysis ($*$: p<0.05, $**$: p<0.01, $***$: p<0.001, $****$: p<0.0001).

A. Performance Evaluation

Table. I introduced four measures: true rejection (TR), false acceptance (FA), false rejection (FR), and true acceptance (TA). As indicated in the Table, the wrong classification results of TRCA/SSCOR can be divided into TR and FA. The correct results of TRCA/SSCOR can be divided into FR and TA. The significance of the proposed method is that low-confidence decisions can be detected and rejected so the system can be more robust and reliable. Therefore, the accuracy of the confusion matrix (i.e., Table. I) [42], also expressed as the recognition reliability (%) of the proposed method, can be defined as follows:

$$\text{Reliability} = \frac{TA + TR}{TA + FA + TR + FR} \times 100 \qquad (16)$$

Fig. 4 shows the average classification reliability of SSCOR, TRCA, SSCOR+CCValue, and TRCA+CCValue on (a) Dataset I and (b) Dataset II. The sampling rates are different in the two datasets, so different data lengths were used to keep the number of samples without decimals. To depict the improvement more intuitively, a pairwise comparison was performed between SSCOR and SSCOR+CCValue, as well as TRCA and TRCA+CCValue. The proposed method can attain higher reliability across a wide range of data lengths. Specifically, SSCOR+CCValue improved SSCOR by 2.90% $\sim$ 27.74% and TRCA+CCValue increased TRCA by 2.04% $\sim$ 21.07% in Dataset I. The classification reliability of SSCOR+CCValue is greater than that of SSCOR by 0.30% $\sim$ 26.37% in Dataset II. Similarly, TRCA+CCValue improved TRCA by 1.37% $\sim$ 17.42%. The paired t-test was conducted to explore the similarity of reliability between the basic recognition method and the corresponding proposed method. Statistical analysis shows that the reliability of SSCOR is significantly different from that of SSCOR+CCValue for almost all data lengths. This conclusion also applies to TRCA and TRCA+CCValue.

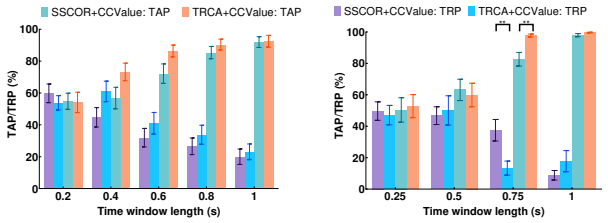
In Fig. 4, the ITRs of the proposed methods are slightly higher than those of SSCOR and TRCA. This is reasonable because SSCOR+CCValue and TRCA+CCValue aim to accept results with high confidence and discard results with low confidence, which would lead to trials corresponding to unconfident results not being classified. It is advantageous to allow the method to leave some trials unclassified since this can prevent the classifier from making errors when the classification results are not confident enough.

Table. II provides the intuitional numerical results for comparing methods more clearly [43]. As shown in Table. II, TRCA+CCValue always achieves the best performance with various data lengths in each dataset. SSCOR consistently performs worse than TRCA, whereas the performance of SSCOR+CCValue was improved after accounting for classification confidence estimation, and finally, SSCOR+CCValue outperforms TRCA at some TWs. Two popular recognition methods, i.e., CCA and Msetcca, were included for comparison. It is obvious that the proposed method achieves much higher recognition reliability than the two methods. A One-way repeated-measures ANOVA was conducted to investigate the similarity of classification reliability among these methods. The P-value is always < 0.0001, indicating statistically significant differences between the reliability of these methods at each TW.

The proposed method enhances recognition performance by accepting highly-trustworthy results and rejecting unconfident ones. Therefore, the method was further assessed in terms of two other indicators, i.e., the true accept proportion (TAP) and the true reject proportion (TRP). TAP is defined as the proportion of correct target identification method decisions to be accepted by $D_{final}(\hat{\boldsymbol{F}})$. TRP is defined as the proportion of wrong decisions rejected by the proposed method. Therefore, TRP indicates the rejection efficiency, whereas TAP relates to the cost [25]. Fig. 5 displays the TAP and TRP of SSCOR+CCValue and TRCA+CCValue on (a) Dataset I and (b) Dataset II using different data lengths. In Dataset I, with increasing data lengths, SSCOR+CCValue's TAP increases from 54.81% to 91.98%, while its TRP decreases

