A Hybrid BCI Integrating EEG and Eye-Tracking for Assisting Clinical Communication in Patients With Disorders of Consciousness

Ziyi Yi, Jiahui Pan, *Member, IEEE*, Zerong Chen, Dehua Lu, Honghua Cai, Jianfeng Li, and Qiuyou Xie

***Abstract*— Assessing communication abilities in patients with disorders of consciousness (DOCs) is challenging due to limitations in the behavioral scale. Electroencephalogram-based brain-computer interfaces (BCIs) and eye-tracking for detecting ocular changes can capture mental activities without requiring physical behaviors and thus may be a solution. This study proposes a hybrid BCI that integrates EEG and eye tracking to facilitate communication in patients with DOC. Specifically, the BCI presented a question and two randomly flashing answers (yes/no). The subjects were instructed to focus on an answer. A multimodal target recognition network (MTRN) is proposed to detect P300 potentials and eye-tracking responses (i.e., pupil constriction and gaze) and identify the target in real time. In the MTRN, the dual-stream feature extraction module with two independent multiscale convolutional neural networks extracts multiscale features from multimodal data. Then, the multimodal attention strategy adaptively extracts the most relevant information about the target from multimodal data. Finally, a prototype network is designed as a classifier to facilitate small-sample data classification. Ten healthy individuals, nine DOC patients and one LIS patient were included in this study. All healthy subjects achieved 100% accuracy. Five patients could communicate with our BCI, with 76.1±7.9% accuracy. Among them, two patients who were noncommunicative on the behavioral scale exhibited communication ability via our BCI. Additionally, we assessed the performance of unimodal BCIs and compared MTRNs with other methods. All the results suggested that our BCI can yield more sensitive outcomes than the CRS-R and can serve as a valuable communication tool.**

***Index Terms*— Hybrid brain–computer interface (BCI), disorder of consciousness (DOC), BCI communication, P300, eye-tracking.**

I. Introduction

DISORDERS of consciousness (DOCs) manifest as varying degrees of arousal states and abnormalities in cognition, including coma, vegetative state/unresponsive wakefulness syndrome (VS/UWS), and minimally conscious state (MCS). Patients with VS/UWS may emerge from a coma without consciousness [1]. Patients with MCS retain some level of consciousness and may exhibit reproducible non-reflexive movements (e.g., visual tracking) [2]. Patients with locked-in syndrome (LIS) maintain near-normal cognitive abilities but have significant sensory and motor deficits [3]. LIS is not a DOC but may be mistaken for it. Some studies have shown that LIS is misdiagnosed as VS/UWS at a rate of approximately 10% [4]. Additionally, some LIS patients may suffer additional brain damage beyond the brainstem, leading to cognitive deficits [5]. Misdiagnosing the consciousness of a patient may have significant medical and ethical ramifications. The command-following ability is a key diagnostic marker for DOC [6]. Currently, doctors predominantly use the Coma Recovery Scale-Revised (CRS-R) to clinically diagnose DOC patients [7], [8]. In the CRS-R, physicians assess patients' functions by scoring their behavioral responses to various stimuli. However, this method is relatively subjective, lacks quantifiable metrics, and harbors inherent contradictions that are difficult to resolve. These contradictions may lead to false negatives in patients with residual consciousness, with a misdiagnosis rate of approximately 40% [9], [10], [11]. Obviously, the behavior-based diagnostic method cannot meet the needs of accurately assessing consciousness levels. Exploring non-behavior-based objective methods to assist in the diagnosis of DOC patients is necessary and urgent. Additionally, the ethical justification for life-sustaining treatment (LST) in these patients is a matter of intense ethical and social debate. Some neurologists mostly favor limiting LST, but their attitudes

Manuscript received 1 April 2024; revised 13 June 2024 and 15 July 2024; accepted 22 July 2024. Date of publication 29 July 2024; date of current version 5 August 2024. This work was supported in part by Science and Technology Innovation (STI) 2030-Major Projects under Grant 2022ZD0208900, in part by the National Natural Science Foundation of China under Grant 62076103 and Grant 62306120, and in part by the Major Projects of Colleges and Universities in Guangdong Province under Grant 2023ZDZX2021. *(Ziyi Yi and Jiahui Pan contributed equally to this work.) (Corresponding authors: Jiahui Pan; Qiuyou Xie.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Medical Ethics Committee of Zhujiang Hospital, Southern Medical University under Application No. 2023-KY-174-01.

Ziyi Yi, Jiahui Pan, Dehua Lu, and Honghua Cai are with the School of Artificial Intelligence, South China Normal University, Guangzhou 510631, China (e-mail: panjiahui@m.scnu.edu.cn).

Zerong Chen and Qiuyou Xie are with the Department of Rehabilitation, Zhujiang Hospital, Guangzhou 510260, China, and also with the School of Rehabilitation Sciences, Southern Medical University, Guangzhou 510282, China (e-mail: xqy7018@163.com).

Jianfeng Li is with Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Foshan 528225, China.

Digital Object Identifier 10.1109/TNSRE.2024.3435016

toward LST vary greatly depending on the patient's willingness to receive treatment [12]. Improving patients' living standards and reducing the nursing burden on their families also require further exploration. The development of a communication system for DOC patients has significant implications for patients, their families, and doctors.

Brain computer interfaces (BCIs), capable of detecting brain activities via electroencephalography (EEG), may thus offer a reliable method for assessing DOC patients. Xiao et al. [13] used a BCI for simulating sound localization assessment in the CRS-R and successfully identified sound localization ability in 11 patients. A study even revealed [14] that over 75% of DOC patients showed evidence of command following via the BCI. The BCI not only exhibited significant potential in detecting residual cognition in patients but also holds promise as a tool for communication. Lulé et al. [15] presented a four-choice BCI to assess the responses of 18 DOC patients to commands and observed that one patient could communicate via the BCI (accuracy: 60%). Wang et al. [16] introduced an audio-visual BCI to augment the assessment of communication ability in the CRS-R. In their study, patients communicated by performing an active task (counting the number of stimuli for the answer). In addition to visual and auditory paradigms, vibrotactile-based BCIs can also facilitate communication. Guger et al. [17] instructed 12 DOC patients to communicate by counting the vibrating stimuli on their left and right wrists. The patients chose their answers by counting the stimuli on their wrists denoting "yes" or "no". The results showed that 2 patients could communicate via the vibrotactile BCI, achieving 70% accuracy. Notably, the above BCIs all rely on a single modality, which may limit the information available for decision-making. One improvement option is to combine multiple modalities to gather more information for judgment. Although Huang et al. [18] proposed a communication BCI combining two types of brain signals and demonstrated the superiority of the hybrid system, drawbacks remain. Low-frequency visual stimuli that induce steady-state visual evoked potentials (SSVEPs) carry the risk of triggering seizures in patients [19], [20]. Moreover, these BCIs require continuous cognitive tasks, which are challenging for patients with limited attention spans. Communication is a basic ability that DOC patients generally lack and urgently need. However, limited research exists on communication, with only 24% of BCIs for DOCs being used for communication [21]. BCIs for communication with DOC patients are still in their infancy, and designing a stable and easy-to-use multimodal communication BCI is promising for improving this situation.

Previous studies on detecting and communicating with DOC patients have primarily focused on traditional machine learning. For instance, all the above studies on BCI [13], [15], [16], [17], [18] used support vector machines (SVMs) to process the data. This approach may not effectively mine deeper information from the data. Deep learning methods may offer a solution. The neural networks BN3 [22] and SCNN [23] for P300 detection and the classical network EEGNet [24] for processing EEG signals have shown satisfactory performance in EEG detection tasks. However, their accuracy depends on the quantity and quality of available data, which may not be suitable for DOC patients. DOC patients are prone to fatigue and inattention, making it difficult to collect sufficient usable data from them. Despite these methods employing batch normalization and dropout [25] strategies to improve generalization, overfitting may still be unavoidable on patient datasets with limited data. Different components may coexist in EEG data over a period of time, reflecting various stages of neural activity [26], [27]. This suggests that information captured from various scales of EEG may contain distinct components and contextual features, yet few studies have noted this. In addition, most multimodal BCIs for DOC patients adopt decision-level fusion strategies [18], [28], which may struggle to integrate rich semantic information from various modalities and depict the associations among them. The development of reliable BCI techniques for clinical use in DOC patients remains a challenge.

In addition to communication ability, visual ability is an important diagnostic indicator of DOC. However, visual assessment (e.g., visual tracking, visual localization) in the CRS-R mainly relies on physicians to make manual discriminations, which lacks reliability. Small eye movements of patients may go unnoticed by physicians. Eye-tracking technology, which detects various eye movements (e.g., gaze localization, saccadic movements, and pupil changes), may be a useful adjunct. Eye movements are influenced by the subject's cognitive processes, behavior, and external stimuli. For example, pupil size oscillates in response to the luminance of visual stimuli (pupillary light reflex; PLR) [29]: higher luminance leads to pupil constriction, while lower luminance results in pupil dilation. The use of eye movement responses can also facilitate stable communication applications.

