A Cascade xDAWN EEGNet Structure for Unified Visual-Evoked Related Potential Detection

Hongtao Wang, *Senior Member, IEEE*, Zehui Wang, Yu Sun, *Senior Member, IEEE*, Zhen Yuan, Tao Xu, and Junhua Li, *Senior Member, IEEE*

***Abstract*— Visual-based brain-computer interface (BCI) enables people to communicate with others by spelling words from the brain and helps professionals recognize targets in large numbers of images. P300 signals evoked by different types of stimuli, such as words or images, may vary significantly in terms of both amplitude and latency. A unified approach is required to detect variable P300 signals, which facilitates BCI applications, as well as deepens the understanding of the P300 generation mechanism. In this study, our proposed approach involves a cascade network structure that combines xDAWN and classical EEGNet techniques. This network is designed to classify target and non-target stimuli in both P300 speller and rapid serial visual presentation (RSVP) paradigms. The proposed approach is capable of recognizing more symbols with fewer repetitions (up to 5 rounds) compared to other models while possessing a better information transfer rate (ITR) as demonstrated on Dataset II (17.22 bits/min in the second repetition round) of BCI Competition III. Additionally, our approach has the highest unweighted average recall (UAR) performance for both 5 Hz (0.8134±0.0259) and 20 Hz (0.6527±0.0321) RSVP. The results show that the cascade network structure has better performance between both the P300 Speller and RSVP paradigms, manifesting that such a cascade structure is robust enough for dealing with P300-related signals (source code is available at https://github.com/embneural/Cascade-xDAWN-EEGNet-for-ERP-Detection).**

Manuscript received 15 December 2023; revised 5 May 2024; accepted 10 June 2024. Date of publication 17 June 2024; date of current version 25 June 2024. This work was supported in part by the Special Projects in Key Fields Supported by the Technology Development Project of Guangdong Province under Grant 2020ZDZX3018, in part by Wuyi University and Hong Kong and Macao Joint Research Project under Grant 2019WGALH16, and in part by the Projects for International Scientific and Technological Cooperation of Guangdong Province under Grant 2023A0505050144. *(Corresponding authors: Hongtao Wang; Tao Xu.)*

Hongtao Wang and Zehui Wang are with the School of Electronics and Information Engineering, Wuyi University, Jiangmen 354300, China (e-mail: hongtaowang@wyu.edu.cn; wzhwzhwzhwzh99@163.com).

Yu Sun is with the Key Laboratory for Biomedical Engineering of Ministry of Education of China, Department of Biomedical Engineering, Zhejiang University, Hangzhou, Zhejiang 310027, China.

Zhen Yuan is with the Faculty of Health Sciences and the Center for Cognitive and Brain Sciences, University of Macau, Taipa, Macau.

Tao Xu is with the Department of Biomedical Engineering, Shantou University, Shantou 515063, China (e-mail: wanxiao0756@gmail.com).

Junhua Li is with the School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ Colchester, U.K.

Digital Object Identifier 10.1109/TNSRE.2024.3415474

***Index Terms*— Brain-computer interface (BCI), P300, xDAWN, EEGNet, rapid serial visual presentation (RSVP).**

I. Introduction

BCIs are systems that enable people to communicate with their surroundings using brain waves [1], [2]. They are particularly helpful for individuals with disabilities, such as those with amyotrophic lateral sclerosis (ALS), who may rely on a BCI to interact with others [3], [4]. Electroencephalogram (EEG)-based BCIs are becoming more popular in both commercial and academic settings because they are non-invasive, portable, and provide fast responses and over 80% of BCI publications rely on EEG [5], [6], [7].

Various components of EEG signals can be appliable for BCIs, such as sensorimotor rhythms (SMRs), steady-state visual evoked potentials (SSVEPs), code-modulated visual evoked potentials (c-VEPs), miniature asymmetric visual evoked potentials (aVEPs) and event-related potentials (ERPs). Specifically, ERP, which is known as electrical potentials time-locked to events, has gained attention in BCI studies because ERPs relevant to different modalities like visual, auditory, or tactile events provide the best results for the control of a BCI system [8], [9], [10]. P300, which is a response that occurs about 300-600 milliseconds after the onset of a stimulus, has been extensively investigated in ERP-based BCIs [11]. In visual P300 paradigms like row/column spellers, differences between target and non-target ERPs are used to generate characters by flashing corresponding rows and columns so that the P300-based speller allows the user to communicate letter by letter [7]. Because the EEG collection for the P300 speller is non-invasive, the resulting signal-to-noise ratio (SNR) is low [12]. To improve SNR, repeated stimuli are used and multiple trials of EEG collections are averaged [13]. However, this averaging method is time-consuming and inefficient [14], and may negatively impact BCI real-time performance by incorporating past information. As a result, it is still very challenging to accurately classify P300 by using a single trial. Apart from the row/column speller explored in BCI with P300, RSVP-based BCI detects and recognizes objects, scenes and events in static images and videos via P300. Different from the P300 speller, the RSVP-based BCI can be beneficial to more applications where a large number of images need to be reviewed by professionals but are unable to be well

analyzed by computers [15], [16], [17]. In the RSVP-based P300 experiment, the participants need to distinguish between target and non-target stimuli where the target should be odd stimuli and account for 5%-10% of all stimuli [15].

Although both BCIs with the P300 speller and BCIs with the RSVP use the P300 component for classification, there are some differences between the two paradigms. Firstly, in the P300 speller, the operator sets the sequence repetition number to improve SNR, while in RSVP, the target or non-target stimuli are set in a single trial. Secondly, the stimulus frequency is different for the two paradigms - the RSVP usually displays with a refresh rate of 5-10 Hz, while the P300 speller flashes for 125 ms and turns off for 62.5 ms until the next sequence flashes [18]. Thirdly, due to the repetition of targeted stimuli elicited by the P300 speller, the average amplitude of the target P300 produced is weaker than that produced by the single-trial RSVP paradigm. Lastly, the P300 speller is more stable in EEG, whereas the RSVP paradigm is prone to noise induction as it requires button presses to determine the target stimuli. As mentioned, notable ERPs emerge at 315 ms in the RSVP paradigm whereas occur at 262 ms in the row-column-based P300 speller paradigm during target events compared to non-target events [18]. As mentioned in [19] and [20], an experimental protocol consisted of two sessions performed on two days, day one of which is the first session for the P300 speller experiment and day two of which is the second session for the RSVP experiment. The results showed that the temporal filtering capacity in the RSVP task can be a predictor of both the P300-based BCI accuracy and the amplitude of the P300 elicited when performing the BCI task. Moreover, Won et al. reported a significant positive correlation between P300 amplitude in RSVP and P300 speller performance. They observed a strong negative correlation between the variation in P300 latency across trials in RSVP and P300 speller performance [21]. These findings indicate that there exists a relationship between RSVP task characteristics and P300 speller performance.

To develop a sophisticated BCI system, a unified framework for processing ERP signals is essential, as it can simplify system maintenance and upgrades, and reduce overall development costs. Additionally, a sophisticated BCI system must meet the diverse requirements of its users. For instance, the standard row-column paradigm used in the P300 speller may not be appropriate for patients who lack gaze control [22]. In such cases, the RSVP paradigm presenting the stimuli in the same position may serve as a useful alternative. While the ERPs obtained from these two paradigms exhibit similarities, a unified decoding approach is necessary to develop a BCI system that can be tailored to different users [22]. To target the above characteristics of the P300 speller and RSVP paradigms, various methods have been proposed for P300 component detection. Traditional machine learning methods, such as support vector machines (SVM), discriminant analysis, and common spatial pattern algorithms, were first applied [42], [43], [46], [49], [50]. These methods detect ERP signals with manual feature extraction, and the quality of the extracted features plays a big role in how well the algorithm performs. Recently, with the emergence of deep learning models, ERP features can now be automatically learned from data without any manual intervention. Several convolutional neural networks (CNN), such as EEGNet, DeepConvNet, ShallowConvNet, and BN3, have been developed for detecting ERP signals [5], [13], [23], [24]. However, deep learning models often require a large number of samples for better learning, as they lack domain knowledge about the data [25]. An important domain knowledge regarding ERP signals is that the SNR can be enhanced through trial averaging [26], [27]. To take advantage of the domain information, it might be more suitable to process the ERP signals by xDAWN spatial filtering before sent to neural networks, as the xDAWN estimates the spatial filters by maximizing the difference between the averaged signals of the corresponding category and the whole EEG data [28]. Research has extensively explored the combination of xDAWN with classification algorithms, yielding promising results. For instance, Cecotti et al. successfully integrated xDAWN with MLP, BLDA, and linear SVM, resulting in performance enhancement [29]. In a similar vein, the xDAWN-based algorithm emerged victorious in the Kaggle BCI competition NER 2015, leveraging Riemannian geometry, channel subset selection, L1 regularization, and elastic net regression [30], [31]. Moreover, Wu et al. achieved remarkable progress by aligning ERP data with Euclidean alignment and enriching features with xDAWN and tangent space mapping which secured the top spot in the RSVP detection competition at the World Robot Contest 2021 [32], [33]. Meanwhile, in our previous work, Zhang et al. achieved second place in the same competition by combining xDAWN with EEGNet [34]. These studies showcase the potential of integrating xDAWN with classification algorithms to achieve superior performance across diverse applications. However, in the existing literature, there is no single algorithm that can be well-suited for both P300 speller and RSVP target detection. To tackle this challenge, we leverage our prior research [34] and extend the methodologies of xDAWN and EEGNet to encompass both the P300 speller and RSVP paradigms.

