bioengineering

MDPI

Article

Continual Learning of a Transformer-Based Deep Learning Classifier Using an Initial Model from Action Observation EEG Data to Online Motor Imagery Classification

Po-Lei Lee [1,2], Sheng-Hao Chen [1], Tzu-Chien Chang [1], Wei-Kung Lee [3], Hao-Teng Hsu [1,2] and Hsiao-Huang Chang [4,5,*]

[1] Department of Electrical Engineering, National Central University, Taoyuan 320, Taiwan
[2] Pervasive Artificial Intelligence Research Labs, Hsinchu 300, Taiwan
[3] Department of Rehabilitation, Taoyuan General Hospital, Taoyuan 330, Taiwan
[4] Division of Cardiovascular Surgery, Department of Surgery, Taipei Veterans General Hospital, Taipei 112, Taiwan
[5] Department of Surgery, School of Medicine, College of Medicine, Taipei Medical University, Taipei 110, Taiwan
[*] Correspondence: shchang@vghtpe.gov.tw; Tel.: +886-937-919-107

Abstract: The motor imagery (MI)-based brain computer interface (BCI) is an intuitive interface that enables users to communicate with external environments through their minds. However, current MI-BCI systems ask naïve subjects to perform unfamiliar MI tasks with simple textual instruction or a visual/auditory cue. The unclear instruction for MI execution not only results in large inter-subject variability in the measured EEG patterns but also causes the difficulty of grouping cross-subject data for big-data training. In this study, we designed an BCI training method in a virtual reality (VR) environment. Subjects wore a head-mounted device (HMD) and executed action observation (AO) concurrently with MI (i.e., AO + MI) in VR environments. EEG signals recorded in AO + MI task were used to train an initial model, and the initial model was continually improved by the provision of EEG data in the following BCI training sessions. We recruited five healthy subjects, and each subject was requested to participate in three kinds of tasks, including an AO + MI task, an MI task, and the task of MI with visual feedback (MI-FB) three times. This study adopted a transformer-based spatial-temporal network (TSTN) to decode the user's MI intentions. In contrast to other convolutional neural network (CNN) or recurrent neural network (RNN) approaches, the TSTN extracts spatial and temporal features, and applies attention mechanisms along spatial and temporal dimensions to perceive the global dependencies. The mean detection accuracies of TSTN were 0.63, 0.68, 0.75, and 0.77 in the MI, first MI-FB, second MI-FB, and third MI-FB sessions, respectively. This study demonstrated the AO + MI gave an easier way for subjects to conform their imagery actions, and the BCI performance was improved with the continual learning of the MI-FB training process.

Keywords: brain computer interface; electroencephalography (EEG); action observation; motor imagery; transformer network

check for updates

Citation: Lee, P.-L.; Chen, S.-H.; Chang, T.-C.; Lee, W.-K.; Hsu, H.-T.; Chang, H.-H. Continual Learning of a Transformer-Based Deep Learning Classifier Using an Initial Model from Action Observation EEG Data to Online Motor Imagery Classification. *Bioengineering* **2023**, *10*, 186. https://doi.org/10.3390/bioengineering10020186

Academic Editor: Larbi Boubchir

Received: 4 December 2022
Revised: 20 January 2023
Accepted: 24 January 2023
Published: 1 February 2023

1. Introduction

Brain computer interface (BCI) is the technology for people to directly communicate with external devices through the acquisitions and translations of their brain activities. A BCI system measures the neural activities from the central nervous system (CNS) and converts them into artificial outputs in order to replace, restore, enhance, supplement, or improve natural CNS output [1]. Several BCI systems have been developed based on different brain imaging modalities, such as electroencephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), and functional near-infrared spectroscopy (fNIRS) [2]. In contrast to the fMRI and fNIRS which measure the slow changes of hemodynamic responses, the EEG/MEG records fast changes of electrophysiological signals with a high sampling rate and enables the possibility of timely observation for the neural activities inside the brain. In particular, the EEG has the advantages of easy preparation, inexpensive equipment cost, and high temporal resolution. It has been chosen for a wide variety of clinical applications, such as sleep disorder diagnosis [3], seizure detection [4], emotion classification [5], etc. Advanced signal processing techniques for EEG have also been developed to avoid the interference of external artifacts (e.g., electromyography, motion artifact, etc.) [6,7]. Owing to the aforementioned reasons, EEG is the most popular choice to implement a BCI system [7].

An EEG-based BCI system requires an elaborately designed task for generating reliable neural responses, a credible signal processing method for extracting particular signal features, and an efficient translation algorithm to produce output signals [8]. Current BCIs can be categorized into endogenous and exogeneous BCIs, according to the necessity of an external stimulus for generating brain activities. Endogenous BCIs utilize spontaneously generated brain patterns (e.g., motor imagery [9], speech imagery [10], slow cortical potential [11], etc.), whereas exogeneous BCIs detect brain patterns induced by external stimuli (e.g., steady-state visual evoked potential [12], flash visual evoked potential [13], steady-state auditory response [14], etc.). Though the exogeneous BCIs have the benefits of less training and a higher information transfer rate (ITR), the indispensable external stimuli, such as flickering LEDs or auditory beeps, are required to evoke discriminative patterns. Patients with locked-in syndrome (LIS) usually have weak muscle activities as well as deficits in their sensations [15], so that limits the feasibility of exogeneous BCI in severe disabilities. On the other hand, endogenous BCIs are independent of external stimuli which utilize neural signals generated from designated mental imagery tasks. Two BCIs, the motor imagery (MI) and speech imagery BCIs, are treated as the two most promising systems because they are operated in more natural ways of our daily life. In contrast to the speech imagery BCI, the motor imagery BCI has been most studied, owing to the cross-individual generality in its task execution and the abundant data resources in the previous literature. For well-trained users, the MI-based BCI system, denoted as the MI-BCI system, can be a straightforward, convenient, and fast way to communicate with external devices.

The MI-BCI requests users to perform extensive training of imagination actions before they are capable of generating desired brain patterns to achieve acceptance performances. Traditional MI-BCIs requested users to imagine the movements of their limbs (e.g., legs or hands) in response to a visual cue presented on a screen with or without feedback [16]. Khare et al. (2022) developed an automatically tuning algorithm to optimize the selection of tunable Q wavelet transform (TQWT) parameters and a high detection accuracy was achieved [17]. Khare et al. (2020) utilized flexible variational mode decomposition (F-VMD) to extract meaningful EEG components. They calculated hjorth, entropy, and quartile features from F-VMD decomposed components and a flexible extreme learning machine (F-ELM) was used to classify the EEG features from different MI tasks [18]. Filho et al. (2022) studied the event-related desynchronization (ERD) occurrence and classification accuracy under the conditions of no feedback and actual feedback in MI tasks. With feedback, they found not only the detection accuracies but also the ERD values in both contralateral and ipsilateral sensorimotor cortexes were significantly improved [19]. Alimardani et al. (2018) discussed the training effects of MI-BCI with different visual feedback representations. They concluded that humanlike visual feedback (e.g., human hand action) achieved a better MI learning performance, compared to non-humanlike visual feedback (e.g., robot gripper or direction bar) [20]. Frideman et al. (2008) studied the performance of a visual-feedback MI-BCI in a virtual reality (VR) environment. They found that a highly-immersive visual environment can strengthen the subjective impression and result in a better training performance [21]. In addition to visual feedback, other BCI systems are designed to train user's MI responses with different feedback modalities, such as auditory or tactile feedbacks. McCreadie et al. (2013) studied the feasibility of replacing visual feedback with stereo auditory feedback [22]. They demonstrated that stereophonic feedback could effectively achieve comparable results with the use of visual feedback. Nijboer et al. (2008) compared the performances of auditory feedback training with visual feedback training in an MI-BCI system [23]. No difference in the BCI performances was found between the uses of the two feedback modalities. Ishihara et al. (2020) reviewed the effectiveness of feedback modalities in 184 BCI studies [24], in which the visual feedback was the most effective approach for neurofeedback in MI-BCI training. Owing to the advantages of an intuitively, highly immersive sensation and a better flexibility for experimental design, the visual feedback is most widely chosen to implement a BCI system.

Though visual feedback has been demonstrated as an effective and efficient way for neural training in MI-BCI system [25], user′s training performance is highly-related to the immersive feeling and self-perception [26]. Ono et al. (2013) compared the values of event-related desynchronization (ERD) in the conditions of no feedback, bar feedback, and incongruent/congruent feedback. In the incongruent and congruent feedbacks, a real hand open/grasp picture was displayed on a screen. They found the participants in a congruent condition achieved higher ERD values and better detection accuracies. They also demonstrated the visual feedback of real-hand picture-induced higher ERD values than bar feedback or no feedback. Achanccaray et al. (2018) designed an MI-based BCI system to control a 3D arm in a VR environment [27]. Ziadeh et al. (2021) [26] studied the influence of embodiment in the VR environment on a subject's BCI performance. They found both the senses of the ownership and agency can influence the MI training performance [28].

Most MI-BCI systems request subjects to participate in multiple experiment sessions before they are able to operate the MI-BCI with an acceptable detection rate, including MI sessions and MI-feedback (MI-FB) sessions [26,29]. In the MI-FB sessions, participants performed the same MI tasks but received online classification results via feedbacks. However, previous research has indicated that a nonnegligible portion of subjects (about 15~30%) were not able to control the BCI systems through the aforementioned BCI training process [30]. Vidaurre et al. (2011) hypothesized two reasons for the problems related to the successful operation of an MI-BCI system, in which the first reason is the lack of a generic pattern for initial BCI training, and the second reason is the difficulty of transition from offline calibration to online feedback. For the first point, the creation of a generic pattern for a classifier setup is difficult since most BCI training lacks a clear instruction for movement imagination which could result in large inter-subject difference in the induced brain wave patterns. It has also been difficult to obtain a consistent brain wave pattern in previous studies due to the diversity of MI task design in different BCI systems. For example, some studies requested participants to imagine the simple moving of an object on the screen [22], while the others claimed a more complicated MI task, such as imagining a serial movement of a hand or a foot. Recent studies have also demonstrated that the training efficacies can be enhanced by creating the illusions of ownership and agency [20,28,30] which can be conveniently achieved by wearing a VR head-mount device (HMD) [21,26,27,31,32]. For the second point, most BCI studies usually created their initial classifiers from the EEG data in MI tasks with no feedback, and tried to modify the initial classifier to fit online MI-FB applications [33]. Nevertheless, the gap between the initial model and the final classifier could be huge, so that subjects usually had to participate in a large amount of MI-FB training until acceptable results were obtained.