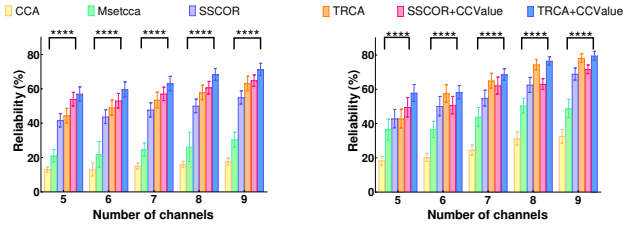
TABLE II
RELIABILITY COMPARISON BETWEEN FOUR METHODS

Methods	Averaged Recognition Reliability ± SEM (%)								
	Dataset I					Dataset II			
	0.2 s	*0.4s*	*0.6 s*	*0.8s*	*1 s*	*0.25 s*	*0.5 s*	*0.75 s*	*1 s*
CCA	3.76 ± 0.21	9.1 ± 0.8	17.8 ± 2.1	30 ± 3	41 ± 4	13.2 ± 1.2	29 ± 4	49 ± 5	72 ± 5
Msetcca	8.0 ± 1.2	19 ± 3	31 ± 4	46 ± 5	55 ± 5	10.5 ± 1.8	20 ± 4	33 ± 4	51.7 ± 2.4
SSCOR	15.7 ± 1.6	37 ± 3	55 ± 4	70 ± 4	79 ± 3	18.3 ± 1.8	44 ± 4	69 ± 4	83.8 ± 2.4
TRCA	29 ± 3	48 ± 4	63 ± 4	79 ± 4	84 ± 3	29.7 ± 2.1	57 ± 4	78.2 ± 2.5	89.4 ± 1.7
SSCOR+CCValue	43.5 ± 2.6	52 ± 4	65 ± 3	75 ± 3	82.0 ± 2.8	45 ± 4	56 ± 3	71.7 ± 2.6	84.1 ± 2.3
TRCA+CCValue	45 ± 4	61 ± 4	71 ± 4	81 ± 3	85.6 ± 2.7	47.1 ± 1.8	60 ± 4	79.6 ± 2.6	91.2 ± 1.6
P-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001



(a) Dataset I (b) Dataset II

Fig. 5. TAP and TRP of SSCOR+CCValue and TRCA+CCValue on (a) Dataset I and (b) Dataset II with different data lengths. The error bars represent SEM. The asterisks indicate a significant difference between methods obtained by t-test analysis.



(a) Dataset I (b) Dataset II

Fig. 6. Barchart of the classification reliability of six methods with different numbers of electrodes on (a) Dataset I and (b) Dataset II. The error bars represent SEM. The asterisks indicate significant differences between the four methods obtained by one-way repeated-measures ANOVA.

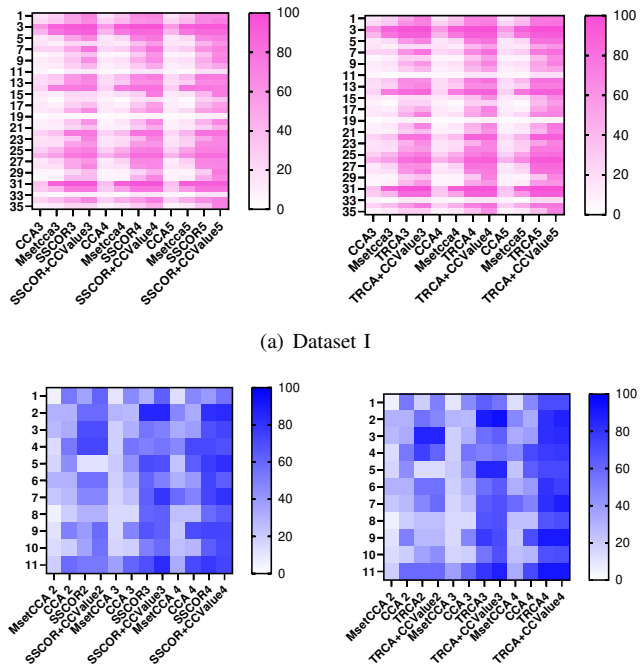
from 53.84% to 23.06%. For TRCA+CCValue, TAP rises from 54.08% to 92.53%, and its TRP changes from 59.79% to 20.01%. Dataset II exhibits similar results as well. Similarly, in Dataset II, SSCOR+CCValue's TAP increases from 50.39% to 97.96%, while its TRP changes from 49.66% to 8.82% with longer TWs. For TRCA+CCValue, TAP rises from 52.80% to 99.64%, and its TRP changes from 47.10% to 17.78%. The underlying reason is that SSCOR/TRCA provides more correct classification results as the TW increases. So SSCOR+CCValue/TRCA+CCValue is more inclined to accept the results of SSCOR/TRCA. It is worth noting that although TRP has dropped, the False Acceptance in TABLE. I generally did not increase due to a decrease in the number of wrong classifications from SSCOR/TRCA. The t-test was used to perform statistical analysis between TAP or TRP of different methods. The result shows that there is no significant difference in almost all data lengths. It indicates that the proposed method has similar effectiveness for both TRCA and SSCOR on datasets of different scales.

The above experiment results were carried out on a DELL laptop with a 1.8GHz quad-core CPU, and 8 GB RAM, using Matlab 2022a and running on Windows 10. The averaged recognition time per time window for performing the TRCA+CCValue and TRCA is 0.0057 s and 0.0020 s on Dataset I, and 0.0043 s and 0.0015 s on Dataset II, respectively. For SSCOR+CCValue and SSCOR, the averaged recognition time is 0.0043 s and 0.0022 s on Dataset I, and 0.0036 s and 0.0014 s on Dataset II. The proposed method incorporates the classification confidence estimation based on the basic method, so the detection time increases slightly, but it is still around an acceptable value.

