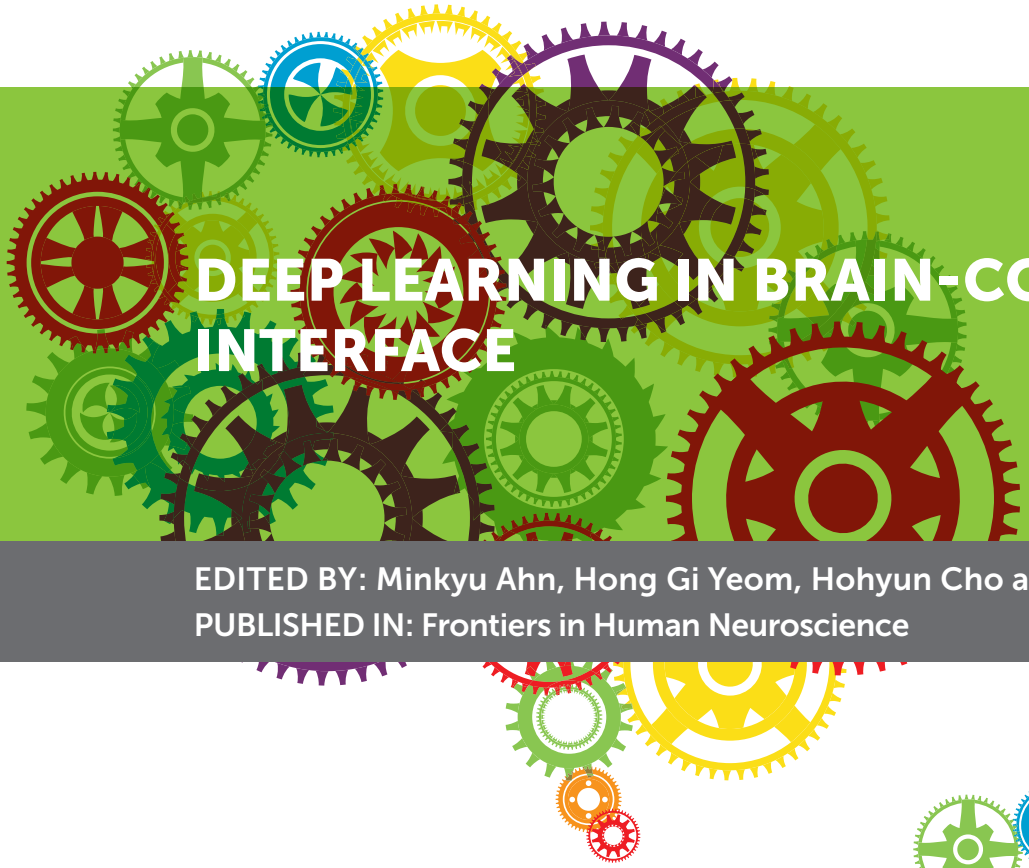

DEEP LEARNING IN BRAIN-COMPUTER INTERFACE

EDITED BY: Minkyu Ahn, Hong Gi Yeom, Hohyun Cho and Sung Chan Jun

PUBLISHED IN: Frontiers in Human Neuroscience

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88976-328-3
DOI 10.3389/978-2-88976-328-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

DEEP LEARNING IN BRAIN-COMPUTER INTERFACE

Topic Editors:
Minkyu Ahn, Handong Global University, South Korea
Hong Gi Yeom, Chosun University, South Korea
Hohyun Cho, Washington University in St. Louis, United States
Sung Chan Jun, Gwangju Institute of Science and Technology, South Korea

Citation: Ahn, M., Yeom, H. G., Cho, H., Jun, S. C., eds. (2022). Deep Learning in Brain-Computer Interface. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-328-3

Table of Contents

frontiers | Frontiers in Human Neuroscience

EDITORIAL
published: 19 May 2022
doi: 10.3389/fnhum.2022.927567

Editorial: Deep Learning in Brain-Computer Interface

Minkyu Ahn[1]*, *Sung Chan Jun*[2], *Hong Gi Yeom*[3] and *Hohyun Cho*[4]

[1] School of Computer Science and Electrical Engineering, Handong Global University, Pohang-si, South Korea, [2] School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, South Korea, [3] Department of Electronics Engineering, Chosun University, Gwangju, South Korea, [4] Department of Neurosurgery, Washington University School of Medicine, St. Louis, MO, United States

Keywords: brain-computer interface, deep learning, machine learning, transfer learning, data augmentation, explainable artificial intelligence

Editorial on Research Topic

Editorial: Deep Learning in Brain-Computer Interface

INTRODUCTION

Recent advancements in deep learning with the support of large-scale datasets and computational power have led many studies to adopt deep neural networks (DNNs) to extract features from brain signals and decode brain states, which is an important element in brain-computer interface (BCI). However, several issues remain to be resolved for BCIs to be applicable in the real world. Brain signals are high-dimensional, noisy, and highly nonstationary. In addition, the datasets are limited substantially compared to image data in computer vision fields. Thus, further research that focuses on deep learning (DL) in applications to BCI and a thorough evaluation of the way this application can be used in practice to implement the interface would be beneficial. The primary goal of this Research Topic is to provide an assorted and complementary collection of contributions that show new advancements and review deep learning methods or approaches in BCIs, as well as create a forum for discussion that brings together researchers' contributions to allow progress in deep learning-based BCIs.

RESEARCH TOPIC COVERAGE

We collected two reviews and seven research papers on this Research Topic. The authors of the publications accepted presented articles that cover a DL-based BCI for specific applications and a novel model for high BCI performance or transfer learning, data augmentation.

Gutierrez-Martinez et al. conducted a systematic review that covers the current state-of-the-art in visual evoked potential-based BCIs (e.g., P300 or SSVEP-based BCIs) for motor rehabilitation applications and artificial intelligence (AI) algorithms used for detection and classification by analyzing many recent articles. The authors provided an overview of the topic of interest, from traditional machine learning (ML) techniques to cutting-edge DL trends and discussed future challenges in the field.

Ko et al. surveyed the recent advances in short/zero calibration methods in the field of DL-based BCIs. In particular, they provided a good overview of generative model-based and geometric manipulation-based data augmentation methods, and transfer learning techniques that use explicit or implicit methods in DL-based BCIs. The trend in short/zero calibration methods is discussed in detail, and recommendations for the practical use of DL for short or zero calibration BCIs are made for potential users.

OPEN ACCESS

Edited and reviewed by:
Gernot R. Müller-Putz,
Graz University of Technology, Austria

***Correspondence:**
Minkyu Ahn
minkyuahn@handong.edu

Specialty section:
This article was submitted to Brain-Computer Interfaces, a section of the journal Frontiers in Human Neuroscience

Received: *24 April 2022*
Accepted: *04 May 2022*
Published: *19 May 2022*

Citation:
Ahn M, Jun SC, Yeom HG and Cho H (2022) Editorial: Deep Learning in Brain-Computer Interface. Front. Hum. Neurosci. 16:927567. doi: 10.3389/fnhum.2022.927567

Bao et al. proposed a model of a two-level domain adaptation neural network to construct a transfer model for electroencephalography (EEG)-based emotion recognition. The first level uses the maximum mean discrepancy to minimize the distribution discrepancy in deep features from the topological graph of EEG signals, while the second uses the domain adversarial neural network to force the deep features closer to the center of their corresponding class.

Aldayel et al. presented a study on preference detection of a neuromarketing dataset using different combinations of EEG features and different algorithms. The comparison of the algorithms revealed that the deep neural network outperforms k-nearest neighbor (KNN) and support vector machine (SVM) in accuracy, precision, and recall, while Random Forest (RF) achieved similar performance to that of the DNN.

Dolmans et al. presented a novel deep learning model that deals with multimodal data (Galvanic skin response, photoplethysmograms, functional near-infrared spectrograms (fNIRS) and eye movements) to classify perceived mental workload, which is also an interesting and important area in the BCI field.

Borra et al. proposed a novel convolutional neural network (CNN) model, which has a lightweight multi-scale design and guarantees high performance for P300-based BCIs. This model merges the multi-scale temporal learning, which allows a greater decrease in the number of trainable parameters than conventional models and learns multi-scale features as well.

Pei et al. presented a data augmentation method that uses channel-level recombination for motor imagery BCI. To obtain an augmented training set, they divided each sample into two according to the brain region to which the channel belongs and then recombined those samples. Based upon a simulation study, they reported that a CNN model trained with these augmented samples outperforms the typical decoding algorithms.

Kwon and Im proposed a subject-independent CNN-based model for fNIRS-based BCI. This model is designed to be relatively simple, since it uses one-dimensional CNN, but shows reasonable performance in decoding a mental arithmetic state from idle state. The authors tested their model with the typical Linear Discriminant analysis (LDA) and typical CNN model [e.g., EEGNET (Lawhern et al., 2018)] in subject-dependent and independent settings.

Although the typical Transfer Learning approach, which uses known (or labeled) data has shown good performance, the way to build a model that works for unseen data is also of interest. To address this issue, Kostas et al. investigated a self-supervised training approach. Specifically, they adapted techniques and architectures used for language model that show the ability to ingest amounts of data, to EEG analysis. In the study, arbitrary EEG segments were encoded as a sequence of learned vectors, referred to as "BErt-inspired Neural Data Representations", to determine whether the model is transferable to unseen EEG datasets recorded from unseen subjects, different hardware, and different tasks. The results showed high potential to make an excellent contribution to the BCI field.

All of the articles presented showed important ideas in AI and deep learning approaches in BCIs. The editors are pleased to present this collection of articles to the BCI field and related scientific communities, and hope that it will help researchers advance BCI and its applications.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural Eng.* 15, 056013. doi: 10.1088/1741-2552/aace8c

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 11 January 2021
doi: 10.3389/fnhum.2020.609096

Perceived Mental Workload Classification Using Intermediate Fusion Multimodal Deep Learning

Tenzing C. Dolmans[1,2], Mannes Poel[1], Jan-Willem J. R. van 't Klooster[2] and Bernard P. Veldkamp[3]*

[1]Data Management and Biometrics, University of Twente, Enschede, Netherlands, [2]Behavioural, Management and Social Sciences Lab, University of Twente, Enschede, Netherlands, [3]Research Methodology, Measurement, and Data Analysis, University of Twente, Enschede, Netherlands

OPEN ACCESS

Edited by:
Sung Chan Jun, Gwangju Institute of Science and Technology, South Korea

Reviewed by:
Valentin Vielzeuf, Orange (France), France
Andrey Eliseyev, Columbia University, United States

***Correspondence:**
Tenzing C. Dolmans
t.c.dolmans@gmail.com

Specialty section:
This article was submitted to Brain-Computer Interfaces, a section of the journal Frontiers in Human Neuroscience

Received: *22 September 2020*
Accepted: *01 December 2020*
Published: *11 January 2021*

Citation:
Dolmans TC, Poel M, van 't Klooster J-WJR and Veldkamp BP (2021) Perceived Mental Workload Classification Using Intermediate Fusion Multimodal Deep Learning. Front. Hum. Neurosci. 14:609096. doi: 10.3389/fnhum.2020.609096

A lot of research has been done on the detection of mental workload (MWL) using various bio-signals. Recently, deep learning has allowed for novel methods and results. A plethora of measurement modalities have proven to be valuable in this task, yet studies currently often only use a single modality to classify MWL. The goal of this research was to classify perceived mental workload (PMWL) using a deep neural network (DNN) that flexibly makes use of multiple modalities, in order to allow for feature sharing between modalities. To achieve this goal, an experiment was conducted in which MWL was simulated with the help of verbal logic puzzles. The puzzles came in five levels of difficulty and were presented in a random order. Participants had 1 h to solve as many puzzles as they could. Between puzzles, they gave a difficulty rating between 1 and 7, seven being the highest difficulty. Galvanic skin response, photoplethysmograms, functional near-infrared spectrograms and eye movements were collected simultaneously using LabStreamingLayer (LSL). Marker information from the puzzles was also streamed on LSL. We designed and evaluated a novel intermediate fusion multimodal DNN for the classification of PMWL using the aforementioned four modalities. Two main criteria that guided the design and implementation of our DNN are modularity and generalisability. We were able to classify PMWL within-level accurate (0.985 levels) on a seven-level workload scale using the aforementioned modalities. The model architecture allows for easy addition and removal of modalities without major structural implications because of the modular nature of the design. Furthermore, we showed that our neural network performed better when using multiple modalities, as opposed to a single modality. The dataset and code used in this paper are openly available.