In this study, we intend to study the feasibility of building the initial classifier for a BCI system using the EEG data obtained from an action observation (AO) task. Action observation has been proposed as an effective rehabilitation therapy approach based on the role of a mirror neuron system in motor learning [34]. The mirror neuron system is activated during the observation of an action execution. In contrast to mirror therapy (MT) which requests semi-paralyzed patients to activate the mirror neuron system by watching the reflection image of moving unaffected sides in a mirror [35], the AO therapy conducts the formation of motor memory by viewing the limb movements which can provide a more flexible way in experimental design. Hsieh et al. (2020) compared the rehabilitation outcomes between the interventions of AO therapy and MT in stroke patients [36], and they concluded the AO therapy can lead to better improvements in patients. Vogt et al. (2013)

reviewed neuroimaging studies and claimed neural activities in motor cortices can be enhanced by performing concurrent AO and MI [37]. Though the role of AO in motor activities has been studied in the previous literature [27,38,39], most BCI studies utilized AO as a feedback approach for improving a subject's MI performance rather than using the brain activities in AO for constructing an initial BCI classifier. Since the poor setting of initial BCI classifiers could frustrate the subject and impede training efficiency [40], the AO, whose brain activation areas are located in conjunction with the brain regions recruited in an MI task [38], could be a possible way to generate similar signal features for the initial setup of a BCI classifier. This study aims to initiate the parameter settings of BCI classifiers based on the collection of EEG data from AO plus the MI (AO + MI) task. We discuss the feasibility of tuning this initial classifier to achieve better classification results by means of using continual learning with MI-FB data. The AO + MI provides a convenient way for subjects to follow the experimental designer's instruction. It also gives an easier way for subjects to conform their imagery actions with the expected task design.

The present study intended to build a training method for MI-BCI based on AO + MI in the VR environment. We aimed to answer the following questions: (1) Is the classifier trained from AO + MI data good enough for MI classifications? (2) Can we generate a classifier from AO + MI data and tune the classifier with the continual learning of MI data? (3) Can we adopt a transformer-based deep learning classifier in our BCI training? (4) Can the detection accuracy of the classifier which is trained from cross-subject AO + MI data be improved by the continual learning of individual MI data? Because both the AO and MI are two motor tasks with no actual motor executions but sharing overlapped motor areas in their activations, we wonder if the requisite of large amount of EEG data in tedious MI task could be replaced by EEG data from AO + MI task in an easier way.

2. Materials and Methods

2.1. Subjects and EEG Recordings

Five normal subjects (5 males, all right-handed subjects; mean age = 24 ± 4.2 years) were recruited to participate in our study. All participants were requested to sit in a comfortable armchair in a dimly illuminated electro-magnetic shielded room. EEG signals were recorded using an eight-channel dry-electrode wireless EEG system (bandpass 0.05–250 Hz; 55–65 Hz bandstop; 24-bits data resolution; digitized at 1 kHz; InMex EEG system, WellFulfill Co., Taoyuan, Taiwan). The EEG electrodes were placed in accordance with the international 10–20 system. Electrodes were located at FC1, FC2, C3, Cz, C4, P1, and P2 positions, with respect to a reference electrode placed at the right mastoid and a ground electrode placed at the left mastoid. Each EEG dry electrode was constructed by 10 spring-loaded copper pins, and its biocompatibility was certificated by ISO 10993. Immersive scenarios were created in the virtual reality environment using Unity 3D Engine (Unity Technologies Inc., San Francisco, CA, USA) and provided to the participants by wearing a Head Mounted Display system (HTC VIVE, HTC Co., Taipei, Taiwan) (see Figure 1).

2.2. Experimental Task

Participants came to participate in our experiment for four weeks, one time in each week. In the first week, subjects were requested to participate in an AO + MI session, and then they were asked to join three MI-FB sessions in the following three weeks. The VR environment was programmed using Unity 3D and displayed on a head-mount display (HMD) device to provide an immersive experience for the subjects. Subjects were asked to imagine they were looking into a mirror and the virtual character was the reflection image of themselves. In the AO + MI session, each trial contained two AO + MI time blocks and one MI time block (without feedback). In each trial, the designated hand, either the left or right hand, was chosen randomly. During the AO + MI time block, the virtual character raised its designated hand once, and subjects were requested to perform an MI movement of the same hand concurrently with the action of the virtual character. During the MI time block, subjects were asked to perform an MI movement of the same designated hand just after they saw the presence of the direction sign (see Figure 1). Subjects were requested to perform 300 trials in the AO + MI session, 150 trials for left hands and 150 trials for right hands, arranged in a random order. For every twenty trials, subjects had a five-minute break to prevent them from mental fatigue. Each AO + MI time block was the concatenation of a resting period (5~7 s) and an AO + MI period (2 s). In the resting period of the AO + MI time block, the virtual character was kept still with no movement. In the AO + MI period, the virtual character performed a front arm raise of the designated hand once. Subjects were instructed to imagine the virtual character was the reflection images of themselves, similar to looking into a mirror. In the MI time block, subjects were requested to perform a mental rehearsal of the arm raise action which they viewed in the previous AO + MI time block. One resting period (5~7 s) was concatenated with an MI period (2 s). In the resting period of the MI time block, a cross sign was presented, and subjects were asked to relax without movement intention. In the MI period, a direction sign was presented, and subjects were requested to perform a mental rehearsal of the arm raise action of the designed arm which they viewed in the AO + MI time block.

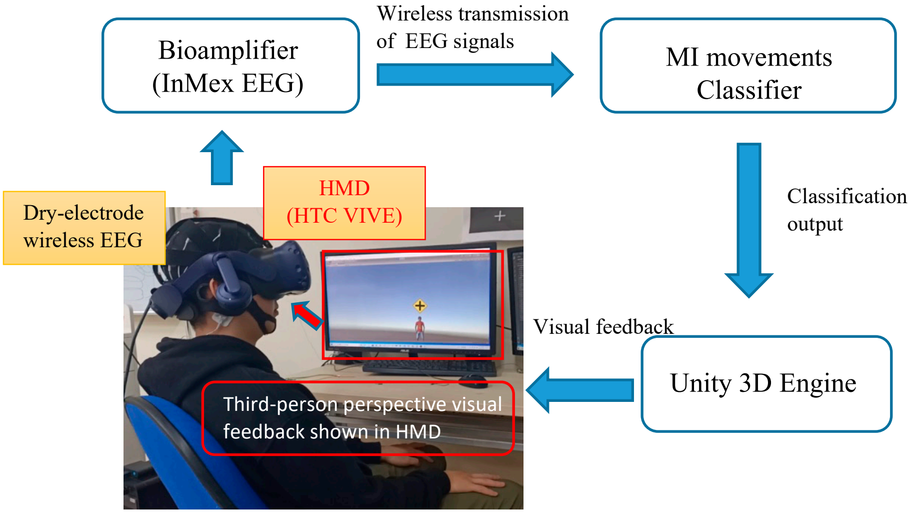


Figure 1. The system architecture of the proposed BCI system with immersive VR feedback.

After the AO + MI sessions, subjects were requested to participate in three MI-FB sessions in the following three weeks, one session in a week. Each MI-FB session contained 150 trials of MI movements (with visual feedback), one-third of the trials (50 trials) for left-hand imagery movements, one-third of the trials (50 trials) for right-hand imagery movements, and one-third of the trials (50 trials) for no movement, arranged in a random order. In each trial, a preparation sign was presented before the movement indication sign. Subjects were initially instructed to keep themselves relaxed in a preparation time block (5~7 s), and then a left or right direction sign was presented to instruct subjects which hand should be performed in the MI movement time block (3 s). For each MI movement, subjects were requested to perform a mental rehearsal of the arm raise action of the designated hand that they viewed in the AO + MI session. The EEG signals in each MI time block were classified into one of the three classes (i.e., the left-arm movement, the right-arm movement, and the resting state), and the classification result was feedbacked to the subject by driving the virtual character to show the arm raise movement of the classified result in the feedback time block (2 s), accompanied with a circle or cross sign to respond to the correctness of the MI classification in this trial. The experimental paradigms of the AO + MI and the MI-FB sessions are shown in Figures 2a and 2b, respectively. All participants gave informed consent, and the study was approved by the Ethics Committee of the Institutional Review Board (IRB) (TYGH 107055), Tao-Yuan General Hospital, Taiwan. All measurements were noninvasive, and the subjects were free to withdraw at any time.

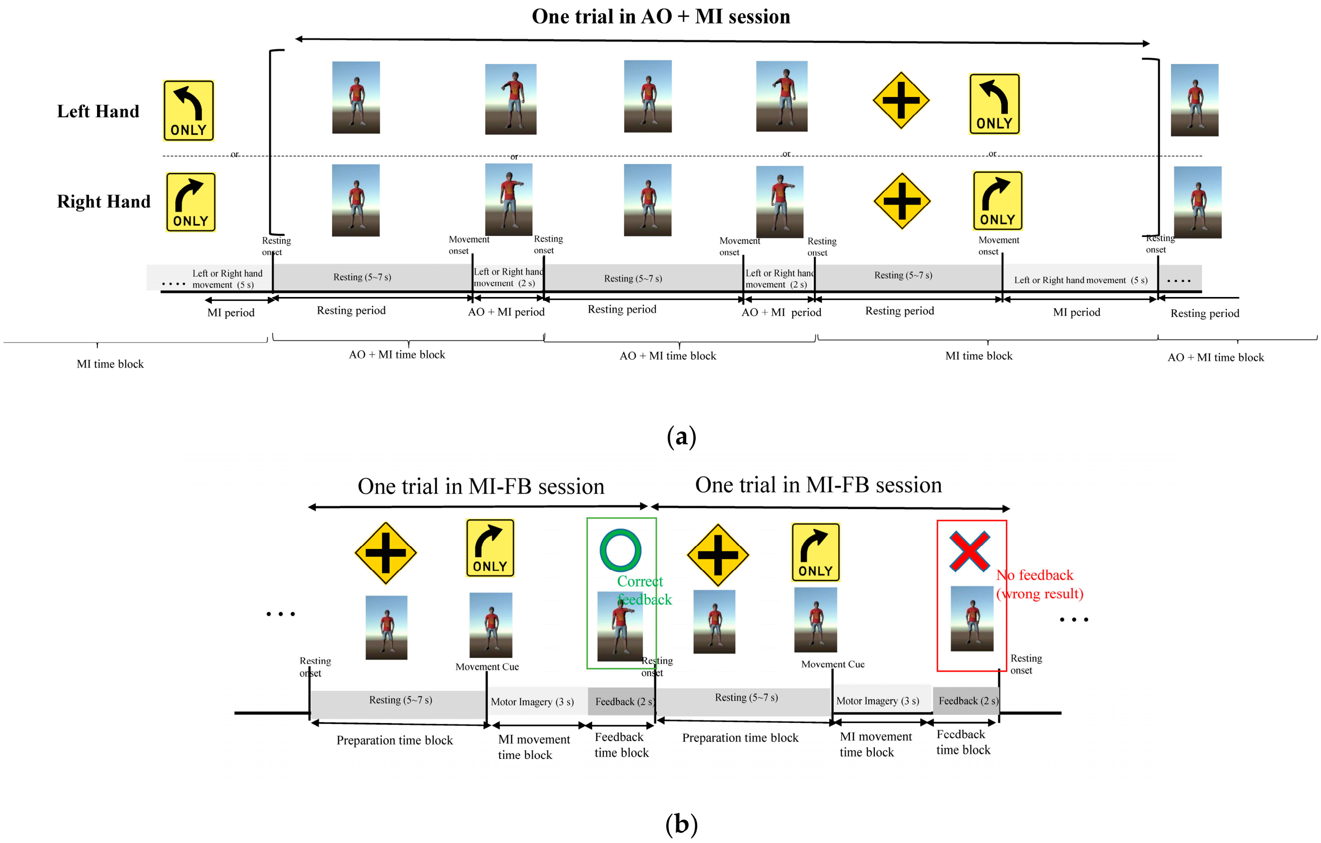


Figure 2. The experimental paradigms of the AO + MI and the MI − FB sessions. Subjects were asked to imagine they were looking into a mirror and the virtual character was the reflection image of themselves. (**a**) The paradigm of the AO + MI session, including two AO + MI time blocks and one MI time block in each trial; (**b**) the paradigm of the MI session, including one preparation time block, one MI movement time block, and one feedback time block.