In the rest of this work, we describe the dataset in Section II and detail the methods applied in Section III. We present corresponding results in Section IV and compare them with prevailing ones in the discussion part. Finally, we conclude our findings.

II. Method and Materials

A. Dataset Description

The study utilized two datasets: the P300 speller-based dataset and the RSVP dataset. The P300 speller-based dataset was derived from two datasets, Dataset IIb from BCI Competition II, which involved one participant, and Dataset II from BCI Competition III, which consisted of subjects A and B. The datasets were recorded using the BCI2000 system, which employed 64 electrodes at a sampling rate of 240 Hz. During the experiment, participants were shown a 6×6 symbol matrix and were instructed to pay attention to specific target symbols. The intensity of all the rows and columns in the symbol matrix was randomly increased at a frequency of 5.7 Hz. Each intensification lasted for 100 ms. After each

intensification, the matrix remained blank for a duration of 75 ms, followed by the next intensification of a row or column. Each symbol presentation consisted of 15 rounds, with each round containing 12 stimuli. In these 12 stimuli, only two stimuli (rows/columns) corresponded to the desired symbols and the elicited responses were labeled as P300 samples while the responses elicited by the remaining 10 stimuli were labeled as non-P300 samples. Thus, each symbol contains 150 non-P300 samples and 30 P300 samples. In Dataset IIb, there were 42 symbols for training and 31 symbols for testing. In Dataset II, for each subject, 85 symbols were used for training and 100 symbols for testing. The dataset is available at https://www.bbci.de/competition/.

On the other hand, the RSVP dataset [35] employed a stimulus set of 200 visual objects from different categories, presented to 16 adult participants (5 females; age range 18–38 years) who were instructed to count target stimuli (boats or geometric star shapes) randomly inserted into the sequence, with a maximum of 4 targets per sequence. Each sequence lasted between 40.2 and 40.8 seconds, with a presentation rate of 5 Hz in the first session and 20 Hz in the second session. There were a total of 40 sequences for each session. The EEG data were recorded at a sampling rate of 1000 Hz using a BrainVision ActiChamp system and international standard 10-10 system for 64-electrode placement. During recording, all scalp electrodes were referenced to Cz. Then the recorded data were filtered with a Hamming windowed FIR filter (0.1 Hz high pass and 100 Hz lowpass filters) and down-sampled to 250 Hz for further processing. This dataset is available at https://osf.io/a7knv/.

B. Data Preprocessing

To preprocess the P300 speller-based dataset, we first cropped an 800 ms segment after the stimulus onset, and then detrended the data to remove linear trends. Next, we applied a 30 Hz low-pass Chebyshev filter with zero phase to filter out high-frequency noise, while preserving the desired signals. Additionally, we down-sampled the data by half, resulting in a 96-time sample segment. For the RSVP-based datasets, we retrieved a 0-1000 ms segment after the stimulus onset, detrended the data, and applied the same 30 Hz low pass Chebyshev filter with zero phase, resulting in a 250-time sample segment. Here, $X_i \in R^{C\times T}$ denotes the i^{th} trial EEG data sample, where C represents the number of EEG electrodes and T is the time samples of one trial data. We set T to 96 and 250 for the P300 speller and RSVP datasets, respectively. Since both datasets were collected using a 64-electrode system, C was set to 64.

C. xDAWN Spatial Filtering

The EEG signals are noisy and have low SNR because they record complete brain activity, including areas of the brain that are not relevant to the task, which leads to a difficult classification problem [30]. In addition, some channels carry more valuable information than others, such as the channels around the parietal lobe in the P300 speller and RSVP paradigms. To effectively enhance the task-relevant information in channels, we used the xDAWN spatial filtering method to refine the original EEG signals. The xDAWN algorithm is defined as follows:

Compute the averaged pattern of class k. $X_i^{(k)}$, $P^{(k)} \in \mathbb{R}^{C\times T}$ and m_k, are the i^{th} trial EEG data in class k, the averaged pattern for class k, the number of trials of class k, respectively.

$$P^{(k)} = \frac{1}{m_k}\sum X_i^{(k)} \tag{1}$$

Estimate spatial filters for class k. The spatial filter is a vector $w\in\mathbb{R}^{C\times 1}$. $w^* \in \mathbb{R}^{C\times 1}$ represents an estimated spatial filter. $X \in \mathbb{R}^{C\times(T\sum m_k)}$ is the concatenation of all the trials (from all classes). Because (3) is a generalized Rayleigh quotient, the solution could be given by calculating the eigenvectors of the matrix $\left[P^{(k)}{P^{(k)}}^T\left(XX^T\right)^{-1}\right]$. The top n eigenvectors (ordered by eigenvalues) were selected as spatial filters.

$$w^* = \text{argmax}\frac{w^T P^{(k)}{P^{(k)}}^T w}{w^T XX^T w} \tag{2}$$

Apply spatial filters to raw EEG data. $W^{(k)} \in \mathbb{R}^{C\times n}$ is the spatial filters for class k. $W = \left[W^{(1)}, W^{(2)}\ldots, W^{(N)}\right] \in \mathbb{R}^{C\times M}$ is the spatial filters of all classes (N classes in total) where $M = n\cdot N$. n is the number of xDAWN spatial filters and N is the number of classes. $Z_i \in \mathbb{R}^{M\times T}$ is the enhanced EEG data.

$$Z_i = W^T X_i \tag{3}$$

Finally, we applied the Z-score normalization to each individual enhanced EEG segment:

$$Z_i = \frac{Z_i - \mu}{\sigma} \tag{4}$$

where μ and σ are the mean and standard deviation of each channel of the enhanced EEG data respectively. This approach ensured that the data were standardized and consistent across participants and electrodes.

D. Network Architecture

The xDAWN filtering approach estimates a set of spatial filters by maximizing the difference between the averaged patterns of each class and the overall EEG signals. Such estimated spatial filters make full use of strong domain knowledge (e.g., the SNR of ERP signals can be improved by trial averaging and the averaged trial signals are representative of typical ERP signals for respective tasks [26]). Deep learning models are hard to learn specific domain knowledge as they often require a large amount of data to learn a certain inductive bias [25]. Therefore, we linked the xDAWN with the classic EEGNet to maximize the use of the prior information and further improve the model's detection performance on ERP signals.

