Eliminating or Shortening the Calibration for a P300 Brain–Computer Interface Based on a Convolutional Neural Network and Big Electroencephalography Data: An Online Study

Wei Gao, Weichen Huang, Man Li, Zhenghui Gu, *Member, IEEE*, Jiahui Pan, *Member, IEEE*, Tianyou Yu, *Member, IEEE*, Zhu Liang Yu, *Member, IEEE*, and Yuanqing Li, *Fellow, IEEE*

***Abstract*— A brain-computer interface (BCI) measures and analyzes brain activity and converts it into computer commands to control external devices. Traditional BCIs usually require full calibration, which is time-consuming and makes BCI systems inconvenient to use. In this study, we propose an online P300 BCI spelling system with zero or shortened calibration based on a convolutional neural network (CNN) and big electroencephalography (EEG) data. Specifically, three methods are proposed to train CNNs for the online detection of P300 potentials: (i) training a subject-independent CNN with data collected from 150 subjects; (ii) adapting the CNN online via a semisupervised learning/self-training method based on unlabeled data collected during the user's online operation; and (iii) fine-tuning the CNN with a transfer learning method based on a small quantity of labeled data collected before the user's online operation. Note that the calibration process is eliminated in the first two methods and dramatically shortened in the third method. Based on these methods, an online P300 spelling system is developed. Twenty subjects participated in our online experiments. Average accuracies of 89.38%, 94.00% and 93.50% were obtained by the subject-independent CNN, the self-training-based CNN and the transfer learning-based CNN, respectively. These results demonstrate the effectiveness of our methods, and thus, the convenience of the online P300-based BCI system is substantially improved.**

***Index Terms*— Brain–computer interface (BCI), electroencephalography (EEG), P300, convolutional neural network (CNN), transfer learning, semisupervised learning/self-training, big data.**

I. Introduction

A BRAIN-COMPUTER interface (BCI) provides a direct human-machine interaction pathway between the brain and external devices without relying on the peripheral nervous system and muscles [1]. It acquires brain signals and translates them into computer commands to control external devices. Electroencephalography (EEG)-based BCIs are some of the most commonly used BCIs. They mainly include P300-based BCIs, steady-state visual evoked potential (SSVEP)-based BCIs, and motor imagery (MI)-based BCIs. In this study, we mainly focus on P300-based BCIs.

A BCI usually requires a subject-specific calibration phase, during which the user is required to perform a specific task while labeled EEG data are recorded for training a subject-specific EEG decoding model. However, the calibration phase is generally tedious and time-consuming, making BCIs inconvenient to use. Some attempts have been made to completely eliminate the calibration phase and build BCIs with instant operation. Such BCIs are usually called zero-calibration/training BCIs or calibration-free BCIs. To build a zero-calibration BCI, a natural idea is to employ a subject-independent model for EEG decoding. Researchers have conducted various offline studies to build subject-independent P300 detection models. These models are usually obtained using two approaches, i.e., the pooled approach and the ensemble approach [2]. The pooled approach involves training a model such as a convolutional neural network (CNN) [3], [4], [5], [6] or a hierarchical recurrent network [7] on a pool of data derived from multiple subjects to extract invariant patterns across the subjects and then using the

Manuscript received 30 December 2022; revised 26 February 2023; accepted 17 March 2023. Date of publication 21 March 2023; date of current version 28 March 2023. This work was supported in part by the STI 2030—Major Projects under Grant 2022ZD0208900; in part by the Key Realm Research and Development Program of Guangzhou, China, under Grant 202007030007; in part by the National Natural Science Foundation of China under Grant 62276102, Grant 61836003, and Grant 62076103; and in part by the Technology Innovation 2030 under Grant 2022ZD0211700. *(Corresponding author: Yuanqing Li.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted the Ethics Committee of Guangzhou First People's Hospital, China, under Application No. S-2022-234.

Wei Gao, Weichen Huang, Man Li, Zhenghui Gu, Tianyou Yu, Zhu Liang Yu, and Yuanqing Li are with the School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China, and also with the Research Center for Brain–Computer Interface, Pazhou Laboratory, Guangzhou 510330, China (e-mail: augaow@mail.scut.edu.cn; huangwch96@gmail.com; aumanli@mail.scut.edu.cn; zhgu@scut.edu.cn; auyuty@scut.edu.cn; zlyu@scut.edu.cn; auyqli@scut.edu.cn).

Jiahui Pan is with the School of Software, South China Normal University, Guangzhou 510631, China, and also with the Research Center for Brain–Computer Interface, Pazhou Laboratory, Guangzhou 510330, China (e-mail: panjiahui@m.scnu.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3259991

obtained model to directly predict for new users. The ensemble approach combines a committee of weak models learned from the EEG data of a pool of subjects or a single subject to create a subject-independent model [8], [9]. Previous studies on building subject-independent models generally achieved accuracies of approximately 60%–90% in offline analyses. In addition to these offline models, in [10], an online zero-calibration P300 spelling system was developed based on a CNN trained with a large dataset containing EEG data from 55 subjects. Another idea for building a zero-calibration BCI is to apply semisupervised learning methods. Researchers first trained a subject-independent model, for example, one based on a support vector machine (SVM) [11] or a linear discriminant analysis (LDA) classifier [12], [13], and then adapted the model based on EEG data recorded from the user and the corresponding labels predicted by the model. In this way, the labeled data for model pretraining are entirely collected from other users, and a subject-specific calibration phase for the user is not needed. However, to our knowledge, such online P300 systems are rarely implemented.