2.3. *Transformer-Baed Spatial-Temporal Network (TSTN) for MI Classification*

In this study, we adopted the work proposed by Song et al. (2021) [41], which applied attention mechanisms on the spatial and temporal features of EEG data. The eight-channel EEG signals were prefiltered within 4~40 Hz (3rd-order Butterworth IIR filter) and then downsampled to 250 Hz. The EEG data were then segmented into two-second epochs and spatially filtered using a common spatial pattern (CSP) to extract the discriminant features. The spatial features in the feature-channel data were further enhanced by applying spatial transforming with an attention mechanism. The enhanced feature-channel data were segmented across time and segmented into embedded patches using two convolution layers. The relationship among different temporal patches was further perceived using multi-head transforming to obtain distinguishable representations. The features processed by the temporal transformer were averaged and a fully connected layer was used to classify the feature-segment representations into one of the three categories (i.e., the left-arm movement, the right-arm movement, and the resting state). The detailed signal processing is described as follows.

2.3.1. Spatial Filtering Using Common Spatial Pattern (CSP)

CSP is an effective feature extraction method which maximizes the discriminability of two classes by constructing a set of optimized spatial filters [41]. In our study, the CSP was constructed based on a one-versus-rest (OVR) strategy, i.e., rest vs. left/right, left vs. rest/right and right vs. rest/left), to extract feature-channel signals. For each CSP, the covariance metrices R_1 and R_2 were calculated from two sets of EEG signals X_1 and X_2, in which C is the number of EEG channel (C = 8), T is the length of sampled data

(T = 2000), and $X_1 \in \Re^{C\times T}$ and $X_2 \in \Re^{C\times T}$ are the EEG data of the classes that we want to identify. The CSP seeks to find a matrix that contains eigenvectors $P = \begin{bmatrix} p_1 & p_2 & \ldots & p_c \end{bmatrix}$ for simultaneously diagonalizing the covariance matrixes R_1 and R_2, which can be represented as:

$$D = P^T R_1 P \tag{1}$$

And

$$I = P^T R_2 P \tag{2}$$

in which D is the diagonal matrix with eigen values $\{\lambda_1, \lambda_2, \ldots, \lambda_c\}$ sorted in descending order and I is the identify matrix. The first two columns p_1 and p_2 in P were chosen, and the transpose of the first two column vectors p_1^T and p_2^T in the subfilters of the three CSPs were stacked as the spatial filter Z $\left(Z \in \Re^{6\times c}\right)$. For an EEG data matrix X ($X \in \Re^{c\times T}$), the filtered EEG data S can be represented as:

$$S = ZX \tag{3}$$

where $S \in \Re^{6\times T}$ contains the feature-channel signals filtered by the six subfilters obtained from the three discriminations, with two subfilters for each discrimination.

2.3.2. Spatial Transforming for the Enhancement of Feature-Channel Signals

The feature information in the feature-channel data S was enhanced by applying the self-attention mechanism. The input vectors were multiplied with three weighted matrices W_q, W_k, and W_v to obtain the key, query, and the value vectors for each of the input vectors, represented as:

$$q_i = W_q a_i \tag{4}$$

$$k_i = W_k a_i \tag{5}$$

and

$$v_i = W_v a_i, \tag{6}$$

in which a_i is the ith input vector, and q_i, k_i, and v_i are the query, key, and value vectors. The similarity between q_i and k_j is then estimated by calculating the scaled dot-product:

$$S(q_i, k_j) = \left(q_i \cdot k_j\right) / \sqrt{d_k} \tag{7}$$

where d (d = 500) is the dimension of $\mathrm{k_j}$.

The value of $S(q_i, k_j)$ is then normalized using the softmax function to obtain $\hat{\alpha}_{i,j}$ and then multiplied by vector v_j to obtain the ith output vector b_i. The calculation can be represented as the following equation:

$$b_i = \sum_j \mathrm{Softmax}\left(\frac{q_i \cdot k_j}{\sqrt{d_k}}\right) \cdot v_j = \sum_j \hat{\alpha}_{i,j} v_j \tag{8}$$

The whole process can be represented as:

$$B = \mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{9}$$

where the output matrix B contains the enhanced feature-channel data.

2.3.3. Patch Embedding of Feature-Channel Signals

In order to reduce computation complexity, the feature-channel data were segmented into data patches and the dependencies among different patches in a trial were learned using a temporal transforming network (see below). The enhanced feature-channel data B were passed through a one-dimension convolution layer (number of filter = 2; kernel size = 51) and then processed by a two-dimension convolution layer (number of filter = 10; kernel size = 6 × 5) to creat 10 signal patches, in which each patch has a data length of 90 samples. The 10 patches were used as inputs for the following temporal transforming network.

2.3.4. Temporal Transforming for Embedded Patches

Multi-head attention (MHA) was employed to explore the dependencies among different patches in our temporal transforming step. The MHA allows the network to attend parts of the sequence differently. The embedded patches (10 signal patches) were divided into five data sets $\mathrm{E_i}$ ($\mathrm{E_i} \in \Re^{2\times 90}; i = 1, \cdots, 5$) passed through h (h = 5) independent attention networks. The outputs of these attention networks were concatenated and combined together with a final weight matrix W^o. The MHA can be represented as follows:

$$\mathrm{F} = \mathrm{Multihead}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{Concat}(\mathrm{head_1}, \mathrm{head_2}, \cdots, \mathrm{head_h})W^o \tag{10}$$

where F $\left(\mathrm{F} \in \Re^{10\times 90}\right)$ is the output of multihead attention, $\mathrm{head_i} = \mathrm{Attention}(\mathrm{Q_i}, \mathrm{K_i}, \mathrm{V_i})$ is the output of *i*th head, Concat (.) is the function to concatenate the output of each head together, $Q_i = W_i^Q E_i$, $K_i = W_i^K E_i$, $V_i = W_i^V E_i$, and $(Q_i, K_i, V_i) \in \Re^{2\times 90}$, $\left(W_i^Q, W_i^K, W_i^V\right) \in \Re^{2\times 2}$ and $W^o \in \Re^{10\times 10}$.

2.3.5. Classifier

The output of multi-head attention F was then passed to an averaging pooling layer, normalized, and then the classification of imagery movements was achieved by a fully-connected layer, in which cross-entropy was chosen as its loss function. The architecture of the TSTN network is shown in Figure 3.

2.4. Training of the TSTN Classifier

It has been demonstrated that both the AO and the MI have significant effects on the modulation of sensorimotor Mu rhythms [42]. The clinical literature in rehabilitation studies found that the AO + MI is an effective way to restore motor function in stroke patients [38]. Since AO and MI activate overlapped motor-related brain areas, it is reasonable to construct an initial classifier based on AO + MI data and then use MI-feedback data to continually train the network for better MI classification results.

Figure 4 shows the training and the testing procedure of our TSTN classifier. The classifier was initially trained using all the AO + MI data (900 trials for each subject; 300 trials for each class). The classifier obtained from the AO + MI data, denoted as $\mathrm{TSTN_{AO+MI}}$, was used to test the classification performance in the MI (without feedback) task. Because inaccurate trials in MI training could impede the user's BCI performance [43], only those trials in MI data (without feedback) which were correctly identified were chosen for the continual learning to obtain $\mathrm{TSTN_{MI}}$. The $\mathrm{TSTN_{MI}}$ was then applied to test the classification performance in the 1st MI-FB data (with VR feedback), and the correctly classified trials were collected to train a new classifier $\mathrm{TSTN_{MI\text{-}FB\_1}}$. The $\mathrm{TSTN_{MI\text{-}FB\_1}}$ was again used to test the classification performance in the 2nd MI-FB data (with VR feedback), and the correctly classified trials were used to train a final classifier $\mathrm{TSTN_{MI\text{-}FB\_2}}$. The final $\mathrm{TSTN_{MI\text{-}FB\_2}}$ was applied to test the BCI performance in the 3rd MI-FB training dataset and the test performances among different steps were compared to see how the BCI training improved the classifier performances. For the model training, Adam with a learning rate of 0.0002 was utilized and batch size was set as 50. The dropout rate was 0.3 in spatial transforming and 0.5 in temporal transforming [40,44]. Ten-fold cross validation was applied to evaluate the final results, with each fit being performed on a training set consisting of 90% of the total training set selected at random, with the remaining 10% used as a hold out set for validation.