B. The Influence of Parameters

1) The Number of Channels: Fig. 6 shows the average classification reliability rate of four methods with different numbers of electrodes using 0.6 s-long data on (a) Dataset I and 0.75 s-long data on (b) Dataset II. The number of training blocks is set to five for Dataset I and four for Dataset II, respectively. Generally, the performance of each method improved as the number of electrodes increased. For N_c = 5, 6, 7, 8, and 9, it is obvious that the proposed SSCOR+CCValue always outperforms SSCOR. TRCA+CCValue shows higher recognition reliability compared with TRCA. Meanwhile, these four methods all achieve better performance than CCA and Msetcca. A one-way repeated-measures ANOVA showed significant differences between the six methods at each TW on two datasets. The results in Fig. 6 demonstrate that, to some extent, our method is superior to some existing advanced methods, irrespective of the number of electrodes. Specifically, SSCOR+CCValue improved SSCOR by 9.31% $\sim$ 12.32% and TRCA+CCValue increased TRCA by 8.12% $\sim$ 12.65% in Dataset I. The classification reliability of SSCOR+CCValue is greater than that of SSCOR by 0.45% $\sim$ 7.27% in Dataset II. Similarly, TRCA+CCValue improved TRCA by 0.76% $\sim$ 15.00%.

2) The Number of Training Blocks: It is also investigated how the number of training blocks affects the classification reliability of six different methods. The heat map is a valuable



(a) Dataset I

(b) Dataset II

Fig. 7. Heat maps of the classification reliability of four methods under different number of training blocks on (a) Dataset I and (b) Dataset II.

data visualization tool for displaying an indicator in color in two dimensions. It offers a method for understanding numerical numbers visually. The heat maps in Fig. 7 show the reliability comparison between SSCOR and SSCOR+CCValue, as well as between TRCA and TRCA+CCValue, on two datasets using 0.6 s or 0.75 s data length. For the sake of comparison, the performance of two other algorithms, namely CCA and Msetcca, were also included. In a heat map, the x-axis indicates recognition methods with varying numbers of training blocks, and the y-axis represents the index of the subject. The range of the number of training blocks was $[3, 5]$ for Dataset I and $[2, 4]$ for Dataset II. The shade of color indicates the level of classification reliability. The darkest color is always displayed at its maximum value. As demonstrated in Fig. 7, with varying numbers of training blocks, the color squares generated by SSCOR+CCValue and TRCA+CCValue are generally more profound than those created by SSCOR and TRCA, and notably darker than those generated by CCA and Msetcca. This indicates that the proposed method produces more reliable and consistent results compared to the other methods. Furthermore, the color squares generally get darker as the number of training blocks increases.

Table. III shows the numerical classification reliability of SSCOR and SSCOR+CCValue and the corresponding paired t-test analysis results. Similarly, Table. IV shows the outcome of TRCA and TRCA+CCValue. The average classification reliability of SSCOR+CCValue is higher than that of SSCOR by 10.33% across different numbers of training blocks in Dataset I, and by 8.55% in Dataset II. TRCA+CCValue improved TRCA by 4.52% in Dataset I, and by 2.44% in Dataset II. The paired t-test analysis results revealed that a statistically

TABLE III
RELIABILITY COMPARISON BETWEEN SSCOR AND SSCOR+CCVALUE WITH DIFFERENT NUMBERS OF TRAINING BLOCKS

Methods	Reliability with different number of training blocks					
	Dataset I			Dataset II		
	3	*4*	*5*	*2*	*3*	*4*
SSCOR	45.23	51.01	55.00	46.21	60.04	68.94
SSCOR+CCValue	55.36	62.07	64.80	51.94	65.15	71.67
P-value	<0.0001	<0.0001	<0.0001	0.0843	0.1848	0.2255

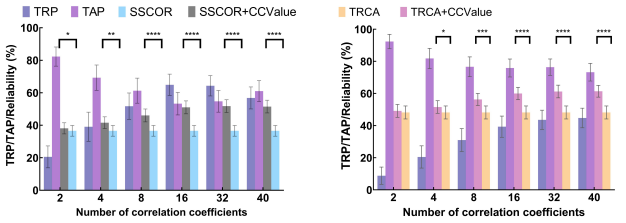
TABLE IV
RELIABILITY COMPARISON BETWEEN TRCA AND TRCA+CCVALUE WITH DIFFERENT NUMBERS OF TRAINING BLOCKS

Methods	Reliability with different number of training blocks					
	Dataset I			Dataset II		
	3	*4*	*5*	*2*	*3*	*4*
TRCA	49.98	57.67	63.17	45.45	67.42	78.18
TRCA+CCValue	57.54	67.64	71.29	47.98	70.83	79.55
P-value	<0.0001	<0.0001	0.0002	0.4825	0.1145	0.1698

significant difference (i.e., $P < 0.0001$) between the compared methods with all numbers of training blocks for Dataset I. For Dataset II, although the significant difference is not as large as for Dataset 1, the proposed method still provides higher recognition reliability than SSCOR and TRCA. In conclusion, the two tables further demonstrate the effectiveness of the proposed method by providing more quantitative evidence.