Mathôt et al. [30] explored the possibility of human-computer interaction through the response of pupillary oscillations following the allocation of attention. Sato and Nakatani [31] utilized the PLR to control an external device with 83.4% accuracy. Stoll et al. [32] tested the possibility of pupil response as a tool for communication. The results showed significantly greater decoding performance than chance (50%) in 3 LIS patients. Villalobo et al. [33] successfully communicated with an LIS patient via the pupillary modulation response. The relatively poor results of Stoll and Villalobos compared to those of other studies may be attributed to the inclusion of patients as experimental subjects, and it also implies that eye-tracking signals alone cannot precisely categorize data from patients. Recent studies [34], [35] have shown the potential for enhancing system performance by fusing EEG data and eye-tracking data. Mannan et al. [34] designed a hybrid speller that incorporates SSVEP and eye-tracking signals, achieving 90.35% accuracy. However, several shortcomings remain. First, multimodal BCIs that integrate eye tracking and EEG primarily focus on patients with motor impairments, neglecting DOC patients whose conditions are more specific. Second, these BCIs mostly involve healthy subjects and lack tests on patient populations. Although these findings increase confidence in communication by fusing EEG and eye-tracking

signals, the application of these multimodal BCIs in DOC has yet to be verified.

Given the above limitations, we propose a hybrid BCI system that combines eye-tracking signals and EEG signals to assist physicians in clinical communication assessment and to address the fundamental communication needs of patients. Patients engage in binary communication via attention and gaze. Specifically, the system presented patients with a situation question and displayed 'yes' and 'no' options beneath it. Randomly flashing options and corresponding audio files were used as visual and auditory stimuli. Patients were instructed to focus on the target stimulus. P300 and eye-tracking signals were collected from the flashing options. By detecting and analyzing these signals, the system can identify patient choices. To enhance feature extraction and maximize the benefits of diverse features, we propose a multimodal target recognition network (MTRN) consisting of three modules. The dual-stream feature extraction module employs multiple convolutions with varying receptive fields to capture multiscale features from EEG and eye-tracking data. The multimodal attention module uses a cross-channel soft attention mechanism to adaptively capture valuable information from multimodal data. Finally, considering the data characteristics of patients, we introduced a prototype network to classify small-sample data. Ten healthy subjects and ten patients (1 VS, 8 MCS, and 1 LIS) participated in this study. The experimental results demonstrated that the system enhances the likelihood of patients providing evidence of residual brain function to the examiner and can serve as an adjunctive tool for communicating with DOC patients. The innovations of this study can be summarized as follows:

1. We designed a novel BCI paradigm that integrates P300 potential and eye tracking. P300 detection and eye-tracking detection are irreplaceable and mutually complementary, and multimodal fusion facilitates clinical binary communication.

2. We propose a new multimodal target recognition network. The dual-stream feature extraction module maximizes the extraction of rich multiscale features of EEG and eye-tracking signals. The multimodal attention module integrates the outputs from the dual streams, adaptively emphasizing important features while eliminating redundancy. To address overfitting and enhance small-sample data classification, we utilize a prototype network based on the cosine distance as the classifier.

3. We developed a hybrid BCI system to facilitate the clinical communication of DOC patients. The experimental results demonstrated the feasibility and efficiency of our hybrid BCI system. To our knowledge, this study is the first attempt to communicate with DOC patients using a hybrid BCI based on P300 potential and eye tracking.

II. Paradigms and Methods

A. Data Acquisition System

This study utilized a 32-electrode EEG cap (including two reference electrodes) based on the International 10-20 system and a SynAmps^2 amplifier (Compumedics, Neuroscan, Inc., Australia) to record scalp EEG signals at a sampling frequency of 250 Hz. The reference electrode was placed on the right mastoid. To remove the noise, 50 Hz trap filtering was performed. We used eye-tracking glasses (SMI ETG 2w, Germany) to record eye movement data. The scene view and eye view were sampled at 60 Hz with resolutions of 1280×960 and 320×240, respectively.

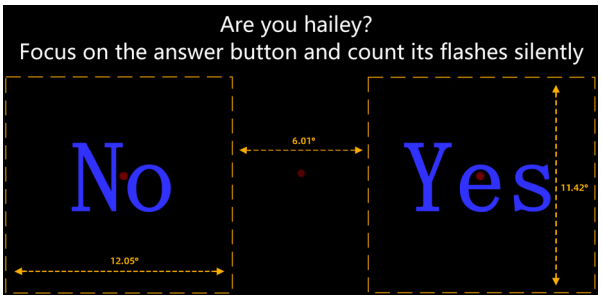

Fig. 1. The GUI of the BCI system.

B. Graphical User Interface and BCI Paradigm

1) GUI: Fig. 1 illustrates the graphical user interface (GUI) of our hybrid BCI system. A question and an instruction are displayed at the top of the screen. All questions were designed based on the CRS-R Communication subscale. Examples include "Is your name hailey?", "Is this a cup/comb?", and "Is the doctor clapping now?". The instructions instructed patients to focus on the correct answer (target). Below the instructions, the text blocks "Yes" and "No" are displayed on the left and right sides, respectively. Each text block is a rectangle with dimensions of 12.05° in width and 11.42° in height. The two text blocks are spaced approximately 6.01° apart in the field of view. To assist participants in gaze, a semitransparent red dot is positioned at the center of each text block and between them.

2) Stimulation Mode: Two text blocks served as stimulus sources to induce P300 signals and eye-tracking signals in the form of random flashes. Upon stimulus presentation, the text color of the corresponding text block changed from blue to green, while the background color shifted from black to white, with luminance increasing from 1.27 cd/m^2 to 90.76 cd/m^2. Simultaneously, the audio corresponding to the word was played at 65 dB. Variations in the luminance of the text block can alter the subject's pupil size, and the magnitude of this alteration increases with focused attention. Randomized audiovisual stimuli can evoke P300 potentials. Text blocks chosen by subjects are identifiable by detecting P300 potentials and analyzing eye-tracking data.

3) BCI Paradigm: The BCI paradigm is shown in Fig. 2. The trial started with a question and instruction stage. Within this phase, the computer displayed and vocalized a predetermined question and an instruction. For example, "Is this a cup? Focus on the answer button and count its flashes silently". Following an 8-second question and instruction period, two text blocks flashed in random order, with accompanying corresponding audio broadcasts. Each flash lasted 300 ms, separated by an interval of 700 ms. Consequently, the duration of a stimulation

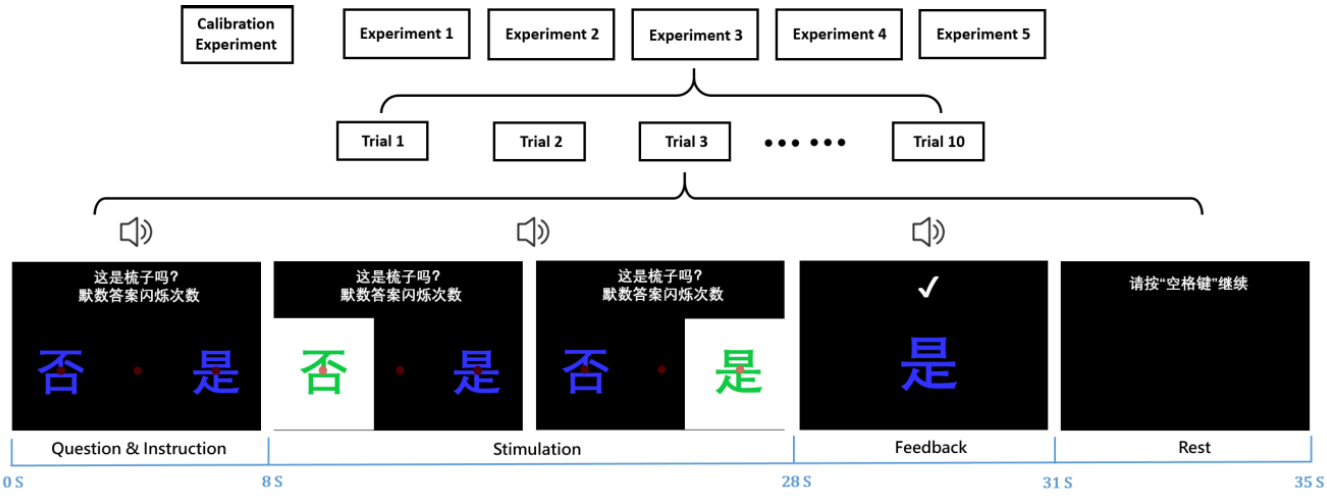

Fig. 2. The BCI paradigm for the online communication experiment.

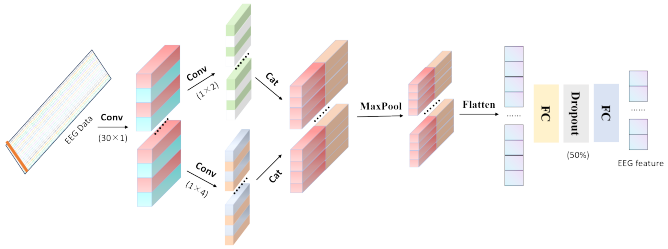

Fig. 3. The architecture of the EEG feature extraction module.

round totals 2000 ms, defined as a cycle wherein both the left and right text blocks flash once. The duration of the stimulus stage was set to 20 s. A total of 10 rounds of flashing were performed during the stimulus stage. Afterwards, the results of this trial were returned to the screen. If the target was recognized by our network, an applause sound was used as encouraging feedback to provide positive reinforcement. After a short break lasting 4 s, the next trial started.

Each experiment consisted of 10 trials, with one trial corresponding to one question. To prevent positional bias, the text blocks containing the correct answers were displayed an equal number of times on both the left and right sides. Before the experiment began, the participants were told to refrain from any form of muscle movement (e.g., blinking) during the trial.