Fig. 1 shows the structure and detailed description of our proposed architecture. The xDAWN spatial filters were first applied to the input EEG data X_i, followed immediately by a temporal convolution operation and BatchNorm (with a convolution kernel of size 1×64, stride of 1 and padding of 'same') to produce F1 feature maps (F1 was set to 8 in this study). We then manipulated these feature maps using depth-wise convolution, with depth D set to 2, a depth convolution

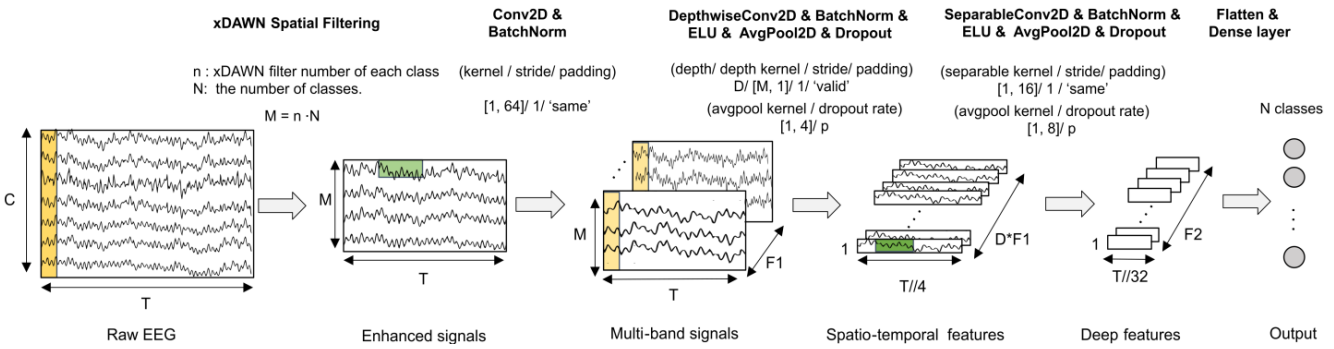

Fig. 1. Network structure of the proposed cascade xDAWN-EEGNet. The network first applies xDAWN spatial filters to the input EEG signals and then feeds the enhanced signals into the EEGNet. The max norm Was used to constrain the weights in the DepthwiseConv2D and dense layers, where it was set to 1 and 0.25 respectively.

kernel size of $M \times 1$ and 'valid' padding. Then the BatchNorm was used, and the results were activated by ELU operation. Next, the average pooling operation was used to reduce the size of feature maps and dropout was used to avoid overfitting. The average pool kernel size and dropout value p were set to be 1×4 and 0.25, respectively. We then used separable convolution (with a kernel size of 1×16, stride of 1 and 'same' padding) to extract deeper features. The BatchNorm and ELU activation were also used. Separable convolution is composed of depthwise convolution and pointwise convolution, to reduce the number of model parameters. Next, we apply an average pooling layer of size 1×8 for dimension reduction and a dropout layer with the p value equal to 0.25. In the dense layer, N neurons are densely linked to the features of the previous layer and activated by Softmax activation.

E. Training

We used the pyRiemann package [36] to estimate xDAWN spatial filters and reproduced EEGNet with PyTorch [37]. The proposed model was trained on GeForce RTX 2080 Ti. The batch size was set to 64. The Adam optimizer with default parameters was used, and the learning rate was initially set to 0.001 with an exponential decay rate of 0.96. The P300 speller dataset consisted of two categories: non-P300 and P300 samples. In contrast, the RSVP dataset encompassed three categories: non-target, boat (target 1), and geometric star (target 2) samples. Both datasets exhibit class imbalance, with the P300 speller dataset having a category ratio of 5:1 and the RSVP dataset showing a more pronounced class imbalance with a category ratio of 145:1:1. To reduce the effect of imbalance, we employed focal loss [38] with weights.

$$L_{FL} = -\sum_{k=1}^{N} w_k t_k \left(1 - p_k\right)^{\gamma} \log\left(p_k\right) \tag{5}$$

where p_k is the probability of class k, t_k is the truth label (a value of 0 or 1), w_k is the assigned weight for class k (see (6)), and γ is a hyperparameter to tune the loss of well-classified samples. We set γ to 2 as recommended in [38].

$$w_k = \frac{\max\left\{m_k\right\}_{k=1}^{N}}{m_k} \tag{6}$$

In addition, we used Mixup [39] defined in (7) for data augmentation. In Mixup, a random variable λ is sampled from a Beta distribution, $\lambda \sim$Beta (α, α), where α is the hyperparameter that controls the degree of interpolation. It picked two random samples x and $\tilde{x}$ and the corresponding one-hot labels y and $\tilde{y}$, and then simply added them together linearly to generate a new sample x_{mix} and label y_{mix}.

$$x_{mix} = \lambda x + \left(1 - \lambda\right)\tilde{x},\ y_{mix} = \lambda y + \left(1 - \lambda\right)\tilde{y} \tag{7}$$

We set up different configurations for the P300 speller and RSVP paradigm. Specifically, for the P300 speller, we set the number of xDAWN filters to 8 and the Mixup alpha value to 0.3. In contrast, for the RSVP paradigm, it was found that a different configuration yielded better performance. Hence, we set the number of xDAWN filters to 4 and the Mixup alpha value to 0.4 for the RSVP paradigm. Afterwards, for the P300 speller task, we trained the model in 80 epochs to achieve optimal results. For the RSVP task, where a dedicated test set was not available, we employed a three-fold cross-validation approach to evaluate the model. This involved dividing the dataset into three folds. Each fold was used as a validation set, while the remaining two folds were used for training. This process was repeated three times. For each subject, we computed the unweighted average recall (UAR) for all three folds during the cross-validation process and averaged the three validation score curves across epochs resulting in a single averaged validation score curve for each subject. Finally, the epoch with the highest UAR was selected from the averaged validation score curve as the RSVP result for each subject.

F. Symbol Decision Function for P300 Speller

In the case of the P300 Speller dataset, once the model is trained, the next step is to determine the position of the desired symbol based on the model's output. This involves detecting the row and column in which the symbol is located. Each symbol in the dataset is repeated 15 times, with each repetition consisting of 12 stimuli represented by a stimulus code value ranging from 1 to 12. Let $q_j^{(i)}$ denote the Softmax output (with temperature 10) of the output neuron, which represents P300 probability when the stimulus code value is j in the i^{th} repetition. $Q_j^{(z)}$ is the sum of those probabilities from the first to the z^{th} repetition under stimulus code value j.

$$Q_j^{(z)} = \sum_{i=1}^{z} q_j^{(i)} \tag{8}$$

Then, on the z^{th} repetition, we may identify the target symbol's column c and row r by:

$$c = argmax\, Q_j^{(z)},\ j \in [1, 6]\,;\ r = argmax\, Q_j^{(z)},\ j \in [7, 12] \tag{9}$$

G. Method of Evaluation

We used symbol recognition accuracy and ITR to evaluate the performance of different models in the P300 speller paradigm. We referred to the formula for calculating ITR in the i^{th} repetition in the paper [40], defined as follows:

$$ITR^{(i)} = \frac{60\left((1-A)\,log_2\frac{1-A}{G-1} + A log_2 A + log_2 G\right)}{2.5 + 2.1i} \tag{10}$$

where A is the symbol recognition accuracy, and G (i.e. 36) denotes the number of symbols presented in the P300 speller paradigm. For the RSVP paradigm, we adopted the unweighted average recall (UAR) [41] to evaluate the accuracy of the imbalanced dataset which is defined as follows:

$$\text{UAR} = \sum_{r=1}^{N} w_r \frac{c_r}{t_r} \tag{11}$$

where N denoted the total number of categories, w_r denotes the weight factor applied for each category which was currently set to [0.33, 0.33, 0.33], t_r denoted the number of images per category, and c_r was the number of correct predictions per category.

H. Models for Comparison

A series of prevailing models are proposed for comparison with the proposed one and we have provided a brief description of the characteristics of these models:

1) Spatially weighted fisher linear discriminant-PCA (SWFP) is a method designed for single-trial ERP detection. It utilizes fisher linear discriminant (FLD) to estimate spatial filters at each time point, which are then applied to an EEG sample for spatial filtering. PCA is then used for dimensionality reduction, with six principal components retained to explain over 70% of the variance [42].
2) Ensemble Support Vector Machines (ESVMs) [43] is a machine learning method that combines multiple SVM models to improve classification performance for P300 detection. This approach has been shown to achieve good classification.
3) DeepConvNet [24] is a deeper convolutional model for end-to-end EEG analysis, utilizing temporal convolution, spatial convolution and pooling operations and is a general approach for EEG decoding tasks in the BCI domain.
4) ShallowConvNet [24] is a simpler model with fewer convolutional layers and has also been successful in EEG decoding tasks in the BCI domain.
5) BN3 [13] is a deep learning model specifically designed for P300 detection, which uses Batch Normalization for EEG signals and a conventional CNN model for classification. This approach has been shown to achieve good classification accuracy.
6) CNN-RG-MINMA is a hybrid feature extraction method using CNN and Riemannian geometry for analyzing ERP data. It aims to enhance feature discriminability by extracting low-dimensional features with CNN and constructing a Riemannian graph [44].
7) ST-CapsNet is an enhanced version of ERP-CapsNet [45] that combined spatial-temporal attention and Capsule Network for improved P300 detection [40].
8) Discriminative Canonical Pattern Matching (DCPM) is a machine learning algorithm that is highly robust in detecting ERP components from different paradigms with excellent performance. It is especially useful when there is limited training data available [46].
9) EEGNet [23] is a lightweight end-to-end CNN network that incorporates temporal convolution, spatial convolution, separable convolution, and classification layers. It has demonstrated good robustness and has been widely used as a benchmark in EEG analysis.