Keywords: brain-computer interface (BCI), deep learning, multimodal deep learning architecture, device synchronisation, fNIRS (functional near infrared spectroscopy), GSR (galvanic skin response), PPG (photoplethysmography), eye tracking (ET)

INTRODUCTION

Mental workload (MWL) has gained a lot of attention in a variety of fields, such as neuroscience (Toppi et al., 2016; Lim et al., 2018), human factors and ergonomics (Schmalfuß et al., 2018) and human factors in computing systems (Duchowski et al., 2018). In the context of this work, MWL depends on two variables: available cognitive resources and required cognitive resources. Determining the available cognitive resources requires information about prior knowledge, ability and task experience and is thus highly personal. The required cognitive resources depend on task difficulty. In a state of "flow," as described by Csikszentmihalyi (1975), one experiences full emersion with the task at hand. In such a state, the ratio between the available and required cognitive resources, or α, is between 0.8 and 1.2 (Csikszentmihalyi, 1997). The ability to approximate α is interesting, since it would yield insight into MWL and allow for adaptations of tasks. Typically, participants are actively involved in (self)assessing their MWL. The NASA Task Load Index (NASA-TLX) questionnaire is often used to retrieve information about the magnitude and sources of six workload-related factors (Hart and Staveland, 1988). Explicitly acquired information about MWL through retrospection is subjective and results in a measure of perceived mental workload (PMWL). The mere act of performing a measurement on a phenomenon can interfere with the phenomenon (Mahtani et al., 2018). Hence, requiring subjects to extensively reflect and report on their PMWL during experimentation will impact objectivity, not to mention interrupt their state of flow. Physiological measurements can provide an alternative to repeated self-assessment; an advantage of such bio-signals is that they can be measured implicitly. They can objectively be acquired in real-time without explicitly asking participants to provide this data.

The classification of PMWL has been attempted in a unimodal setting using various physiological signals, such as functional near-infrared spectroscopy (fNIRS; Shin et al., 2018), galvanic skin response (GSR; Nourbakhsh et al., 2017) and heart rate (HR), through photoplethysmography (PPG; Schmalfuß et al., 2018). All the aforementioned modalities have individually proven to be useful for the classification of PMWL. This research sought to use an advantageous approach to the classification of PMWL by leveraging both information inherent in individual modalities, as well as cross-modality information. Fusion-based approaches have been surveyed in Baltrušaitis et al. (2018), covering a.o. multi-layer multimodal fusion (Vielzeuf et al., 2018), attention-based methods (Hori et al., 2017) and correlation neural networks (Chandar et al., 2016). Our primary objective in this study is, however, not to give a literature overview, but to actually classify PMWL. The secondary objective is to determine what physiological signals provide valuable information about PMWL. First, we formulated design principles that are relevant and effective within the context of multimodal signal classification that makes use of deep learning. To achieve the primary objective, these design principles were used in the formulation of an intermediate fusion multimodal network (IFMMoN).

MATERIALS AND METHODS

Our goal was to classify PMWL using a deep neural network (DNN) that flexibly makes use of multiple modalities. During the design of such a multimodal brain–computer interface, the principles used to design the end-to-end data path, or pipeline, guide the outcome. Two key aspects of the pipeline were modularity and generalisability (MG). To be modular, new devices should be easy to add to the setup, and their data (collection and processing) should fit within the pipeline with minimal structural implications. Two important libraries that aided modularity throughout the research were used: LabStreamingLayer (LSL) and TensorFlow. LSL provided modularity by allowing device-specific data streams to be easily added (Kothe, 2014). The TensorFlow API allowed for modularity in deep learning model creation (Abadi et al., 2016). Generalisability implies that the additional data that become available from added modalities contribute to classification accuracy. To further improve generalisability and thus applicability, the pipeline should also function well in the classification of other topics besides PMWL. The MG criteria require our methods to be circumstance and device independent where possible. Serendipitously, they served as a way of reducing human error by automating much of the data gathering and analysis pipelines. All methods and designs applied in this project were formulated and executed with the MG criteria in mind.

In the first part of this section, we look to previous works in the field to determine approaches for each of our modalities, as well as fusion options of the DNN. From there, we discuss stimulus presentation, participants and data collection and synchronisation. Lastly, model optimisation with the help of the TensorFlow and Optuna toolboxes is discussed (Abadi et al., 2016; Akiba et al., 2019).

Related Work

We combined a total of four modalities to classify PMWL using this novel approach. To record brain activity, we opted for fNIRS to measure change in (de)oxygenation in the brain (Villringer et al., 2013). Our method is based on Shin et al. (2018). Other bio-signals that we measure are GSR, based on Nourbakhsh et al. (2017), and HR using PPG, on the basis of Schmalfuß et al. (2018). Lastly, eye tracking (ET) was also done, as inspired by Duchowski et al. (2018). In the following subsections, we discuss these modalities in more detail. Then, we discuss what deep learning methods have previously been used to process their data. Fusion options for the combination of data from various devices are finally discussed.

Functional Near-Infrared Spectroscopy

Through fNIRS, relative changes in (de)oxyhaemoglobin concentrations in the brain can be measured. During activation of brain function, energy use and thus the distribution of haemoglobin change (Villringer et al., 1993). This change can be measured using near-infrared light and then be correlated with activation in specific regions of tissue. It seems that there exists no clear consensus about the "best" deep learning-based analysis method for fNIRS data in MWL detection.

Literature can be divided into two main categories: Multilayer Perceptrons (MLPs), consisting of several densely connected layers, and Convolutional Neural Networks (CNNs). Though less common, Recurrent Neural Networks (RNNs) were also used for processing fNIRS (Zhao et al., 2019). Some authors who opted for generic MLPs (e.g., Naseer et al., 2016; McDonald and Solovey, 2017) show great accuracy on binary problems, as well as on more complicated problems. The papers report 63% accuracy on user identification (n = 30; McDonald and Solovey, 2017) and over 91% in binary classification of mental arithmetic vs. rest (Naseer et al., 2016). While the former accuracy is seemingly low, the objective of classification in the work of McDonald and Solovey (2017) is much more natural since the authors' objective was to do user identification on the basis of recorded data. They reported that among 30 subjects, they could determine what data the participant belonged to with a 63% accuracy, whereas chance level is 3.3%.

Tanveer et al. (2019) used two models in their work: one for Beer–Lambert modified optode densities and another for heatmaps of channels over time. Their first network was a DNN with six fully connected dense layers, and their second was a CNN with two convolutional layers and two dense layers. Binary cross-entropy loss was used as loss measure. They report an accuracy of 99.3% on binary classification, achieving the best result with the CNN. Dargazany et al. (2019) showed that an accuracy of over 80% can be reached in 5-class motor imagery problem using a MLP. The benefit of their approach is that they did not perform any pre- or post-processing to the data. This makes their solution very scalable in terms of required human attention, since the majority of the time invested by future users of the system is spent on the collection of data, rather than the (pre)processing of it. However, to facilitate this, their network used two fully connected layers with 10,000 neurons each, leading to quite serious computational complexity.

PPG and GSR

PPG is an optical method for measuring blood volume changes in microvascular tissues and is directly related to cardiac activity (Selvaraj et al., 2008). As such, it can be used to measure HR and compute measures, such as HR variability and inter-beat intervals. Biswas et al. (2019) demonstrated that an accuracy of over 95% can be reached on a HR classification task where the goal was to perform biometric identification of users. They propose using two convolutional layers in conjunction with two long short-term memory (LSTM) layers, followed by a dense output layer. GSR is an electrodermal response that is associated with the innervation of the sympathetic nervous system that is often used to measure affective and cognitive arousal (Venables and Christie, 1980). Sun et al. (2019) showed that a LSTM–CNN hybrid network can reach up to 74% accuracy in a six-class emotion recognition problem using GSR. The use of LSTM is attractive in GSR for several reasons: the time domain and temporal nature of the data enables the extraction of metrics, such as peak frequency and amplitude (Nourbakhsh et al., 2017).

Both PPG and GSR can be processed using methods that are focussed on feature extraction. The benefit of working with such features is that they are easy and cheap to compute. However, such feature extraction removes hidden features that may be found by a DNN and negates the possibility of serendipitous findings when combined with other modalities. Besides the above described methods, both modalities can also conveniently be processed with fully connected layers due to their unidimensional shape.

Eye Tracking

ET is used to gain information about where a person is looking at any given time, which can help understand visual- and display-based information processing (Poole and Ball, 2006). The training and evaluation of ET data are highly task dependent; therefore, this section does not contain any statements about achieved accuracies and will only discuss the types of networks that are used in the literature. Louedec et al. (2019) use a CNN to predict saliency maps in chess games. Their model is based on VGG16, which was first introduced by Simonyan and Zisserman (2014). Furthermore, their model comprises several deconvolutional layers and fusion layers. Krafka et al. (2016) also use convolutional layers and combine them with fully connected layers. In their work, they classified gaze based on an input face-grid that contains the location of the face, the right and left eyes as well as the full face. Generally, the consensus is to use convolutional layers for the classification of ET data, regardless of objective. Intuitively, this makes sense since we are interested in spatial features in the data.

Fusion Options

There exist many strategies to tackling the multimodal problem in deep learning. Given that most neural networks are highly task dependent, the design of a multimodal DNN follows this same trend. Ramachandram and Taylor formulated several key considerations to be made for deep learning with multiple modalities in their overview of deep multimodal learning (Ramachandram and Taylor, 2017). The first key consideration is when to fuse the modalities. In general terms, there exist three options for the time of fusion. The first is early fusion, or data level fusion. This can, for example, be achieved by concatenating features or raw data and feeding said data into a neural network. The second is intermediate fusion. This involves mapping input to a lower dimension using various types of layers and fusing somewhere along the way between the input and output layers. The third option is late fusion, through e.g., majority vote of several smaller networks. The choice of where fusion takes place is flexible and immensely impactful on model performance, as demonstrated by Karpathy et al. (2014). The second key consideration is which modalities to fuse, since not all data contribute to solving a problem equally. The third and final consideration to make is what to do with missing modalities or data. The absence of data can be prohibitively problematic, especially in real-time applications.

In light of MG, early fusion is an unattractive option: it requires the input data to be "stitched" together, which leads to multiple problems in our application. First, we are working with vastly different sampling rates ranging from 10–256 Hz. Furthermore, the dimensionality differs between devices, requiring us to devise a strategy that would guarantee

equal share of data in each sample without losing any features that are present in either temporal or spatial dimensions. Lastly, instead of working with an MG network, all concatenated data would be fed into the same network, regardless of what devices are featured in the data. This would entail tuning early layers and shapes of the network when the parameters of the data change. Late fusion through majority vote aligns with the modularity requirement, but not the generalisability requirement. Adding or removing modality networks, or MNets, does not require the adaption of other MNets. However, separated networks are unable to learn from multiple modalities simultaneously, since there is no information exchange between them. Intermediate fusion allows for the creation of several modular MNets that develop "expertise" in their respective domain. This expertise can then be shared with an overarching network. Furthermore, adding or removing modalities is as simple as "clipping on" MNets, or switching them off in the head class, respectively. Hence, intermediate fusion satisfied the MG criteria best.