3) The Number of Correlation Coefficients Incorporated in the Feature Vector: In this study, it is also explored how the number of correlation coefficients used for constructing the feature vector affects the classification performance. The aforementioned performance evaluation figures were all generated by $(N_f - 1)$-dimensional feature vectors. It implies that the feature vector was constructed using N_f correlation coefficients via (3). In this subsection, a 40-class benchmark dataset was used to evaluate more types of coefficient numbers. The correlation coefficients were sorted in descending order. The top two, four, eight, sixteen, thirty-two, or forty values were chosen to construct the feature vector via (3). Here, the number of electrodes and training trials are set to be nine and five, respectively. Fig. 8(a) shows TRP, TAP, and classification reliability of SSCOR+CCValue for various numbers of correlation coefficients. The reliability of SSCOR (blue bars) was also incorporated into the performance comparison. The number of correlation coefficients does not affect the performance of SSCOR. Hence, the corresponding reliability remains constant (i.e., SSCOR: 36.58%). Similarly, Fig. 8(b) shows the evaluation results of TRCA and TRCA+CCValue. The reliability of TRCA is represented by the orange bars at 48.18%.

With an increasing number of correlation coefficients, TRP generally climbs fast and then lowers slightly. TAP gradually decreased and then increased. TRP and TAP are both critical indicators for a classification confidence evaluation model. TRP indicates the model's rejection effectiveness, whereas



(a) SSCOR and SSCOR+CCValue (b) TRCA and TRCA+CCValue

Fig. 8. Barchart of the TRP, TAP and classification reliability of SSCOR+CCValue and TRCA+CCValue with different numbers of correlation coefficients. The error bars represent SEM. The asterisks indicate significant differences between methods obtained by paired t-test analysis. The reliability of SSCOR and TRCA were used as a comparison.

TAP relates to the cost. As a result, it is preferable to keep them both at a relatively high level. TAPs have achieved the largest values for two coefficients in Fig. 8(a) and Fig. 8(b), but TRPs reached relatively low values. Therefore, two is not an ideal number of coefficients for this dataset. It is worth mentioning that TRP and TAP for forty coefficients were both within a satisfactory range, and their difference was not as large as that of other coefficients. Moreover, the reliability also reached superior values for forty coefficients (i.e., TRCA+CCValue: 61.39% and SSCOR+CCValue: 51.48% with 0.4 s-long data). The TRCA+CCValue and SSCOR+CCValue provide consistently higher reliability than TRCA and SSCOR, regardless of the number of correlation coefficients. TRCA+CCValue improves TRCA by 3.84% $\sim$ 15.84%, and SSCOR+CCValue increases SSCOR by 1.42% $\sim$ 12.71%. Moreover, paired t-test analysis showed that statistical differences between the compared algorithms become more significant as the number of coefficients increases.

IV. DISCUSSION

A. Performance of SSCOR+CCValue and TRCA+CCValue

Almost all existing advanced SSVEP recognition methods determine the signal triggered by which stimulus via the largest correlation coefficient, such as CCA [18], MsetCCA [21], SSCOR [24], and TRCA [20]. It can easily lead to erroneous results when the maximum coefficient is slightly larger than the other values. In this study, a classification confidence estimation method based on Bayesian theory was proposed to improve the SSVEP recognition performance. The feature vector was constructed by differences between the largest coefficient and the remaining values. This kind of design can make full use of all the coefficient information. As a consequence, the proposed method can accept high-confidence classification results while rejecting results with low confidence. As shown in Fig. 4(a), TRCA+CCValue and SSCOR+CCValue obtained the highest reliability of 85.57% and 81.98% for the data length of 1 s. TRCA+CCValue and SSCOR+CCValue both improved the performance of the basic target recognition methods.

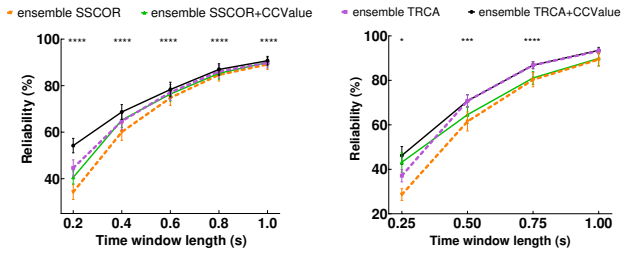
In Fig. 4, the performance of TRCA+CCValue is slightly better than that of TRCA at 1 s TW. Besides, a similar situation is reflected in SSCOR+CCValue and SSCOR. The underlying reason is that long-length signals generally contain more EEG information and are thus more likely to lead to correct classification results. The proposed methods can accept results with high confidence and reject results with low confidence. Therefore, the proposed method accepted more reliable results and achieved reliability comparable to basic methods at 1 s TW.