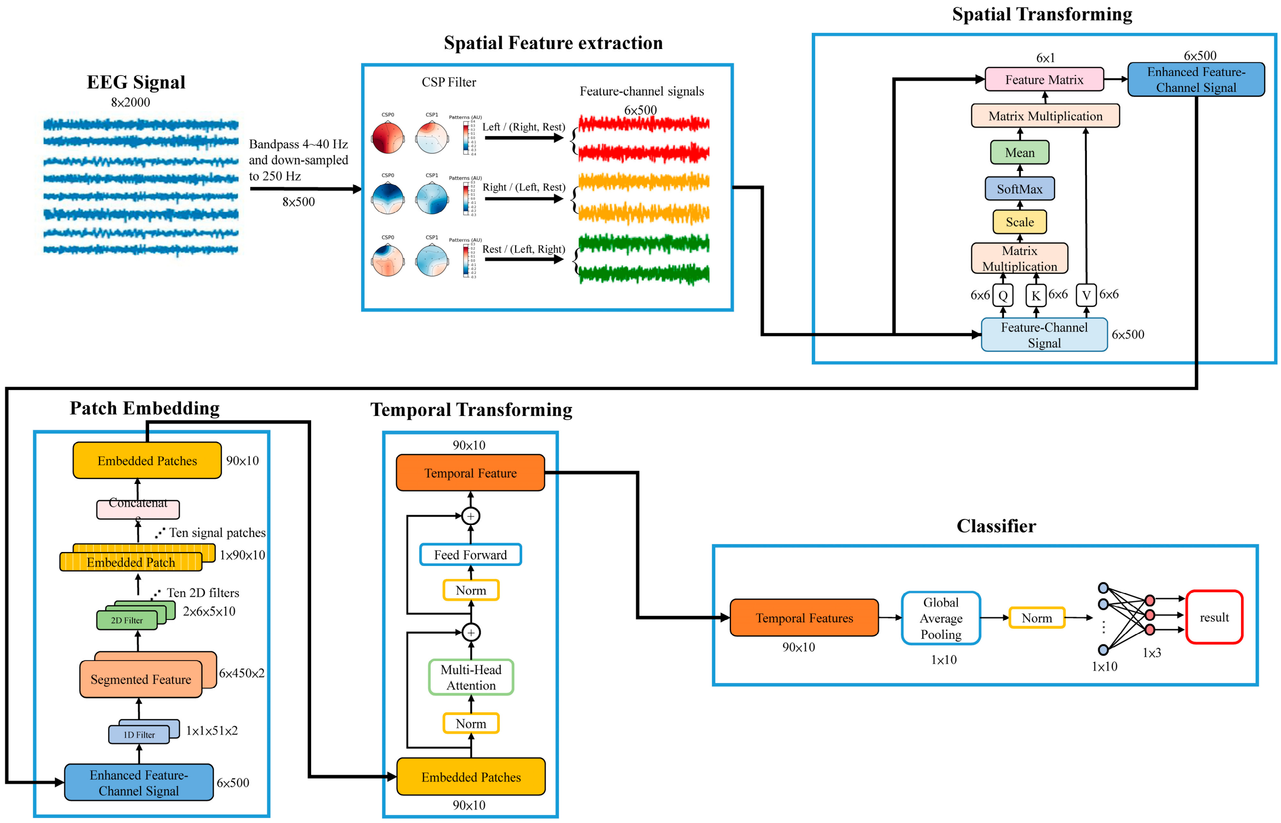


Figure 3. The architecture of the TSTN network. The EEG signals were prefiltered within 4~40 Hz and then downsampled to 250 Hz. The feature-channel signals were extracted using CSP, and then further enhanced by means of applying spatial transforming with an attention mechanism. The enhanced feature-channel data were segmented into ten embedded patches and the relationships among different temporal patches were perceived using multi-head transforming to obtain distinguishable representations.

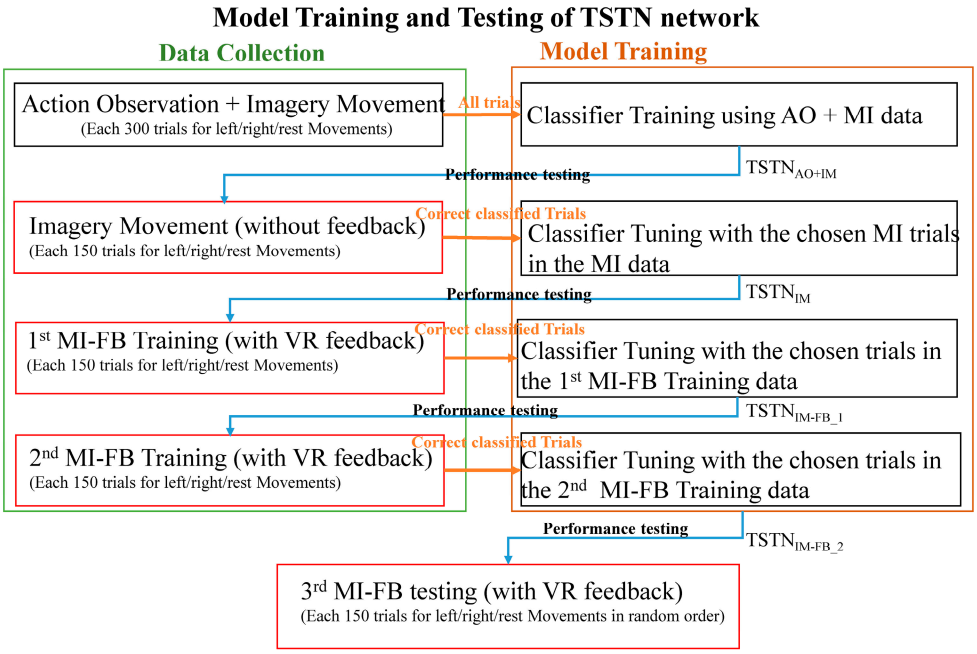

Model Training and Testing of TSTN network
Data Collection
Model Training
Action Observation + Imagery Movement
(Each 300 trials for left/right/rest Movements)
All trials
Classifier Training using AO + MI data
Performance testing
TSTN$_{AO+IM}$
Imagery Movement (without feedback)
(Each 150 trials for left/right/rest Movements)
Correct classified Trials
Classifier Tuning with the chosen MI trials in the MI data
Performance testing
TSTN$_{IM}$
1st MI-FB Training (with VR feedback)
(Each 150 trials for left/right/rest Movements)
Correct classified Trials
Classifier Tuning with the chosen trials in the 1st MI-FB Training data
Performance testing
TSTN$_{IM\text{-}FB\_1}$
2nd MI-FB Training (with VR feedback)
(Each 150 trials for left/right/rest Movements)
Correct classified Trials
Classifier Tuning with the chosen trials in the 2nd MI-FB Training data
Performance testing
TSTN$_{IM\text{-}FB\_2}$
3rd MI-FB testing (with VR feedback)
(Each 150 trials for left/right/rest Movements in random order)

Figure 4. The training and testing procedure for the AO + MI, MI, 1st MI-FB, 2nd MI-FB, and 3rd MI-FB datasets.

2.5. Comparing the Detection Performance with Other Classifiers

In order to compare the effectiveness of the TSTN with other classifiers, two convolutional neural network (CNN) based classifiers, the EEGNet [45] and the DeepConvNet [46], and three support vector machines (SVMs) with kernel functions of linear, radial basis, and polynomial functions, were tested and compared with TSTN. The training and testing procedures were the same as our training procedures shown in Figure 4.

3. Results

In this study, we built the initial classifier $\text{TSTN}_{\text{AO+MI}}$ using AO + MI data, and then performed continual learning with chosen trials in MI, first MI-FB, and second MI-FB datasets to obtain TSTN_{MI}, $\text{TSTN}_{\text{MI-FB\_1}}$, and $\text{TSTN}_{\text{MI-FB\_2}}$, respectively.

Table 1 shows the classification results of using TSTN in our five participants. The test targets for the $\text{TSTN}_{\text{AO+MI}}$, TSTN_{MI}, $\text{TSTN}_{\text{MI-FB\_1}}$, and $\text{TSTN}_{\text{MI-FB\_2}}$ were the MI (without feedback), first MI-FB, second MI-FB, and third MI-FB datasets, respectively. The classification accuracies were increased after each continual learning step, in which the classification accuracies averaged over the five participants were 0.63, 0.68, 0.75, and 0.77 for the $\text{TSTN}_{\text{AO+MI}}$, TSTN_{MI}, $\text{TSTN}_{\text{MI-FB\_1}}$, and $\text{TSTN}_{\text{MI-FB\_2}}$, respectively. The specificities were 0.81, 0.84, 0.87, and 0.89 for the $\text{TSTN}_{\text{AO+MI}}$, TSTN_{MI}, $\text{TSTN}_{\text{MI-FB\_1}}$, and $\text{TSTN}_{\text{MI-FB\_2}}$, respectively. The F1 scores averaged over the five subjects were 0.63, 0.68, 0.75, and 0.77, respectively. It is worth noting that the initial $\text{TSTN}_{\text{AO+MI}}$ had already achieved an accuracy higher than 60% and the detection accuracies kept increasing with the classifier tuning in each continual learning step. This demonstrated the feasibility of using AO + MI data to create an initial classifier model for MI classification.

Table 1. Classification results of the TSTN network generated from individual data.

Classifier **(Test Target)/Subject**	$\textbf{TSTN}_{\textbf{AO+MI}}$ **(IM Data)**	$\textbf{TSTN}_{\textbf{MI}}$ **(1st IM-FB Data)**	$\textbf{TSTN}_{\textbf{MI-FB\_1}}$ **(2nd IM-FB Data)**	$\textbf{TSTN}_{\textbf{MI-FB\_2}}$ **(3rd IM-FB Data)**
	Acc/Spec/F1	Acc/Spec/F1	Acc/Spec/F1	Acc/Spec/F1
S1	0.73/0.88/0.73	0.73/0.87/0.74	0.78/0.90/0.78	0.83/0.92/0.83
S2	0.58/0.79/0.59	0.66/0.83/0.67	0.72/0.87/0.72	0.73/0.87/0.73
S3	0.61/0.80/0.61	0.62/0.87/0.62	0.73/0.87/0.73	0.75/0.89/0.75
S4	0.61/0.80/0.61	0.65/0.83/0.65	0.73/0.84/0.74	0.73/0.87/0.74
S5	0.60/0.80/0.61	0.73/0.87/0.74	0.78/0.89/0.78	0.80/0.90/0.80
Averaged accuracy	0.63/0.81/0.63	0.68/0.84/0.68	0.75/0.87/0.75	0.77/0.89/0.77

%Remark: Acc: Accuracy; Spec: Specificity; F1: F1 score.

Figure 5 plots the detected accuracies of the three MI classes in the five subjects by applying the $\text{TSTN}_{\text{AO+MI}}$, TSTN_{MI}, $\text{TSTN}_{\text{MI-FB\_1}}$, and $\text{TSTN}_{\text{MI-FB\_2}}$ to the MI (without feedback), first MI-FB, second MI-FB, and 3rd MI-FB datasets, respectively. The detected accuracies obtained from S1 to S5 are shown in Figure 5a–e, and the averages of detection accuracies are shown in Figure 5f. The detected accuracies of individual $\text{TSTN}_{\text{AO+MI}}$ on MI dataset in each subject were 0.74, 0.58, 0.61, 0.61, and 0.60 for S1, S2, S3, S4, and S5, respectively. For individual TSTN_{MI} on the first MI-FB data in each subject, the detected accuracies were 0.73, 0.66, 0.62, 0.65, and 0.73 for S1, S2, S3, S4, and S5, respectively. For individual $\text{TSTN}_{\text{BCI\_1}}$ on the second MI-FB data in each subject, the detected accuracies were 0.78, 0.72, 0.73, 0.73, and 0.78 for S1, S2, S3, S4, and S5, respectively. For individual $\text{TSTN}_{\text{MI-FB\_2}}$ on the third MI-FB data in each subject, the detected accuracies were 0.83, 0.73, 0.75, 0.73, and 0.80 for S1, S2, S3, S4, and S5, respectively. The averages of the detection accuracies over the five subjects are shown in Figure 5f.