In addition to eliminating the calibration process, other approaches propose shortening the calibration time. When a training set containing EEG signals collected from a pool of subjects is available, researchers typically use this training set along with a small quantity of subject-specific calibration data to build models based on transfer learning methods. For instance, a classifier based on an xDAWN filter [14], a CNN [15], [16], [17], and a reinforcement learning model [18] were previously trained on data acquired from a pool of subjects and then adapted with subject-specific labeled data via incremental training or model fine-tuning. By applying probabilistic frameworks, each parameter of the subject-specific models shared the prior learned from a pool of subjects and was optimized using subject-specific data [19], [20]. Riemannian geometry methods affine transform the covariance matrices of different subjects to center them with respect to a reference covariance matrix and then classify them using minimum-distance-to-mean (MDM) classifiers [21], [22], [23], [24]. In [25], a small quantity of user data was added to the training datasets, each of which contained data from one subject, and the model was trained by an ensemble method. These methods generally obtain accuracies of approximately 75%–90% in offline analyses. For online implementation purposes, several P300 spelling systems [26], [27], [28] and a robot control system [29] based on transfer learning were proposed, and accuracies of approximately 80%–90% were achieved. When a training set containing EEG signals collected from a pool of subjects was unavailable, some other studies trained their models based on a small quantity of subject-specific labeled data as well as unlabeled data recorded during use. These studies were mainly based on semisupervised learning algorithms. For instance, a model was initially trained with a small quantity of subject-specific data and then adapted with unlabeled data [30], [31]. In [32], two models were first trained with a small quantity of labeled calibration data, and then the models taught each other to build a final classifier with unlabeled data using a cotraining algorithm. In [33], the relationship between unlabeled data and labeled data was used to define a penalty term for a regularized discriminant analysis model. Most of these semisupervised learning-based models have achieved accuracies of approximately 80%. In addition to the above offline analyses, Gu et al. pushed the related research to online practices, and accuracies above 85% were achieved [34], [35]. Although existing studies have shown that various methods can build models or develop BCI systems with zero-calibration or shortened calibration processes, such studies are still in their infancy. Most previous studies did not adopt large training datasets, which are more likely to contain individual diversity and provide the possibility to learn invariant brain patterns across subjects. Additionally, most studies only established their models via offline analyses, which require online validations. The performance of the existing online BCI systems with zero-calibration or shortened calibration processes needs further improvement. Therefore, most existing studies can hardly meet the practical requirements of this task.

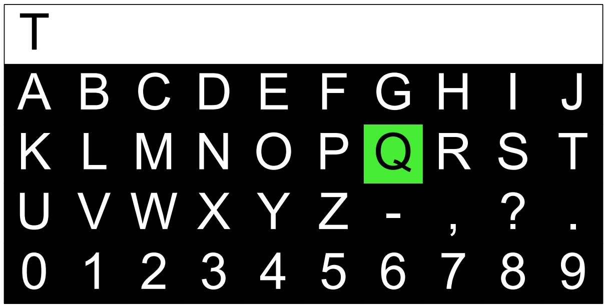
In this study, based on a CNN and big EEG data, an online P300 BCI spelling system with zero-calibration or shortened calibration is developed. Specifically, three methods for training cross-subject P300 detection models are proposed, including (i) training a subject-independent CNN with a dataset containing EEG signals collected from 150 subjects; (ii) adapting the CNN trained in (i) online via a self-training algorithm based on unlabeled data collected during the user's online operation; and (iii) fine-tuning the CNN trained in (i) through a transfer learning method with a small quantity of labeled data, which are collected during a calibration phase before the user's online operation. Based on these methods, an online P300 BCI spelling system is developed. Twenty healthy subjects participated in our online experiments. The experimental results demonstrated that with the help of a CNN and a training dataset collected from a large pool of subjects, an online P300 BCI with zero-calibration or shortened calibration can be established, which will substantially improve the convenience of the use of P300 BCIs.

The remainder of this paper is organized as follows. Section II presents the utilized methods, including those for data acquisition, P300 detection model establishment, and online decision making. The experimental implementation and results are presented in Section III, and a discussion is provided in Section IV. Finally, the conclusion in Section V reviews the approach developed in this paper.

II. Methods

A. Equipment

During the experiment, EEG signals were collected at a sampling rate of 1,000 Hz with a 30-channel EEG cap (LT 37) following the extended 10-20 system and referenced to the right mastoid. A SynAmps2 amplifier (Compumedics, Neuroscan, Inc., Australia) was used to collect EEG signals. All electrode impedances were maintained below 5 kΩ during the experiment.



T

A B C D E F G H I J

K L M N O P 🟩Q R S T

U V W X Y Z - , ? .

0 1 2 3 4 5 6 7 8 9

Fig. 1. GUI of the P300 speller. The buttons flash in green (such as the button "Q" on the virtual keyboard), and the predicted target characters are displayed in the textbox at the top of the GUI (such as the text "T" in the textbox).

B. Subjects

Twenty healthy subjects (14 males and 6 females, aged between 21 and 41 years, average age 25.85 years) participated in all online experiments, which are detailed in Section III-A. The study was approved by the Ethics Committee of Guangzhou First People's Hospital, China. Written informed consent was obtained from all subjects.

C. Graphical User Interface

The graphical user interface (GUI) of the proposed P300 spelling system is shown in Fig. 1. A 4 × 10 button matrix of characters was presented to each subject for stimulus presentation. The paradigm was the same as that employed in our previous study [5]. Specifically, for each trial corresponding to one character input, to prepare the subject, during the 3 s before the stimulus onset, all buttons were not intensified. Upon onset, all buttons started to flash successively in a random order. Each flash lasted for 100 ms, and the interval between the onsets of two successive flashes was 30 ms, which meant that there was an overlap of 70 ms between any pair of successive flashes. Each of the 40 buttons flashed once in each round, and 10 rounds of button flashes formed a trial. No pause occurred between adjacent rounds. Therefore, it took $[(400-1)\times 30+100]\,\mathrm{ms} = 12.07\mathrm{s}$ to complete 400 flashes in a trial.

During each trial, to input a character, the subject was instructed to focus his/her attention on the flashes of the character he/she intended to input (i.e., the target) and to keep a running mental count of the number of flashes.

D. A Subject-Independent CNN Model

In this study, a subject-independent CNN, which was established for an offline analysis in our previous study [5], was applied as one of the three P300 detection models. We briefly review the method for training the CNN model in this section for the sake of the completeness of this paper.

1) Training Set Construction: We applied a large EEG dataset collected in our previous study [5] as a training set. To build this dataset, we recruited 150 subjects (128 males and 22 females between 18 and 32 years of age) in an experiment to collect training data. Each subject performed 60 character input trials. During this phase, the target of each trial was randomly specified by the system rather than freely determined by the subject.

2) Data Preprocessing: The EEG signals were first bandpass filtered at 0.5–10 Hz using a fourth-order Butterworth filter. After that, epochs corresponding to each button flash from 0 to 600 ms after the onset of the stimulus were extracted and then downsampled at a rate of 24. Consequently, in each trial, there were $N_\mathrm{c} \cdot N_\mathrm{r}$ epochs, and in each epoch, there were $1,000\ \mathrm{Hz} \times 600\ \mathrm{ms} \times \frac{1}{24} = 25$ sampling points for each channel. Here, N_c and N_r are the numbers of buttons (40 in this study) and rounds (10 in this study), respectively. Finally, the signals of each epoch were normalized as follows:

$$\tilde{f}_{i,j} = \frac{f_{i,j} - \bar{f}_i}{\sigma_i}, \tag{1}$$

where $f_{i,j}$ and $\tilde{f}_{i,j}$ are the unnormalized and normalized signals of channel i at sampling point j, respectively, and $\bar{f}_i$ and σ_i are the average and standard deviation of the signal of channel i in the epoch, respectively.