Although the experiment was conducted in a relatively quiet environment and the subjects were typically requested to avoid movements during signal recording, complete elimination of environmental and body noises is difficult to achieve. Noise is usually present due to a variety of factors, including muscle movements, eye blinks, and external sources such as traffic or other environmental factors. In this study, the fundamental target recognition methods employed for feature extraction are TRCA and SSCOR. These two methods can reduce background EEG activities in different ways [20], [24]. For example, TRCA is a spatial filtering method, in which weight coefficients are optimized to maximize inter-trial covariance of brain activities. It can be used for removing background EEG activities from scalp recordings [20]. TRCA+CCValue and SSCOR+CCValue can improve classification performance via confidence estimation and take advantage of TRCA and SSCOR to decrease background noises. To evaluate the performance of the proposed method, CCA and Msetcca were used in this study for extensive comparison, and the numerical results of six methods are shown in Table II. Additional experiments were also conducted to compare the performance of the six methods under various parameters, such as the number of electrodes and the number of training blocks, as shown in Fig. 6 and Fig. 7. The evaluation results indicate that the proposed method provides better recognition performance than the other four methods across a range of different parameter settings. For example, SSCOR+CCValue improved CCA and Msetcca by 40.00% $\sim$ 47.10% and 31.17% $\sim$ 34.48% with the different number of channels. For different number of training blocks, TRCA+CCValue increased them by 39.88% $\sim$ 53.59% and 31.34% $\sim$ 40.84%, respectively.

In the presented method, leave-one-block-out cross-validation was performed in the experiments. The detailed process was shown in Fig. 3. Cross-validation is a widely used technique in machine learning and statistical modelling to estimate the performance of a model and prevent over-fitting. Cross-validation provides an accurate evaluation of the performance of the proposed method because it uses all the available data for both training and testing. Therefore, it helps improve the reliability and generalization of the experimental results.

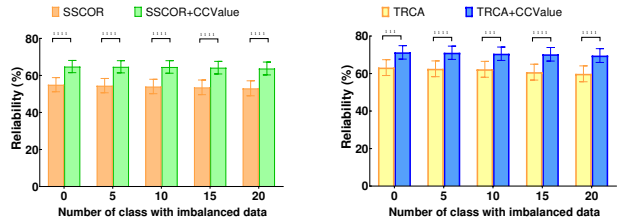
B. Ensemble-based methods comparison

In the previous sections, the effectiveness and superiority of the proposed method were demonstrated by comparing TRCA+CCValue and SSCOR+CCValue with the basic target recognition methods. In this part, the performance comparison of ensemble-based methods was carried out. Specifically, the target recognition method was enhanced by utilizing an ensemble approach, in which N_f spatial filters were concatenated to



(a) Dataset I (b) Dataset II

Fig. 9. Comparison of average recognition reliability among ensemble methods on (a) Dataset I and (b) Dataset II. The asterisks indicate significant differences between the four methods obtained by one-way repeated-measures ANOVA.



(a) SSCOR and SSCOR+CCValue (b) TRCA and TRCA+CCValue

Fig. 10. Barchart of the classification reliability of SSCOR, TRCA, SSCOR+CCValue and TRCA+CCValue with different numbers of classes considering imbalanced data. The error bars represent SEM. The asterisks indicate significant differences between methods obtained by paired t-test analysis.

create an ensemble spatial filter $\boldsymbol{W} \in \mathbb{R}^{N_c \times N_f}$:

$$\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_1, ..., \boldsymbol{w}_{N_f}] \tag{17}$$

The correlation coefficient in (1) can be re-defined as follows:

$$r_i = \rho(\widetilde{\boldsymbol{X}}^{\mathsf{T}}\boldsymbol{W}, \overline{\boldsymbol{X}}_i^{\mathsf{T}}\boldsymbol{W}),\ i = 1, 2, ..., N_f \tag{18}$$

The feature extraction, classification confidence evaluation, and decision-making steps are the same as described in the previous section. Fig. 9 shows the classification reliability comparison between several ensemble-based methods on (a) Dataset I and (b) Dataset II. As shown by the black line and the purple dotted line, the ensemble TRCA+CCValue achieved higher reliability than the ensemble TRCA, with almost TWs on two datasets. The ensemble SSCOR+CCValue also exhibits a superior performance than SSCOR at all TWs. For example, ensemble TRCA+CCValue improved ensemble TRCA by 3.96%, and ensemble SSCOR+CCValue increased ensemble SSCOR by 5.13% with 0.4s data length in Dataset I. Similarly, the classification reliability of SSCOR+CCValue is greater than that of SSCOR by 14.55%, and TRCA+CCValue improved TRCA by 9.09% with 0.25s data length in Dataset II. A one-way repeated measures ANOVA revealed a statistically significant difference between the compared methods with various data lengths. As a result, the proposed method can improve the performance of both basic and ensemble-based SSVEP detection methods.