Stimulus Presentation

To simulate MWL, participants were asked to solve several zebra puzzles. Zebra puzzles are verbal logic puzzles that are solved by connecting attributes to objects on the basis of hints. The difficulty of the puzzle was modulated by the number of hints that were given and the average number of hints required before an attribute could be chosen. Hints could be ticked off when used. **Figure 1** provides an example of a zebra puzzle. In total, there were five different puzzles, each with their own difficulty ranging from "very low" to "very high" difficulty. All puzzles were retrieved from Brainzilla (2020), and initial difficulty indications were also based on the content of Brainzilla. Between every puzzle, participants were asked to take a moment to relax. Furthermore, they indicated how difficult they perceived the puzzle to be on a scale of one to seven, seven being the highest difficulty. These ratings were later used as labels during training. The order in which the puzzles were presented was completely randomised. An LSL stream was active during the entirety of the stimulus presentation. This stream sent a marker at every action.

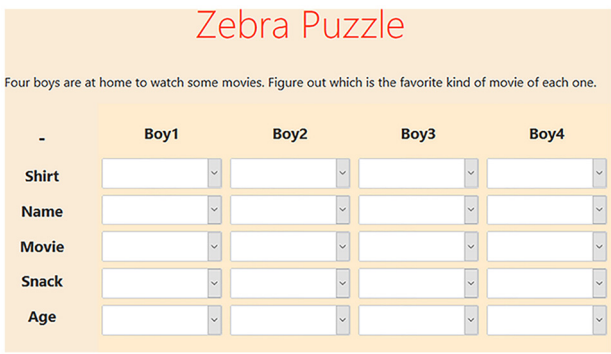

FIGURE 1 | Example zebra puzzle. Below the puzzle, several hints are given that allow participants to connect all attributes (vertical) to each boy (horizontal). An example hint is: "Joshua is in one of the ends." Clicking the arrow of a cell drops down all options for that cell.

Actions were (un)selecting hints and (un)selecting answers. Markers contained the participant ID, action timestamp, type of action, the id of the action and status of the action (correct, incorrect or checked). The timestamps of this stream are later used to segment the data.

Participants

In total, data were collected from 23 participants (11 males, 12 females, mean age = 24.7, SD = 9.8, min = 20, max = 57). Of these participants, one was excluded from the dataset because of poor data quality. Participants were recruited using the Sona system, a cloud-based participant management software (SonaSystems, 2020) that is used at the University of Twente. Recruitment was also done in social circles. Prior to experimentation, the study was approved by the ethics committee of the BMS faculty of the University of Twente. All participants granted written informed consent for the collection and open sourcing of data.

Data Collection and Synchronisation

All data were streamed and recorded on a Dell Precision 3530 Laptop with an Intel i7–8750H CPU, 16 GB RAM and an NVIDIA Quadro P600 GPU. Three different devices measured four modalities. The Shimmer3 GSR+ was used to measure GSR and PPG (Shimmer GSR3+; Shimmer, Dublin, Ireland), the Tobii Pro X3–120 was used for ET (Tobii X3–120; Tobii Group, Stockholm, Sweden) and the Brite^{24} was used to collect fNIRS (Brite^{24}; Artinis Medical Systems, Elst, The Netherlands). Since participants were aware of sensors that were attached to their bodies, measurements were not unobtrusive. Each device was set up such that data streams were sent to LSL in real-time. The LabRecorder app was used to record data from all streams into a single XDF file per participant (Kothe, 2014). Data were then imported into Python using PyXDF (Boulay, 2020), which automatically performs checks on the indicated vs. received sampling rates and de-jitters the data where necessary. Data synchrony was also checked manually to ensure that all streams were aligned throughout the recording. Several checks for synchrony were also implemented during data selection and processing, which are documented in the "Data Selection" section.

Raw GSR and PPG were directly streamed to LSL from the Shimmer3 GSR+ using an application that was written by the HBA Lab of Thales (Groot de, 2020). The sampling rate of this stream was 256 Hz. The data of the Tobii Pro X3–120 were streamed using a custom python application that was made with the Tobii Pro SDK and PyLSL (Kothe, 2014; TobiiProAB, 2019). ET data were streamed at 120 Hz and contained x and y coordinates for both eyes. For the collection of fNIRS, Oxysoft 3.2.51.4 × 64 was used (OxySoft; Artinis Medical Systems, Elst, The Netherlands) with the Brite^{24} in the available 27 channels optode arrangement. Two wavelengths (756 and 853 nm) were sampled at 10 Hz, and the Beer–Lambert modified optode densities of O_2Hb and HHb were mapped to LSL directly from Oxysoft. See **Figure 2** for a detailed view of the optode template. For a complete overview of the data pipeline, please refer to **Figure 3**.

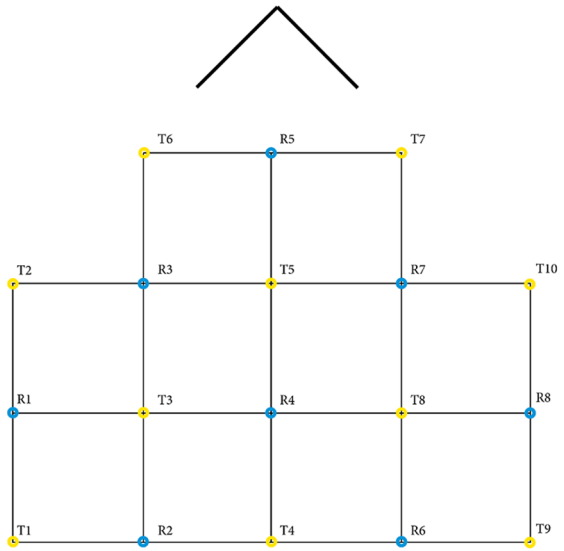

FIGURE 2 | Optode template for fNIRS collection. Between each transmitter (T) and receiver (R), one channel exists. In total, there are 10 transmitters, eight receivers and 27 channels. The arrow represents the nose of the participant. This Brite^{24} 27 channels optode arrangement is available in Oxysoft 3.2.51 (OxySoft; Artinis Medical Systems, Elst, the Netherlands).

Model Optimisation

Data Selection

Data were selected on the basis of markers that were present in the Zebra Puzzle's data stream. These markers were sent through LSL at a variable rate that depended on the participant's actions. The nature of the stimulus presentation contributed to several key things to pay attention during data selection. For one, markers could be close together when participants selected multiple answers in quick succession. Hence, selections made around these markers contained some overlap in data. Due to software issues and practical shortcomings, some parts of data were missing. To overcome these problems, several Boolean masks determined which markers were fit for usage. First, the nearest index to the time of a marker was identified in the data of each device. When said indices were identical for multiple markers, the samples were removed from the dataset. Such exactly matching indices were likely the result of drifting device timestamps and/or missing data and were hence excluded. Segmented selections were inspected for noise by means of computing simple statistics of samples, such as mean, variance, max, min, etc., to gain an overview of data quality. However, noisy samples exposed the network to 'realistic' data and were not removed thusly.

Once a full selection of the markers was made, a segment of 8 s of data before the marker was selected; 8 s, because the haemodynamic response function shows a peak after 5–8 s of neuronal activity onset (Zhang et al., 2005); before, because the participant's contemplation takes place prior to knowing and selecting the correct answer. A CSV file that contained the final selection of the markers was created for each participant. Each sample in our dataset consists of four synchronised measurements: fNIRS, GSR, PPG and ET. The difficulty rating for the sample's respective puzzle served as the label. These samples were added to a TFRecord file, which

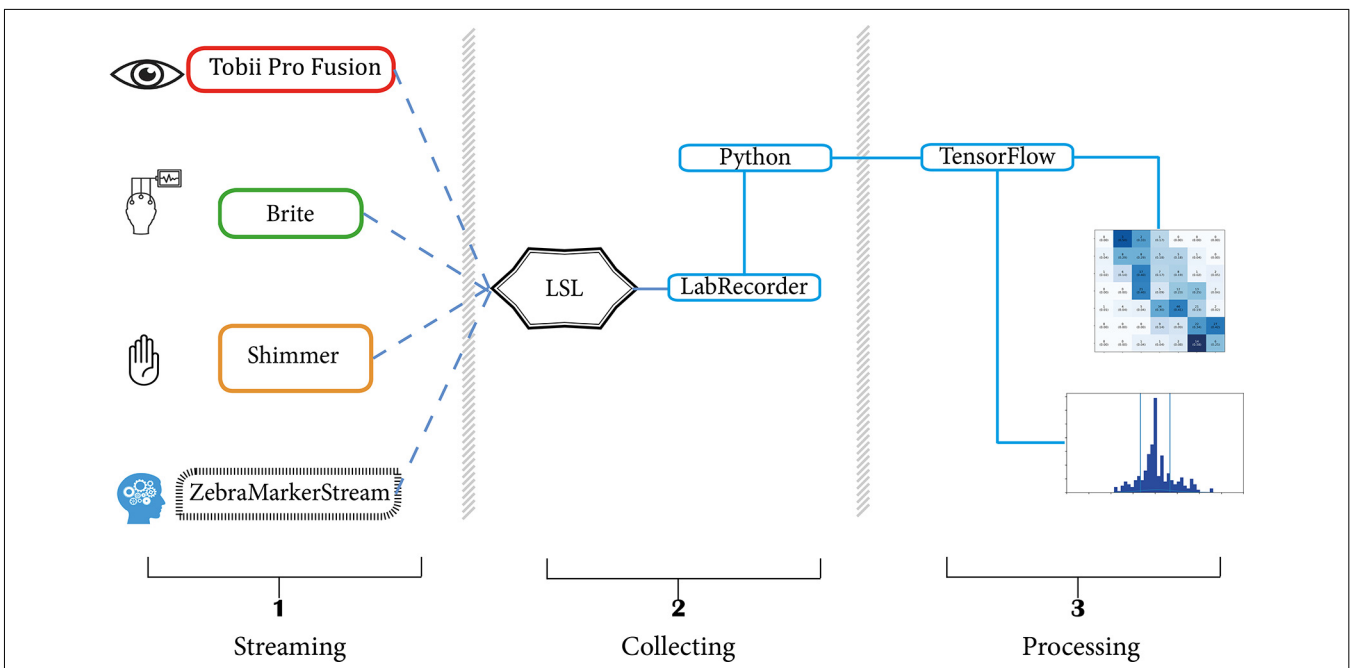

FIGURE 3 | Experimental setup. The left section contains an overview of all active data streams, being: Tobii Eye Tracking data, Brite fNIRS data, Shimmer GSR and PPG data, and the Zebra Puzzle marker stream. Dotted lines indicate a streamed connection to LSL. The LabRecorder in the middle section records the data in XDF format. The right section shows example outputs of processed data.

allowed many useful methods, such as shuffling, batching and splitting, to be applied to all the selected data simultaneously (TensorFlow, 2020).

The dataset that was generated and analysed for this study can be found in the 4TU.ResearchData repository under the following doi: 10.4121/12932801 (Dolmans et al., 2020).

Models

To maximise usability and allow for feature sharing of our multimodal data while also adhering to the MG criteria, we opted for intermediate fusion. This resulted in a model architecture that adheres to the general structure of one base network, or MNet, for each modality and one Head network that integrates all MNets. Two concrete routes were chosen for the implementation of both the MNets and the Head networks: one model based on literature and one model that contains only densely connected layers. The model based on previous work as discussed in the "Related Work" to "Participants" sections had four custom MNets, one for each modality and one custom Head network. The PPG MNet consisted of two convolutional layers; the GSR MNet consisted of two convolutional and two LSTM layers; the ET MNet consisted of four convolutional layers; finally, the fNIRS MNet consisted of two convolutional and two dense layers. All MNets were represented in a lower-dimensional space with the help of a single densely connected layer before fusion in the Head. **Figures 4**, **5** contain the structure and layers of the model based on literature and the densely connected model, respectively. **Table 1** details the models' layers and the number of units/filters for each layer. Batch normalisation and max pooling were utilised as a means of stabilisation. For both models, a smaller alternative model was created that contained exactly half of the units and filters in each layer in order to gain an initial idea of the effect of reduced network size. This brought the total number of models to four, which will be referred to as MLP (only dense), S_MLP (small, only dense), LIT (literature) and S_LIT (small literature). All models were trained on the same dataset.