After preprocessing, the data of each epoch formed a 30 × 25 matrix denoted as $\mathbf{F}_{n_\mathrm{s},n_\mathrm{t},n_\mathrm{r},n_\mathrm{c}}$, where n_s represents the index of the subject (ranging from 1 to N_s), n_t represents the index of the trial (ranging from 1 to N_t), n_r represents the flash round index (ranging from 1 to N_r), and n_c represents the character index (ranging from 1 to N_c). Herein, $N_\mathrm{r} = 10$, and $N_\mathrm{c} = 40$.

To reduce the influence of the low EEG signal-to-noise ratios (SNRs) and the short interstimulus intervals (ISIs) of the stimuli, we averaged the preprocessed signals corresponding to the first n_r ($n_\mathrm{r} = 1, 2, \ldots, N_\mathrm{r}$) rounds in a trial as follows:

$$\mathbf{X}_{n_\mathrm{s},n_\mathrm{t},n_\mathrm{r},n_\mathrm{c}} = \frac{1}{n_\mathrm{r}} \sum_{m=1}^{n_\mathrm{r}} \mathbf{F}_{n_\mathrm{s},n_\mathrm{t},m,n_\mathrm{c}}. \tag{2}$$

In our online study, only $\mathbf{X}_{n_\mathrm{s},n_\mathrm{t},N_\mathrm{r},n_\mathrm{c}}$ ($n_\mathrm{s} = 1, 2, \ldots, N_\mathrm{s}, n_\mathrm{t} = 1, 2, \ldots, N_\mathrm{t},\ n_\mathrm{c} = 1, 2, \ldots, N_\mathrm{c}$) were used for both model training and online prediction.

A sample $\mathbf{X}_{n_\mathrm{s},n_\mathrm{t},n_\mathrm{r},n_\mathrm{c}}$ was labeled as a positive sample if and only if its corresponding character n_c was the target of the current trial. Otherwise, it was labeled as a negative sample.

3) CNN Architecture: We built a CNN with the architecture shown in Fig. 2 for cross-subject P300 detection. This network architecture is similar to the one used in [36]. It contains three convolutional layers and two fully connected layers. All layers except FC5 use the rectified linear unit (ReLU) function as the activation function, while FC5 uses the logistic sigmoid function as its activation function. The network takes preprocessed data $\mathbf{X}_{n_\mathrm{s},n_\mathrm{t},n_\mathrm{r},n_\mathrm{c}}$ as its inputs, and the output can be regarded as the modeled probability of the presence of a P300 potential $P\left(y = 1 \mid \mathbf{X}_{n_\mathrm{s},n_\mathrm{t},n_\mathrm{r},n_\mathrm{c}}; \mathcal{M}\right)$, where y is the binary label indicating the presence or absence of a P300 potential with values of 1 or 0, respectively, and $\mathcal{M}$ is the model for P300 potential detection.

4) Subject-Independent CNN Model Training: The subject-independent CNN model was the same as the model employed in our previous offline study [5]. It was established by training a CNN with the architecture described in Section II-D.3 offline using the large training set described in Section II-D.1. The convolutional kernels and weights of the network were initialized with the Xavier initialization method [37]. The model

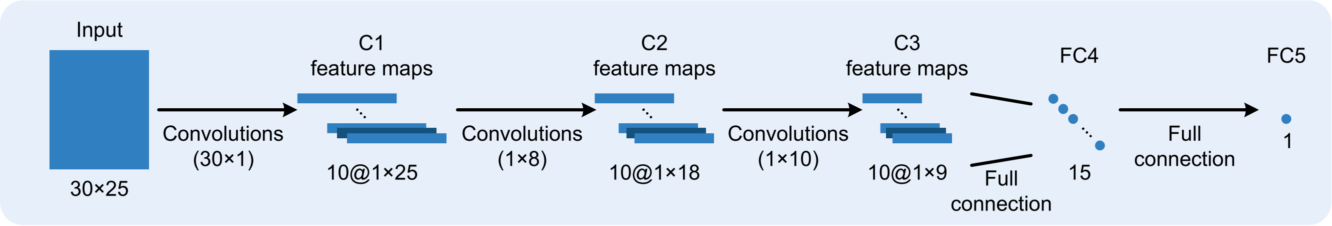


Fig. 2. The architecture of the CNN used for cross-subject P300 detection.

was trained using adaptive moment estimation (Adam) [38] to optimize the mean-squared error (MSE). Since the ratio of positive and negative samples in the training set was 1 : 39, the loss function was weighted by multiplying the positive samples by 39. The model was trained on an NVIDIA GeForce GTX 1080 Ti GPU with CUDA 9.0 and cuDNN v7 using TensorFlow [39].

5) Online Decision Making: In this study, the subject-independent CNN was employed online as a P300 detection model for the proposed system. Specifically, in each trial, once the system stopped the stimulus presentation process, each preprocessed signal segment was input into the model, whose output was regarded as the probability of the presence of a P300 potential $P\left(y=1 \mid \mathbf{X}_{n_s,n_t,n_r,n_c}; \mathcal{M}\right)$. The system output the character with the maximum probability of P300 potential presence as the predicted target, i.e.,

$$\begin{aligned}&\hat{c}_{\mathrm{T}}(n_s, n_t, n_r; \mathcal{M})\\&\quad = \underset{n_c \in \{1,2,\ldots,N_c\}}{\operatorname{argmax}} \; P\left(y=1 \mid \mathbf{X}_{n_s,n_t,n_r,n_c}; \mathcal{M}\right).\end{aligned} \quad (3)$$

With the subject-independent model, users operated the system instantly without subject-specific calibration.

E. A Subject-Specific CNN Model Adapted Online by Self-Training

In the following, we propose a semisupervised learning/self-training method to adapt the CNN model online and improve its performance. Specifically, the user operated the BCI at the beginning without calibration, and the subject-independent CNN was employed as the P300 detection model. After 10 character input trials, the model was automatically adapted online based on the subject-independent model and the data derived from the 10 trials by using the self-training algorithm presented in Algorithm 1. In the next 10 trials, the updated model was employed instead of the subject-independent model for P300 detection and target character identification. Then, the model was adapted online once again based on data recorded in trials 11–20 using the self-training method, and the obtained model was used in the remaining trials.

F. A Subject-Specific CNN Model Fine-Tuned by Transfer Learning

We further propose a transfer learning method to adapt the CNN model and improve its performance. Specifically, before the online operation, the user performed a calibration task containing five character input trials. During the calibration process, the target character for each trial was cued by the computer. The subject-independent CNN was fine-tuned using the calibration data with labels. The fine-tuned CNN was used for online prediction. As described in Section II-C, in each trial, 12.07 s of stimulus presentation was employed. Therefore, it took approximately 1 min to perform the calibration task for each user, which is much shorter than the full calibration process.