C. Data Processing and Algorithms

The overall data processing procedure is shown in Fig. 4.

1) Data Processing:

a) EEG data preprocessing: The scalp EEG signals recorded during the experiment were first filtered through a 0.1–20 Hz bandpass filter. Subsequently, epochs corresponding to each stimulus were extracted for each channel within the timeframe of 0 to 800 ms after stimulus onset. These epochs were baseline-corrected with a baseline of 100 ms before stimulus onset, followed by channel filtering to remove the reference electrodes, and finally downsampled at a rate of 4. To prevent scaling effects between different modal data, we normalized the epochs at the channel level. In addition, we captured electrooculograms from two pairs of electrodes, "HEOR" and "HEOL" and "VEO" and "VEOL", to filter out eye movement artifacts from the EEG data. After preprocessing, the EEG data are structured as $[N_c \times N_t]$. Here, $N_c = 30$ is the number of electrode channels, and $N_t = 50$ is the temporal dimension.

b) Eye-tracking data preprocessing: Cubic spline interpolation was employed to compensate for the data gaps caused by blinking or gaze positioning outside the detection range, ensuring the integrity of the eye-tracking data (if insufficient eye-tracking data were available for fitting, the data were discarded for this trial). We extracted epochs of eye-tracking data from 0 to 800 ms for each flash of the text block. Considering the differences in data scale and individual pupil size, we downsampled and normalized each epoch. Additionally, the sizes of the left and right pupils, along with the distance between the gaze point and the currently flashing text block, were selected as eye-tracking features. The distance features were calculated from the gaze coordinates and the center coordinates of the flashing text block. After preprocessing, the eye-tracking data were shaped as $[N_c \times N_t]$, where $N_c = 3$ denotes 3 features and $N_t = 50$ represents the temporal dimension.

2) Dual-Stream Feature Extraction Module:

a) EEG feature extraction: The dual-stream feature extraction module comprises two independent multiscale CNNs for EEG and eye-tracking feature extraction. The EEG feature extraction network consists of a 6-layer architecture, as shown in Fig. 3.

L1-Spatial Convolutional Layer: Common spatial filtering and weighted superposition averaging are utilized to eliminate redundant spatial information and enhance the signal-to-noise ratio of the signal. The convolutional layer has a kernel size of (30, 1), corresponding to the number of electrodes, with a step size of (1, 1). The computational procedure is as follows:

$$a_n^{[1]}(i) = f_{\tanh}\left(\sum_{Nc}^{c=1} I_{c,i} \times w_n^{[1]}(i) + b_n^{[1]}\right) \quad (1)$$

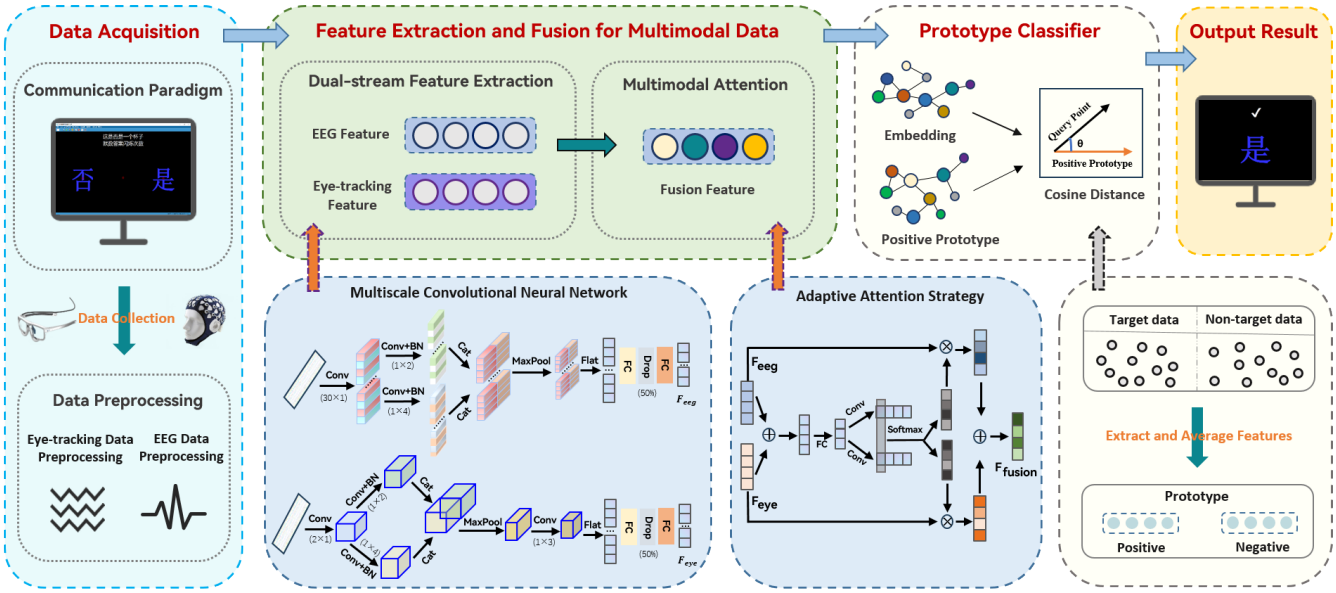

Fig. 4. The overall data processing procedure and the architecture of the MTRN. EEG and eye-tracking data collected from the communication paradigm are first preprocessed. The MTRN then processes and categorizes these data. The predicted positive results are displayed on the screen. The MTRN comprises a dual-stream feature extraction module, a multimodal attention module, and a prototype classifier. The dual-stream feature extraction module, which is based on a multiscale CNN, outputs the spatiotemporal features extracted from the data of each modality. Learned multimodal features undergo feature integration via an adaptive attention strategy. The task of the prototype classifier is to predict data labels by calculating the cosine distance between the data's nonlinear mapping (fusion feature) in the embedding space and the positive sample prototype, defined as the average of all positive sample features in the training data.

where $f_{\tanh}$ is the hyperbolic tangent activation function. $I_{c,i}$ is the i-th element in the c-th channel of the input data. $w_n^{[1]}$ and $b_n^{[1]}$ denote the convolution kernel and bias, respectively.

L2-Temporal Convolution Layer: We introduce two parallel convolution operations and batch normalization operations to capture diverse temporal features. Both convolutional layers share a convolutional step size of (1, 1) and employ different-sized convolutional kernels [(1, 4), (1, 2)]. This strategy improves feature effectiveness and mitigates saturation issues. The formulas are outlined as follows:

$$a_n^{[2,1]} = f_{BN}\left(f_{\tanh}\left(\sum_{c=1}^{N_c} a_c^{[1]}(i) \times w_n^{[2,1]}(i) + b_n^{[2,1]}\right)\right) \quad (2)$$

$$a_n^{[2,2]} = f_{BN}\left(f_{\tanh}\left(\sum_{c=1}^{N_c} a_c^{[1]}(i) \times w_n^{[2,2]}(i) + b_n^{[2,2]}\right)\right) \quad (3)$$

where $a_n^{[2,1]}$ and $a_n^{[2,2]}$ denote the feature maps obtained from the receptive fields of different scales of different convolutional kernels, which contain the contextual features of varying scales in the EEG data. f_{BN} denotes batch normalization. $w_n^{[2,1]}$, $w_n^{[2,2]}$, $b_n^{[2,1]}$, and $b_n^{[2,2]}$ refer to the convolutional kernel matrices and biases, respectively, for the two different kernels.

L3-Integration Layer: The temporal features extracted from the previous layer are cascaded to integrate the local details of the EEG data with broader global information.

L4-Maxpooling Layer: Pooling the integrated features with a kernel of size (1, 2) to expand the receptive field and reduce redundancy, thus improving the performance of the network.

L5-Fully Connected Layer: The multidimensional features are flattened in preparation for subsequent fully connected (FC) operations. Furthermore, we employ a dropout strategy to enhance network generalization [25]. The formulas for this layer are outlined as follows:

$$v = Bernoulli(P) \quad (4)$$

$$a^{[5]} = f_{\tanh}\left(\left(v \times a^{[4]}\right) \times w^{[5]} + b^{[5]}\right) \quad (5)$$

The Bernoulli function randomly generates a vector composed of 0 and 1 with a probability of P. Vectwise multiplication of this vector with the input feature map can suppress P% of the neurons. $f_{\tanh}$ is the tanh activation function, $a^{[4]}$ represents the feature output from L4, and $w^{[5]}$ and $b^{[5]}$ denote the convolutional kernel matrix and bias, respectively.

L6-Fully Connected Layer: This layer executes FC operations, generating representative EEG features for all the input samples. The formula is outlined as follows:

$$a^{[6]} = f_{\tanh}\left(a^{[5]} \times w^{[6]} + b^{[6]}\right) \quad (6)$$

where $f_{\tanh}$ is the activation function and $w^{[6]}$ and $b^{[6]}$ are the convolutional kernel matrix and bias, respectively.

b) Eye-tracking feature extraction: The CNN designed for eye-tracking feature extraction consists of a 7-layer architecture. The structure of this network is similar to that of the EEG feature extraction network, except that the parameter settings are different and a new convolutional layer is added for fine-grained feature extraction. This newly added layer incorporates a convolutional kernel of size (1, 3) and a batch normalization operation. The formula for this layer is as follows:

$$a_n^{[5]} = f_{\tanh}\left(f_{BN}\left(\sum_{c=1}^{N_c} a_c^{[4]}(i) \times w_n^{[5]}(i) + b_n^{[5]}\right)\right) \quad (7)$$

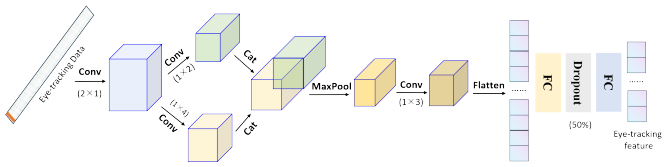

Fig. 5. The architecture of the eye-tracking feature extraction module.