C. Feature Vector Construction

Recently, some studies have also focused on estimating classification confidence based on correlation coefficients for SSVEP-based BCI. These works usually use the largest and the second-largest values or their difference, such as [27]–[29]. In this study, N_f correlation coefficients were incorporated, and then a $(N_f - 1)$-dimensional feature vector was formed by calculating the differences between the maximum value and the other values. The higher-dimensional features are beneficial to improving SSVEP detection, which was confirmed in Fig. 8. The N_f of the benchmark dataset is forty. Compared with other numbers of correlation coefficients, TAP, TRP, and classification reliability generated by the feature vector with forty correlation coefficients achieve high values. For example, TRCA+CCValue reached the highest reliability of 61.39%, and SSCOR+CCValue reached the reliability of 51.49% (highest value: 51.79% with thirty-two correlation coefficients) with 0.4 s TW. For those cases with similar reliability, the gap between TAP and TRP provided by the proposed method is relatively smaller. For example, the gap is 4.19% for forty coefficients but 9.56% for thirty-two coefficients for SSCOR+CCValue. Therefore, it indicates that the proposed method can achieve high classification reliability while maintaining a better balance between the model's rejection efficiency (TRP) and the cost (TAP).

D. Data Imbalance

Data imbalance is a common issue in real-world datasets, and it occurs when the distribution of classes in a dataset is uneven. Therefore, it is important to evaluate the effectiveness of the proposed method on unbalanced datasets to further validate its reliability in real SSVEP-based BCI systems. In Fig. 10, the classification reliability of four methods is shown under different numbers of classes with imbalanced data. For instance, when the x-axis is five, it means that five classes are randomly selected with insufficient training data (i.e., four training blocks), while the other thirty classes have sufficient training data (i.e., five training blocks). The evaluation results indicate that the proposed method achieves consistently better performance. The paired t-test was used to perform statistical analysis of the recognition performance of different methods. Statistical analysis shows that the reliability of SSCOR is significantly different from that of SSCOR+CCValue regardless of the number of imbalanced classes. The same conclusion applies to TRCA and TRCA+CCValue. In addition, the performance of the proposed method does not show much difference between datasets with many imbalanced classes and those without any imbalanced classes. For example, the recognition reliability of TRCA+CCValue and SSCOR+CCValue are 69.59% and 63.76% when tested on a dataset with twenty imbalanced classes, while on a dataset with zero imbalanced class, the recognition reliability of TRCA+CCValue is 71.29% and SSCOR+CCValue is 64.80%. This suggests that TRCA+CCValue and SSCOR+CCValue are robust to the number of imbalanced classes in the dataset, indicating their potential for handling imbalanced datasets in practical situations. Additionally, Fig. 7, TABLE. III, and TABLE. IV in Section III-B show the experimental evaluation results after balancing the dataset. It involves adjusting the class distribution so that each class has an equal number of examples.

E. Future Work

Although the proposed method enhanced the recognition performance of some popular methods in the SSVEP-based BCI field, it still has some potential directions for further improvement. First, the presented method focused on the fixed data length which means that the same amount of data was collected and analyzed for every trial. This fixed data length approach may not be optimal because it could include redundant data. To address this limitation, future work could aim to develop an adaptive time segment approach that can dynamically adjust the length of data collected for each trial. It has the potential to improve system performance in practical BCI applications. Furthermore, due to individual differences commonly observed in BCI systems, it can be challenging to achieve satisfactory classification results when transmitting data directly between individuals. Therefore, future work could focus on incorporating transfer learning techniques to improve the reusability and generalization of models [44]. By leveraging pre-trained models and knowledge from the source domain, transfer learning has the potential to enhance the performance of the BCI system, even with limited training data.

V. Conclusion

In this study, a Bayesian-based classification confidence estimation method was proposed for enhancing the SSVEP recognition performance. The differences between the largest correlation coefficient and the other values were used to define the feature vector. The probability density functions of feature vectors given correct and wrong classifications were then estimated using the GMM model. In the test process, the posterior probabilities of an accurate and wrong recognition can be calculated using Bayesian inference with the newly obtained feature vector. The CCValue, the difference between two posterior probabilities, was applied to evaluate the confidence of the classification result. Eventually, the decision-making process can determine whether to accept trustworthy results or reject unconfident results. Our method was evaluated on a publicly available benchmark dataset and a self-collected dataset. The experimental results demonstrated the effectiveness and feasibility of the proposed method in the SSVEP-based BCIs.

Author Contributions: Conceptualization, Z.Q.Z., and Y.Z.; Methodology, Z.Q.Z., and Y.Z.; Validation, Y.Z., and C.Y.S.; Writing-original draft preparation, Z.Q.Z., Y.Z., and C.Y.S.; Writing-review and editing, Z.Q.Z., Y.Z., and H.W.; Visualization, Z.Q.Z., Y.Z., and H.W.; Supervision, Z.Q.Z., and S.Q.X.; Funding acquisition, Z.Q.Z., and S.Q.X. All authors have read and agreed to the published version of the manuscript.

Data Access Statement: The data presented in this study are available from the corresponding author upon request.