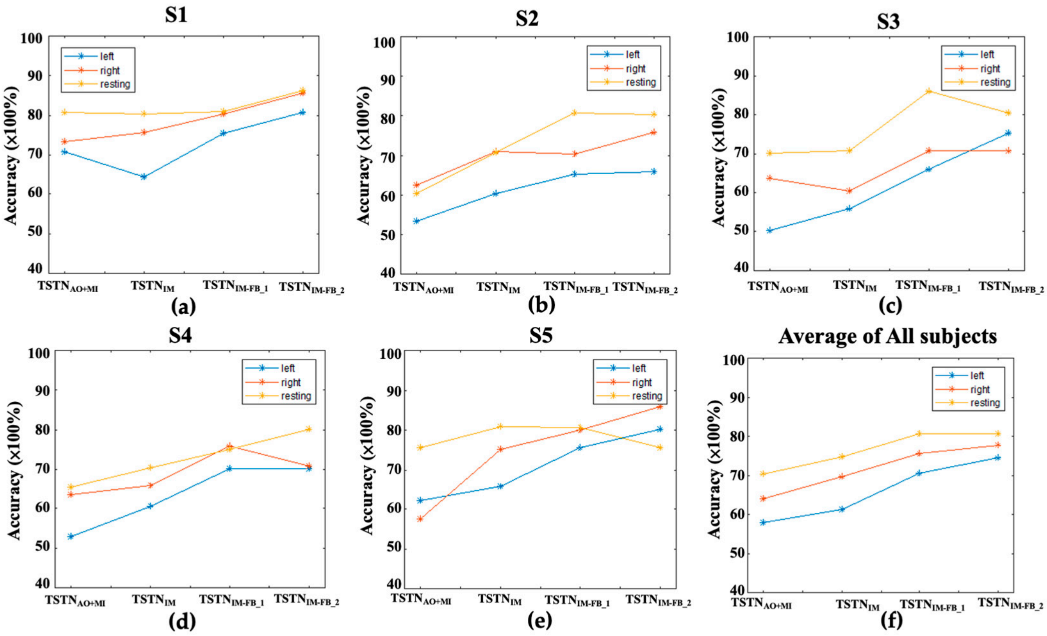


Figure 5. The detected accuracies by applying the TSTN$_{\text{AO}}$, TSTN$_{\text{MI}}$, TSTN$_{\text{MI-FB\_1}}$, and TSTN$_{\text{MI-FB\_2}}$ to the MI data (without feedback), first MI-FB data, second MI-FB data, and third MI-FB data in (**a**) S1, (**b**) S2, (**c**) S3, (**d**) S4, (**e**) S5, and (**f**) the average of all the five subjects.

In addition to the model built from individual data, we were interested in testing whether the data collected from different individuals could be pooled to train a BCI classifier. The TSTN$_{\text{AO+MI\_all}}$, TSTN$_{\text{MI\_all}}$, TSTN$_{\text{MI-FB\_1\_all}}$, and the TSTN$_{\text{MI-FB\_2\_all}}$ were obtained, following the training procedure described in Figure 4, while the measured data over the five subjects were pooled for classifier training. The classification accuracies, specificities, and F1 scores in the five subjects are listed in Table 2. The averaged accuracies in the five subjects were 0.61, 0.63, 0.68, and 0.70 for the TSTN$_{\text{AO+MI\_all}}$, TSTN$_{\text{MI\_all}}$, TSTN$_{\text{MI-FB\_1\_all}}$, and the TSTN$_{\text{MI-FB\_2\_all}}$, respectively. The averaged specificities in the five subjects were 0.78, 0.82, 0.84, and 0.84 for the TSTN$_{\text{AO+MI\_all}}$, TSTN$_{\text{MI\_all}}$, TSTN$_{\text{MI-FB\_1\_all}}$, and the TSTN$_{\text{MI-FB\_2\_all}}$, respectively. The averaged F1 scores in the five subjects were 0.61, 0.64, 0.67, and 0.70, respectively.

Table 2. Classification results of the TSTN network using the training data pooled for all subjects.

Classifier (Test Target)/Subject	**TSTN$_{\text{AO+IM\_All}}$ (MI Data)**	**TSTN$_{\text{IM\_All}}$ (1st MI-FB)**	**TSTN$_{\text{IM-FB\_1\_All}}$ (2nd MI-FB)**	**TSTN$_{\text{IM-FB\_2\_All}}$ (3rd MI-FB)**
	Acc/Spec/F1	Acc/Spec/F1	Acc/Spec/F1	Acc/Spec/F1
S1	0.65/0.77/0.64	0.660.83/0.67	0.71/0.85/0.71	0.73/0.85/0.73
S2	0.59/0.79/0.59	0.65/0.82/0.66	0.65/0.83/0.65	0.67/0.83/0.67
S3	0.62/0.78/0.62	0.62/0.83/0.62	0.70/0.85/0.70	0.71/0.85/0.71
S4	0.63/0.81/0.63	0.63/0.81/0.63	0.65/0.84/0.65	0.69/0.83/0.69
S5	0.58/79/0.58	0.61/0.81/0.62	0.65/0.83/0.65	0.68/0.84/0.68
Averaged accuracy	0.61/0.78/0.61	0.63/0.82/0.64	0.67/0.84/0.67	0.70/0.84/0.70

Figure 6a–e presents the detection accuracies using the classifiers trained from the pooled data of the five subjects. The averaged accuracies over the five subjects are shown in Figure 6f. It can be observed that the detection accuracies were increased with the continual training, which might indicate the feasibility of pooling subjects' data to obtain a generally applicable model. The detected accuracies of TSTN$_{\text{AO+MI\_all}}$ for MI data were

0.65, 0.59, 0.62, 0.63, and 0.58 for S1, S2, S3, S4, and S5, respectively. For $\text{TSTN}_{\text{MI\_all}}$, the detected accuracies for the first MI-FB data were 0.66, 0.65, 0.62, 0.63, and 0.61 for S1, S2, S3, S4, and S5, respectively. For $\text{TSTN}_{\text{MI-FB\_1\_all}}$, the detected accuracies for the second MI-FB data were 0.71, 0.65, 0.70, 0.68, and 0.65 for S1, S2, S3, S4, and S5, respectively. For $\text{TSTN}_{\text{MI-FB\_2\_all}}$, the detected accuracies for the third MI-FB data were 0.73, 0.67, 0.71, 0.69, and 0.68 for S1, S2, S3, S4, and S5, respectively.

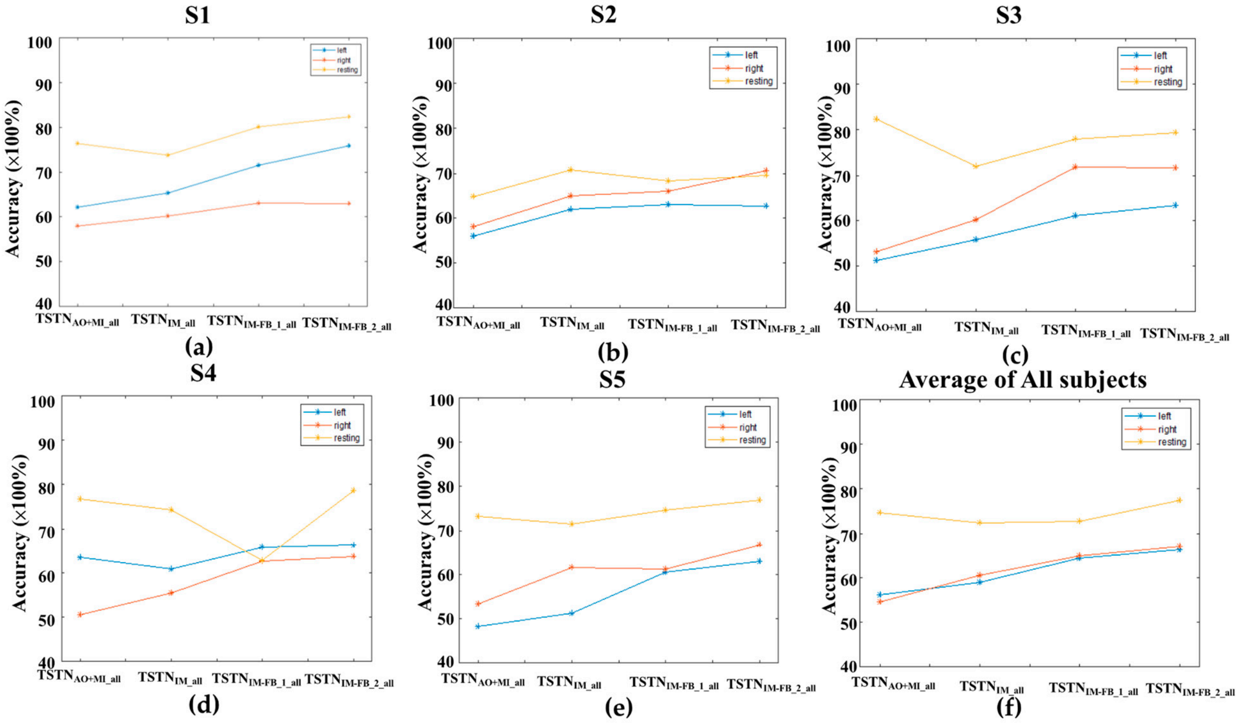


Figure 6. The detected accuracies by applying the $\text{TSTN}_{\text{AO+MI\_all}}$, $\text{TSTN}_{\text{MI\_all}}$, $\text{TSTN}_{\text{MI-FB\_1\_all}}$, and $\text{TSTN}_{\text{MI-FB\_2\_all}}$ to the MI (without feedback), first MI-FB, second MI-FB, and third MI-FB datasets in (**a**) S1, (**b**) S2, (**c**) S3, (**d**) S4, (**e**) S5 and (**f**) the average of all the five subjects.

To demonstrate the effectiveness of the TSTN network in this study, the EEGNet [45], DeepConvNet [46], and three SVMs with kernel functions of linear, polynomial, and radial basis functions were compared. All the classifiers were built according to the training procedure listed in Figure 4. The detected accuracies obtained from the TSTN network were compared with the five classifiers and the detection accuracies are listed in Table 3. It can be observed that all the classifiers showed increased detection accuracies in our continual learning process. The three deep learning networks (i.e., the TSTN, EEGNet, and DeepConvNet) showed comparable detection accuracies in the third MI-FB data, which were 0.77, 0.74, and 0.75 for TSTN, EEGNet, and DeepConvNet, respectively. For the initial model, the $\text{TSTN}_{\text{AO+MI}}$ had the highest detection accuracy (0.63) in the detection of MI data, compared to other classifiers. This might have been due to the substantial differences between the AO + MI and MI/MI-FB datasets. The AO + MI task activated both the neural circuitries of action observation and motor imagery, while the MI/MI-FB activated the neural circuitry of motor imagery only. Because the transformer has the greater ability to model long-distance dependencies among temporal embedded patches, it could provide more possibilities to find the dependency between different datasets.

Table 3. The comparisons of the detected accuracies between TSTN and SVMs with different kernel functions.

Classifier/Test Task	TSTN	$\text{SVM}_{\text{linear}}$	SVM_{poly}	SVM_{RBF}	EEGNet [44]	DeepConvNet [45]
MI	**0.63**	0.55	0.48	0.52	0.53	0.58
1st MI-FB	**0.68**	0.60	0.53	0.59	0.59	0.64
2nd MI-FB	**0.75**	0.67	0.56	0.67	0.69	0.72
3rd MI-FB	**0.77**	0.68	0.59	0.70	0.74	0.75

%Remark: $\text{SVM}_{\text{linear}}$: SVM with linear kernel function; SVM_{poly}: SVM with polynomial kernel function; SVM_{RBF}: SVM with radial basis kernel function; EEGNet: Compact convolutional network for EEG; DeepConvNet: Deep learning with convolutional neural networks.