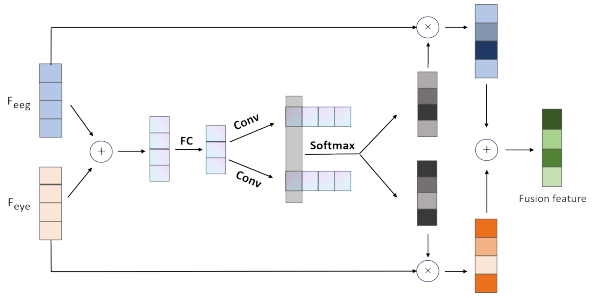

Fig. 6. The architecture of the multimodal attention module.

where $f_{\tanh}$ is the activation function, f_{BN} is the batch normalization, and $a^{[4]}$ refers to the feature map output from the L4 maximum pooling layer. $w^{[5]}$ and $b^{[5]}$ are the convolutional kernel matrix and bias, respectively.

Fig. 5 illustrates the architecture of the eye-tracking feature extraction network. The network comprises the following 7 layers: the L1-spatial convolutional, L2-temporal convolutional, L3-integration, L4-max pooling, L5-fine-grained feature extraction, L6-fully connected, and L7-fully connected layers. The L7 layer outputs the representative eye-tracking features for all the input samples.

3) Multimodal Attention Module: To integrate the multimodal signal features learned from the dual-stream feature extraction module and achieve feature complementarity, we design a feature fusion network based on adaptive attention, whose structure is shown in Fig. 6. First, elementwise summation is employed to initially integrate the features extracted by the dual-stream feature extraction module:

$$F_{\text{fuse}} = F_{eeg} + F_{eye} \tag{8}$$

where $F_{eeg} \in R_{1\times X}$ and $F_{eye} \in R_{1\times X}$ are the representative features of the EEG data and eye-tracking data, respectively.

For precise and adaptive feature selection, we introduce a compact feature $z \in R_{1\times d}$. z is realized through an FC layer that we designed to improve efficiency via dimensionality reduction. The computation for z is as follows:

$$z = FC\left(F_{fuse}\right) = f_{relu}\left(f_{BN}\left(F_{fuse} \times w^{[2]}\right)\right) \tag{9}$$

where f_{relu} denotes the linear rectification function. f_{BN} is batch normalization, and $w^{[2]}$ represents the kernel.

We achieve cross-channel soft attention to information at different spatial scales by means of convolution and Softmax functions. Specifically, two convolutions are conducted guided by the compact feature z to obtain the corresponding weights for each feature. Subsequently, the Softmax function is applied at the channel level to dynamically attend to information at different spatial scales. The calculations are outlined below:

$$W^{[3,1]} = f_{relu}\left(z \times w^{[3,1]} + b^{[3,1]}\right) \tag{10}$$

$$W^{[3,2]} = f_{relu}\left(z \times w^{[3,2]} + b^{[3,2]}\right) \tag{11}$$

$$v_x^{[4,1]} = \frac{e^{W_x^{[3,1]}}}{e^{W_x^{[3,1]}} + e^{W_x^{[3,2]}}}, v_x^{[4,2]} = \frac{e^{W_x^{[3,2]}}}{e^{W_x^{[3,1]}} + e^{W_x^{[3,2]}}} \tag{12}$$

where $W^{[3,1]}$ and $W^{[3,2]}$ are the feature maps computed with different convolution kernels. $w^{[3,1]}$ and $w^{[3,2]}$ denote the convolution kernels, and $b^{[3,1]}$ and $b^{[3,2]}$ are the biases. $W_x^{[3,1]}$ and $v_x^{[4,1]}$ refer to the x-th elements of $W^{[3,1]}$ and $v^{[4,1]}$, respectively, where $v^{[4,1]} \in R_{1\times X}$ is the feature map (soft attention vector) derived from the Softmax operation. Similarly, $W_x^{[3,2]}$, $v_x^{[4,2]}$, and $v^{[4,2]}$ follow a similar pattern.

Finally, the final fusion feature is obtained by weighting each original feature with the soft attention vector.

$$F = v^{[4,1]} \cdot F_{eeg} + v^{[4,2]} \cdot F_{eye}, v^{[4,1]} + v^{[4,2]} = 1 \tag{13}$$

The fusion feature F is the final representative feature of the input sample. In other words, this layer computes the embeddings for all samples (nonlinear mapping of inputs into the embedding space).

4) Prototypical Classifier: To enhance the system's classification performance on small sample data, we employ a prototype network based on the cosine distance as the classifier. The classifier operates on the principle that the embeddings of samples are clustered around their class prototypes [36]. Our work can be described as a binary classification task, distinguishing between positive (target, labeled 1) and negative (non-target, labeled -1) classes. We classify the embedding F of a query sample by identifying the most similar class prototype, which is derived by averaging the features of all samples from the same class in the training set:

$$F_{positive} = \sum_{i=1}^{n} \frac{F_i}{n} \tag{14}$$

Here, $F_{positive}$ denotes the positive class prototype, F_i refers to the embedding of the i-th positive sample, n is the number of positive samples in the training set, and $F_{negative}$ follows a similar definition. The similarity is judged by the cosine distance D, which is calculated by the following formula:

$$D_{positive} = \cos\left(F_{positive}, F\right) = \frac{F_{positive} \cdot F}{||F_{positive}|| \cdot ||F||} \tag{15}$$

where $D_{positive} \in [-1, 1]$ denotes the cosine distance between the embedding F and $F_{positive}$. $D_{negative}$ is similarly defined. A value of $D_{positive}$ close to 1 indicates high similarity between the shape of the input sample and that of the positive sample. In other words, the input sample is more likely to be positive. Conversely, if the $D_{positive}$ value approaches -1, the shape of the input sample is extremely dissimilar to that of the positive sample, indicating a greater probability of the sample being negative. Thus, the prediction for a sample can be outlined as:

$$\text{Pr}ed\,(S) = \begin{cases} -1, & D_{positive} \le 0 \\ 1, & D_{positive} > 0 \end{cases} \tag{16}$$

TABLE I
SUMMARY OF PATIENT INFORMATION

Patient	Gender	Clinical diagnosis	Age	Time since Injury (months)	CRS-R (subscores)
P1	F	MCS	56	12	17 (3-4-5-2-1-2)
P2	M	MCS	14	7	9 (1-3-2-1-0-2)
P3	F	MCS	17	8	9 (1-3-2-1-0-2)
P4	M	LIS	43	3	17(4-5-2-1-2-3)
P5	F	MCS	54	2	12 (3-3-2-1-0-2)
P6	M	VS/UWS	66	2	6 (0-1-2-1-0-2)
P7	F	MCS	18	19	10 (1-3-2-2-0-2)
P8	M	MCS	31	6	17 (3-4-5-2-1-2)
P9	M	MCS	17	9	10 (1-3-2-2-0-2)
P10	F	MCS	72	5	13 (2-3-5-1-0-2)

CRS-R subscales: auditory, visual, motor, oromotor, communication, and arousal functions.

where Pred represents the label prediction function, S is the sample to be predicted, and $D_{positive}$ denotes the cosine similarity between the positive prototype and the embedding F of sample S. Note that the label prediction for the sample relies solely on the cosine distance between the embedding of the input sample and the positive prototype.

The entire network is trained with the mean square error (MSE) loss, and the loss function is defined by:

$$loss = f_{MSE}\left(D_{positive}, y\right) + f_{MSE}\left(D_{negative}, -y\right) \quad (17)$$

where y represents the true label of the sample.

III. MATERIALS AND EXPERIMENT

A. Subjects

Our study included 10 patients (mean age ± SD = 38.8±22.2 years; 5 males) and 10 healthy volunteers (mean age ± SD = 24.0±1.2 years; 5 males). None of the subjects had high myopia (≥ 600 degrees) or impairments in vision or hearing, except for P4 (inability to move the eyeball horizontally, limited to slight vertical movement), P6 and P7 (abnormal visual evoked potentials). All patients underwent a CRS-R assessment one week before the experiment (see Table I for details). This study was approved by the Medical Ethics Committee of Zhujiang Hospital, Southern Medical University, and complied with the Code of Ethics of the World Medical Association. The ethical number is '2023-KY-174-01'.

B. Experiment

The main screen of the experiment, a 24-inch monitor with a 60 Hz refresh rate, was positioned approximately 1 meter from the subjects. A 20-inch secondary screen was utilized for monitoring and calibration. To obtain more accurate gaze data, eye-tracking calibration should be performed for all subjects. However, manual calibration was not performed for patients because of the lack of assurance that patients would gaze at the calibration points as instructed, thus maintaining uniform gaze deviation. We synchronized the EEG data and eye-tracking data by sending event triggers to two servers via the parallel port of the computer. Since patients are easily fatigued and unable to stay awake for long periods, the experiments were scheduled across two days with a two-day break in between.