References

[1] X. Yu, M. Z. Aziz, M. T. Sadiq, Z. Fan, and G. Xiao, "A new framework for automatic detection of motor and mental imagery EEG signals for robust BCI systems," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[2] J. Jin, H. Sun, I. Daly, S. Li, C. Liu, X. Wang, and A. Cichocki, "A novel classification framework using the graph representations of electroencephalogram for motor imagery based brain-computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 20–29, 2021.

[3] J. Huang, P. Yang, B. Xiong, B. Wan, K. Su, and Z.-Q. Zhang, "Latency aligning task-related component analysis using wave propagation for enhancing SSVEP-based BCIs," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 851–859, 2022.

[4] Y. Zhou, Z. Xu, Y. Niu, P. Wang, X. Wen, X. Wu, and D. Zhang, "Cross-task cognitive workload recognition based on EEG and domain adaptation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2022.

[5] Y.-C. Jiang, R. Ma, S. Qi, S. Ge, Z. Sun, Y. Li, J. Song, and M. Zhang, "Characterization of bimanual cyclical tasks from single-trial EEG-fNIRS measurements," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 146–156, 2022.

[6] Y. Zhang, S. Q. Xie, H. Wang, and Z. Zhang, "Data analytics in steady-state visual evoked potential-based brain–computer interface: A review," *IEEE Sensors J.*, vol. 21, no. 2, pp. 1124–1138, 2020.

[7] J. Jin, Z. Chen, R. Xu, Y. Miao, X. Wang, and T.-P. Jung, "Developing a novel tactile P300 brain-computer interface with a cheeks-stim paradigm," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 9, pp. 2585–2593, 2020.

[8] X. Deng, Z. L. Yu, C. Lin, Z. Gu, and Y. Li, "A bayesian shared control approach for wheelchair robot with brain machine interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2019.

[9] L. Angrisani, P. Arpaia, A. Esposito, and N. Moccaldi, "A wearable brain–computer interface instrument for augmented reality-based inspection in industry 4.0," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1530–1539, 2019.

[10] P. Arpaia, L. Duraccio, N. Moccaldi, and S. Rossi, "Wearable brain–computer interface instrumentation for robot-based rehabilitation by augmented reality," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 9, pp. 6362–6371, 2020.

[11] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "EEG-based emotion recognition in music listening," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1798–1806, 2010.

[12] P. P. Roy, P. Kumar, and V. Chang, "A hybrid classifier combination for home automation using EEG signals," *Neural Comput. Appl.*, vol. 32, no. 20, pp. 16 135–16 147, 2020.

[13] L. Angrisani, P. Arpaia, D. Casinelli, and N. Moccaldi, "A single-channel SSVEP-based instrument with off-the-shelf components for trainingless brain-computer interfaces," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 10, pp. 3616–3625, 2018.

[14] A. Liu, Q. Liu, X. Zhang, X. Chen, and X. Chen, "Muscle artifact removal towards mobile SSVEP-based BCI: A comparative study," *IEEE Trans. Instrum. Meas.*, 2021.

[15] Y. Chen, C. Yang, X. Ye, X. Chen, Y. Wang, and X. Gao, "Implementing a calibration-free SSVEP-based BCI system with 160 targets," *J. Neural Eng.*, vol. 18, no. 4, p. 046094, 2021.

[16] J. Huang, P. Yang, B. Xiong, Q. Wang, B. Wan, Z. Ruan, K. Yang, and Z.-Q. Zhang, "Incorporating neighboring stimuli data for enhanced SSVEP-Based BCIs," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022.

[17] P. Gaur, H. Gupta, A. Chowdhury, K. McCreadie, R. B. Pachori, and H. Wang, "A sliding window common spatial pattern for enhancing motor imagery classification in EEG-BCI," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[18] M. Nakanishi, Y. Wang, Y.-T. Wang, and T.-P. Jung, "A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials," *PloS one*, vol. 10, no. 10, 2015.

[19] T. Tanaka and M. Arvaneh, *Signal processing and machine learning for brain-machine interfaces*. Institution of Engineering & Technology, 2018.

[20] M. Nakanishi, Y. Wang, X. Chen, Y.-T. Wang, X. Gao, and T.-P. Jung, "Enhancing detection of SSVEPs for a high-speed brain speller using task-related component analysis," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 1, pp. 104–112, 2017.

[21] Y. Zhang, G. Zhou, J. Jin, X. Wang, and A. Cichocki, "Frequency recognition in SSVEP-based BCI using multiset canonical correlation analysis," *Int. J. Neural Syst.*, vol. 24, no. 04, p. 1450013, 2014.

[22] Y. Jiao, Y. Zhang, Y. Wang, B. Wang, J. Jin, and X. Wang, "A novel multilayer correlation maximization model for improving CCA-based frequency recognition in SSVEP brain–computer interface," *Int. J. Neural Syst.*, vol. 28, no. 04, p. 1750039, 2018.

[23] J. Jin, Z. Wang, R. Xu, C. Liu, X. Wang, and A. Cichocki, "Robust similarity measurement based on a novel time filter for SSVEPs detection," *IEEE Trans. Neural Netw. Learn. Syst.*, 2021.