4. Discussion

This study aimed to study the feasibility of using EEG data induced from AO + MI to build an initial model for MI-BCI, and the model could be continually improved with the provision of MI data in continual learning. Compared to the MI training in previous BCIs [47], those BCIs requested subjects to perform imagery movements by providing a simple visual or auditory cue. The lack of clear instructions makes it difficult to conform the ways for subjects to perform imagery movements. Therefore, if we can give subjects clear instructions to perform imagery movements, we might have the chance to reduce inter-individual difference and obtain generic brain wave patterns for training BCI classifiers. Alimardani et al. (2022) surveyed the impacts of different demographic and psychological variables [48] on the performance of imagery movement BCI in 54 subjects. They suggested the needs of prior BCI training and clear instructions for the experimental protocol are two of the most important factors which can influence a subject's BCI performance. In our previous study [49], the lack of explicit MI instructions would make it difficult for subjects to follow. Subjects may try to develop their own strategies to achieve better BCI accuracies. The distinct strategies in imagery movements could result in large inter-individual variations in the induced brain wave patterns, which makes the group analysis of MI data difficult.

In our research, we used the data obtained from the AO + MI task to create an initial model for MI-BCI control. The AO + MI task requested users to passively watch the action of the characters in the VR environment and concurrently imagine themselves performing the same movements. According to [50], Cengiz et al. (2018) studied ERD in the sensorimotor Mu rhythm when the subjects were performing an AO task. Zhanget et al. (2018) also found that AO can promote the effect of motor relearning [51], and the AO task can be performed concurrently with MI (AO + MI) to achieve more effective motor learning or a rehabilitation setting [38]. In our experiment, we asked the subjects to perform imagery movements the same as the actions of the virtual character, in order to achieve a better motor learning effect. It has been reported in the previous literature that the task of AO + MI can be performed either from the first person visual perspective [52] or from the third-person visual perspective [53,54]. Some studies have claimed the motor activities in AO + MI from the third-person visual perspective might involve rotation actions to transform the third-person visual perspective into a first-person imagery perspective [54]. However, it has been reported that the rotation and the transforming effects can be removed with the provision of a clear instruction, by asking subjects to imagine that the observed movements in the third-person perspective is similar to viewing the images in a mirror [38]. In our study, we instructed subjects to imagine that the virtual character is the reflection image of themselves in a mirror. The imagery actions in the following MI and MI-feedback tasks were the same as the actions they viewed in the AO task. Therefore, subjects recalled the images of virtual characters' actions in their MI and MI-FB tasks. Since the imagery movements in the AO + MI, MI, and MI-FB tasks were the same, this could be the reason why the detection accuracy is able to be improved in the continual learning process.

In this paper, we adopted the transformer-based deep learning network proposed by Song et al. (2021) [40] to classify the EEG signals. Compared to traditional ML, traditional ML assumes the data points are dispersed independently and identically. However, in many cases, the acquisition of these data is related to a subject's physiological state (e.g., cognitive state, neural learning, language learning, emotion, electrocardiogram (ECG)). The artificial neural network-based approaches usually have better flexibility to adapt themselves in the learning of sequential data [55]. There are three salient features in the use of TSTN. First, the TSTN utilizes CSP to create discriminable features by designing a set of spatial filters. Since distinct brain areas are responsible for different brain functions [56], the use of CSP can create a specific spatial filter to extract brain activities induced from a particular task. Second, the TSTN utilizes self-attention to enhance the feature-channel data extracted from CSP. In contrast to other popular approaches using convolutional neural networks (CNNs) [45,46], the classification of CNN is related to the selection of kernel, in which large convolutional kernels are used to capture high-resolution details and small convolutional kernels are used to extract low-resolution features. The TSTN does not have to decide the kernel size. It applies self-attention on the feature-channel data in order to weight those channels which are relevant to the performing of a subject's task. Moreover, the CNN does not consider the dependency of time-series information. Third, the TSTN slices the enhanced feature-channel data into signal patches and uses the multi-head transforming to perceive the dependencies along the temporal dimension. Compared to the previous literature using recurrent neural network (RNN) or long short-term memory networks (LSTMs) [57], both RNN and LSTM have the problems of vanishing gradients [58], so that RNN or LSTM-based solutions might have problem of dealing with EEG data induced from an MI task with a longer execution time. For example, stroke patients have slower information processing in moving the affected hand and require a longer execution time for performing mental tasks [59]. One noteworthy disadvantage of the transformer-based network is its huge model size. The TSTN requires considerable computing power and training time to achieve good accuracy. Implementation of the proposed model on wearable devices might be difficult.

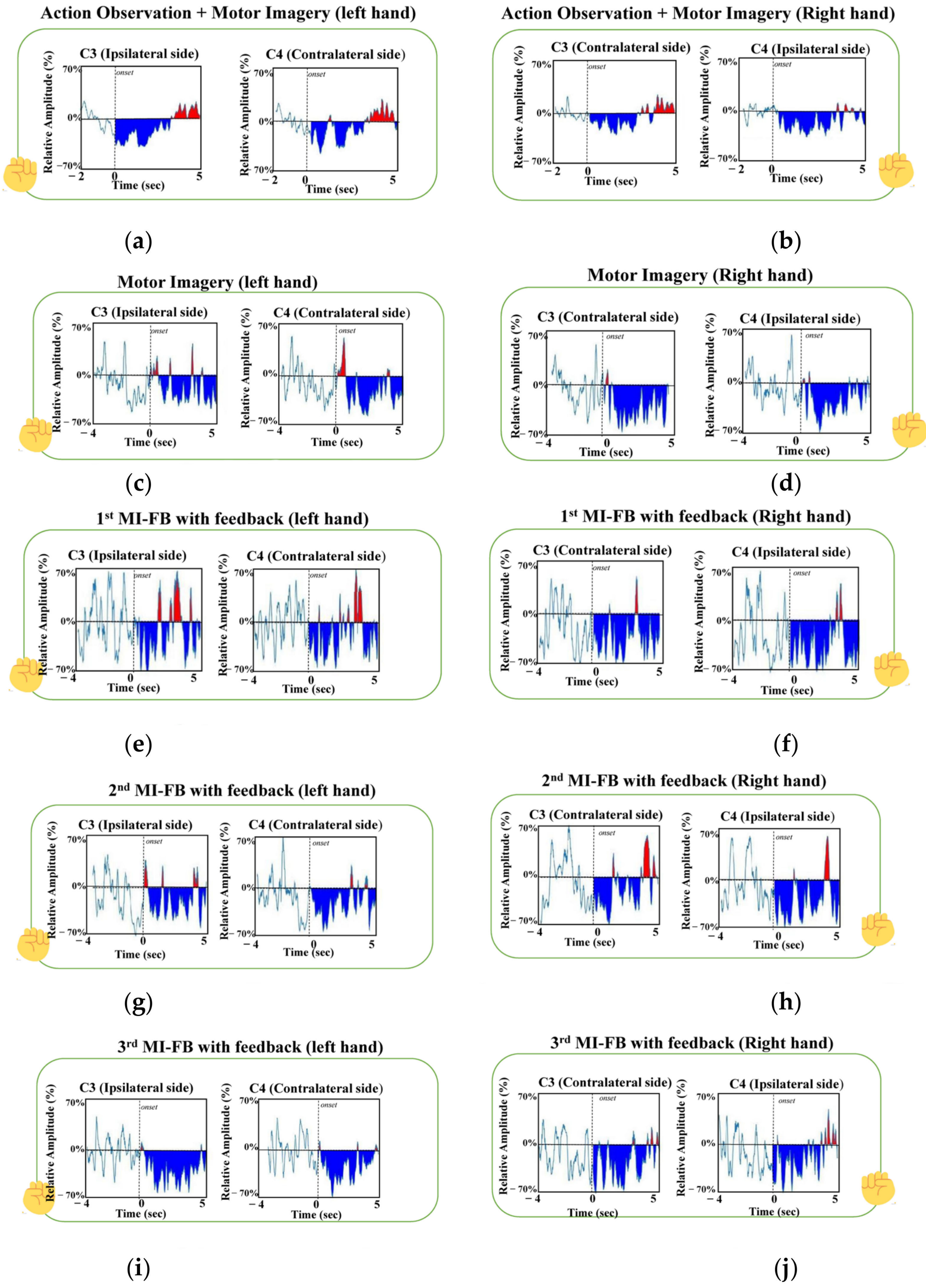
In order to show the model interpretability of the TSTN neural network, the correctly classified trials were used to improve the classifiers in the continual learning processes. The relative power change of each correctly classified trial was calculated, in which the relative event-related power change was calculated using the ERD/ERS technique, formulated by Pfurtscheller and Lopes da Silva (1999) [60]. The ERD/ERS describes that the power decrease/increase within a specific frequency band anchored to a given event is calculated relative to the power of a reference period. The ERD/ERS represent the percentages of power changes according to the reference period, in which the relative power change can be represented as follows:

$$\mathrm{RP}\ (\%) = (\mathrm{A} - \mathrm{R})/\mathrm{R}\ \times\ 100\%, \tag{11}$$

where RP% is the percentage of a relative power change within alpha band (8~13 Hz) [61], A is the power of event-induced signal power, and R is the power in the reference period.

In Equation (11), the A is calculated from the period of the AO + MI or MI time block, and the reference period R is calculated from the resting period, from −2 s to 0 s, preceding the action of a virtual character or the presence of the visual cue for imagery movement. Figure 7 shows the RPs obtained from the average of the five subjects in the four datasets. In the MI and MI-FB datasets, only the RPs of the correctly classified trials were calculated. The RP of the AO + MI tasks in viewing the left-hand and right-hand movements are shown in Figures 7a and 7b, respectively. The RP changes in left-hand and right-hand MI tasks are shown in Figures 7c and 7d, respectively. The RP changes of the first MI-FB, second MI-FB, and third MI-FB datasets in the left-hand and right-hand movements are shown in Figures 7e and 7f, Figures 7g and 7h, and Figures 7i and 7j, respectively. It can be observed that the ERD (the blue area of each plot in Figure 7) was enhanced with the continual learning process. This demonstrated that the TSTN network can effectively pick up those

EEG trials with prominent ERDs. In Figure 7, only the ERD within the alpha band was demonstrated, because the ERD in the alpha band has been reported as the most indicative parameter in operating an MI-BCI system [61].



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**) (**j**)

Figure 7. The RPs obtained from the average of the five subjects in the four datasets. The ERD and ERS are marked in blue and red colors, respectively. (**a**) The RP changes in the left-hand AO + MI task. (**b**) The RP changes in the right-hand AO + MI task. (**c**) The RP changes in the left-hand MI task. (**d**) The RP changes in the right-hand MI task. (**e**) The RP changes of the left-hand imagery movement in the first MI-FB task. (**f**) The RP changes of right-hand imagery movement in the first MI-FB task. (**g**) The RP changes of left-hand imagery movement in the second MI-FB task. (**h**) The RP changes of right-hand imagery movement in the second MI-FB task. (**i**) The RP changes of left-hand imagery movement in the third MI-FB task. (**j**) The RP changes of right-hand imagery movement in the third MI-FB task.