This experiment comprises two phases, offline training and online testing, as depicted in Fig. 2. Offline training: Before conducting the online experiment, each subject completed a training experiment comprising 10 trials to collect data to train the MTRN. Note that offline training cannot provide feedback on results. Online testing: Each participant performed 5 online tests, each consisting of 10 trials. However, due to discomfort, patients P1, P4, and P8 completed only three and four online tests, respectively. The setup and procedure of the online test are similar to those of offline training, with the difference that during the online experiment, the collected data are processed in real time by the MTRN to provide feedback on the test results at the end of the stimulation.

In each online experiment, the representative epochs for the text blocks were calculated as the average from 10 rounds of flash stimulus epochs. The trained model was applied to the representative epochs of the text blocks, and the predicted results were the highest scoring text blocks.

In our study, the communication accuracy for each subject was determined by the ratio of correct responses to the total number of trials. To evaluate the significance of the accuracy, a χ^2 statistical test was performed with the following formula:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(f_{o_i} - f_{e_i}\right)^2}{f_{e_i}} \quad (18)$$

where f_{o_i} and f_{e_i} are the observed and expected frequencies of the i-th class, respectively. Here, the observations were divided into two classes (hit or miss), so the degree of freedom was 1. For example, in two-choice paradigms, the chances of hitting and missing were both 25 for 50 trials. Differences were deemed statistically significant at P≤0.05. Thus, the calculated χ^2 value was 3.84, indicating a 64% accuracy rate for 50 trials.

IV. RESULTS

To validate the design and superiority of our multimodal system, we conducted offline tests. These tests assessed the accuracy of the MTRN in unimodal mode and the performance of three representative models (SVM, EEGNet, and EyeNet) in each modality. SVM is a classification algorithm commonly used in communication BCIs. EEGNet [24] is a classical network for EEG recognition. As no network specializes in eye-tracking data processing and classification, we designed EyeNet, a feature extraction network referencing the classical convolutional neural network AlexNet [37]. EyeNet comprises three convolutional blocks and one fully connected block. Each convolutional block contains convolution, normalization, activation, and pooling operations. The fully connected block comprises a dropout operation and three fully connected operations. Additionally, to verify the effectiveness of the multimodal attention module and prototype classifier, we tested the MTRN-A and MTRN-B under multimodal conditions.

TABLE II
ACCURACY OF THE ONLINE AND OFFLINE EXPERIMENTS FOR THE TEN HEALTHY SUBJECTS (50 TRIALS)

Subject	P300			Eye-tracking			P300 + Eye-tracking			
	SVM	EEGNet	MTRN	SVM	EyeNet	MTRN	SVM	MTRN-A	MTRN-B	MTRN
H1	66%	72%	84%	100%	100%	100%	100%	100%	100%	100%
H2	86%	98%	100%	99%	100%	100%	100%	100%	100%	100%
H3	86%	92%	94%	100%	100%	100%	100%	100%	100%	100%
H4	96%	100%	100%	100%	100%	100%	100%	100%	100%	100%
H5	70%	82%	86%	100%	100%	100%	100%	100%	100%	100%
H6	82%	88%	92%	100%	100%	100%	100%	100%	100%	100%
H7	88%	84%	94%	100%	100%	100%	100%	100%	100%	100%
H8	84%	84%	96%	100%	100%	100%	100%	100%	100%	100%
H9	92%	94%	94%	100%	100%	100%	100%	100%	100%	100%
H10	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Mean ± SD	85.0 ± 10.5%	89.4 ± 9.1%	94.0 ± 5.5%	99.9 ± 0.3%	100%	100%	100%	100%	100%	100%

For the feature extraction and classification of the MTRN under unimodal conditions, we eliminated the multimodal attention module and used only the prototype classifier and the feature extraction network corresponding to this modality in the dual-stream feature extraction module. MTRN-A uses a data-layer fusion strategy, and MTRN-B uses a Softmax classifier instead of a prototype classifier.

TABLE III
ACCURACY OF THE ONLINE AND OFFLINE EXPERIMENTS FOR THE TEN PATIENTS

Patient	Trials	P300			Eye-tracking			P300 + Eye-tracking				P value
		SVM	EEGNet	MTRN	SVM	EyeNet	MTRN	SVM	MTRN-A	MTRN-B	MTRN	
P1	30	56.7%	63.33%	66.7%	**70%**	**70%**	**80%**	68%	**73.33%**	**70%**	**80%**	0.001
P2	50	50%	58%	56%	52%	60%	60%	54%	58%	60%	56%	0.396
P3	50	50%	56%	60%	46%	60%	62%	50%	62%	62%	62%	0.089
P4	40	55%	65%	65%	58%	65%	62.5%	60%	65%	**70%**	**72.5%**	0.004
P5	50	**64%**	**66%**	**72%**	56%	62%	**74%**	**64%**	**70%**	**74%**	**84%**	<0.001
P6	50	60%	58%	62%	46%	58%	62%	48%	60%	60%	62%	0.089
P7	50	58%	60%	58%	44%	62%	58%	52%	60%	60%	58%	0.257
P8	40	62.5%	**67.5%**	**77.5%**	52.5%	55%	62.5%	62.5%	**72.5%**	**75%**	**80%**	<0.001
P9	50	44%	56%	58%	48%	56%	60%	48%	58%	60%	62%	0.089
P10	50	56%	60%	60%	60%	60%	62%	62%	62%	**64%**	**64%**	0.047
Mean	/	55.6%	61.0%	63.6%	53.3%	60.8%	64.3%	56.9%	64.1%	65.5%	68.1%	/

Note that the P values in the table correspond to the online accuracy obtained through the MTRN (multimodal) for each patient. The accuracies exceeding the significance threshold (P≤0.05) are highlighted in bold.

MTRN-A and MTRN-B are variants of the MTRN network we designed. MTRN-A employs a data-layer fusion strategy (i.e., the collected multimodal data are directly concatenated after preprocessing), and MTRN-B discards the prototype classifier and uses a Softmax classifier instead.

Table II details the accuracies of the experiments for the ten healthy subjects. The results showed that all healthy subjects achieved 100% online accuracy, which was much greater than the significance threshold (P≤0.05). For all healthy subjects, the accuracy of the unimodal-based systems surpassed the significance level of 64%. Moreover, the multimodal-based system, whether employing SVM, MTRN-A, MTRN-B or MTRN, achieved higher accuracy than the system based only on P300 or eye tracking. This finding suggests that multimodal systems outperform unimodal systems, confirming the effectiveness of the MTRN and multimodal communication design. The multimodal-based SVM, MTRN-A, MTRN-B and MTRN all attained identical accuracies of 100%, rendering them incomparable. However, under unimodal conditions, the MTRN outperformed the SVM and EEGNet/EyeNet, indicating that the MTRN has superior feature extraction and classification capabilities.

Table III presents the accuracy rates of the online and offline experiments for the 10 patients. Among them, 5 patients (4 MCS and 1 LIS) achieved significant results in the online experiment, with an average accuracy of 76.1±7.9%. We divided the patients into a response group (P1, P4, P5, P8, and P10) and a nonresponse group (the other 5 patients) based on whether their online accuracy reached the significance threshold. For the response group, the multimodal-based MTRN achieved the highest accuracy. Moreover, the overall accuracy of the MTRN was significantly greater than that of the SVM, EEGNet/EyeNet, MTRN-A, and MTRN-B under both unimodal and multimodal conditions, further validating the effectiveness of the multimodal attention module. For the nonresponse group, the mean accuracy was 60.0±2%, which was slightly above the 50% chance level.

Interestingly, P5 and P10 achieved 84% and 64% accuracy, respectively, in the online communication experiment, which far exceeded the threshold of significance. This suggests that P5 and P10 were deemed communicative and capable of binary communication in the BCI assessment. However, they scored 0 on the communication subscale of the CRS-R, implying a potential lack of communicative ability.

By averaging the EEG signals for each stimulus type (target and nontarget) across all online trials, we extracted P300 event-related potential (ERP) waveforms from 0 to 800 ms for the response group and the healthy control group (H4) and calculated the standard deviation of their ERPs for all experiments. Fig. 7 displays the average EEG