Two variations of labels were used. The first variation contained samples that were labelled with the indicated difficulty of their respective puzzle and participant, thus containing a total of seven different 'individual' labels. e.g., participant 1 indicated a difficulty of 6 for puzzle 3; hence, all samples in puzzle 3 have label 6 for participant 1. Models under this labelling variation were evaluated by their ability to predict what level of workload (LoW) the participant indicated. Participants rated their PMWL on a 7-point scale, seven being the highest PMWL. These ratings were converted to values between 0 and 1 using the formula: (rating 1)/7, such that a rating of 1 corresponded to a label of "0," a rating of 2 corresponds to a label of "0.1667," etc. It follows that, in order to be within-level accurate, the average difference between predicted and true labels must be lower than 0.1667. All models used a single output unit with a Sigmoid activation function, resulting in predicted labels between 0 and 1.

The second variation contained samples that were labelled with the average indicated difficulty of all participants over the respective puzzle. This resulted in five different "group" labels, one for each of the puzzles. These ratings were mapped between 0 and 1, where the lowest average rated difficulty corresponded to 0, etc. This labelling variation was used to assess the difference in classification accuracies for individual vs. group labelling schemes. Like the first labelling variation, all models used a single output unit with a Sigmoid activation function, resulting in predicted labels between 0 and 1. An intuitive way of visualising performance was through histograms. By removing the true labels from the predictions, we created a distribution of the error; the ideal result would be a slim Gaussian distribution around zero.

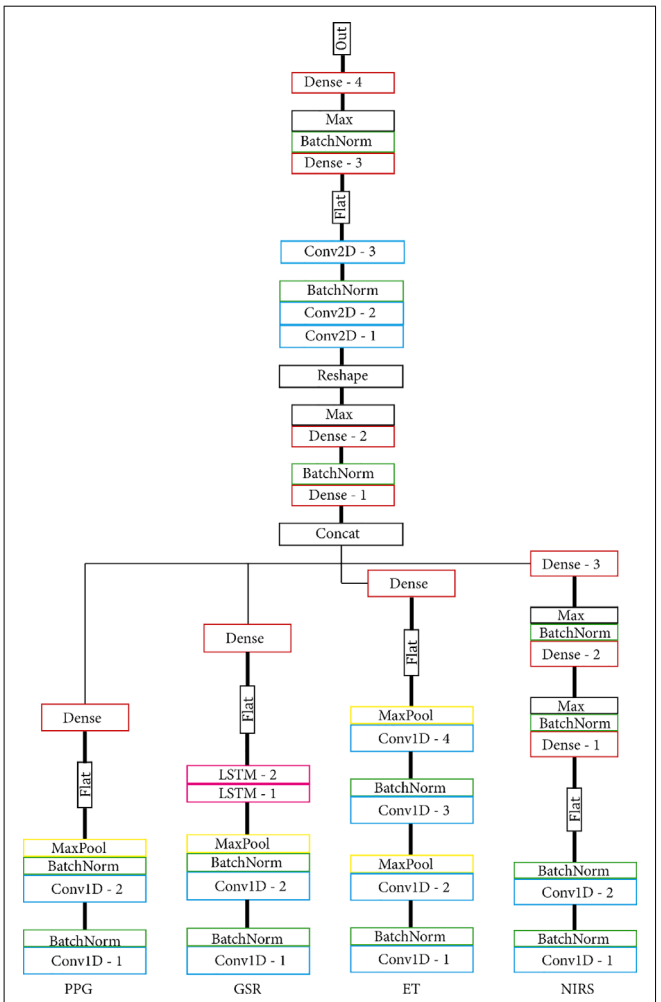

FIGURE 4 | (S_) LIT model. Each modality is indicated at the bottom. All MNets feed into one Head network and are based on what is commonly used in the literature. Before fusion in the Head network, MNets are flattened and represented in a lower-dimensional space with the help of a single densely connected layer.

Hyperparameter Optimisation

The models were optimised with the help of hyperparameter tuning, or hyperparameter optimisation (HPO). The combination of hyperparameters greatly impacted model performance. In this work, we made use of Optuna, an open-source define-by-run API that allowed us to flexibly and quickly set up a parameter search space (Akiba et al., 2019). We used the default Tree-structured Parzen Estimator to sample

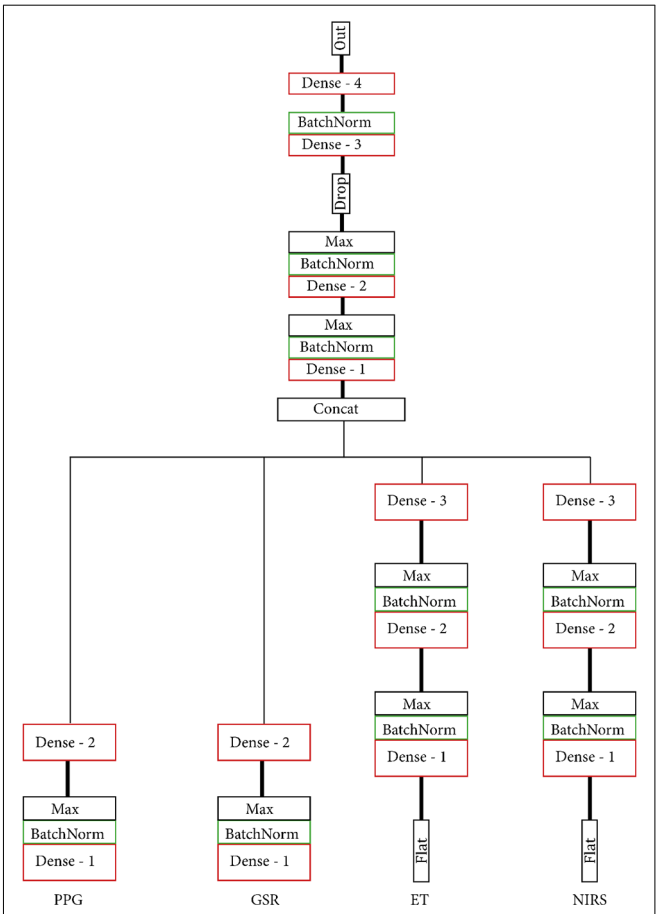

FIGURE 5 | (S_) MLP model. Each modality is indicated at the bottom. All MNets feed into one Head network. Both the MNets and the Head consist solely of densely connected layers.

values for learning rate, dropout rate and momentum. All losses were calculated as a mean squared error. We opted to use mean squared error because we have ordinal labels ranging from 1–7. Classifying a data sample with label 1 as 7 is a larger error (6 off) than classifying it as a 2 (1 off). Cross-entropy does not take the distance in misclassification into account, whereas MSE does. Furthermore, we provided four different models as categorical suggestions, and these models are discussed in the "Models" section. In total, we ran 20 "trials"; each trial contained a 5-fold cross-validation, in which the total dataset was split into four training and one testing part for every fold. On every fold, a different split was made, and a new model was trained to prevent exposing any trained model to the entire dataset simultaneously. The objective of the HPO was to minimise the average difference between predicted and true labels. To further optimise the learning, we used a learning rate policy that is based on the "1Cycle Policy" as described in Smith (2018). This policy slowly increases and decreases the learning rate in a pyramidal shape as a network cycles through a dataset. This helps prevent getting stuck in local minima. For the creation of visualisations, the best performing models were trained separately using the hyperparameters that were found during HPO. During training,

TABLE 1 | Overview of models, their layers and the number of units/filters for each layer.

LIT	Full-sized units/filters	MLP	Full-sized units
PPG	Conv1: 128	GSR	Dense1: 256
	Conv2: 128		Dense2: 256
	Dense: 256		
GSR	Conv1: 128	PPG	Dense1: 256
	Conv2: 128		Dense2: 256
	LSTM1: 256		
	LSTM2: 256		
	Dense: 256		
ET	Conv1: 256	ET	Dense1: 1,024
	Conv2: 256		Dense2: 1,024
	Conv3: 256		Dense3: 1,024
	Conv4: 256		
	Dense: 1,024		
NIRS	Conv1: 512	NIRS	Dense1: 2,048
	Conv2: 512		Dense2: 2,048
	Dense1: 2,048		Dense3: 2,048
	Dense2: 2,048		
	Dense3: 2,048		
HEAD	Dense1: 3,584	HEAD	Dense1: 3,584
	Dense2: 4,096		Dense2: 2,048
	Conv1: 512		Dense3: 1,024
	Conv2: 512		Dense4: 512
	Conv3: 256		
	Dense3: 512		
	Dense4: 256		

LIT refers to the model that was based on literature, and MLP refers to the model that only contains densely connected layers. Table contains information about the full-sized models, and half-sized models contain exactly half the number of units/filters per layer.

the dataset was split 90–10% train-test randomly, in order to prevent testing the network on previously seen samples. All training was done on a single NVIDIA GeForce GTX 1080-Ti GPU. For further details on the HPO and its implementation, kindly refer to the code that can be found on the following doi: 10.5281/zenodo.4043058 (Dolmans, 2020), or on GitHub[1].

RESULTS

In this section, we discuss the retrieved data and the achieved results for the various configurations of models.

Data

The total number of samples that were selected is 4,082, with an average of 185.5 samples per participant (max: 345, min: 77). The distribution of samples across LoWs can be seen in **Table 2**. Participants most commonly indicate that the puzzles have a 5/7 difficulty, followed by a 6/7 difficulty. Participants rarely indicate that the puzzles are the easiest (1/7) or hardest (7/7) possible difficulties: 2.1 and 9.3%, respectively. The internal consistency of labels was assessed using Cronbach's alpha (Tavakol and Dennick, 2011). There were two perspectives from which an alpha could be calculated for the labels. The first was how internally consistent the puzzles are as a predictor of PMWL under the assumption that the puzzles are test items; the resulting alpha was 0.74. The second perspective was how internally consistent the participants are in self-assessing MWL under the

[1]https://github.com/Tech4People-BMSLab/mwl-detection

assumption that participants are test items; the resulting alpha was 0.97.

Model Performance—Individual Labels

As discussed previously, models were evaluated by their ability to predict the LoW the participant indicated. Two sets of 10 trails were done using the Optuna toolbox, once with the MLP and LIT models and once with the S_MLP and S_LIT models. **Table 3** contains an overview of the trials that were done on personal labels. The best result that was achieved with the MLP model is an average absolute difference 0.1892 between predicted and true labels. This corresponded to 1.13 LoW when translated back to the 7-point scale that participants rated their PMWL on. Training times per 5-fold cross-validation with 25 epochs per fold were around 40 min. The best result that was achieved with the LIT model is 0.1978, or an average of 1.19 LoW. Training times lay around 70 min. The best result that was achieved with the S_MLP model is 0.1642, or an average of 0.985 LoW. This is also the best result that was achieved within our search space. Training times lay around 25 min. The best result that was achieved with the S_LIT model is 0.1681, or an average of 1.009 LoW. Training times lay around 43 min.