Algorithm 1 Adapting the CNN Based on a Semisupervised Learning/self-Training Method

```
Input: The data obtained online during $N$ trials ($N = 10$ in
    this study), and the CNN model currently used for P300
    detection.
Output: An updated CNN model for P300 detection.
repeat
    Apply the CNN model to the data from $N$ trials. For
    each trial, we obtain a predicted label as well as a
    probability showing confidence of the prediction.
    Select the $2n$ ($n$ is the index of the current iteration)
    trials with the largest probabilities.
    Retrain the CNN model using the data from the selected
    $2n$ trials with the predicted labels.
until The maximum number of iterations (5 in this study)
    is reached.
```

III. Experiments and Results

A. Experiments

Twenty subjects participated in three online experiments. The order of the experiments was random for each subject. Experiments I, II and III correspond to spelling tasks in which the subject-independent CNN, the self-training-based CNN and the transfer learning-based CNN, respectively, were employed.

Experiment I: An online test was conducted for the system with the subject-independent model. Specifically, each subject performed a spelling task involving the spelling of the following 40 characters: “THE FIVE BOXING WIZARDS JUMP QUICKLY. - 510641?”.

Experiment II: An online test was conducted for the system with the self-training-based model. Each subject spelled the same characters as those in Experiment I. The experiment containing 40 character spelling trials was divided into three stages. The first stage containing trials 1–10 employed the subject-independent model, while the second stage containing trials 11–20 and the third stage containing trials 21–40 respectively employed the models adapted once (using the data from trials 1–10) and twice (first using the data from trials 1–10 and then using the data from trials 11–20). We calculated the performance achieved for each stage, and the performance attained during the last stage was regarded as the performance of the self-training-based model.

Experiment III: An online test was conducted for the system with the transfer learning-based model. Specifically, each subject performed a calibration task involving the spelling of five characters cued by the computer. The model was fine-tuned with the data recorded during the calibration process and was then employed for online decision making. After that, each subject spelled the same characters as those in Experiment I.

B. Results of the Online Experiments

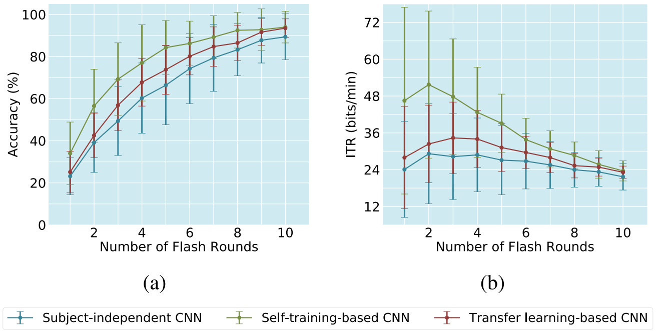
In this study, accuracy, defined as the ratio of the number of correctly spelled characters to the total number of spelled characters, was adopted as a performance metric. Moreover, the information transfer rate (ITR) was also applied to evaluate the ability of the system to balance accuracy and spelling speed. The ITR is defined by

$$\mathrm{ITR} = \frac{1}{T}\left(\log_2 N_c + a\log_2 a + (1-a)\log_2\left(\frac{1-a}{N_c-1}\right)\right), \tag{4}$$

where a is the accuracy of target character prediction, N_c is the number of characters in the GUI, and T is the time needed to spell one character. Herein, $N_c = 40$, and $T = \frac{1}{60}\,(1.2\ N_r + 0.07)$ min.

The results of online Experiments I–III are presented in Table I. As shown in the table, with the subject-independent CNN, the self-training-based CNN and the transfer learning-based CNN, average accuracies of 89.38%, 94.00% and 93.50% were achieved, respectively. These results demonstrated that with the subject-independent CNN, the system was able to achieve satisfactory performance. The performance was further improved when the self-training or transfer learning method was applied.

It is worth noting that the results of Experiment II in Table I were obtained from the last 20 online trials, where the updated CNN model was applied. To explore the difference between the performance achieved before and after the online adaptation process based on self-training, we present the average accuracies obtained across all subjects in the three stages of Experiment II, as shown in Table II. Note that the results of trials 1–10, trials 11–20, and trials 21–40 were obtained with the subject-independent CNN model, the updated CNN model based on the data of trials 1–10, and the updated CNN model based on the data of trials 1–20, respectively. It follows from Table II that the average accuracies increased gradually. With an online adaptation based on the unlabeled data collected from 20 character input trials, the system performed significantly better in trials 21–40 than in trials 1–10 ($p = 0.034$), with the average accuracy improved from 87.00% to 94.00%.



(a)

(b)

Fig. 3. Average accuracies and ITRs with standard deviations across 20 subjects with respect to the number of flash rounds.

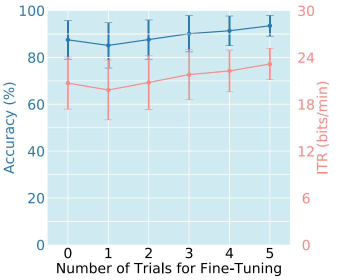


Fig. 4. Average accuracy and ITR with standard deviations across 20 subjects with respect to the number of trials used to fine-tune the CNN.

C. Results of the Offline Analyses

1) The Change in Performance With Respect to the Number of Flash Rounds: In the online experiments, the number of flash rounds N_r for each trial was 10. In order to explore the relationship between N_r and the accuracy as well as the ITR, we conducted an offline test on the changes in the accuracy and the ITR with respect to N_r. As shown in Fig. 3, the average accuracy monotonically increased as the number of flash rounds increased for all models. However, the average ITRs increased at first, reached maximum values at approximately 2–3 rounds, and then gradually decreased. The best average ITR was 51.72 bits/min, achieved at 2 rounds of button flashes when the self-training-based CNN was applied.

2) The Performance of the Transfer Learning-Based Models With Respect to the Number of Calibration Trials: In Experiment III, for each subject, the CNN model was fine-tuned using five trials of calibration data and consequently outperformed the subject-independent CNN validated in Experiment I. We further conducted an offline analysis to explore the relationship between the model performance and the quantity of calibration data used to fine-tune the CNN model. Specifically, for each subject, the CNN model was fine-tuned based on the subject-independent CNN, with the number of calibration trials varying from one to five, and the fine-tuned models were validated with data collected in Experiment III. The average accuracy and ITR are shown in Fig. 4. As seen in the figure, as the number of trials used to fine-tune the CNN increases from one to five, both the average accuracy and the ITR increase gradually.