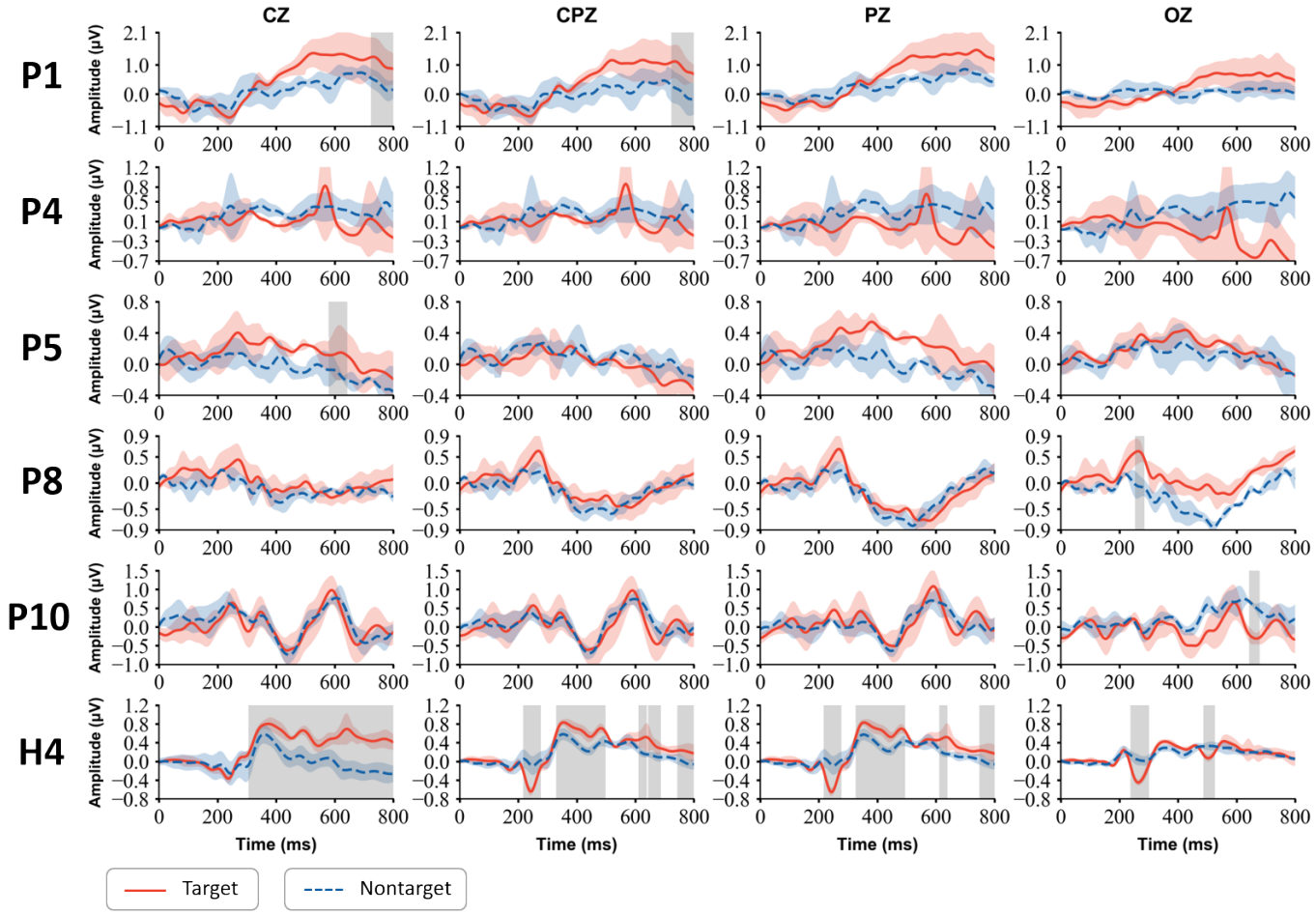

Fig. 7. ERP waveforms in selected channels for patients in the response group (P1, P4, P5, P8, and P10) and healthy controls (H4). The solid red and dashed blue lines correspond to the EEG waveforms during target and nontarget stimuli, respectively. Shading in red/blue indicates the standard deviation of the corresponding ERPs across all experiments. Significant differences between the two stimulus conditions (t-test, FDR correction, $p \leq 0.05$) are denoted by gray shaded areas.

signal amplitude and standard deviations across the selected channels ('Pz', 'CPz', 'Cz', and 'Oz') for these five patients and H4. The EEG waveforms of the five patients and H4 clearly showed P300 responses to the target stimuli. Additionally, we performed pointwise t-test statistical analyses of ERPs under different stimulation conditions and corrected for multiple comparisons using the false discovery rate (FDR) procedure ($P \leq 0.05$).

To analyze the eye-tracking data, we obtained representative pupil sizes for the response group and the healthy controls from 0 to 1 s following stimulus onset by averaging the pupil data for each stimulus type across all online trials. Additionally, we averaged the distance data obtained from 10 trials in each experiment to determine the average distance of the gaze point from the target and nontarget points across these 10 trials. Fig. 8 displays these visualized data. Approximately 300 ms after the target stimulation, the pupils of both the response group and healthy subjects exhibited varying degrees of contraction. Specifically, the left pupil of the response group/healthy subjects decreased by an average of 0.11 mm/0.5 mm, and the right pupil decreased by an average of 0.11 mm/0.54 mm.

V. Discussion

We developed a hybrid BCI to assist DOC patients in consciousness detection and communication. In this study, we proposed a novel communication paradigm and a multimodal target recognition network. The BCI analyzes EEG and eye-tracking data collected during paradigm stimulation via the MTRN to identify patient answers. Ten healthy volunteers and 10 patients participated in the communication experiment. All healthy participants achieved a high accuracy rate of 100%, demonstrating the system's potential for communication. Moreover, 5 out of 10 patients (4 DOC and 1 LIS) significantly exceeded chance level, reaching 76.1±7.9%, which indicates their ability to communicate. Most notably, two DOC patients, previously deemed noncommunicative in CRS-R assessments, succeeded in communicating via the BCI, suggesting that the hybrid BCI holds promise for improving the detection of brain function among this challenging patient group.

We integrated eye tracking into a communication paradigm for three primary reasons: (i) Pupil constriction, mediated by the autonomic nervous system (ANS), does not require a functional somatomotor system in principle. Furthermore,

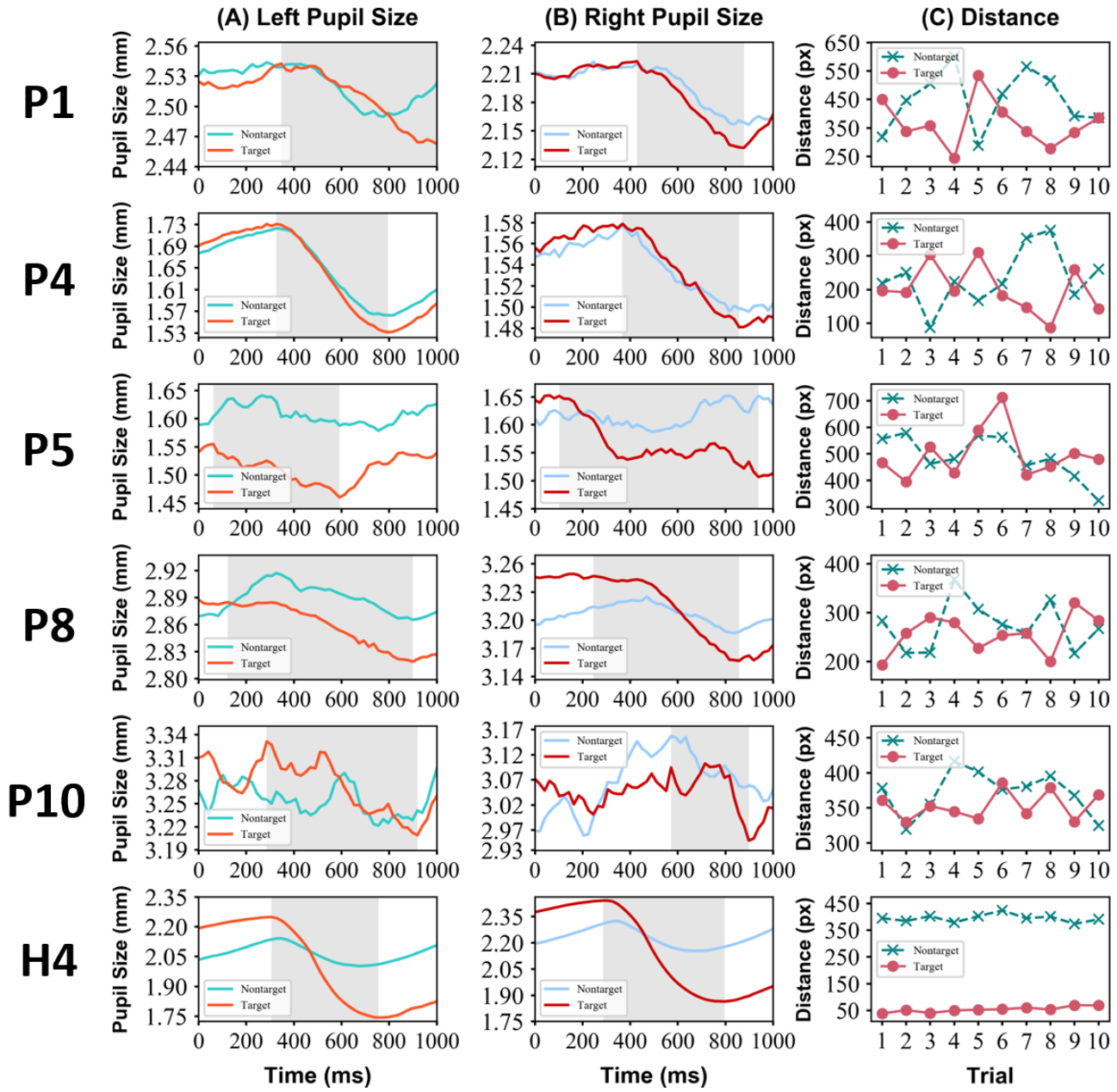

Fig. 8. (A)/(B): Variations in left/right pupil size upon stimulus presentation. The blue line indicates the pupil size during nontarget stimulation. The red line denotes the pupil size during target stimulation. The gray shaded area signifies the occurrence of the PLR under the target stimulus. (C): Average distance of the gaze point from the target and nontarget text blocks. The pink line depicts the distance from the center of the target text block, while the green line depicts the distance of the gaze point from the nontarget text block.

despite the extensive neurological damage in patients, the ANS can be largely spared [38], [39] and therefore may constitute an output pathway for patient communication. (ii) Eye tracking requires only that patients open their eyes and possess vision, aligning with the prerequisites for eliciting the visual P300 ERP and thereby not increasing the threshold for BCI use. Moreover, eye tracking does not demand that subjects perform extra tasks beyond gaze (e.g., counting), reducing training and execution difficulty. Even if the patient is unable to maintain sustained gaze behavior (i.e., intermittent attention) during the stimulus phase, it does not affect the final recognition result. This is because the MTRN can recognize the patient's choices by detecting eye movement responses (e.g., pupil constriction, gaze) produced during the stimulus process. (iii) Eye tracking is more user-friendly than the methods used in other studies [18], [40]. Recent studies have demonstrated the effectiveness of monitoring hemodynamic signals related to yes/no responses in re-establishing communication with LIS patients [40]. However, this method requires costly equipment.