III. Results

A. Performance of Symbol Recognition for P300 Speller

Dataset II consists of two subjects A and B, while Dataset II-b only contains one subject. Table I and Table II present the number of symbols correctly recognized per repetition by each model on Dataset II and Dataset II-b, respectively. Table III presents the results of paired t-tests (i.e., 30 pairs for Dataset II and 15 pairs for IIb) which compare symbol accuracy according to [45]. The authors did not report the performance of the P300 speller for DeepConvNet, ShallowConvNet, EEGNet and DCPM. Hence, we ran these models (except for DCPM as it is a traditional machine learning algorithm) with the same training strategy (see Section E Training). We also took the reported results of other models (SWFP, ESVMs, BN3, CNN-RG-MINMA and ST-CapsNet) for comparison. In Dataset II, our proposed model shows the ability to recognize more symbols with fewer repetition rounds. Notably, our model has better performance than other models ($p < 0.05$), except for the DeepConvNet ($p > 0.05$). In Dataset IIb, our model exhibits superior performance compared to SWFP, BN3 and DCPM, achieving statistically significant results ($p < 0.05$). While the other methods may show better performance than our model in Dataset IIb, these differences are not statistically significant ($p > 0.05$). These findings underscore the effectiveness of our model as a viable option for P300 signal detection, particularly in scenarios where there are limitations on the number of repetition rounds and a need for a balance between high accuracy and repetitions.

B. Effect of xDAWN Number on Symbol Recognition

Extensive experiments were conducted to investigate the effect of the number of xDAWN filters on symbol recognition rates, with results presented in Fig. 2. The averaged symbols under repetitions (ASUR) metric [40] was used to compare the performance of models with different numbers of xDAWN filters more intuitively.

$$\text{ASUR}_\text{k} = \frac{1}{k}\sum_{i=1}^{k} C_i, \tag{12}$$

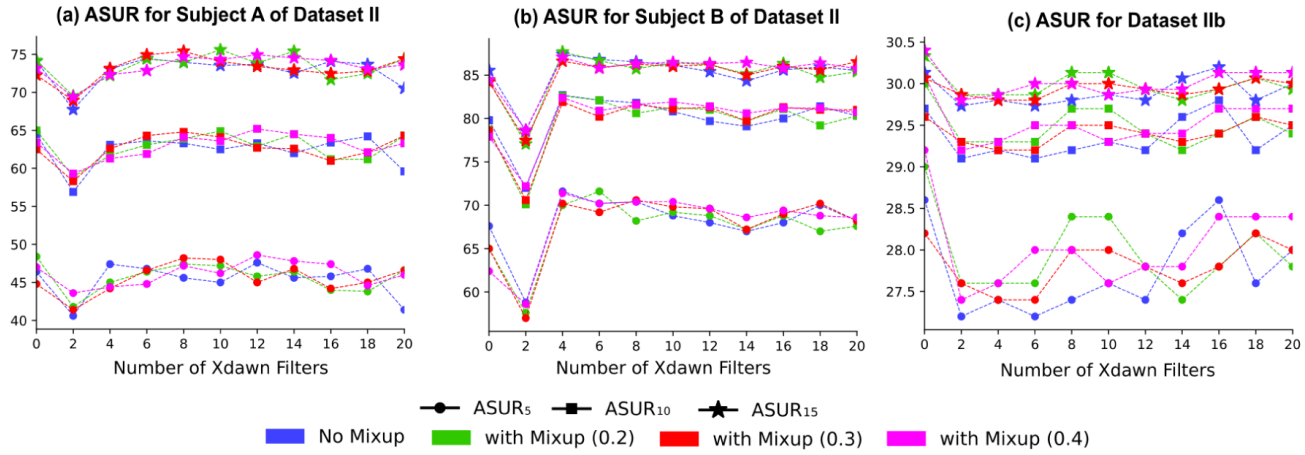

Fig. 2. Effect of different number of xDAWN filters on the symbol recognition rate. Subplots (a), (b), and (c) represent the symbol recognition rates of subjects A and B in Dataset II and Dataset IIb, respectively.

TABLE I
NUMBER OF SYMBOLS CORRECTLY RECOGNIZED PER REPETITION FOR EACH MODEL ON DATASET II

Model	Subject	Repetition														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
SWFP [40]	A	16	28	48	58	67	71	78	81	84	87	89	92	91	95	97
	B	40	60	69	71	80	82	86	89	89	92	92	93	92	94	94
	Mean	28	44	58.5	64.5	73.5	76.5	82	85	86.5	89.5	90.5	92.5	91.5	94.5	95.5
ESVMs [43]	A	17	34	55	64	69	80	74	79	85	89	92	94	96	98	99
	B	42	62	70	71	80	83	87	90	94	95	95	97	96	96	97
	Mean	29.5	48	62.5	67.5	74.5	81.5	80.5	84.5	89.5	92	**93.5**	95.5	96	**97**	**98**
DeepConvNet	A	19	31	47	61	68	76	83	83	80	88	88	93	92	95	96
	B	45	67	79	78	83	90	95	97	97	98	98	98	99	99	99
	Mean	32	49	63	69.5	75.5	**83**	**89**	**90**	88.5	**93**	93	95.5	95.5	**97**	97.5
ShallowConvNet	A	17	20	37	42	42	56	59	58	65	68	78	82	82	85	88
	B	48	61	68	71	82	85	91	93	96	98	98	96	96	96	97
	Mean	32.5	40.5	52.5	56.5	62	70.5	75	75.5	80.5	83	88	89	89	90.5	92.5
BN3 [13]	A	22	39	58	67	73	75	79	81	82	86	89	92	94	96	98
	B	47	59	70	73	76	82	84	91	94	95	95	95	94	94	95
	Mean	**34.5**	49	**64**	70	74.5	78.5	81.5	86	88	90.5	92	93.5	94	95	96.5
CNN-RG-MINMA [44]	A	22	33	49	57	63	66	72	77	79	86	89	94	92	94	97
	B	40	57	65	75	80	81	88	87	91	95	93	94	95	96	97
	Mean	31	45	57	66	71.5	73.5	80	82	85	90.5	91	94	93.5	95	97
ST-CapsNet [40]	A	18	31	53	56	68	79	82	85	84	88	89	92	92	95	98
	B	41	61	66	78	85	86	92	90	91	95	96	96	95	97	96
	Mean	29.5	46	59.5	67	**76.5**	82.5	87	87.5	87.5	91.5	92.5	94	93.5	96	97
DCPM	A	15	24	38	42	51	60	63	72	77	79	82	83	83	88	90
	B	29	48	55	66	69	73	83	83	87	91	94	93	93	94	97
	Mean	22	36	46.5	54	60	66.5	73	77.5	82	85	88	88	88	91	93.5
EEGNet	A	23	29	47	64	65	77	78	77	83	88	90	92	95	95	96
	B	39	58	70	71	81	89	91	93	94	96	96	98	97	96	97
	Mean	31	43.5	58.5	67.5	73	**83**	84.5	85	88.5	92	93	95	96	95.5	96.5
Our Method	A	19	42	52	61	67	75	76	81	85	90	92	96	99	98	98
	B	44	70	76	80	83	91	89	93	95	95	95	96	96	95	96
	Mean	31.5	**56**	**64**	**70.5**	75	**83**	82.5	87	**90**	92.5	**93.5**	**96**	**97.5**	96.5	97

where $ASUR_k$ stands for the average correctly recognized symbols per repetition when we take k repetitions into account. C_i stands for the correctly recognized symbols in the i^{th} repetition. The number of xDAWN filters ranged from 0 to 20 with an interval of 2, where 0 means xDAWN filter was excluded (i.e. EEGNet). The interval of the alpha value of Mixup was 0.2, 0.3, and 0.4. From Fig. 2, we observed that in Dataset II, subject A displayed an upward trend in the average symbol recognition rate as the number of xDAWN filters increased from 2 to 8, eventually reaching a stable level. Conversely, subject B exhibited a noticeable improvement in symbol recognition from 2 to 4 filters, followed by stabilization with a slight decline. Notably, when the xDAWN filter number was set to 8, both subjects A and B demonstrated a performance improvement. In addition to examining the influence of xDAWN filters, we also investigated the impact of the Mixup alpha value on the symbol recognition rate. We found that the Mixup alpha value had a relatively minor