The BCI performance of our TSTN network was improved with the continual learning process; not only was the detection accuracy improved (see Tables 1 and 2), but the value of ERD was also enhanced (see Figure 7). The implication between ERD and the classification accuracy echoes the findings in the previous literature that ERD is an important parameter in MI-BCI training [61]. Nevertheless, we did not see the phenomenon of ERD lateralization as mentioned in previous MI studies [62]. This might have been due to the difference in experiment design between our study and the other previous literature. In our study, we requested subjects to perform AO + MI and recall the action imagination in the following MI and MI-FB tasks. The memory recall of action imagery in MI and MI-FB tasks could involve the mirror neuron system, which presents the activations in the bilateral primary motor cortex, the primary somatosensory cortex, and the middle frontal cortex [51]. According to the study proposed by Rizzolatti et al. (2010) [63], the bilaterally distributed parietofrontal network in the mirror neuron system serves as a neural substrate to achieve the transformation of visual information into motor execution (i.e., visuomotor transformation).

Previous BCI studies only gave subjects a simple visual or auditory cue (e.g., visual or auditory cues, etc.) to perform imagery movements [64,65]. Unclear instruction could make it difficult for subjects to follow the experimenter's guidance to achieve the desired task [65]. Since the AO task has been reported as an effective way to activate the human motor cortex with brain wave patterns similar to those induced by the MI task, we found it was easier for subjects to understand the experimenter's instruction by requesting them to watch and follow the virtual character's actions, instead of just providing them with simple textual instruction or visual/auditory cues. To the best of our knowledge, our research is the first study to train the BCI classifiers using EEG signals induced from AO + MI and MI tasks. Current BCI databases (e.g., BNCI Horizon; http://bnci-horizon-2020.eu/database/data-sets, accessed on 12 October 2022) do not collect both the EEG signals of AO + MI and MI tasks for the same subjects. The collected EEG data in this study could inspire future research on studying the relevance between observation and imagery movements.

One limitation of our current study is the small sample size. Because we wanted to demonstrate the training transition of a BCI classifier from AO + MI to MI-FB tasks, the subject had to join one experiment in one week and it took four weeks to complete the data collection for a subject. The small amount of data was not sufficient to have an effective comparison for the detection performances among different classifiers (see Table 3). Nevertheless, this paper aimed to study the feasibility of continual learning in a BCI classifier, from AO + MI data to MI-FB data. We observed that the detection accuracies were improved in all classifiers, and all the deep learning classifiers (i.e., TSTN, EEGNet, and DeepConvNet) showed superior performances than those in the use of traditional SVMs. The second limitation of this study is the huge computation load of our transformer framework. The transformer-based classifier has a large model size which has difficulties in coping with fast updating or fluctuation conditions.

5. Conclusions

In this study, we designed a BCI training method in the virtual reality (VR) environment. Subjects wore a head-mounted device (HMD) [32] and started MI training from the AO + MI task in VR environments. Unlike other MI-BCI studies which asked naïve subjects to perform unfamiliar MI tasks, the training procedure provided an easier way for subjects to conform their imagery actions. The AO-oriented training procedure also provided a more flexible design for BCI training. Subjects were only requested to follow the movements of the virtual characters and simultaneously performed imagery actions. The utilization of the AO + MI data to build the initial classifier model had several advantages. First, the instruction of AO + MI was much easier for subjects to follow [36] compared to the traditional MI task which prompts subjects to perform imagery actions by providing a simple visual/auditory cue. Second, the AO + MI provided a convenient and standardized experimental protocol. Third, the AO + MI elicited increased neural activities in motor-related brain areas, relative to the use of AO or MI only [38]. Fourth, the AO + MI has been proved as an effective way for motor learning and rehabilitation in clinics [66]. This study has answered the following issues: (1) the effectiveness of using AO + MI data to build an initial model for MI classification was validated; (2) the use of a continual learning process for the improvement of classifier performance was demonstrated; (3) the feasibility of a transformer-based deep learning model in MI-BCI classification was demonstrated; (4) the interpretability of the proposed TSTN model was shown in the analysis of alpha ERD in different BCI training steps. Our current study achieved a mean detection accuracy of 77% over the five participants in the three-class classification. In future applications, experimenters will be able to change the actions of the virtual character to train wanted imagery actions, which is important for the use of BCI in metaverse applications.

Author Contributions: Conceptualization, P.-L.L.; methodology, P.-L.L., S.-H.C. and T.-C.C.; validation, W.-K.L. and H.-T.H.; formal analysis, P.-L.L. and H.-H.C.; investigation, W.-K.L.; resources, H.-H.C.; writing—original draft preparation, P.-L.L. and H.-H.C.; writing—review and editing, S.-H.C. and H.-H.C.; project administration, P.-L.L.; funding acquisition, P.-L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science and Technology Council under Grant Number: 11-2634-F-A49-013 through Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan. This study was also funded by the National Central University, National Science and Technology Council under Grant Number: 110-2221-E-008-095-MY3, 10-2923-E-008-002-MY3, 111-2811-E-008-002-MY3, 111-2221-E-008-102-MY3, 111-2217-E-008-001-MY3, 111-2622-E-008-013.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Tao-Yuan General Hospital, Taiwan. (IRB number: TYGH 107055).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wolpaw, J.R. Brain-Computer Interfaces (BCIs) for Communication and Control. In Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility, New York, NY, USA, 15–17 October 2007; pp. 1–2.
2. Chen, C.-H.; Ho, M.-S.; Shyu, K.-K.; Hsu, K.-C.; Wang, K.-W.; Lee, P.-L. A noninvasive brain computer interface using visually-induced near-infrared spectroscopy responses. *Neurosci. Lett.* **2014**, *580*, 22–26. [CrossRef] [PubMed]
3. Behzad, R.; Behzad, A. The Role of EEG in the Diagnosis and Management of Patients with Sleep Disorders. *J. Behav. Brain Sci.* **2021**, *11*, 257–266. [CrossRef]
4. Follis, J.L.; Lai, D. Modeling Volatility Characteristics of Epileptic EEGs using GARCH Models. *Signals* **2020**, *1*, 26–46. [CrossRef]
5. Ahmed, M.Z.I.; Sinha, N.; Phadikar, S.; Ghaderpour, E. Automated Feature Extraction on AsMap for Emotion Classification Using EEG. *Sensors* **2022**, *22*, 2346. [CrossRef]
6. Phadikar, S.; Sinha, N.; Ghosh, R.; Ghaderpour, E. Automatic Muscle Artifacts Identification and Removal from Single-Channel EEG Using Wavelet Transform with Meta-Heuristically Optimized Non-Local Means Filter. *Sensors* **2022**, *22*, 2948. [CrossRef]
7. Jiang, X.; Bian, G.-B.; Tian, Z. Removal of artifacts from EEG signals: A review. *Sensors* **2019**, *19*, 987. [CrossRef]
8. Wolpaw, J.R.; Birbaumer, N.; Heetderks, W.J.; McFarland, D.J.; Peckham, P.H.; Schalk, G.; Donchin, E.; Quatrano, L.A.; Robinson, C.J.; Vaughan, T.M. Brain-computer interface technology: A review of the first international meeting. *IEEE Trans. Rehabil. Eng.* **2000**, *8*, 164–173. [CrossRef]
9. Ahn, M.; Jun, S.C. Performance variation in motor imagery brain–computer interface: A brief review. *J. Neurosci. Methods* **2015**, *243*, 103–110. [CrossRef]
10. Kaongoen, N.; Choi, J.; Jo, S. Speech-imagery-based brain–computer interface system using ear-EEG. *J. Neural Eng.* **2021**, *18*, 016023. [CrossRef]
11. Dornhege, G.; Blankertz, B.; Curio, G. Speeding up classification of multi-channel brain-computer interfaces: Common spatial patterns for slow cortical potentials. In Proceedings of the First International IEEE EMBS Conference on Neural Engineering, Capri, Italy, 20–22 March 2003; IEEE: Piscataway, NJ, USA, 2003; pp. 595–598.
12. Lee, P.-L.; Sie, J.-J.; Liu, Y.-J.; Wu, C.-H.; Lee, M.-H.; Shu, C.-H.; Li, P.-H.; Sun, C.-W.; Shyu, K.-K. An SSVEP-actuated brain computer interface using phase-tagged flickering sequences: A cursor system. *Ann. Biomed. Eng.* **2010**, *38*, 2383–2397. [CrossRef]