The best performing model predicted 63.6% of the samples within one LoW and 72.7% within 1.5 LoW. The distribution of the difference between the predicted and true labels had $\mu = 0.033$ and $\sigma = 0.233$, see **Figure 6**. The mean of the distribution was slightly larger than zero, meaning that the model was prone to overestimating the workload of the participant. The confusion matrix shows that the model most frequently correctly classified samples with a label of 0.6667, corresponding to a difficulty rating of 5. See **Figure 7** for the confusion matrix. From this confusion matrix, it can be deduced that the accuracy of the classifier is 32%, which is considerably above chance level: a random classifier for seven target labels would correctly classify 14% of the samples. If one considers a difference of one label to also be correct, the accuracy of the classifier is 77%. In this case, a random classifier would have a performance of 3/7 that equals 43%.

Model Performance—Group Labels

As discussed, a second labelling variation was assessed. This variation leads to five different labels, and their distribution can be viewed in **Table 4**. Since these labels depended on the average rating of participants, they were not equidistant. There existed a gap of 0.33 from the lowest difficulty puzzle to the next puzzle. The puzzles thereafter only had 0.05 LoW between them. Similar gaps are present between the third and fourth and fourth and fifth difficulties. Similar to model training with individual labels, a total of 20 trials of HPO were done using the Optuna toolbox. **Table 5** contains the results of these trials. The best result was achieved with the S_LIT model, with a mean difference between true label and predicted label of 0.2386. The distribution of prediction vs. label had $\mu = -0.055$ and $\sigma = 0.284$, see **Figure 8** for clarification. The confusion matrix shows that the model most frequently classifies data into the fourth difficulty, regardless of true label, see **Figure 9** for details. In this case, the accuracy of the classifier is 27%, which is just above chance level (accuracy of 20%). If 1 label off is also correct, the performance of the classifier is 72%. Hence, one can conclude that using group labels decreases the performance of the classifier, probably due to the introduction of noise in the labels.

TABLE 2 | Drift, sample distribution, and puzzle difficulties.

Drift	Difficulties: number of samples	Average puzzle difficulties
Avg: 548 ms	1: 87 (2.131%)	VLow: 2.73
Std: 590 ms	2: 350 (8.574%)	Low: 3.73
	3: 479 (11.73%)	Mid: 4.7
	4: 611 (14.97%)	High: 3.89
	5: 1,275 (31.23%)	VHigh: 5.76
	6: 902 (22.10%)	
	7: 378 (9.260%)	
	Total: 4,082	

The first column details the recorded drift in terms of the average and standard deviation. The second column summarises the samples and their distribution within each level of difficulty. The third column contains the average difficulty rating for each puzzle on a 7-point scale, as rated by the participants.

TABLE 3 | Overview of hyperparameter optimisation (HPO) trials and their respective scores, ran on individual labels.

Trial	Model	Difference	LoW	Duration (min)
1	MLP	0.5996	3.60	39:09
2	MLP	0.3556	2.13	39:04
3	MLP	0.3685	2.21	39:14
4	MLP	0.2208	1.32	39:18
5	LIT	0.2440	1.46	1:15:49
6	LIT	0.3772	2.26	1:10:32
7	**MLP**	**0.1892**	**1.14**	38:50
8	LIT	0.5672	3.40	1:09:05
9	LIT	0.3798	2.28	1:09:38
10	**LIT**	**0.1978**	**1.19**	1:09:33
11	S_LIT	0.2957	1.77	43:03
12	S_MLP	0.1840	1.104	25:30
13	S_LIT	0.5930	3.558	47:39
14	S_MLP	0.1772	1.063	25:30
15	S_LIT	0.1715	1.029	47:30
16	S_MLP	0.1701	1.021	25:31
17	**S_MLP**	**0.1642***	**0.985***	25:21
18	**S_LIT**	**0.1681**	**1.009**	47:09
19	S_MLP	0.1808	1.085	25:13
20	S_LIT	0.4534	2.720	47:14

Trial indicates the trial number of the HPO. Model indicates which neural network was used in the trial. Difference indicates the mean difference between true label and predicted label. LoWs (levels of workload) indicate the number of LoW the mean difference translates to, one LoW being 0.1667. Duration is time needed for a five-fold cross-validation. Bolded numbers are the best performances for each model type. An asterisk indicates best overall performance.

Unimodal Performance

To investigate the additional value of additional modalities, individual modalities were also evaluated. All models were trained using the hyperparameters that proved most effective for each model type, using individual labels. In other words, HPO was not run for each of the unimodal problems, but relied on earlier optimisations for the respective models. **Table 6** details the results of these tests. The best performance was achieved using the PPG modality and the S_MLP model, with an average absolute difference between predicted vs. true label of 0.1969, or 1.18 LoW. The best overall performance for a single modality was achieved with the S_LIT model and the GSR; a difference

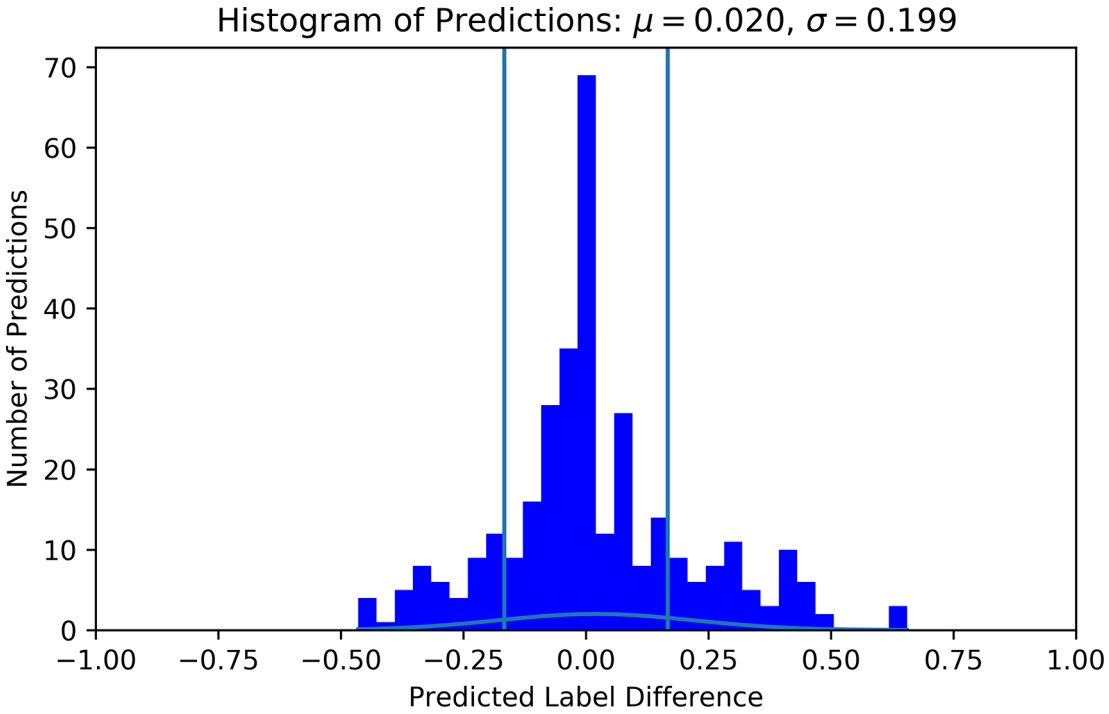

FIGURE 6 | Histogram of predicted vs. true label difference using individual labels. Labels and predictions have been offset such that all true labels are 0 and their respective predictions are deviations away from 0. The vertical lines indicate −1/6 and 1/6 bounds, or one LoW on each side. Predictions outside these lines are more than one workload level inaccurate. For this histogram, the S_MLP was used with HP corresponding to trial 17 in **Table 3**.

of 0.1796, or 1.08 LoW. The fNIRS modality performed worst on both models, reaching a difference of 0.2865, or 1.71 LoW, and 0.3188, or 1.91 LoW, for S_MLP and S_LIT, respectively.

DISCUSSION

Performance

Model performance on models trained with individual labels reached sub-level accuracies on a seven-level scale in the classification of PMWL on unseen samples. This showed that the choice of implementation and the IFMMoN were able to learn and generalise from the training data. The histogram shows a normal distribution where the vast majority of points lie around zero, indicating that predicted labels were frequently close to the true labels. The confusion matrix showed that the average prediction of the IFMMoN was slightly higher than the true label, yet a diagonal trend could still be observed. The model performed best in the classification of the fifth LoW, followed by the third and sixth. This followed the trend where more prevalent labels show better performance, with the exception of the third LoW. Furthermore, the fourth LoW, though third-most represented in samples, showed the worst results for unknown reasons. Because of the sample distribution, performance in the relatively underrepresented extremes was hard to assess. This distribution also led to great variability in performance during HPO since the dataset was shuffled differently on each fold. Hence, some folds contained relatively many samples from underrepresented classes in the test set.

The classification of the alternative labelling variation showed no significant diagonal trend; the IFMMoN classified the majority of samples in the fourth LoW, regardless of label. This LoW was the most prevalent label, meaning that the IFMMoN was unable to generalise and minimise loss by classifying into the class that five was the lowest loss. A reason for this could be that the average PMWL did not represent the individual PMWL of participants well enough, rendering the connection between sample and label insignificant. The IFMMoN might learn individual physiology given an individual label but might be unable to generalise in the dataset if all labels are common for vastly different physiologies. In future works, determining the effect of individual differences would be worth pursuing. Model affectivity, as well as performance in general, may vary between subjects, and gaining more insight into these differences can improve the usability of the IFMMoN.

Evaluation of unimodal performance showed that some modalities performed significantly better than others. Though not visually reported in this work, unimodal classification showed a similar trend to the alternative labelling variation: all classifications merge towards one label. This reduces the credibility and value of the classifications based on a single modality. In particular, the fNIRS modality performed poorly. This could mean that the data do not contain valuable signals, or that the implemented processing was inadequate. Another reason for the poor performance of the modality could be the stimulus presentation. The fNIRS modality is often only used in block designs; this optimises the separability of conditions,

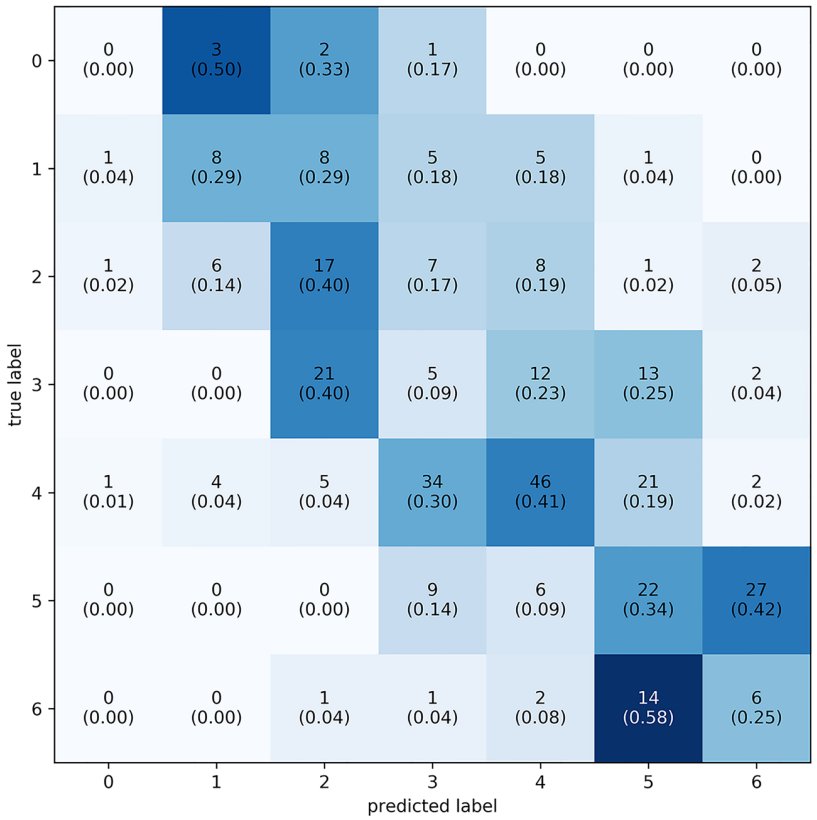

FIGURE 7 | Confusion matrix of the predicted vs. actual classes using individual labels. Labels range from 0–6, totalling seven levels of workload. Every square contains two numbers: the number of times the label was predicted and the relative proportion of predictions in their respective class (in parenthesis). To plot this confusion matrix, predicted labels were placed into the nearest class. The majority of prediction is on or above the diagonal.

which is known from the field of functional MRI (fMRI), since the two modalities essentially measure the same signal (Maus et al., 2010). However, stimulus presentation in our research did not follow such a block design. Instead, our work uses a more naturalistic stimulus in that participants worked on several longer tasks. This 'realness' of the stimulus allowed us to assess the effectiveness of fNIRS in non-lab situations but likely also negatively impacted the distinguishability between conditions. Because HPO was not done for each of the unimodal problems, but the best parameters for the respective network type were used, performance may not be optimal. In order to determine qualities, such as the informativity of the individual modalities, HPO will have to be done.