TABLE II
NUMBER OF SYMBOLS CORRECTLY RECOGNIZED PER REPETITION FOR EACH MODEL ON DATASET IIB

Model	Repetition													
	1	2	3	4	5	6	7	8	10	11	12	13	14	15
SWFP [40]	21	26	27	29	29	30	30	**31**	**31**	**31**	**31**	**31**	**31**	**31**
ESVMs [43]	25	**29**	28	30	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**
DeepConvNet	24	**29**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**
ShallowConvNet	23	26	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**
BN3 [13]	24	23	27	28	29	30	30	30	**31**	**31**	**31**	**31**	**31**	**31**
CNN-RG-MINMA [44]	22	24	30	30	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**
ST-CapsNet [40]	**27**	**29**	30	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**
DCPM	18	22	27	29	28	30	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**
EEGNet	22	28	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**
Our Method	22	27	30	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**	**31**

TABLE III
PAIRED T-TEST COMPARATION ON P300 SPELLER

Our Method paired with	Dataset			
	II (subject A and B)		II-b	
	t-value	p-value	t-value	p-value
SWFP	6.5546	3.5e-7	2.9550	0.0104
ESVMs	2.3803	0.0241	-0.4588	0.6534
DeepConvNet	0.1386	0.8907	-1.7838	0.0961
ShallowConvNet	5.2202	1.4e-5	-0.5641	0.5816
BN3	2.1712	0.0382	2.1624	0.0484
CNN-RG-MINMAX-CNN	6.0166	1.5e-6	1.2929	0.2170
ST-CapsNet	2.2330	0.0334	-1.3331	0.2038
DCPM	9.4582	2.3e-10	2.6706	0.0183
EEGNet	2.6595	0.0126	-1.4676	0.1643

effect compared to xDAWN. Our results indicate that employing 8 xDAWN filters and a Mixup alpha value of 0.3 led to improved performance compared to scenarios where xDAWN and Mixup were not utilized. However, it is important to note that for Dataset IIb, the xDAWN filters resulted in a decrease in the symbol recognition rate. This finding emphasizes the need for careful consideration when selecting the number of xDAWN filters, taking into account the specific dataset and task at hand.

C. Performance of ITR for P300 Speller

The ITRs of each model on Dataset II and IIb were plotted in Fig. 3 to visually compare the speed of symbol spelling. For subject A of Dataset II, BN3 and ESVMs had higher ITRs than other models, while our method and DeepConvNet had faster ITR performance for subject B. Overall, our model had the best ITR performance on Dataset II, particularly in the second repetition where its ITR reached 17.22 bits/min. On Dataset IIb, ST-CapsNet has the best ITR performance, as shown in Fig. 3 (d). These findings suggest that our cascaded xDAWN-EEGNet model, DeepConvNet, ST-CapsNet can be suitable models for achieving high ITRs in P300 speller systems.

D. Performance of UAR for RSVP

To evaluate the model performance on RSVP tasks, a 3-fold cross-validation was implemented for each model. The training strategy (refer to Section E Training for details) was kept consistent across all models (except for SWFP, ESVMs, and DCPM which are traditional machine learning algorithms). The significant difference was analyzed by a paired t-test ($n = 16$). The UAR performance of each model at 5 Hz and 20 Hz RSVP is illustrated in Fig. 4 (a) and (b), respectively. The proposed method achieved the highest UAR performance for 5 Hz RSVP (proposed method: 0.8134 ± 0.0259; EEGNet: 0.7823 ± 0.0201, $p < 0.05$). Moreover, our method exhibited even more significant improvements (p-value < 0.0001) over the other models for both 5 Hz and 20 Hz RSVP. It is worth noting that DeepConvNet exhibits significant variance across multiple subjects. This can be primarily attributed to its large model parameters and sensitivity to hyperparameter selection, such as the γ value in the focal loss function, which leads to convergence difficulty (i.e., failed to learn information from multiple categories). SWFP performed the worst among all models. In conclusion, the cascade xDAWN-EEGNet model demonstrated the best UAR performance on RSVP tasks compared to the other models. These findings suggest that the proposed method shows the potential as an effective approach for analyzing EEG data in RSVP tasks.

E. Effect of xDAWN Number on UAR Performance

To show the effect of xDAWN filter number and Mixup on UAR, several experiments were conducted, and the results are shown in Fig.5. The xDAWN filter number interval ranged from 0 to 14 with an interval of 2, excluding the case of xDAWN filter of 2 due to its known poor effect as shown in Fig. 2. Fig. 5 (a) shows that as the number of xDAWN filters increases, the detection accuracy for 5 Hz RSVP EEG also increases. Furthermore, using larger alpha values for Mixup leads to larger UAR. In Fig. 5 (b), for the detection of 20Hz RSVP EEG, only when the number of xDAWN filters was 4 was there a slight improvement in performance over EEGNet. In other cases, performance decreased, but the improvement of UAR performance by Mixup was still significant. The results suggest that the number of xDAWN filters should be carefully chosen according to the specific RSVP characteristics. In addition, the findings indicate that our model is particularly effective for improving detection accuracy in RSVP tasks at low frequencies.

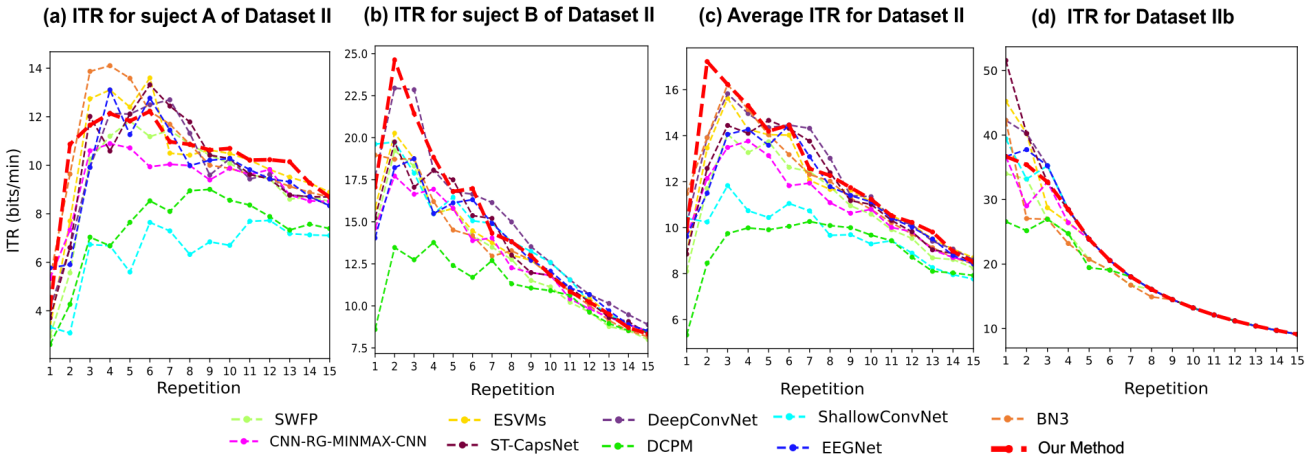

Fig. 3. The ITR of each model on Datasets II and IIb. Subplots (a) and (b) represent the ITR measured on Dataset II for subjects A and B, respectively; Subplot (c) represents the average of the ITR of subjects A and B; Subplot (d) represents the ITR of Dataset IIb.