3) Results of an Offline Analysis Conducted on a Dynamic Stopping Strategy: The above experiments were all based on a system with a consistent number of flash rounds for all trials. To seek a better balance between accuracy and spelling speed, we further conducted an offline analysis where a dynamic

TABLE I
RESULTS OF ONLINE EXPERIMENTS I–III

Subject	Experiment I		Experiment II		Experiment III	
	Accuracy (%)	ITR (bits/min)	Accuracy (%)	ITR (bits/min)	Accuracy (%)	ITR (bits/min)
1	97.50	24.96	95.00	23.72	95.00	23.72
2	80.00	17.61	95.00	23.72	95.00	23.72
3	92.50	22.57	90.00	21.50	97.50	24.96
4	82.50	18.53	100.00	26.46	90.00	21.50
5	100.00	26.46	100.00	26.46	97.50	24.96
6	67.50	13.39	75.00	15.85	90.00	21.50
7	95.00	23.72	90.00	21.50	92.50	22.57
8	85.00	19.48	100.00	26.46	97.50	24.96
9	87.50	20.47	100.00	26.46	95.00	23.72
10	90.00	21.50	100.00	26.46	92.50	22.57
11	90.00	21.50	90.00	21.50	92.50	22.57
12	97.50	24.96	100.00	26.46	82.50	18.53
13	95.00	23.72	100.00	26.46	95.00	23.72
14	100.00	26.46	90.00	21.50	92.50	22.57
15	97.50	24.96	95.00	23.72	90.00	21.50
16	97.50	24.96	90.00	21.50	97.50	24.96
17	72.50	15.01	95.00	23.72	92.50	22.57
18	62.50	11.86	75.00	15.85	85.00	19.48
19	97.50	24.96	100.00	26.46	100.00	26.46
20	100.00	26.46	100.00	26.46	100.00	26.46
Mean ± SD	89.38 ± 10.92	21.68 ± 4.32	94.00 ± 7.52	23.61 ± 3.28	93.50 ± 4.43	23.15 ± 2.01

TABLE II
ONLINE PERFORMANCES OF DIFFERENT CNN MODELS IN EXPERIMENT II

	Trials 1–10	Trials 11–20	Trials 21–40
Average accuracy ± SD (%)	87.00 ± 15.20	89.00 ± 14.11	94.00 ± 7.52
Average ITR ± SD (bits/min)	21.01 ± 5.63	21.75 ± 5.19	23.61 ± 3.28

stopping strategy was used in each trial, i.e., the number of flash rounds was adaptive.

The dynamic stopping strategy was described in our previous study [5], and we briefly review it here. First, we empirically set the minimum and maximum numbers of flash rounds for each trial to 4 and 9, respectively. Second, in each round after the fourth round of each trial, we fed the data into the CNN model and obtained a predicted target character as well as a probability showing the confidence of the prediction. If the probability was larger than a preset threshold or the number of flash rounds reached 9, the system output the predicted target character. Otherwise, the next round of button flashes progressed. The threshold for the probability in each round was set by applying leave-one-subject-out cross-validation to the training set (collected from 150 subjects). Specifically, the data of 149 subjects were used to train a CNN model, whereas the data of the remaining subject were used for testing. To set the threshold for the fourth round, for each trial in the training set, the data from the first to the fourth rounds were averaged and then fed into the CNN, and a predicted target character as well as its probability showing the confidence of the prediction were obtained. The probabilities were averaged over all trials for the test subject. The probabilities of the 150 subjects were obtained through leave-one-subject-out cross-validation, which formed a distribution. A threshold was set for the fourth round such that the top 20% of the probabilities were larger than it (0.9811 in this study). By using the same method, we set the thresholds for the fifth to the eighth rounds according to the top 40%, 60%, 80%, and 100% of the probability values in the distributions obtained after the fifth to the eighth rounds.

The results of the offline analysis based on a dynamic stopping strategy are shown in Table III. Note that the results of the self-training-based CNN were obtained from the last 20 online trials, where the updated CNN model was applied. Comparing the results shown in Tables I and III, we can see that the dynamic stopping strategy improved the spelling speed with acceptable average accuracies and thus improved the ITR.

IV. DISCUSSION

In this study, we developed a CNN- and big EEG data-based online P300 BCI spelling system with zero-calibration or shortened calibration. Specifically, three methods were proposed to train cross-subject P300 detection models, including (i) training a subject-independent CNN using data collected from 150 subjects, (ii) adapting the CNN online based on a self-training method and the unlabeled data collected during the user's online operation, and (iii) fine-tuning the CNN based on a transfer learning method and a small quantity of labeled

TABLE III
RESULTS OF AN OFFLINE ANALYSIS BASED ON A DYNAMIC STOPPING STRATEGY

Model	Subject-independent CNN	Self-training-based CNN	Transfer learning-based CNN
Average accuracy ± SD (%)	81.25 ± 13.36	91.50 ± 12.95	88.12 ± 6.66
Average number of rounds ± SD	7.126 ± 0.146	6.242 ± 0.636	6.607 ± 0.744
Average ITR ± SD (bits/min)	25.97 ± 6.92	36.69 ± 8.71	31.96 ± 5.76

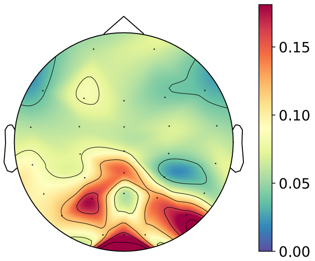


Fig. 5. Spatial filter obtained with the subject-independent CNN. The spatial filters are obtained by averaging the absolute values of the weights in layer C1 across the ten kernels.

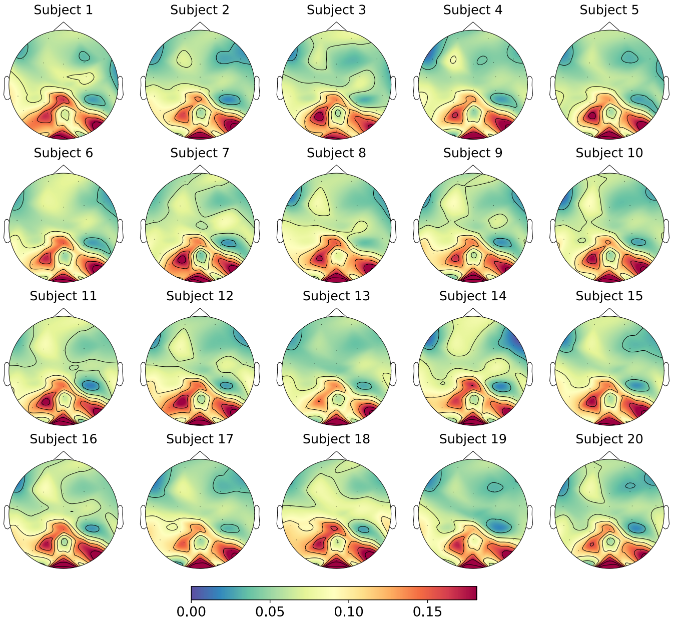


Fig. 6. Spatial filters obtained with the self-training-based CNNs. Compared with the spatial filter obtained with the subject-independent CNN shown in Fig. 5, the spatial filters obtained with the self-training-based CNNs change slightly for most subjects.

data. The experimental results demonstrated the effectiveness of our system.