TABLE 4 | Labels, occurrences, and prevalence percentages.

Label (puzzle)	Occurrences	Percentage
0 (VLow)	424	10.39%
0.33 (Low)	691	16.93%
0.38 (High)	965	23.64%
0.65 (Medium)	1,138	27.88%
1 (VHigh)	864	21.17%
	Total: 4,082	

The first column shows the scaled average rated difficulty for the respective puzzle level between parentheses. The second column indicates the number of occurrences for each LoW. The third column details the percentage of occurrences in the respective difficulties.

To really prove that a multimodal approach outperforms, one needs to validate this on different workload paradigms, such as N-back and visual information overload. Moreover, one needs a sound statistical methodology for proving significance that also includes a false discovery rate correction due to multiple testing. We would not be surprised that for some workload situations, the unimodal approach is on par with a multimodal approach. This however, lies outside of the scope of this research and is a line of further research. While lacking in efficacy, unimodal evaluation allowed us to assess the MG of the implementation. Altering the configuration of the IFMMoN was quick and simple due to the modular design. This indicated that the modularity criterium was adhered to, and that it was practical during research. Furthermore, since we were able to achieve better performance using all modalities, the generalisability criterium was also satisfied. The IFMMoN appeared to generalise better given data from multiple physiological sources. Hence, we consider the MG criteria to be practical and valuable.

Limitations

Our limitation is that we are aware of some in several categories. First, there are hardware limitations, which become most apparent when inspecting device synchrony. The average

TABLE 5 | Overview of HPO trials and their respective scores, ran on group labels.

Trial	Model	Difference	Duration (min)
1	**LIT**	**0.2491**	39:09
2	S_MLP	0.2583	39:04
3	**S_MLP**	**0.2484**	39:14
4	S_MLP	0.2642	39:18
5	S_MLP	0.4670	1:15:49
6	S_MLP	0.2540	1:10:32
7	S_MLP	0.3202	38:50
8	S_MLP	0.5389	1:09:05
9	S_MLP	0.2616	1:09:38
10	**MLP**	**0.2616**	1:09:33
11	S_MLP	0.2594	43:03
12	S_LIT	0.2352	25:30
13	S_LIT	0.2546	47:39
14	S_LIT	0.3129	25:30
15	S_LIT	0.2396	47:30
16	**S_LIT**	**0.2304***	25:31
17	LIT	0.4366	25:21
18	S_LIT	0.2447	47:09
19	MLP	0.5276	25:13
20	S_MLP	0.2804	47:14

Trial indicates the trial number of the HPO. Model indicates which neural network was used in the trial. Difference indicates the mean difference between true label and predicted label. Duration is time needed for a five-fold cross-validation. Bolded numbers are the best performances for each model type. An asterisk indicates best overall performance.

recorded drift across all participants was 548 ms (SD = 590 ms, max = 2, 827 ms, min = 58 ms). In total, four participants had a drift of larger than 1 s. For two of these, the reason is known: crashing software and device shutdown before recording end. The reason for the drift in the remaining two is unknown. An average recorded drift of 548 ms is quite large for some modalities, such as ET, and not so much for fNIRS. If this system is to be used with a modality that is even more sensitive to drift, such as electroencephalography (EEG), significant improvements need to be made. One way of doing this is by performing data collection on a more powerful computer, or by distributing the data collection over multiple computers. Dedicating more CPU to each device and stream will likely yield better results. Device-specific hardware limitations may also play a role in the drift of data streams. Finally, the recording software may also be looked to when investigating the drift further.

Second is the limited number of participants. A common way of improving performance is by gathering more data. In total, 4,082 samples were collected from 22 participants. To put this dataset in perspective, ImageNet, a large image database that is commonly used, has over 14 million images (Deng et al., 2009). Of course, gathering physiological data is significantly more time-consuming, especially when using multiple devices. Nonetheless, an almost-guaranteed way of improving the performance is to gather data from more participants.

Third is the choice of modalities. Currently, only gaze data (X and Y coordinates of both eyes) are used in this work. Duchowski et al. (2018) demonstrate the efficacy of pupillary activity with regard to the assessment of cognitive load. The inclusion of pupillary data into this work may have led to different results. The same can be said for our measurements of the brain. Currently, fNIRS is used to measure the relative changes in (de)oxyhaemoglobin. However, EEG can also be used to predict cognitive load, as demonstrated by Friedman et al. (2019). This train of thought can be extended to other measures, to the extent where this same research can be performed with a different set of modalities to achieve vastly different results.

A fourth limitation is the model architecture and optimisation. Currently, two variations of IFMMoN were used, each with a small and large version. Smaller versions showed better performance while also being more efficient, possibly due to the small size of the dataset. No further exploration into where model performance stops improving with the reduction of model size and/or complexity was done. Furthermore, HPO was done only on momentum, learning rate and dropout rate. This can be improved by also varying the number of hidden layers and neurons, as demonstrated by Akiba et al. (2019). On the other hand, the performance estimates given in the results could be a little optimistic due to the fact that HPO applied was on the total dataset instead of using nested cross-validation (i.e., applying HPO on each train fold in the cross-validation approach). But, since HPO was used to optimise only some learning parameters, the presented performances are a good reflection of the actual performances.

Data selection around markers can be changed and customised for each of the modalities. For example, selecting fNIRS data around marker can be done differently when compared to ET data, given that the haemodynamic response is very 'slow' compared to eye movements. For this reason, data after the markers could also contain valuable information for some modalities. Worth noting, however, is that changes to the dataset or modalities would require the network to be retrained and HPO would need to be redone. If a modality is added, then the Head network and the sub-network corresponding to the new modality need to be (re-)trained. If a modality is removed, then only the Head network needs to be re-trained. This is a time- and resources-consuming process. Testing on additional subjects that contain the same modalities does not require retraining. The latter is something that was not tested in this research and is thus considered one of its shortcomings. Lastly, network outputs could be encoded in a 7-dimensional vector where each output gives the probability of this the respective label, rather than outputting a single number between 0 and 1.

Finally, some complications arose during collection and upon inspection of the retrieved data. Data of Participants 3, 8, 13 and 16 were partially excluded due to software crashes and poor device connectivity. Participants 6, 9, 15 and 21 sporadically show minor artefacts likely related to movement or dark hair. However, this data was included in the dataset with the intention to expose the system to a certain, perhaps more realistic, degree of noise.

Labelling

A more psychometric point of discussion lies in our labelling scheme. Participants will give different ratings for the

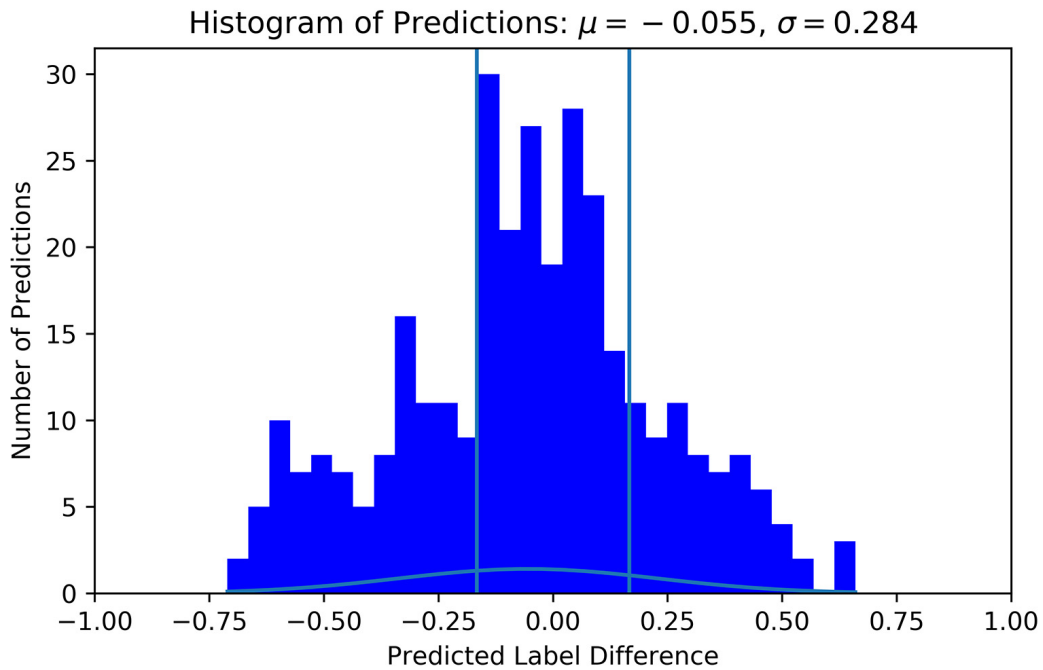

FIGURE 8 | Histogram of predicted vs. true label difference using group labels. Labels and predictions have been offset such that all true labels are 0 and their respective predictions are deviations away from 0. μ indicates the mean, and σ indicates the standard deviation. For this histogram, the S_LIT was used with HP corresponding to trial 16 in **Table 5**. Vertical lines show the original bounds of 1 LoW for reference.

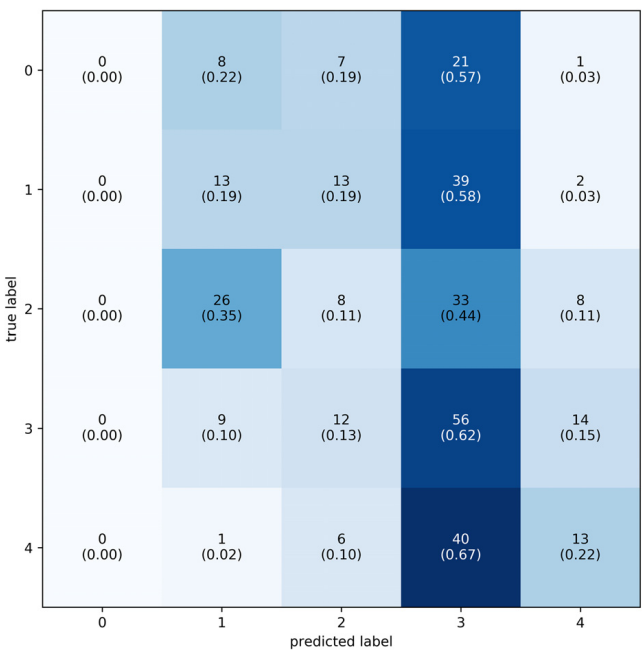

FIGURE 9 | Confusion matrix of the predicted vs. actual classes using group labels. Labels range from 0–4, totalling 5, one for each puzzle difficulty. Every square contains two numbers: the number of times the label was predicted and the relative proportion of predictions in their respective class (in parenthesis). To plot this confusion matrix, predicted labels were placed into bins that represent their class. The majority of predictions are in the second highest workload level.

TABLE 6 | Overview of results for individual modalities.