Observing P300 responses in patients indicates the presence of residual cognitive function. Despite the inconsistency in P300 waveforms and latencies between the patient response group and healthy controls, we still observed that their P300 components were evoked to varying degrees (Fig. 7). The

TABLE IV
COMPARISON OF AVERAGE CLASSIFICATION PERFORMANCE FOR TEN HEALTHY SUBJECTS

Methods	Modality	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	F-measure	Precision
MTRN	EEG + Eye-tracking	**499.7**	**499.7**	**0.3**	**0.3**	**99.94%**	**99.94%**	**99.94%**	**99.93%**	**99.93%**
MTRN-A		499.1	499.3	0.7	0.9	99.84%	99.82%	99.86%	99.83%	99.86%
MTRN-B		499.3	499.6	0.4	0.7	99.89%	99.86%	99.92%	99.89%	99.91%
SVM		496.9	498.4	1.6	3.1	99.53%	99.38%	99.68%	99.53%	99.69%
MTRN	EEG	**452.2**	**428.8**	**71.2**	**47.8**	**88.10%**	**90.44%**	**85.76%**	**88.37%**	**86.39%**
EEGNet		426.8	409.2	90.8	73.2	83.60%	85.36%	81.84%	83.88%	82.46%
SVM		381.5	359.4	140.6	118.5	74.09%	76.30%	71.88%	74.65%	73.07%
MTRN	Eye-tracking	**499.2**	**499.8**	**0.2**	**0.8**	**99.90%**	**99.84%**	**99.96%**	**99.89%**	**99.96%**
EyeNet		493.5	498.9	1.1	6.5	99.24%	98.70%	99.78%	99.22%	99.78%
SVM		496.9	498.5	1.5	3.1	99.54%	99.38%	99.70%	99.54%	99.71%

For the feature extraction and classification of the MTRN under unimodal conditions, we eliminated the multimodal attention module and used only the prototype classifier and the feature extraction network corresponding to this modality in the dual-stream feature extraction module. MTRN-A and MTRN-B are variants of the MTRN. MTRN-A uses a data-layer fusion strategy, and MTRN-B uses a Softmax classifier instead of a prototype classifier. TP, TN, FP and FN indicate the number of true positives, true negatives, false positives and false negatives, respectively.

inability to clearly observe the P300 potential of P1 in the EEG waveform graph may result from fluctuations in the consciousness levels of DOC patients over time, hindering the effective evocation of P300. The LIS patient (P4) had near-normal cognitive abilities, enabling him to induce a more pronounced P300 potential. However, P4's ERP did not significantly differ between the two stimulus conditions. This may be because P4 experienced psychomotor agitation during most of the experiments, preventing him from performing the test. This agitation may also account for P4's poor accuracy in communication using only the EEG-based BCI. In addition to the average ERP waveforms, the standard deviations of the ERPs for all experiments have been calculated and are included in Fig. 7. Patients P1, P4, and P5 exhibit markedly higher variability in their ERPs, particularly in response to target and nontarget stimuli. This increased dispersion is indicative of the substantial inter-experimental differences encountered, largely attributed to the patients' variable fluctuating conditions and states of consciousness during the communication trials. These observations are consistent with the inherent challenges in interpreting EEG data from DOC patients, highlighting the importance of considering individual variability and physiological states when analyzing neurological responses.

The observed PLR responses to the target stimulus in the response group and healthy controls (Fig. 8) reflected their attention to the target. Although the pupil of P4 also constricted significantly during nontarget stimulation, this does not imply that P4 necessarily actively gazed at the nontarget. Limited eye movements may expose P4 to target and nontarget stimuli of equal intensity, thus potentially inducing PLR in both cases. Indeed, we discovered that patients were unable to fixate on the target as precisely as we anticipated. Throughout the experiment, due to body tension or other reasons, some patients (e.g., P5) exhibited uncontrollable eye movements, preventing a sustained gaze. Fig. 8(c) shows that the patient's gaze did not consistently remain on or near the target, aligning with our observations. This may be the reason for the limited accuracy of communicating with DOC patients solely through eye tracking.

It should be noted that the absence of significant accuracy in BCI assessment does not conclusively indicate a lack of awareness in DOC patients. Intact cognition, including language comprehension, memory, and attention, is necessary for patients to effectively use command-following-based BCIs. The absence of any of the aforementioned cognitive abilities may result in misdiagnosis of the patient's condition. A hybrid BCI integrating eye tracking and P300 potentials could enhance the detection of covert awareness in patients with low cognition. For instance, some patients (e.g., P1) may communicate via eye tracking but not via the P300 ERP. Enhanced BCI performance can boost patient confidence and motivation, fostering greater engagement in treatment [41]. Moreover, patients with motor-cognitive dissociation (showing awareness on neuroimaging but no detectable command-following behavior) tend to have a better prognosis [42]. Detecting potential awareness in patients can foster family positive expectations.

We computed commonly used metrics to further evaluate the categorization performance of the MTRN and the contributions of its different components. The true positivity rate (also known as the sensitivity) is often used to estimate the sensitivity of a test. A higher sensitivity indicates a lower probability of missed detection. Our model primarily identifies targets by detecting P300 and eye movement responses, so emphasis should be placed on the sensitivity of the method. In addition to sensitivity, accuracy, specificity, F1 score, and precision are also crucial metrics. Table IV displays the average classification performance of each model for ten healthy subjects. The sensitivity of the MTRN in each modality surpassed that of other models, illustrating its ability to classify P300 and eye movement responses and demonstrating the validity of multiscale and prototype-based classification techniques. The multimodal-based MTRN outperformed the MTRN-A without the multimodal attention module and the MTRN-B without the prototype classifier in all the metrics, suggesting that the adaptive attention strategy and the prototype classifier can facilitate the fusion and classification of small-sample data. Additionally, we statistically compared the accuracy of the

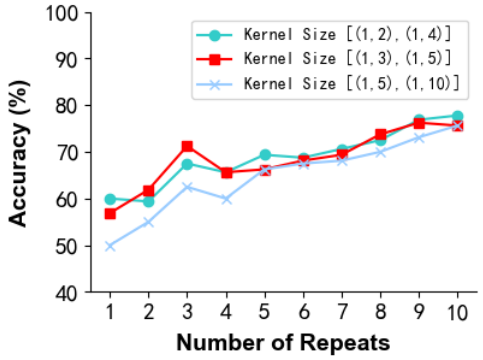

Fig. 9. The impact of the convolution kernel size in the L2 temporal feature extraction layers on target identification and repetition.

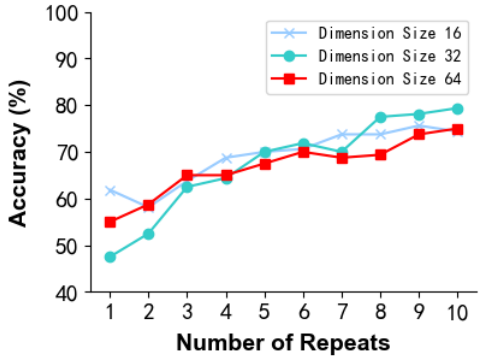

Fig. 10. The impact of the compact feature's dimension size (d) on target recognition and repetition.

MTRN under unimodal and multimodal conditions. Given the small number of subjects, we used the Wilcoxon signed-rank test for the analysis. The results showed a significant difference ($P \leq 0.05$) in accuracy between multimodal-based MTRN and unimodal-based MTRN for both healthy subjects and patients, suggesting that communication through the integration of eye tracking and EEG is effective.

The results of the ablation and comparison experiments confirmed that the hybrid system outperformed the unimodal system based solely on P300 or eye tracking, validating the effectiveness of the MTRN. In the MTRN, the feature extraction module captures differentiating temporal features through two parallel convolutions, and the multimodal attention module guides adaptive feature selection through compact features z. To ascertain the optimal kernel size and assess the impact of z size on model performance, we conducted tests with several sets of common convolution kernels and dimension sizes in patient data, as illustrated in Fig. 9 and Fig. 10. With an equal number of kernels, the kernel sizes [(1,2),(1,4)] outperform the kernels [(1,3),(1,5)] and [(1,5),(1,10)] in terms of accuracy. The optimal performance was achieved with a z size of 32. The reason may be the short duration of the P300 and PLR responses. Larger convolution kernels expand the receptive field, potentially introducing redundant and irrelevant information. Moreover, overly large or small compact features may lead to information loss or redundancy during feature selection. The task of prototype networks is to perform classification by calculating the distance between embeddings and class prototypes in metric space. Numerous studies have used the Euclidean distance as a distance metric with good results. However, Pan et al. [43] suggested that the cosine distance metric is more sensitive to small-sized data and may yield better results for classifying such datasets.

Despite these encouraging results, several limitations should be overcome. First, the limited sample of this study, which included only 10 patients, affects the generalizability of the results. Second, communication is constrained to yes/no answers, which limits interaction for patients with high cognitive abilities. Future work should optimize the communication model to enhance the utility of the system.