[24] K. K. GR and R. Reddy, "Designing a sum of squared correlations framework for enhancing SSVEP-based BCIs," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 2044–2050, 2019.
[25] T. Bao, S. A. R. Zaidi, S. Q. Xie, P. Yang, and Z.-Q. Zhang, "CNN confidence estimation for rejection-based hand gesture classification in myoelectric control," *IEEE T. Hum-Mach. Syst.*, vol. 52, no. 1, pp. 99–109, 2021.
[26] J. Zhao, W. Zhang, J. H. Wang, W. Li, C. Lei, G. Chen, Z. Liang, and X. Li, "Decision-making selector (DMS) for integrating CCA-based methods to improve performance of SSVEP-based BCIs," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1128–1137, 2020.
[27] Y. Chen, C. Yang, X. Chen, Y. Wang, and X. Gao, "A novel training-free recognition method for SSVEP-based BCIs using dynamic window strategy," *J. Neural Eng.*, vol. 18, no. 3, p. 036007, 2021.
[28] H. Cecotti, "Adaptive time segment analysis for steady-state visual evoked potential based brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 3, pp. 552–560, 2020.
[29] J. Jiang, E. Yin, C. Wang, M. Xu, and D. Ming, "Incorporation of dynamic stopping strategy into the high-speed SSVEP-based BCIs," *J. Neural Eng.*, vol. 15, no. 4, p. 046025, 2018.
[30] P. Yuan, X. Chen, Y. Wang, X. Gao, and S. Gao, "Enhancing performances of SSVEP-based brain–computer interfaces via exploiting inter-subject information," *J. Neural Eng.*, vol. 12, no. 4, p. 046006, 2015.
[31] J. Zhao, W. Li, and M. Li, "Comparative study of SSVEP-and P300-based models for the telepresence control of humanoid robots," *PLoS One*, vol. 10, no. 11, p. e0142168, 2015.
[32] M. Wang, R. Li, R. Zhang, G. Li, and D. Zhang, "A wearable SSVEP-based BCI system for quadcopter control using head-mounted device," *IEEE Access*, vol. 6, pp. 26 789–26 798, 2018.
[33] Y. Wang, X. Chen, X. Gao, and S. Gao, "A benchmark dataset for SSVEP-based brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1746–1752, 2016.
[34] T. Tsoneva, G. Garcia-Molina, and P. Desain, "SSVEP phase synchronies and propagation during repetitive visual stimulation at high frequencies," *Sci. Rep-UK.*, vol. 11, no. 1, pp. 1–13, 2021.
[35] N. Zhuang, L. Jiang, B. Yan, L. Tong, J. Shu, K. Yang, D. Yao, P. Xu, and Y. Zeng, "Neural mechanism of affective perception: Evidence from phase and causality analysis in the cerebral cortex," *Neuroscience*, vol. 461, pp. 44–56, 2021.
[36] R. Zhang, Z. Xu, L. Zhang, L. Cao, Y. Hu, B. Lu, L. Shi, D. Yao, and X. Zhao, "The effect of stimulus number on the recognition accuracy and information transfer rate of SSVEP-BCI in augmented reality," *J. Neural Eng.*, 2022.
[37] G. Zhang, Y. Cui, Y. Zhang, H. Cao, G. Zhou, H. Shu, D. Yao, Y. Xia, K. Chen, and D. Guo, "Computational exploration of dynamic mechanisms of steady state visual evoked potentials at the whole brain level," *NeuroImage*, vol. 237, p. 118166, 2021.
[38] Y. Zhang, Z. Li, S. Q. Xie, H. Wang, Z. Yu, and Z.-Q. Zhang, "Multi-objective optimisation-based high-pass spatial filtering for SSVEP-based brain-computer interfaces," *IEEE Trans. Instrum. Meas.*, 2022.
[39] Y. Yang, W. Wu, B. Wang, and M. Li, "Analytical reformulation for stochastic unit commitment considering wind power uncertainty with gaussian mixture model," *IEEE Trans. Power Syst.*, vol. 35, no. 4, pp. 2769–2782, 2019.
[40] G. Xuan, W. Zhang, and P. Chai, "EM algorithms of Gaussian mixture model and hidden Markov model," in *Proc. Int. Conf. Image. Process.*, vol. 1. IEEE, 2001, pp. 145–148.
[41] X.-Y. Zhou, J. S. Lim, I. Kwon *et al.*, "EM algorithm with GMM and naive Bayesian to implement missing values," *Adv. Sci. Technol. Lett.*, vol. 46, pp. 1–5, 2014.
[42] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating trust prediction and confusion matrix measures for web services ranking," *IEEE Access*, vol. 8, pp. 90 847–90 861, 2020.
[43] J. Taylor, *Introduction to error analysis, the study of uncertainties in physical measurements*. University Science Books, 1997.
[44] Y. Zhang, S. Q. Xie, C. Shi, J. Li, and Z.-Q. Zhang, "Cross-subject transfer learning for boosting recognition performance in SSVEP-Based BCIs," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1574–1583, 2023.