Model	Labels	Difference	LoW
MLP	Individual	0.1892	1.14
S_MLP	Individual	0.1642	0.985**
LIT	Individual	0.1978	1.19
S_LIT	Individual	0.1681	1.09
MLP	Group	0.2616	–
S_MLP	Group	0.2616	–
LIT	Group	0.2484	–
S_LIT	Group	0.2304	–
	Modality		
S_MLP	PPG	0.1969	1.18
	GSR	0.2160	1.30
	NIRS	0.2865	1.71
	ET	0.2159	1.30
S_LIT	PPG	0.2400	1.44
	GSR	0.1796	1.08*
	NIRS	0.3188	1.91
	ET	0.2224	1.33

The first column contains the name of the model. The second column describes which labels were used. The third column reports the achieved mean difference between predicted vs. true label. The last column shows the mean difference in level of workload. An asterisk indicates the best performance in the unimodal setting; a double asterisk indicates the overall best performance.

same PMWL. Ratings are entirely subjective and volatile because one can only assess one's PMWL relative to oneself. Furthermore, one person might feel confident and calm while experiencing high workload, whereas another might feel stressed while experiencing low workload. Hence, one should always expect to see a high degree of error and variance when assessing PMWL, or any human emotion for that matter. Since the objective of our system is to eventually classify PMWL in naturalistic environments in real-time, we chose to work with such naturalistic stimuli from the start. Comparing classification results between our labelling variations that accuracy is highly sensitive to which labelling scheme is used. Based on these observations, our recommendation is to use individual labels to train the IFMMoN.

CONCLUSION

The goal of this research was to use PMWL using a multimodal DNN. While participants were solving verbal logic puzzles, GSR, PPGF, fNIRS and ET data were collected simultaneously using LSL. We proposed a novel IFMMoN; the best model was able to classify PMWL with a 0.985 LoW accuracy on a 7-level scale. This result allows us to conclude that the IFMMoN can use the provided four modalities to classify PMWL. The MG criteria were guiding in all stages of the research: data collection, data selection and model design. The modularity criterium was satisfied through streaming of data from various separate applications into one collection software, as well as the choice of intermediate fusion using MNets that feed into one Head model. Generalisability was satisfied through improved model performance when adding multiple modalities. We showed that smaller models achieved better results in our classification task, while experiencing a speedup factor roughly equivalent to the size-difference factor. A critical discussion highlights the strong and weak points of this work, and we highlight clear avenues for improvement. Future works will work towards the classification of PMWL in real-time so that applications can be adapted to their users.

FUTURE WORKS

Currently, two variations of labels were trained on: one with individual difficulty ratings and one with averaged difficulty ratings over all participants. However, the output of our models is always a number between 0 and 1. Different objectives for classification can be interesting to pursue. Given a known option space, a vector containing probabilities of a participant's next move can be outputted. This could be interesting because it would allow the prediction and even interception of mistakes. A different route would be to train on data that originate in an alternative task. This would yield insight into the generalisation capabilities of the pipeline and networks and would thus likely benefit the overall robustness of the system.

The long-term outlook of this line of research is to create a system that can classify user PMWL in real-time and eventually can do so for multiple users simultaneously. Users can then be steered to improve the overall efficacy in their task, whatever it may be. This can, for example, be done by adapting the environment's intensity to trigger a state of flow. However, the participant can also be adapted to the environment by modulating the participant. Such modulation can be done with the help of visual, audial or even olfactory stimulation (Hughes, 2004; Weinbach et al., 2015). This research takes several relevant steps in the direction of such a system since it shows that PMWL can be classified accurately using multiple modalities. Additional modalities and users can easily be added due to the employed design principles. Furthermore, the size of the network allows for real-time implementation.

Moreover, such a system might even be able to detect which person is currently using it. This can improve the system's adaptability, as well as clusters usage patterns, similar to what is done in unsupervised problems. The ability to cluster users together may prove especially valuable in collaborative and team contexts. McDonald and Solovey (2017) demonstrated the potential of using fNIRS to distinguish between 30 different users with 63% accuracy, providing a clear route for implementation.

Data Augmentation

A proven method of increasing model accuracy is simply to supply more data, so that the model is better able to generalise. However, the task of collecting, formatting and labelling data is time-consuming and expensive. Data augmentation allows for the generation of new and unseen data, offering a solution to the data shortage problem. There are several different options for data augmentation. It can be done in the input space, the feature space or in the learned feature space, to name a few. Augmentations in the input space involve

performing several transformations on the original data. In image classification, this often takes the form of rotation or scaling, or by adding noise to the image (Sajjad et al., 2019; Sun et al., 2019). For data augmentation in input and feature space, domain expertise is often required to ensure that newly generated data respects the domain from which it is synthesised. Examples are works by Steven Eyobu and Han (2018) and Schlüter and Grill (2015). Generative models were also proven to be capable of performing such tasks while also overcoming missing data and even modalities (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2012). In the above examples, new feature extraction and augmentation blocks must be designed for each data type and problem specifically. This requires both tailoring, as well as domain expertise, and therefore does not generalise well across domains and problem statements.

Vries and Taylor propose to perform data augmentation in the learned feature space (DeVries and Taylor, 2017). Their approach relies on first learning a representation of the data and then performing data augmentations on those representations. They hypothesise that simple augmentations on encoded data, rather than input data, result in more plausible synthetic data. They propose using a sequence autoencoder on the bases of the proven generalisability of the seq2seq models that were independently devised by Cho et al. (2014) and Sutskever et al. (2014). The approach that DeVries and Taylor (2017) propose has several benefits over the other discussed methods of data augmentation. Similar to a previous work (Sutskever et al., 2014), the feature augmentation is done in reduced dimensionality, making the implementation lightweight. However, instead of designing constraints that are specific to the domain task and input data, more generalised parameters that dictate augmentation can be formulated. These parameters are then also eligible for HPO. Hence, this approach fits best within the MG criteria and would be worthy of pursuing in future works.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: 4TU.ResearchData; doi: https://doi.org/10.4121/12932801.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Twente Behavioural, Management and Social Sciences Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TD: ideation, execution of research, data collection and processing, manuscript writing, dataset and code publication. MP: feedback, supervision and funding. J-WK: feedback and supervision. BV: feedback and supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded under OP Oost of the European Regional Development Fund, as part of the BCI-Testbed Consortium (OP-OOST EFRO PROJ-00900).

ACKNOWLEDGMENTS

Special thanks to Andre Bester for writing the stimulus presentation and Thomas de Groot for writing the LSL App for the Shimmer GSR3+. We express our gratitude to Myrthe Tilleman for her extensive feedback and advice on the implementation details of the IFMMoN.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1603.04467.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). "Optuna: a next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, eds Ankur Teredesai and Vipin Kumar (Anchorage, AK: Association for Computing Machinery), 2623–2631. doi: 10.1145/3292500.3330701

Baltrušaitis, T., Ahuja, C., and & Morency, L. P. (2018). Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 14, 423–443. doi: 10.1109/TPAMI.2018.2798607

Biswas, D., Everson, L., Liu, M., Panwar, M., Verhoef, B.-E., Patki, S., et al. (2019). CorNET: deep learning framework for PPG-based heart rate estimation and biometric identification in ambulant environment. *IEEE Trans. Biomed. Circuits and Syst.* 13, 282–291. doi: 10.1109/TBCAS.2019.2892297

Boulay, C. (2020). *PyXDF*. Available online at: https://pypi.org/project/pyxdf/. Accessed June 20, 2020.

Brainzilla. (2020). *Brainzilla Zebra Puzzles*. Available online at: https://www.brainzilla.com/. Accessed June 20, 2020.

Chandar, S., Khapra, M. M., Larochelle, H., and Ravindran, B. (2016). Correlational neural networks. *Neural Comput.* 28, 257–285. doi: 10.1162/NECO_a_00801

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1406.1078.

Csikszentmihalyi, M. (1975). *Beyond Boredom and Anxiety Jossey-Bass Inc.* San Francisco, CA: Jossey-Bass.

Csikszentmihalyi, M. (1997). *Finding Flow: The Psychology of Engagement With Everyday Life*. Washington, DC: Basic Books. Available online at: https://psycnet.apa.org/record/1997-08434-000

Dargazany, A. R., Abtahi, M., and Mankodiya, K. (2019). An end-to-end (deep) neural network applied to raw EEG, fNIRs and body motion data for data fusion and BCI classification task without any pre-/post-processing. *arXiv* [Preprint]. Available online at: https://arxiv.org/pdf/1907.09523.pdf.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL. doi: 10.1109/CVPR.2009.5206848

DeVries, T., and Taylor, G. W. (2017). Dataset augmentation in feature space. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1702.05538.

Dolmans, T. (2020). Code for: perceived mental workload classification using intermediate fusion multimodal deep learning. *Zenodo* doi: 10.5281/zenodo.4043058

Dolmans, T., Poel, M., van 't Klooster, J.-W., and Veldkamp, B. P. (2020). *Perceived Mental Workload Detection Using Multimodal Physiological Data—Deep Learning, GitHub Linked.* Available online at: https://data.4tu.nl/articles/dataset/Perceived_Mental_Workload_Detection_using_Multimodal_Physiological_Data_-_Deep_Learning_GitHub_Linked/12932801. Accessed July 30, 2020.

Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., et al. (2018). "The index of pupillary activity: measuring cognitive load vis-à-vis task difficulty with pupil oscillation," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, (Montreal, QC: ACM), Paper no. 282. doi: 10.1145/3173574.3173856

Friedman, N., Fekete, T., Gal, Y. A. K., and Shriki, O. (2019). EEG-based prediction of cognitive load in intelligence tests. *Front. Hum. Neurosci.* 13:191. doi: 10.3389/fnhum.2019.00191

Groot de, T. (2020). *Shimmer*. 1st Edn. Hengelo: Thales HBA Lab.

Hart, S. G., and Staveland, L. E. (1988). "Development of NASA-TLX (task load index): results of empirical and theoretical research," in *Advances in Psychology* eds Peter A. Hancock, and Najmedin Meshkati (Amsterdam: Elsevier), 139–183.

Hori, C., Hori, T., Lee, T.-Y., Zhang, Z., Harsham, B., Hershey, J. R., et al. (2017). Attention-based multimodal fusion for video description. *CVPR*. 1, 4193–4202. Available online at: https://openaccess.thecvf.com/content_ICCV_2017/papers/Hori_Attention-Based_Multimodal_Fusion_ICCV_2017_paper.pdf.

Hughes, M. (2004). Olfaction, emotion & the amygdala: arousal-dependent modulation of long-term autobiographical memory and its association with olfaction: beginning to unravel the proust phenomenon? *Impulse* 1, 1–58. Available online at: https://impulse.appstate.edu/sites/impulse.appstate.edu/files/2004_01_01_hughes.pdf.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1725–1732. Available online at: https://openaccess.thecvf.com/content_cvpr_2014/html/Karpathy_Large-scale_Video_Classification_2014_CVPR_paper.html.

Kothe, C. (2014). Lab streaming layer (LSL). Available online at: https://github.com/sccn/labstreaminglayer. Accessed October 26, 2015.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., et al. (2016). "Eye tracking for everyone," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV.

Lim, W., Sourina, O., and Wang, L. (2018). STEW: simultaneous task EEG workload data set. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 2106–2114. doi: 10.1109/TNSRE.2018.2872924

Louedec, J. L., Guntz, T., Crowley, J., and Vaufreydaz, D. (2019). "Deep learning investigation for chess player attention prediction using eye-tracking and game data," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (New York, NY: Association for Computing Machinery), Article no.: 1. doi: 10.1145/3314111.3319827

Mahtani, K., Spencer, E. A., Brassey, J., and Heneghan, C. (2018). Catalogue of bias: observer bias. *BMJ Evid. Based Med.* 23, 23–24. doi: 10.1136/ebmed-2017-110884

Maus, B., van Breukelen, G. J., Goebel, R., and Berger, M. P. (2010). Optimization of blocked designs in fMRI studies. *Psychometrika* 75, 373–390. doi: 10.1007/s11336-010-9159-3

McDonald, D. Q., and Solovey, E. (2017). "User identification from fNIRS data using deep learning," in *The First Biannual Neuroadaptive Technology Conference*, Berlin, Germany.