13. Lee, P.-L.; Hsieh, J.-C.; Wu, C.-H.; Shyu, K.-K.; Wu, Y.-T. Brain computer interface using flash onset and offset visual evoked potentials. *Clin. Neurophysiol.* **2008**, *119*, 605–616. [CrossRef]
14. Hill, N.J.; Schölkopf, B. An online brain–computer interface based on shifting attention to concurrent streams of auditory stimuli. *J. Neural Eng.* **2012**, *9*, 026011. [CrossRef] [PubMed]
15. Smith, E.; Delargy, M. Locked-in syndrome. *BMJ* **2005**, *330*, 406–409. [CrossRef] [PubMed]
16. Padfield, N.; Zabalza, J.; Zhao, H.; Masero, V.; Ren, J. EEG-based brain-computer interfaces using motor-imagery: Techniques and challenges. *Sensors* **2019**, *19*, 1423. [CrossRef] [PubMed]
17. Khare, S.K.; Gaikwad, N.; Bokde, N.D. An Intelligent Motor Imagery Detection System Using Electroencephalography with Adaptive Wavelets. *Sensors* **2022**, *22*, 8128. [CrossRef] [PubMed]
18. Khare, S.K.; Bajaj, V. A facile and flexible motor imagery classification using electroencephalogram signals. *Comput. Methods Programs Biomed.* **2020**, *197*, 105722. [CrossRef]
19. Stefano Filho, C.A.; Attux, R.; Castellano, G. Actual, sham and no-feedback effects in motor imagery practice. *Biomed. Signal Process. Control* **2022**, *71*, 103262. [CrossRef]
20. Alimardani, M.; Nishio, S.; Ishiguro, H. Brain-computer interface and motor imagery training: The role of visual feedback and embodiment. *Evol. BCI Ther.-Engag. Brain State Dyn.* **2018**, *2*, 64.
21. Friedman, D.; Leeb, R.; Pfurtscheller, G.; Slater, M. Human–computer interface issues in controlling virtual reality with brain–computer interface. *Hum. Comput. Interact.* **2010**, *25*, 67–94. [CrossRef]
22. McCreadie, K.A.; Coyle, D.H.; Prasad, G. Sensorimotor learning with stereo auditory feedback for a brain–computer interface. *Med. Biol. Eng. Comput.* **2013**, *51*, 285–293. [CrossRef]
23. Nijboer, F.; Furdea, A.; Gunst, I.; Mellinger, J.; McFarland, D.J.; Birbaumer, N.; Kübler, A. An auditory brain–computer interface (BCI). *J. Neurosci. Methods* **2008**, *167*, 43–50. [CrossRef] [PubMed]
24. Ishihara, W.; Moxon, K.; Ehrman, S.; Yarborough, M.; Panontin, T.L.; Nathan-Roberts, D. Feedback modalities in brain–computer interfaces: A systematic review. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2020**, *64*, 1186–1190.
25. Jo, S.; Choi, J.W. Effective motor imagery training with visual feedback for non-invasive brain computer interface. In Proceedings of the 2018 6th International Conference on Brain-Computer Interface (BCI), Gangwon, Korea, 15–17 January 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–4.
26. Ziadeh, H.; Gulyas, D.; Nielsen, L.D.; Lehmann, S.; Nielsen, T.B.; Kjeldsen, T.K.K.; Hougaard, B.I.; Jochumsen, M.; Knoche, H. "Mine Works Better": Examining the Influence of Embodiment in Virtual Reality on the Sense of Agency During a Binary Motor Imagery Task with a Brain-Computer Interface. *Front. Psychol.* **2021**, *12*, 6174–6184. [CrossRef] [PubMed]
27. Achanccaray, D.; Pacheco, K.; Carranza, E.; Hayashibe, M. Immersive virtual reality feedback in a brain computer interface for upper limb rehabilitation. In Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 7–10 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1006–1010.
28. Braun, N.; Emkes, R.; Thorne, J.D.; Debener, S. Embodied neurofeedback with an anthropomorphic robotic hand. *Sci. Rep.* **2016**, *6*, 1–13. [CrossRef]
29. Alimardani, M.; Nishio, S.; Ishiguro, H. Effect of biased feedback on motor imagery learning in BCI-teleoperation system. *Front. Syst. Neurosci.* **2014**, *8*, 52. [CrossRef]
30. Kübler, A.; Neumann, N.; Wilhelm, B.; Hinterberger, T.; Birbaumer, N. Predictability of brain-computer communication. *J. Psychophysiol.* **2004**, *18*, 121–129. [CrossRef]
31. Montemurro, N.; Condino, S.; Carbone, M.; Cattari, N.; D'Amato, R.; Cutolo, F.; Ferrari, V. Brain Tumor and Augmented Reality: New Technologies for the Future. *Int. J. Environ. Res. Public Health* **2022**, *19*, 6347.
32. Sayadi, L.R.; Naides, A.; Eng, M.; Fijany, A.; Chopan, M.; Sayadi, J.J.; Shaterian, A.; Banyard, D.A.; Evans, G.R.; Vyas, R. The new frontier: A review of augmented reality and virtual reality in plastic surgery. *Aesthetic Surg. J.* **2019**, *39*, 1007–1016. [CrossRef]
33. Buccino, G. Action observation treatment: A novel tool in neurorehabilitation. *Philos. Trans. R. Soc. B Biol. Sci.* **2014**, *369*, 20130185. [CrossRef]
34. Altschuler, E.L.; Wisdom, S.B.; Stone, L.; Foster, C.; Galasko, D.; Llewellyn, D.M.E.; Ramachandran, V.S. Rehabilitation of hemiparesis after stroke with a mirror. *Lancet* **1999**, *353*, 2035–2036. [CrossRef] [PubMed]
35. Hsieh, Y.-W.; Lin, Y.-H.; Zhu, J.-D.; Wu, C.-Y.; Lin, Y.-P.; Chen, C.-C. Treatment effects of upper limb action observation therapy and mirror therapy on rehabilitation outcomes after subacute stroke: A pilot study. *Behav. Neurol.* **2020**, *2020*, e6250524. [CrossRef]
36. Vogt, S.; Di Rienzo, F.; Collet, C.; Collins, A.; Guillot, A. Multiple roles of motor imagery during action observation. *Front. Hum. Neurosci.* **2013**, *7*, 807. [CrossRef]
37. Hardwick, R.M.; Caspers, S.; Eickhoff, S.B.; Swinnen, S.P. Neural correlates of motor imagery, action observation, and movement execution: A comparison across quantitative meta-analyses. *BioRxiv* **2018**, *94*, 31–44.
38. Eaves, D.L.; Riach, M.; Holmes, P.S.; Wright, D.J. Motor imagery during action observation: A brief review of evidence, theory and future research opportunities. *Front. Neurosci.* **2016**, *10*, 514. [CrossRef]
39. Yu, T.; Xiao, J.; Wang, F.; Zhang, R.; Gu, Z.; Cichocki, A.; Li, Y. Enhanced motor imagery training using a hybrid BCI with feedback. *IEEE Trans. Biomed. Eng.* **2015**, *62*, 1706–1717. [CrossRef]

40. Song, Y.; Jia, X.; Yang, L.; Xie, L. Transformer-based spatial-temporal feature learning for eeg decoding. *arXiv* **2021**, arXiv:2106.11170.
41. Wang, Y.; Gao, S.; Gao, X. Common spatial pattern method for channel selelction in motor imagery based brain-computer interface. In Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China, 1–4 September 2005; IEEE: Piscataway, NJ, USA, 2006; pp. 5392–5395.
42. Fox, N.A.; Bakermans-Kranenburg, M.J.; Yoo, K.H.; Bowman, L.C.; Cannon, E.N.; Vanderwert, R.E.; Ferrari, P.F.; Van IJzendoorn, M.H. Assessing human mirror activity with EEG mu rhythm: A meta-analysis. *Psychol. Bull.* **2016**, *142*, 291. [CrossRef] [PubMed]
43. Barbero, Á.; Grosse-Wentrup, M. Biased feedback in brain-computer interfaces. *J. Neuroeng. Rehabil.* **2010**, *7*, 34. [CrossRef] [PubMed]
44. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
45. Lawhern, V.J.; Solon, A.J.; Waytowich, N.R.; Gordon, S.M.; Hung, C.P.; Lance, B.J. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural Eng.* **2018**, *15*, 056013. [CrossRef]
46. Schirrmeister, R.T.; Springenberg, J.T.; Fiederer, L.D.J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; Ball, T. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* **2017**, *38*, 5391–5420. [CrossRef]
47. Choi, D.; Lee, Y.; Jeong, W.; Lee, S.; Kang, D.; Lee, M. Evaluation of Motor Imagery Using Combined Cue Based EEG-Brain Computer Interface. In Proceedings of the 5th Kuala Lumpur International Conference on Biomedical Engineering 2011, Kuala Lumpur, Malaysia, 20–23 June 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 516–518.
48. Alimardani, M.; Gherman, D.-E. Individual Differences in Motor Imagery BCIs: A Study of Gender, Mental States and Mu Suppression. In Proceedings of the 2022 10th International Winter Conference on Brain-Computer Interface (BCI), Gangwon-do, Republic of Korea, 21–23 February 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–7.
49. Hung, C.-I.; Lee, P.-L.; Wu, Y.-T.; Chen, L.-F.; Yeh, T.-C.; Hsieh, J.-C. Recognition of motor imagery electroencephalography using independent component analysis and machine classifiers. *Ann. Biomed. Eng.* **2005**, *33*, 1053–1070. [CrossRef]
50. Cengiz, B.; Vurallı, D.; Zinnuroğlu, M.; Bayer, G.; Golmohammadzadeh, H.; Günendi, Z.; Turgut, A.E.; İrfanoğlu, B.; Arıkan, K.B. Analysis of mirror neuron system activation during action observation alone and action observation with motor imagery tasks. *Exp. Brain Res.* **2018**, *236*, 497–503. [CrossRef]
51. Zhang, J.J.; Fong, K.N.; Welage, N.; Liu, K.P. The activation of the mirror neuron system during action observation and action execution with mirror visual feedback in stroke: A systematic review. *Neural Plast.* **2018**, *2018*, e2321045. [CrossRef] [PubMed]
52. Villiger, M.; Estévez, N.; Hepp-Reymond, M.-C.; Kiper, D.; Kollias, S.S.; Eng, K.; Hotz-Boendermaker, S. Enhanced activation of motor execution networks using action observation combined with imagination of lower limb movements. *PloS ONE* **2013**, *8*, e72403. [CrossRef] [PubMed]
53. Mouthon, A.; Ruffieux, J.; Wälchli, M.; Keller, M.; Taube, W. Task-dependent changes of corticospinal excitability during observation and motor imagery of balance tasks. *Neuroscience* **2015**, *303*, 535–543. [CrossRef] [PubMed]
54. Taube, W.; Mouthon, M.; Leukel, C.; Hoogewoud, H.-M.; Annoni, J.-M.; Keller, M. Brain activity during observation and motor imagery of different balance tasks: An fMRI study. *Cortex* **2015**, *64*, 102–114. [CrossRef]
55. Janiesch, C.; Zschech, P.; Heinrich, K. Machine learning and deep learning. *Electron. Mark.* **2021**, *31*, 685–695. [CrossRef]
56. Kim, J.; Yeon, J.; Ryu, J.; Park, J.-Y.; Chung, S.-C.; Kim, S.-P. Neural activity patterns in the human brain reflect tactile stickiness perception. *Front. Hum. Neurosci.* **2017**, *11*, 445. [CrossRef]
57. Hosman, T.; Vilela, M.; Milstein, D.; Kelemen, J.N.; Brandman, D.M.; Hochberg, L.R.; Simeral, J.D. BCI decoder performance comparison of an LSTM recurrent neural network and a Kalman filter in retrospective simulation. In Proceedings of the 2019 9th International IEEE/EMBS Conference on Neural Engineering (NER), San Francisco, CA, USA, 20–23 March 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1066–1071.
58. Fadziso, T. Overcoming the Vanishing Gradient Problem during Learning Recurrent Neural Nets (RNN). *Asian J. Appl. Sci. Eng.* **2020**, *9*, 207–218.
59. Winkens, I.; Van Heugten, C.M.; Wade, D.T.; Habets, E.J.; Fasotti, L. Efficacy of time pressure management in stroke patients with slowed information processing: A randomized controlled trial. *Arch. Phys. Med. Rehabil.* **2009**, *90*, 1672–1679. [CrossRef]
60. Pfurtscheller, G.; Da Silva, F.L. Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clin. Neurophysiol.* **1999**, *110*, 1842–1857. [CrossRef]
61. Ahn, M.; Cho, H.; Ahn, S.; Jun, S.C. High theta and low alpha powers may be indicative of BCI-illiteracy in motor imagery. *PLoS ONE* **2013**, *8*, e80886. [CrossRef]
62. Lotte, F.; Rimbert, S. How ERD Modulations during Motor Imageries Relate to Users' Traits and BCI Performances. In Proceedings of the 44th International Engineering in Medicine and Biology Conference, Glasgow, UK, 11–15 July 2022.
63. Rizzolatti, G.; Sinigaglia, C. The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nat. Rev. Neurosci.* **2010**, *11*, 264–274. [CrossRef]
64. Benzy, V.; Vinod, A.; Subasree, R.; Alladi, S.; Raghavendra, K. Motor imagery hand movement direction decoding using brain computer interface to aid stroke recovery and rehabilitation. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2020**, *28*, 3051–3062. [CrossRef] [PubMed]

65. Qiu, Z.; Allison, B.Z.; Jin, J.; Zhang, Y.; Wang, X.; Li, W.; Cichocki, A. Optimized motor imagery paradigm based on imagining Chinese characters writing movement. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2017**, *25*, 1009–1017. [CrossRef] [PubMed]
66. Higuchi, S.; Holle, H.; Roberts, N.; Eickhoff, S.B.; Vogt, S. Imitation and observational learning of hand actions: Prefrontal involvement and connectivity. *Neuroimage* **2012**, *59*, 1668–1683. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.