The online P300 spelling system developed in this study achieved good performances, with accuracies near or above 90% for all three models. This is probably due to the following reasons. First, deep neural networks have excellent data-fitting abilities, and our dataset included data from a relatively large number of subjects compared with those in existing works. As demonstrated in [5], these two factors provided the model with the possibility of extracting subject-independent features. Second, we adapted the CNN by performing self-training or transfer learning during or before the online operation to further improve its performance. Third, we implemented the P300 spelling system online, and thus, during the online operation, users received feedback regarding the spelling results from the system and accordingly adjust their mental states in real time to better complete the spelling task. Note that the same subject-independent CNN was employed for both the offline analysis (see our previous study [5]) and the online test, and the average accuracies were 83.74% and 89.38%, respectively. The fact that the online test yielded better performance than the offline analysis is probably due to the effect of the feedback presented to the subjects.

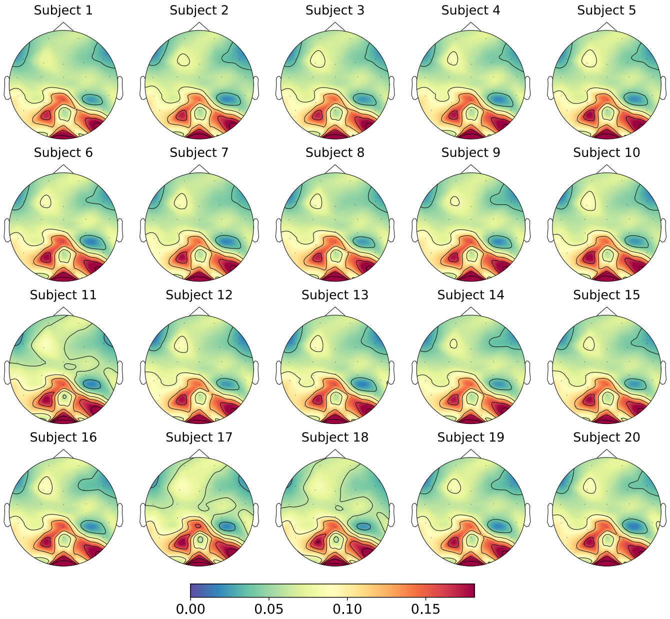


Fig. 7. Spatial filters obtained with the transfer learning-based CNNs. Compared with the spatial filter obtained with the subject-independent CNN shown in Fig. 5, the spatial filters obtained with the transfer learning-based CNNs change slightly for most subjects.

To further explore what spatial and temporal features are important for EEG classification and how does parameter update affect the models, we visualize the models before and after the adaptation from two aspects. (i) We first visualize the convolutional kernels of the first convolutional layer C1, which plays a role in spatial filtering. Specifically, for each trained model, the absolute values of the weights in layer C1 are averaged across the ten kernels, resulting in a 30-dimensional weight vector with each entry representing the discriminant power of the corresponding channel. We use this weight vector to generate a topology map to show the importance of each channel to the classification result. The topology maps of the subject-independent CNN, the self-training-based CNN, and the transfer learning-based CNN are shown

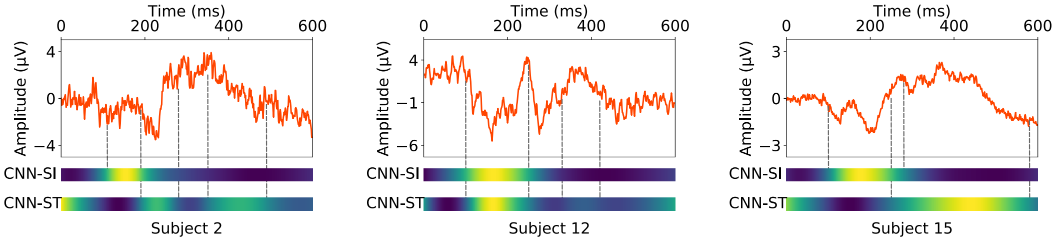


Fig. 8. ERP waveforms and the corresponding Grad-CAM heatmaps obtained with the subject-independent CNN (denoted "CNN-SI") and the self-training-based CNN (denoted "CNN-ST"). Note that in the heatmaps, yellow intervals correspond to the important time intervals, while blue intervals are the opposite.

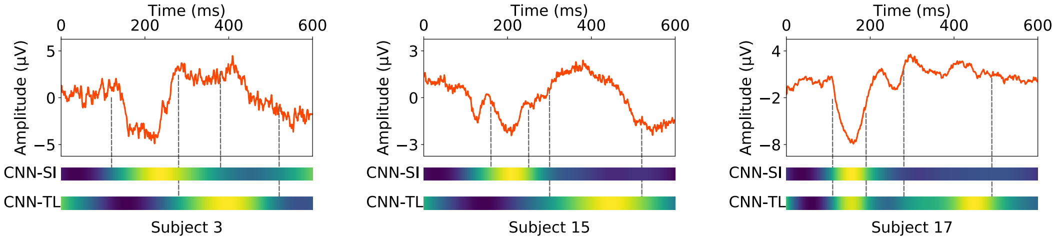


Fig. 9. ERP waveforms and the corresponding Grad-CAM heatmaps obtained with the subject-independent CNN (denoted "CNN-SI") and the transfer learning-based CNN (denoted "CNN-TL"). Note that in the heatmaps, yellow intervals correspond to the important time intervals, while blue intervals are the opposite.