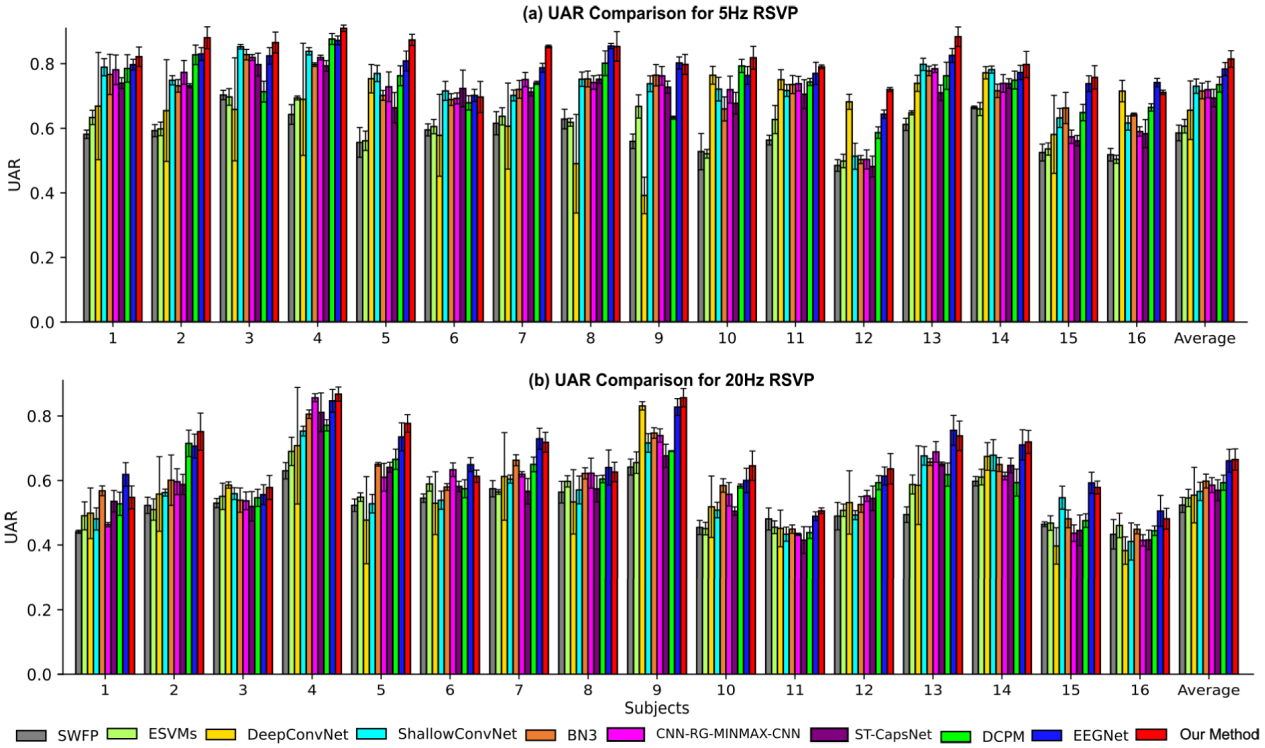

Fig. 4. Comparison of the UAR performance of each model on RSVP. Figures (a) and (b) represent the UAR performance of the models at 5 Hz and 20 Hz RSVP, respectively.

IV. Discussion

In this study, we proposed a cascade structure combining xDAWN and EEGNet for both the P300 speller and RSVP paradigms. Compared with other methods like DeepConvNet or mere EEGNet, the proposed method achieved better ITR with fewer repetition rounds for the P300 speller and gained high UAR performance in the RSVP paradigm.

A. P300 Speller and RSVP Paradigm

Selective attention, as measured by accuracy in an RSVP task, is closely linked to an individual's ability to update changing information over time and may also be connected to performance in P300 speller tasks [20], [47], [48]. This ability relies on attentional filtering capacity, which involves

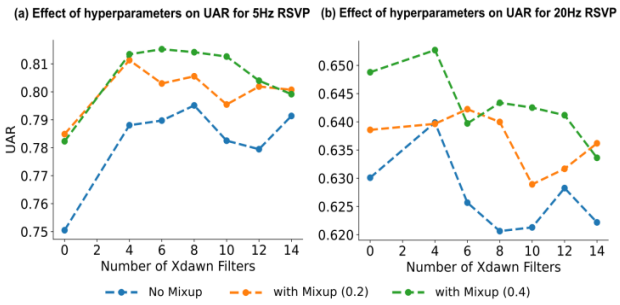

Fig. 5. Effect of different xDAWN filters and different Mixups on UAR. Figures (a) and (b) show the effects on UAR for 5 Hz and 20 Hz RSVP, respectively.

the ability to distinguish the object of interest from distractors and maintain this differentiation over time [21]. This concept

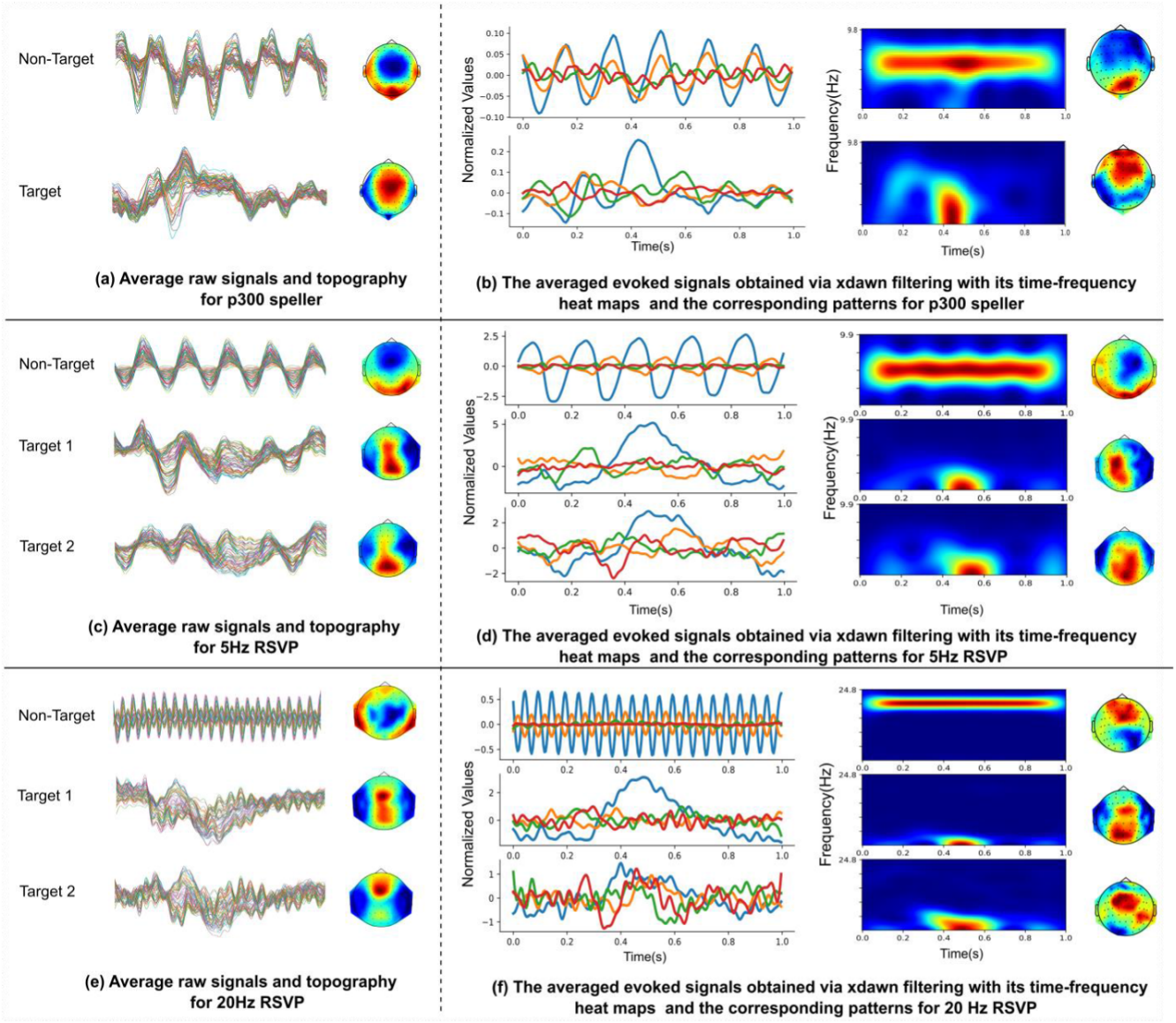

Fig. 6. Comparison of EEG signals of different paradigms. Figures (a), (c), (E) represent the averaged raw signal and its EEG topography; Figures (b), (d), (F) represent the evoked potentials, time-frequency thermograms and xDAWN-extracted patterns after xDAWN filtering.

aligns with the P300 speller, where individuals must filter out non-target rows/columns and focus on the target until a letter is identified. The similarity between selective attention and attentional filtering is reflected in EEG components. Moreover, a multi-feature predictor that includes multiple RSVP features has demonstrated that it can accurately predict P300 speller performance, outperforming single-feature predictors [21]. Furthermore, the P300 speller and RSVP paradigms exhibit shared characteristics, such as the utilization of low-frequency stimuli and the presence of a positive waveform observed within a specific time window following stimulus presentation. These findings demonstrate the commonality between these two paradigms and suggest that a uniform method could be applied across different BCI-related tasks.