REFERENCES

[1] S. Laureys et al., "Unresponsive wakefulness syndrome: A new name for the vegetative state or apallic syndrome," *BMC Med.*, vol. 8, no. 1, pp. 1–4, Dec. 2010.

[2] S. Wannez, L. Heine, M. Thonnard, O. Gosseries, and S. Laureys, "The repetition of behavioral assessments in diagnosis of disorders of consciousness," *Ann. Neurol.*, vol. 81, no. 6, pp. 883–889, Jun. 2017.

[3] J. R. Patterson and M. Grabois, "Locked-in syndrome: A review of 139 cases," *Stroke*, vol. 17, no. 4, pp. 758–764, Jul. 1986.

[4] L. Naci, L. Sinai, and A. M. Owen, "Detecting and interpreting conscious experiences in behaviorally non-responsive patients," *NeuroImage*, vol. 145, pp. 304–313, Jan. 2017.

[5] C. Schnakers et al., "Cognitive function in the locked-in syndrome," *J. Neurol.*, vol. 255, no. 3, pp. 323–330, Mar. 2008.

[6] V. Galiotta et al., "EEG-based brain–computer interfaces for people with disorders of consciousness," *Frontiers Human Neurosci.*, vol. 16, Dec. 2022, Art. no. 1040816.

[7] J. Giacino, T. Kalmar, and J. Whyte, "The JFK coma recovery scale-revised: Measurement characteristics and diagnostic utility," *Arch. Phys. Med. Rehabil.*, vol. 85, no. 12, pp. 2022–2029, 2004.

[8] D. Kondziella et al., "European academy of neurology guideline on the diagnosis of coma and other disorders of consciousness," *Eur. J. Neurol.*, vol. 27, no. 5, pp. 741–756, May 2020.

[9] L. S. M. Johnson and C. Lazaridis, "The sources of uncertainty in disorders of consciousness," *AJOB Neurosci.*, vol. 9, no. 2, pp. 76–82, Apr. 2018.

[10] O. Gosseries, N. D. Zasler, and S. Laureys, "Recent advances in disorders of consciousness: Focus on the diagnosis," *Brain Injury*, vol. 28, no. 9, pp. 1141–1150, Aug. 2014.

[11] W. S. van Erp, J. C. M. Lavrijsen, F. A. van de Laar, P. E. Vos, S. Laureys, and R. T. C. M. Koopmans, "The vegetative state/unresponsive wakefulness syndrome: A systematic review of prevalence studies," *Eur. J. Neurol.*, vol. 21, no. 11, pp. 1361–1368, Nov. 2014.

[12] K. Kuehlmeyer, E. Racine, N. Palmour, E. Hoster, G. D. Borasio, and R. J. Jox, "Diagnostic and ethical challenges in disorders of consciousness and locked-in syndrome: A survey of German neurologists," *J. Neurol.*, vol. 259, no. 10, pp. 2076–2089, Oct. 2012.

[13] J. Xiao et al., "Toward assessment of sound localization in disorders of consciousness using a hybrid audiovisual brain–computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1422–1432, 2022.

[14] W. H. Curley, P. B. Forgacs, H. U. Voss, M. M. Conte, and N. D. Schiff, "Characterization of EEG signals revealing covert cognition in the injured brain," *Brain*, vol. 141, no. 5, pp. 1404–1421, May 2018.

[15] D. Lulé et al., "Probing command following in patients with disorders of consciousness using a brain–computer interface," *Clin. Neurophysiol.*, vol. 124, no. 1, pp. 101–106, Jan. 2013.

[16] F. Wang et al., "Enhancing clinical communication assessments using an audiovisual BCI for patients with disorders of consciousness," *J. Neural Eng.*, vol. 14, no. 4, Aug. 2017, Art. no. 046024.

[17] C. Guger et al., "Assessing command-following and communication with vibro-tactile P300 brain–computer interface tools in patients with unresponsive wakefulness syndrome," *Frontiers Neurosci.*, vol. 12, Jun. 2018, Art. no. 359448.

[18] J. Huang et al., "Hybrid asynchronous brain–computer interface for yes/no communication in patients with disorders of consciousness," *J. Neural Eng.*, vol. 18, no. 5, Oct. 2021, Art. no. 056001.

[19] F.-B. Vialatte, M. Maurice, J. Dauwels, and A. Cichocki, "Steady-state visually evoked potentials: Focus on essential paradigms and future perspectives," *Prog. Neurobiol.*, vol. 90, no. 4, pp. 418–438, Apr. 2010.

[20] X. Duart, E. Quiles, F. Suay, N. Chio, E. García, and F. Morant, "Evaluating the effect of stimuli color and frequency on SSVEP," *Sensors*, vol. 21, no. 1, p. 117, Dec. 2020.

[21] V. Galiotta et al., "EEG-based brain–computer interfaces for people with disorders of consciousness: Features and applications. A systematic review," *Frontiers Human Neurosci.*, vol. 16, Dec. 2022, Art. no. 1040816.

[22] 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.

[23] M. Alvarado-González, G. Fuentes-Pineda, and J. Cervantes-Ojeda, "A few filters are enough: Convolutional neural network for P300 detection," *Neurocomputing*, vol. 425, pp. 37–52, Feb. 2021.

[24] 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.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[26] J. E. Desmedt, N. Tran Huy, and M. Bourguet, "The cognitive P40, N60 and P100 components of somatosensory evoked potentials and the earliest electrical signs of sensory processing in man," *Electroencephalogr. Clin. Neurophysiology*, vol. 56, no. 4, pp. 272–282, Oct. 1983.

[27] J. Polich, "Updating P300: An integrative theory of P3a and P3b," *Clin. Neurophysiol.*, vol. 118, no. 10, pp. 2128–2148, Oct. 2007.

[28] J. Pan et al., "A hybrid brain–computer interface combining P300 potentials and emotion patterns for detecting awareness in patients with disorders of consciousness," *IEEE Trans. Cogn. Develop. Syst.*, vol. 15, no. 3, pp. 1386–1395, Mar. 2022.

[29] L. Stark and P. M. Sherman, "A servoanalytic study of consensual pupil reflex to light," *J. Neurophysiol.*, vol. 20, no. 1, pp. 17–26, Jan. 1957.

[30] S. Mathôt, J.-B. Melmi, L. van der Linden, and S. Van der Stigchel, "The mind-writing pupil: A human–computer interface based on decoding of covert attention through pupillometry," *PLoS ONE*, vol. 11, no. 2, Feb. 2016, Art. no. e0148805.

[31] A. Sato and S. Nakatani, "Noncontact brain–computer interface based on steady-state pupil light reflex using independent bilateral eyes stimulation," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2021, pp. 3386–3390.

[32] J. Stoll, C. Chatelle, O. Carter, C. Koch, S. Laureys, and W. Einhäuser, "Pupil responses allow communication in locked-in syndrome patients," *Current Biol.*, vol. 23, no. 15, pp. R647–R648, Aug. 2013.

[33] A. E. Lorenzo Villalobos et al., "When assistive eye tracking fails: Communicating with a brainstem-stroke patient through the pupillary accommodative response—A case study," *Biomed. Signal Process. Control*, vol. 67, May 2021, Art. no. 102515.

[34] M. M. N. Mannan, M. A. Kamran, S. Kang, H. S. Choi, and M. Y. Jeong, "A hybrid speller design using eye tracking and SSVEP brain–computer interface," *Sensors*, vol. 20, no. 3, p. 891, Feb. 2020.

[35] J. Wang, S. Xu, Y. Dai, and S. Gao, "An eye tracking and brain–computer interface based human-environment interactive system for amyotrophic lateral sclerosis patients," *IEEE Sensors J.*, vol. 23, no. 20, pp. 24095–24106, Oct. 2022.

[36] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[38] D. Lule, V. Diekmann, H.-P. Müller, J. Kassubek, A. C. Ludolph, and N. Birbaumer, "Neuroimaging of multimodal sensory stimulation in amyotrophic lateral sclerosis," *J. Neurol., Neurosurgery Psychiatry*, vol. 81, no. 8, pp. 899–906, Aug. 2010.

[39] W. Sun, S.-H. Liu, X.-J. Wei, H. Sun, Z.-W. Ma, and X.-F. Yu, "Potential of neuroimaging as a biomarker in amyotrophic lateral sclerosis: From structure to metabolism," *J. Neurol.*, vol. 271, no. 5, pp. 2238–2257, May 2024.

[40] V. Johansson, S. R. Soekadar, and J. Clausen, "Locked out: Ignorance and responsibility in brain–computer interface communication in locked-in syndrome," *Cambridge Quart. Healthcare Ethics*, vol. 26, no. 4, pp. 555–576, Oct. 2017.

[41] R. Mane, T. Chouhan, and C. Guan, "BCI for stroke rehabilitation: Motor and beyond," *J. Neural Eng.*, vol. 17, no. 4, Aug. 2020, Art. no. 041001.

[42] J. Pan et al., "Prognosis for patients with cognitive motor dissociation identified by brain–computer interface," *Brain*, vol. 143, no. 4, pp. 1177–1189, Apr. 2020.

[43] J. Pan, H. Cai, H. Huang, Y. He, and Y. Li, "Multiple scale convolutional few-shot learning networks for online P300-based brain–computer interface and its application to patients with disorder of consciousness," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–16, 2023.