in Figs. 5, 6, and 7, respectively. From the figures, we can see that after the model adaptation, for both the self-training-based CNN and the transfer learning-based CNN, the weights in layer C1 change, reflecting interindividual variability. For instance, we find that the self-training-based CNNs for Subjects 13 and 17 and the transfer learning-based CNNs for Subjects 11, 17, and 18 have relatively large weight changes on the spatial filters, while the weight changes for other subjects are slight. (ii) We then use the gradient-weighted class activation mapping (Grad-CAM) algorithm [40] to produce a coarse localization map highlighting the important time intervals of the signals for EEG classification. Specifically, for each trial, the EEG signal corresponding to the target character is fed into the subject-independent CNN and the self-training/transfer learning-based CNN to obtain a heatmap for each model. For each subject and each model, the EEG waveforms and the heatmaps are averaged across the 40 trials. Several averaged waveforms from the EEG channel OZ and the corresponding heatmaps obtained with the subject-independent CNN and the self-training/transfer learning-based CNN are presented in Figs. 8 and 9, respectively. Note that the subject-independent CNN is applicable for all subjects, and its corresponding heatmaps show some consistency. After model adaptations, the models become subject specific, and the important time intervals vary by subject. This is probably because the EEG signals contain different discriminative components effective for the classification, and these components vary by subject. For instance, as shown in Fig. 8, the time intervals where typical event-related potential (ERP) components (such as N200 or P300) occur are coarsely marked as the important time intervals for the subject-independent CNN, while the self-training-based CNNs utilize more components in different time intervals for Subjects 2 and 12 and focus more attention on the P300 component for Subject 15. Similarly, in Fig. 9, the classification with the subject-independent CNN mainly relies on a single time interval with typical ERP components, while the transfer learning-based CNNs additionally utilize the signal at approximately 400 ms after the stimulus onset for Subject 17 and pay more attention to the time interval where a P300 component occurs for Subjects 3 and 15.

Compared with existing online BCI systems, the advantages of the system developed in this study are as follows. First, the calibration phase is completely eliminated or dramatically shortened, and the convenience of the BCI system is thus improved. By applying the subject-independent model or the self-training-based model, the system is plug-and-play, which means that new users can operate this system without a calibration phase. It is worth mentioning that although the self-training-based model needs online adaptation, the required calibration data are entirely unlabeled data collected during user operations. Moreover, the model adaptation process does not suspend the user's operation of the system. For the use of the transfer learning-based model, the system requires users to perform a short calibration task. Specifically, the time needed for the calibration of this system, which is approximately 1 min, is much shorter than that for traditional P300 BCIs, which usually take more than 10 min for subject-specific calibration. Eliminating or shortening the calibration also reduces the mental load for the users; our experiments reflect that most subjects did not feel obvious fatigue. Second, by applying a zero-calibrated CNN (the subject-independent CNN or the self-training-based CNN) as a P300 detection model, this system achieves comparable performances to those of traditional P300 BCIs with full calibration [41], [42], [43]. Additionally, to the best of our knowledge, few studies have implemented zero-calibrated models online. In [10], an online spelling system was developed, and an average accuracy of 85% was achieved after 33 s of button flashes for each trial. Although it was a good attempt to produce BCIs with zero-calibration, this approach still needs further performance improvement. In our study, two zero-calibrated CNNs were implemented online. An average accuracy of 89.38% was achieved after 12 s of button flashes for each trial when the subject-independent model was applied, and the accuracy improved to 94.00% after the model was adapted by self-training. Finally, this study improves the online performance of the BCI system using a transfer learning-based model. Existing studies based on traditional transfer learning have obtained accuracies of approximately 80%–85% in online experiments [26], [28]. Among the several existing P300 BCI studies based on deep transfer learning, almost all of them only provided offline analyses with accuracies of 70%–90%, and their models needed further online validation [15], [16], [17]. Our experimental results showed that by using the transfer learning-based CNN, the online accuracy could be improved from 89.38% (obtained with the subject-independent CNN) to 93.50%. In addition, all subjects achieved accuracies of 80% or above.

Three models are available in our BCI system. The user can select one model for operating the system according to the following strategy. (i) If the available computing resources are insufficient for supporting CNN retraining, the subject-independent CNN can be conveniently applied. (ii) If the available computing resources are sufficient for retraining the CNN, the self-training method can be used when a calibration phase is not allowed, for instance, when the user does not know how to collect calibration data or when the user is unwilling to perform the shortened calibration phase. Note that there needs to be a period of model adaptation (4 min in this study) via self-training that does not suspend the user's operation of the BCI system. During this period, the performance of the CNN model is improved step by step. (iii) If the available computational resources are sufficient for retraining the CNN and a short period of calibration is allowed, the transfer learning-based CNN is a good choice since comparable performance to that of a fully calibrated BCI model can be achieved with a much shorter calibration time period.

V. Conclusion

This study developed an online P300 BCI spelling system with zero-calibration or shortened calibration based on a CNN and big EEG data. Specifically, three methods to train CNNs for the online detection of P300 potentials were proposed: training a subject-independent CNN with data collected from 150 subjects, adapting the CNN online based on a self-training method and unlabeled data collected during the user's online operation, and fine-tuning the CNN based on a transfer learning method and a small quantity of labeled data collected before the user's operation. Based on these methods, an online P300 spelling system was developed. Average accuracies of 89.39%, 94.00% and 93.50% were achieved with the subject-independent CNN, the self-training-based CNN and the transfer learning-based CNN, respectively. These experimental results indicated that based on a CNN and big EEG data, an online P300 BCI with zero-calibration or shortened calibration could be built. In future studies, we will extend this system to patients, such as those with strokes or spinal cord injuries, to help them improve their self-care ability.

References

[1] J. J. Vidal, "Toward direct brain–computer communication," *Annu. Rev. Biophys. Bioeng.*, vol. 2, no. 1, pp. 157–180, Jun. 1973.

[2] F. Lotte, "Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces," *Proc. IEEE*, vol. 103, no. 6, pp. 871–890, Jun. 2015.

[3] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.

[4] D. Kostas and F. Rudzicz, "Thinker invariance: Enabling deep neural networks for BCI across more people," *J. Neural Eng.*, vol. 17, no. 5, Oct. 2020, Art. no. 056008.

[5] W. Gao et al., "Learning invariant patterns based on a convolutional neural network and big electroencephalography data for subject-independent P300 brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1047–1057, 2021.

[6] B. Abibullaev, K. Kunanbayev, and A. Zollanvari, "Subject-independent classification of P300 event-related potentials using a small number of training subjects," *IEEE Trans. Hum.-Mach. Syst.*, vol. 52, no. 5, pp. 843–854, Oct. 2022.

[7] Z. Ni, J. Xu, Y. Wu, M. Li, G. Xu, and B. Xu, "Improving cross-state and cross-subject visual ERP-based BCI with temporal modeling and adversarial training," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 369–379, 2022.

[8] S. Lu, C. Guan, and H. Zhang, "Subject-independent brain computer interface through boosting," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[9] I. Dolzhikova, B. Abibullaev, and A. Zollanvari, "An ensemble of convolutional neural networks for zero-calibration ERP-based BCIs," in *Proc. 10th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2022, pp. 1–4.