B. xDAWN Could Enhance the P300 Pattern

In the proposed approach, we used xDAWN spatial filtering to improve the SNR of the raw EEG signals before feeding them into EEGNet, thus providing a process for incorporating prior domain knowledge into the model. To visualize the EEG signals evoked by different paradigm stimuli, we first plotted the EEG signals evoked by P300 speller, 5 Hz RSVP, and 20 Hz RSVP, along with their EEG spatial distribution (0.25-0.6 s), as is shown in Fig. 6 (a), (c) and (e). From these figures, it can be seen that the spatial distribution of the EEG signals evoked by the target sample in the P300 speller and the EEG signals evoked by the target 1 and target 2 samples in the 5 Hz and 20 Hz RSVP are similar. Afterward, we utilized the xDAWN spatial filter to enhance the EEG signal evoked by the target stimulus. For uniformity, we set the xDAWN spatial filter to 4 for visualization. The evoked signals of each paradigm after the xDAWN filter, the time-frequency thermograms, and their extracted EEG distribution patterns are shown in Fig. 6 (b), (d) and (f). From these plots, it can be observed that, compared to the original signal, the evoked signal corresponding to the target category in the P300 speller and RSVP has a

clear positive wave at 0.33-0.5 s and 0.4-0.6 s, although the amplitudes of the evoked potentials are different. Such characteristics together with an inverse relationship between the target and non-target stimulus imply that BCIs with different paradigms can use uniform methods for different classification applications. The xDAWN spatial filter enhances the P300 component in the original EEG in time and improves the SNR, making the P300 signal more easily to be detected by the classifier.

C. Compare With Other Methods

The key for ERP-based BCI is to distinguish ERP from the background of EEG signals as ERPs have many components, and they are weak and can be influenced by many factors. Linear discriminative analysis (LDA) is a traditional method for the detection of ERPs. However, such a method cannot handle various components of ERPs [42], [49], [50]. Furthermore, as mentioned before, DCPM is a robust method that has excellent performance for the detection of ERPs from various paradigms [46]. To clearly illustrate the difference between DCPM and the proposed method, we compare our work with [46]. In [46], only the classification performance of various models was compared across different ERP paradigms in a single trial. However, important metrics such as symbol recognition and ITR performance in the P300 speller paradigm were not taken into consideration. These metrics are crucial for evaluating the effectiveness of a BCI system. In contrast, we conducted a comprehensive evaluation of our method and compared it with other models on standard BCI competition P300 speller datasets. We considered both ITR and symbol recognition performance, and our results showed that our method outperformed DCPM in both aspects, demonstrating superior performance.

Furthermore, EEGNet has been proven its effectiveness in BCI competitions [41], [51], [52]. Notably, the EEGNet exhibits exceptional generalization capabilities, displaying the ability to perform well on diverse datasets, and also exhibits robustness to noise, enabling reliable performance even in the presence of noisy input signals [23]. Besides, EEGNet stands out for its computational efficiency, allowing for efficient real-time processing. The effectiveness and simplicity of its architecture make it an optimal choice as our basic model. Building upon our previous study [34], which utilized the combination of xDAWN with EEGNet and achieved the second place in the RSVP competition at the BCI Controlled Robot Contest during the 2021 World Robot Contest, we have further extended our previous work [34]. In this extended work, we focus on analyzing the impact of varying xDAWN filter numbers on RSVP classification results, a crucial factor that was not previously investigated [34]. Moreover, we have explored the applicability of our algorithm to the P300 speller and investigated the effectiveness of Mixup data augmentation techniques for both the P300 speller and RSVP tasks. Through our investigations, we discovered that selecting the appropriate number of xDAWN filters and Mixup alpha value can enhance the performance of our model. These findings highlight the capability of our algorithm to address a wider range of BCI applications effectively.

V. Conclusion

This study introduces a cascade structure for unified detection of visual-evoked related potentials. Evaluated on Dataset II of the BCI Competition III, our method exhibited better symbol recognition accuracy and achieved a higher ITR compared to the compared methods, especially for reaching 17.22 bits/min in the second repetition round. Furthermore, the results demonstrated that our method was superior to the compared models in terms of the UAR on the RSVP paradigm (0.8134 ± 0.0259 at 5 Hz and 0.6527 ± 0.0321 at 20 Hz). In addition, we observed that applying xDAWN filters to raw evoked EEG signals effectively enhances the P300 pattern, which partially explains why our method has better performance on both the P300 speller and RSVP paradigms. These results underscored the effectiveness of the proposed cascade structure for detecting P300-related signals across both P300 speller and RSVP paradigms.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, Jun. 2002, doi: 10.1016/s1388-2457(02)00057-3.

[2] U. Chaudhary, N. Birbaumer, and A. Ramos-Murguialday, "Brain–computer interfaces for communication and rehabilitation," *Nature Rev. Neurol.*, vol. 12, no. 9, pp. 513–525, Sep. 2016, doi: 10.1038/nrneurol.2016.113.

[3] N. Birbaumer and L. G. Cohen, "Brain–computer interfaces: Communication and restoration of movement in paralysis," *J. Physiol.*, vol. 579, no. 3, pp. 621–636, Mar. 2007, doi: 10.1113/jphysiol.2006.125633.

[4] N. Birbaumer et al., "A spelling device for the paralysed," *Nature*, vol. 398, no. 6725, pp. 297–298, Mar. 1999, doi: 10.1038/18581.

[5] H. Cecotti and A. Graser, "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, doi: 10.1109/TPAMI.2010.125.

[6] S. G. Mason, A. Bashashati, M. Fatourechi, K. F. Navarro, and G. E. Birch, "A comprehensive survey of brain interface technology designs," *Ann. Biomed. Eng.*, vol. 35, no. 2, pp. 137–169, Feb. 2007, doi: 10.1007/s10439-006-9170-0.

[7] J. Leoni, S. C. Strada, M. Tanelli, A. Brusa, and A. M. Proverbio, "Single-trial stimuli classification from detected P300 for augmented brain–computer interface: A deep learning approach," *Mach. Learn. with Appl.*, vol. 9, Sep. 2022, Art. no. 100393, doi: 10.1016/j.mlwa.2022.100393.

[8] B. Z. Allison, A. Kübler, and J. Jin, "30+ years of P300 brain–computer interfaces," *Psychophysiology*, vol. 57, no. 7, pp. 1–18, Jul. 2020, doi: 10.1111/psyp.13569.

[9] Á. Fernández-Rodríguez, A. Darves-Bornoz, F. Velasco-Álvarez, and R. Ron-Angevin, "Effect of stimulus size in a visual ERP-based BCI under RSVP," *Sensors*, vol. 22, no. 23, p. 9505, Dec. 2022, doi: 10.3390/s22239505.

[10] S. H. Patel and P. N. Azzam, "Characterization of N200 and P300: Selected studies of the event-related potential," *Int. J. Med. Sci.*, vol. 2, no. 4, pp. 147–154, 2005, doi: 10.7150/ijms.2.147.

[11] J. Polich and E. Donchin, "P300 and the word frequency effect," *Electroencephalogr. Clin. Neurophysiol.*, vol. 70, no. 1, pp. 33–45, Jul. 1988, doi: 10.1016/0013-4694(88)90192-7.

[12] J. C. Henry, "Electroencephalography: Basic principles, clinical applications, and related fields, fifth edition," *Neurology*, vol. 67, no. 11, p. 2092, Dec. 2006, doi: 10.1212/01.wnl.0000243257.85592.9a.

[13] M. Liu, W. Wu, Z. Gu, Z. Yu, F. Qi, and Y. Li, "Deep learning based on batch normalization for P300 signal detection," *Neurocomputing*, vol. 275, pp. 288–297, Jan. 2018, doi: 10.1016/j.neucom.2017.08.039.

[14] P. Du, P. Li, L. Cheng, X. Li, and J. Su, "Single-trial P300 classification algorithm based on centralized multi-person data fusion CNN," *Frontiers Neurosci.*, vol. 17, pp. 1–11, Feb. 2023, doi: 10.3389/fnins.2023.1132290.

[15] S. Lees et al., "A review of rapid serial visual presentation-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 2, Apr. 2018, Art. no. 021001, doi: 10.1088/1741-2552/aa9817.

[16] S. Mathan et al., "Rapid image analysis using neural signals," in *Proc. CHI Extended Abstr. Hum. Factors Comput. Syst.*, Apr. 2008, pp. 3309–3314, doi: 10.1145/1358628.1358849.