[10] J. Lee, K. Won, M. Kwon, S. C. Jun, and M. Ahn, "CNN with large data achieves true zero-training in online P300 brain–computer interface," *IEEE Access*, vol. 8, pp. 74385–74400, 2020.

[11] K. Vo, T. Pham, D. N. Nguyen, H. H. Kha, and E. Dutkiewicz, "Subject-independent ERP-based brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 719–728, Apr. 2018.

[12] S. Lu, C. Guan, and H. Zhang, "Learning adaptive subject-independent P300 models for EEG-based brain–computer interfaces," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intelligence)*, Jun. 2008, pp. 2461–2465.

[13] S. Lu, C. Guan, and H. Zhang, "Unsupervised brain computer interface based on intersubject information and online adaptation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 17, no. 2, pp. 135–145, Apr. 2009.

[14] H. Woehrle, M. M. Krell, S. Straube, S. K. Kim, E. A. Kirchner, and F. Kirchner, "An adaptive spatial filter for user-independent single trial detection of event-related potentials," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 7, pp. 1696–1705, Jul. 2015.

[15] S. Kundu and S. Ari, "MsCNN: A deep learning framework for P300-based brain–computer interface speller," *IEEE Trans. Med. Robot. Bionics*, vol. 2, no. 1, pp. 86–93, Feb. 2020.

[16] H. Wang et al., "Performance enhancement of P300 detection by multiscale-CNN," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[17] I. Dağ, L. G. Dui, S. Ferrante, A. Pedrocchi, and A. Antonietti, "Leveraging deep learning techniques to improve P300-based brain computer interfaces," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 10, pp. 4892–4902, Oct. 2022.

[18] Z. Huang, J. Guo, W. Zheng, Y. Wu, Z. Lin, and H. Zheng, "A calibration-free approach to implementing P300-based brain–computer interface," *Cognit. Comput.*, vol. 14, no. 2, pp. 887–899, Mar. 2022.

[19] P.-J. Kindermans, H. Verschore, D. Verstraeten, and B. Schrauwen, "A P300 BCI for the masses: Prior information enables instant unsupervised spelling," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.

[20] P.-J. Kindermans, M. Tangermann, K.-R. Muller, and B. Schrauwen, "Integrating dynamic stopping, transfer learning and language models in an adaptive zero-training ERP speller," *J. Neural Eng.*, vol. 11, no. 3, Jun. 2014, Art. no. 035005.

[21] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu, "Transfer learning: A Riemannian geometry framework with applications to brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 5, pp. 1107–1116, May 2018.

[22] S. Khazem, S. Chevallier, Q. Barthélemy, K. Haroun, and C. Nous, "Minimizing subject-dependent calibration for BCI with Riemannian transfer learning," in *Proc. 10th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, May 2021, pp. 523–526.

[23] F. Li, Y. Xia, F. Wang, D. Zhang, X. Li, and F. He, "Transfer learning algorithm of P300-EEG signal based on xDAWN spatial filter and Riemannian geometry classifier," *Appl. Sci.*, vol. 10, no. 5, p. 1804, Mar. 2020.

[24] P. L. C. Rodrigues, C. Jutten, and M. Congedo, "Riemannian Procrustes analysis: Transfer learning for brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 8, pp. 2390–2401, Aug. 2018.

[25] X. An, X. Zhou, W. Zhong, S. Liu, X. Li, and D. Ming, "Weighted subject-semi-independent ERP-based brain–computer interface," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 2969–2972.

[26] J. Jin et al., "The study of generic model set for reducing calibration time in P300-based brain–computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 1, pp. 3–12, Jan. 2020.

[27] H. Qi, Y. Xue, L. Xu, Y. Cao, and X. Jiao, "A speedy calibration method using Riemannian geometry measurement and other-subject samples on a P300 speller," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 602–608, Mar. 2018.

[28] Y. Zhao et al., "A transplantation of subject-independent model in cross-platform BCI," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 6, pp. 959–967, Jun. 2018.

[29] J. Hou, Y. Li, H. Liu, and S. Wang, "Improving the P300-based brain–computer interface with transfer learning," in *Proc. 8th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, May 2017, pp. 485–488.

[30] Y. Li, C. Guan, H. Li, and Z. Chin, "A self-training semi-supervised SVM algorithm and its application in an EEG-based brain computer interface speller system," *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1285–1294, 2008.

[31] M. Ogino, S. Kanoga, S.-I. Ito, and Y. Mitsukura, "Semi-supervised learning for auditory event-related potential-based brain–computer interface," *IEEE Access*, vol. 9, pp. 47008–47023, 2021.

[32] R. C. Panicker, S. Puthusserypady, and Y. Sun, "Adaptation in P300 brain–computer interfaces: A two-classifier cotraining approach," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 12, pp. 2927–2935, Dec. 2010.

[33] Y. Xin, Q. Wu, Q. Zhao, and Q. Wu, "Semi-supervised regularized discriminant analysis for EEG-based BCI system," in *Proc. Int. Conf. Intell. Data Eng. Automated Learn.* Cham, Switzerland: Springer, 2017, pp. 516–523.

[34] Z. Gu, Z. Yu, Z. Shen, and Y. Li, "An online semi-supervised brain–computer interface," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 9, pp. 2614–2623, Sep. 2013.

[35] J. Wang, Z. Gu, Z. Yu, and Y. Li, "An online semi-supervised P300 speller based on extreme learning machine," *Neurocomputing*, vol. 269, pp. 148–151, Dec. 2017.

[36] H. Cecotti and A. Gräser, "Convolutional neural networks for P300 detection with application to brain–computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 433–445, Mar. 2011.

[37] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[39] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16, Aug. 2016, pp. 265–283.

[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[41] R. K. Chaurasiya, N. D. Londhe, and S. Ghosh, "Binary DE-based channel selection and weighted ensemble of SVM classification for novel brain–computer interface using Devanagari script-based P300 speller paradigm," *Int. J. Hum.-Comput. Interact.*, vol. 32, no. 11, pp. 861–877, Nov. 2016.

[42] W. Speier, A. Deshpande, L. Cui, N. Chandravadia, D. Roberts, and N. Pouratian, "A comparison of stimulus types in online classification of the P300 speller using language models," *PLoS ONE*, vol. 12, no. 4, Apr. 2017, Art. no. e0175382.

[43] L. Bianchi, C. Liti, and V. Piccialli, "A new early stopping method for P300 spellers," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 8, pp. 1635–1643, Aug. 2019.