[17] A. D. Gerson, L. C. Parra, and P. Sajda, "Cortical origins of response time variability during rapid discrimination of visual objects," *NeuroImage*, vol. 28, no. 2, pp. 342–353, Nov. 2005, doi: 10.1016/j.neuroimage.2005.06.026.

[18] K. Won, M. Kwon, M. Ahn, and S. C. Jun, "EEG dataset for RSVP and P300 speller brain–computer interfaces," *Sci. Data*, vol. 9, no. 1, pp. 1–11, Jul. 2022, doi: 10.1038/s41597-022-01509-w.

[19] R. Verleger, K. Śmigasiewicz, and F. Möller, "Mechanisms underlying the left visual-field advantage in the dual stream RSVP task: Evidence from N2pc, P3, and distractor-evoked VEPs," *Psychophysiology*, vol. 48, no. 8, pp. 1096–1106, Aug. 2011, doi: 10.1111/j.1469-8986.2011.01176.x.

[20] A. Riccio et al., "Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis," *Frontiers Hum. Neurosci.*, vol. 7, pp. 1–9, 2013, doi: 10.3389/fnhum.2013.00732.

[21] K. Won, M. Kwon, S. Jang, M. Ahn, and S. C. Jun, "P300 speller performance predictor based on RSVP multi-feature," *Frontiers Hum. Neurosci.*, vol. 13, pp. 1–14, Jul. 2019, doi: 10.3389/fnhum.2019.00261.

[22] Á. Fernández-Rodríguez, M. T. Medina-Juliá, F. Velasco-Álvarez, and R. Ron-Angevin, "Preliminary results using a P300 brain–computer interface speller: A possible interaction effect between presentation paradigm and set of stimuli," in *Proc. Int. Work-Conf. Artif. Neural Netw.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, 2019, pp. 371–381, doi: 10.1007/978-3-030-20521-8_31.

[23] 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, doi: 10.1088/1741-2552/aace8c.

[24] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017, doi: 10.1002/hbm.23730.

[25] S. Roychowdhury, M. Diligenti, and M. Gori, "Regularizing deep networks with prior knowledge: A constraint-based approach," *Knowl.-Based Syst.*, vol. 222, Jun. 2021, Art. no. 106989, doi: 10.1016/j.knosys.2021.106989.

[26] T. Rahne, H. von Specht, and R. Mühler, "Sorted averaging—Application to auditory event-related responses," *J. Neurosci. Methods*, vol. 172, no. 1, pp. 74–78, Jul. 2008, doi: 10.1016/j.jneumeth.2008.04.006.

[27] H. Wang et al., "Performance enhancement of P300 detection by multiscale-CNN," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021, doi: 10.1109/TIM.2021.3067943.

[28] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "XDAWN algorithm to enhance evoked potentials: Application to brain–computer interface," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 8, pp. 2035–2043, Aug. 2009, doi: 10.1109/TBME.2009.2012869.

[29] H. Cecotti, M. P. Eckstein, and B. Giesbrecht, "Single-trial classification of event-related potentials in rapid serial visual presentation tasks using supervised spatial filtering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 11, pp. 2030–2042, Nov. 2014, doi: 10.1109/TNNLS.2014.2302898.

[30] A. Barachant, "MEG decoding using Riemannian geometry and unsupervised classification," Grenoble Univ., Grenoble, France, Tech. Rep., 2014. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=753be0d9cf14014b2e6ac0f9e2a861d9b1468461

[31] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain–computer interface classification by Riemannian geometry," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 4, pp. 920–928, Apr. 2012, doi: 10.1109/TBME.2011.2172210.

[32] H. He and D. Wu, "Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 2, pp. 399–410, Feb. 2020, doi: 10.1109/TBME.2019.2913914.

[33] H. Wu and D. Wu, "Review of training-free event-related potential classification approaches in the world robot contest 2021," *Brain Sci. Adv.*, vol. 8, no. 2, pp. 82–98, Jun. 2022, doi: 10.26599/bsa.2022.9050001.

[34] H. Zhang, Z. Wang, Y. Yu, H. Yin, C. Chen, and H. Wang, "An improved EEGNet for single-trial EEG classification in rapid serial visual presentation task," *Brain Sci. Adv.*, vol. 8, no. 2, pp. 111–126, Jun. 2022, doi: 10.26599/bsa.2022.9050007.

[35] T. Grootswagers, A. K. Robinson, and T. A. Carlson, "The representational dynamics of visual objects in rapid serial visual processing streams," *NeuroImage*, vol. 188, pp. 668–679, Mar. 2019, doi: 10.1016/j.neuroimage.2018.12.046.

[36] A. Barachant, "pyRiemann documentation," Tech. Rep., 2018. [Online]. Available: https://pyriemann.readthedocs.io/_/downloads/en/stable/pdf/

[37] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–12.

[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[39] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "MixUp: Beyond empirical risk minimization," in *Proc. 6th Int. Conf. Learn. Represent.*, 2018, pp. 1–13.

[40] Z. Wang, C. Chen, J. Li, F. Wan, Y. Sun, and H. Wang, "ST-CapsNet: Linking spatial and temporal attention with capsule network for P300 detection improvement," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 991–1000, 2023, doi: 10.1109/TNSRE.2023.3237319.

[41] Z. Wang, H. Zhang, Z. Ji, Y. Yang, and H. Wang, "A review of deep learning methods for cross-subject rapid serial visual presentation detection in world robot contest 2022," *Brain Sci. Adv.*, vol. 9, no. 3, pp. 195–209, Sep. 2023, doi: 10.26599/bsa.2023.9050013.

[42] G. F. Alpert, R. Manor, A. B. Spanier, L. Y. Deouell, and A. B. Geva, "Spatiotemporal representations of rapid visual target detection: A single-trial EEG classification algorithm," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 8, pp. 2290–2303, Aug. 2014, doi: 10.1109/TBME.2013.2289898.

[43] S. Kundu and S. Ari, "P300 detection using ensemble of SVM for brain–computer interface application," in *Proc. 9th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2018, pp. 1–5, doi: 10.1109/ICCCNT.2018.8493903.

[44] S. N. Aghili and A. Erfanian, "A P300-based speller design using a MINMAX Riemannian geometry scheme and convolutional neural network," *IEEE Access*, vol. 11, pp. 98633–98652, 2023, doi: 10.1109/ACCESS.2023.3313260.

[45] R. Ma, T. Yu, X. Zhong, Z. L. Yu, Y. Li, and Z. Gu, "Capsule network for ERP detection in brain–computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 718–730, 2021, doi: 10.1109/TNSRE.2021.3070327.

[46] X. Xiao, M. Xu, J. Jin, Y. Wang, T.-P. Jung, and D. Ming, "Discriminative canonical pattern matching for single-trial classification of ERP components," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 8, pp. 2266–2275, Aug. 2020, doi: 10.1109/TBME.2019.2958641.

[47] V. Di Lollo, J.-I. Kawahara, S. M. Shahab Ghorashi, and J. T. Enns, "The attentional blink: Resource depletion or temporary loss of control?" *Psychol. Res.*, vol. 69, no. 3, pp. 191–200, Jan. 2005, doi: 10.1007/s00426-004-0173-x.

[48] C. Kranczioch, S. Debener, A. Maye, and A. K. Engel, "Temporal dynamics of access to consciousness in the attentional blink," *NeuroImage*, vol. 37, no. 3, pp. 947–955, Sep. 2007, doi: 10.1016/j.neuroimage.2007.05.044.

[49] U. Hoffmann, J.-M. Vesin, T. Ebrahimi, and K. Diserens, "An efficient P300-based brain–computer interface for disabled subjects," *J. Neurosci. Methods*, vol. 167, no. 1, pp. 115–125, Jan. 2008, doi: 10.1016/j.jneumeth.2007.03.005.

[50] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components—A tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814–825, May 2011, doi: 10.1016/j.neuroimage.2010.06.048.

[51] J. Luo, Q. Mao, Y. Wang, Z. Shi, and X. Hei, "Algorithm contest of calibration-free motor imagery BCI in the BCI controlled robot contest in world robot contest 2021: A survey," *Brain Sci. Adv.*, vol. 8, no. 2, pp. 127–141, Jun. 2022, doi: 10.26599/bsa.2022.9050011.

[52] C. Tang, Y. Li, and B. Chen, "Comparison of cross-subject EEG emotion recognition algorithms in the BCI controlled robot contest in world robot contest 2021," *Brain Sci. Adv.*, vol. 8, no. 2, pp. 142–152, Jun. 2022, doi: 10.26599/bsa.2022.9050013.