Naseer, N., Qureshi, N. K., Noori, F. M., and Hong, K.-S. (2016). Analysis of different classification techniques for two-class functional near-infrared spectroscopy-based brain-computer interface. *Comput. Intell. Neurosci.* 2016:5480760. doi: 10.1155/2016/5480760

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning, ICML*, eds L. Getoor and T. Scheffer (Washington, DC: Omnipress), 689–696.

Nourbakhsh, N., Chen, F., Wang, Y., and Calvo, R. A. (2017). Detecting users' cognitive load by galvanic skin response with affective interference. *ACM Trans. Interact. Intell. Syst.* 7, 1–20. doi: 10.1145/1234

Poole, A., and Ball, L. J. (2006). "Eye tracking in HCI and usability research," in *Encyclopedia of Human Computer Interaction*, (Lancaster, UK: IGI Global), 211–219.

Ramachandram, D., and Taylor, G. W. (2017). Deep multimodal learning: a survey on recent advances and trends. *IEEE Signal Process. Mag.*, 34, 96–108. doi: 10.1109/MSP.2017.2738401

Sajjad, M., Khan, S., Muhammad, K., Wu, W., Ullah, A., and Baik, S. W. (2019). Multi-grade brain tumor classification using deep CNN with extensive data augmentation. *J. Comput. Sci.* 30, 174–182. doi: 10.1016/j.jocs.2018.12.003

Schlüter, J., and Grill, T. (2015). "Exploring data augmentation for improved singing voice detection with neural networks," in *Proceedings of the 16th International Society for Music Information Retrieval Conference*, (Málaga, Spain: ISMIR), 121–126. doi: 10.5281/zenodo.1417745

Schmalfuß, F., Mach, S., Klüber, K., Habelt, B., Beggiato, M., Körner, A., et al. (2018). "Potential of wearable devices for mental workload detection in different physiological activity conditions," in *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference*, Florence, Italy, 179–191. doi: 10.13140/RG.2.2.25439.92327

Selvaraj, N., Jaryal, A., Santhosh, J., Deepak, K. K., and Anand, S. (2008). Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography. *J. Med. Eng. Technol.* 32, 479–484. doi: 10.1080/03091900701781317

Shin, J., Kwon, J., and Im, C.-H. (2018). A ternary hybrid EEG-NIRS brain-computer interface for the classification of brain activation patterns during mental arithmetic, motor imagery and idle state. *Front. Neuroinform.* 12:15. doi: 10.3389/fninf.2018.00005

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1409.1556.

Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: part 1–learning rate, batch size, momentum and weight decay. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1803.09820.

SonaSystems. (2020). *Sona Systems*. Available online at: https://www.sona-systems.com/default.aspx. Accessed June 2020.

Srivastava, N., and Salakhutdinov, R. (2012). "Learning representations for multimodal data with deep belief nets," in *International Conference on Machine Learning Workshop*, Scotland, UK.

Steven Eyobu, O., and Han, D. S. (2018). Feature representation and data augmentation for human activity classification based on wearable IMU sensor data using a deep LSTM neural network. *Sensors* 18:2892. doi: 10.3390/s18092892

Sun, X., Hong, T., Li, C., and Ren, F. (2019). Hybrid spatiotemporal models for sentiment classification *via* galvanic skin response. *Neurocomputing* 358, 385–400. doi: 10.1016/j.neucom.2019.05.061

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, eds Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence and K. Q. Weinberger (Montreal, Canada: Neural Information Processing Systems Foundation), 3104–3112.

Tanveer, M. A., Khan, M. J., Qureshi, M. J., Naseer, N., and Hong, K.-S. (2019). Enhanced drowsiness detection using deep learning: an fNIRS study. *IEEE Access* 7, 137920–137929. doi: 10.1109/ACCESS.2019.2942838

Tavakol, M., and Dennick, R. (2011). Making sense of cronbach's alpha. *Int. J. Med. Educ.* 2, 53–55. doi: 10.5116/ijme.4dfb.8dfd

TensorFlow. (2020). *TFRecord and tf.Example*. Available online at: https://www.tensorflow.org/tutorials/load_data/tfrecord. Accessed July 30, 2020.

TobiiProAB. (2019). *Tobii Pro SDK*. Available online at: http://developer.tobiipro.com/python/python-sdk-reference-guide.html. Accessed January 15, 2020.

Toppi, J., Borghini, G., Petti, M., He, E. J., De Giusti, V., He, B., et al. (2016). Investigating cooperative behavior in ecological settings: an EEG

hyperscanning study. *PLoS One* 11:e0154236. doi: 10.1371/journal.pone.0154236

Venables, P. H., and Christie, M. J. (1980). Electrodermal activity. *Tech. Psychophysiol.* 54, 3–67.

Vielzeuf, V., Lechervy, A., Pateux, S., and Jurie, F. (2018). "Multilevel sensor fusion with deep learning," in *IEEE Sensors Lett.* 3:7100304. doi: 10.1109/LSENS.2018.2878908

Villringer, A., Planck, J., Hock, C., Schleinkofer, L., and Dirnagl, U. (1993). Near infrared spectroscopy (NIRS): a new tool to study hemodynamic changes during activation of brain function in human adults. *Neurosci. Lett.* 154, 101–104. doi: 10.1016/0304-3940(93)90181-j

Villringer, A., Margulies, D., Yating, L. V., and Craddock, R. C. (2013). U.S. Patent Application No. 13/673,630.

Weinbach, N., Kalanthroff, E., Avnit, A., and Henik, A. (2015). Can arousal modulate response inhibition? *J. Exp. Psychol. Learn. Mem. Cogn.* 41, 1873–1877. doi: 10.1037/xlm0000118

Zhang, Y., Brooks, D. H., and Boas, D. A. (2005). A haemodynamic response function model in spatio-temporal diffuse optical tomography. *Phys. Med. Biol.* 50, 4625–4644. doi: 10.1088/0031-9155/50/19/014

Zhao, Q., Li, C., Xu, J., and Jin, H. (2019). "FNIRS based brain-computer interface to determine whether motion task to achieve the ultimate goal," in *2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM)*, (Osaka, Japan: IEEE), 136–140.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 13 January 2021
doi: 10.3389/fnhum.2020.604639

Recognition of Consumer Preference by Analysis and Classification EEG Signals

Mashael Aldayel[1,2], *Mourad Ykhlef*[2] *and Abeer Al-Nafjan*[3]*

[1] Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia, [2] Information System Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia, [3] Computer Science Department, College of Computer and Information Sciences, Imam Muhammad ibn Saud Islamic University, Riyadh, Saudi Arabia

Neuromarketing has gained attention to bridge the gap between conventional marketing studies and electroencephalography (EEG)-based brain-computer interface (BCI) research. It determines what customers actually want through preference prediction. The performance of EEG-based preference detection systems depends on a suitable selection of feature extraction techniques and machine learning algorithms. In this study, We examined preference detection of neuromarketing dataset using different feature combinations of EEG indices and different algorithms for feature extraction and classification. For EEG feature extraction, we employed discrete wavelet transform (DWT) and power spectral density (PSD), which were utilized to measure the EEG-based preference indices that enhance the accuracy of preference detection. Moreover, we compared deep learning with other traditional classifiers, such as k-nearest neighbor (KNN), support vector machine (SVM), and random forest (RF). We also studied the effect of preference indicators on the performance of classification algorithms. Through rigorous offline analysis, we investigated the computational intelligence for preference detection and classification. The performance of the proposed deep neural network (DNN) outperforms KNN and SVM in accuracy, precision, and recall; however, RF achieved results similar to those of the DNN for the same dataset.

Keywords: deep learning, feature extraction, customer neuroscience, classification, signal processing, neuromarketing

OPEN ACCESS

Edited by:
Hong Gi Yeom,
Chosun University, South Korea

Reviewed by:
Saugat Bhattacharyya,
Ulster University, United Kingdom
Dalin Zhang,
Aalborg University, Denmark

***Correspondence:**
Abeer Al-Nafjan
nnafjan@imamu.edu.sa

Specialty section:
This article was submitted to Brain-Computer Interfaces, a section of the journal Frontiers in Human Neuroscience

Received: *10 September 2020*
Accepted: *23 November 2020*
Published: *13 January 2021*

Citation:
Aldayel M, Ykhlef M and Al-Nafjan A (2021) Recognition of Consumer Preference by Analysis and Classification EEG Signals. Front. Hum. Neurosci. 14:604639. doi: 10.3389/fnhum.2020.604639

1. INTRODUCTION

Neuromarketing or consumer neuroscience is an emerging disciplinary area that connects the affective and cognitive aspects of customer behavior utilizing neuroimaging tools such as brain-computer interfaces (BCIs). BCIs play the role of a communication tool between humans and computer systems without any external devices or muscle intervention to issue commands, control, or complete an interaction. BCI research and development initially considered as an assistive technology aimed to help individuals with physical disabilities in various aspects such as communication, control, and mobility. In recent times, alternative BCI applications for healthy humans have been developed, and an increasing number of these re-searches target fields such as neuromarketing (Al-Nafjan et al., 2017a). Electroencephalography (EEG) is a practical, versatile, affordable, portable, and non-invasive technique for performing repetitive sessions, tasks, and

observations. EEG-based BCIs have gained increasing interest in the literature from various scientific disciplines (Al-Nafjan et al., 2017a).

In neuromarketing, EEG-based preference detection seeks to provide insights into an individual's experience with a variety of products and media as well as his responses to market stimuli. It is a well-known fact that consumer emotions impact decision-making. On the other hand, consumer's emotions can strongly be influenced by many internal and external factors. The detection and recognition of a consumer's emotional state thus reveal true consumer preferences (Aldayel et al., 2020). Although several studies have been conducted on EEG-based emotion recognition (Ramadan et al., 2015), EEG-based studies for detecting preferences in consumers are in a very early phase. Furthermore, only a few preference-recognition studies have evaluated passive BCIs compared to the number of active BCIs. Additional research that employee BCIs to assess unconscious customer preferences is therefore needed, as opposed to research on BCIs for direct control actions (van Erp et al., 2012).

An EEG-based preferences detection system helps us understanding consumer preferences and behavior to understand how one makes a buying decision. It will help marketers and organizations acting upon them to increase customer satisfaction, positive customer experiences, consumer loyalty, and revenue. (Aldayel et al., 2020).

Although the neuromarketing field has evolved significantly in the last decade; it still has not been fully implemented in the separated academic fields in marketing research. This is because marketing researchers lack training on systematic cognitive practices in neuroscience. In addition, marketing researchers have previously doubted the implications of violating ethical rules and the privacy of consumers when using neuroscience technologies for commercial purposes. However, there are still reservations against the use of neuromarketing to extract specific knowledge of customers (Ait Hammou et al., 2013). Consequently, the potential use of EEG data during passive observations to derive product preferences remains an open debate (Telpaz et al., 2015). Accordingly, only a few neuromarketing research on advertising efficiency (Morin, 2011) were reported. This research aims to thoroughly examine the preference detection in neuromarketing using EEG indices. We chose these EEG indices based on an analysis of neural correlations of the preference that was explained in our previous research (Aldayel et al., 2020). We employed two approaches for the extraction of EEG features, namely, discrete wavelet transform (DWT) and power spectral density (PSD).

These approaches were used to measure the EEG-based preference indices. The preference indices enhance the accuracy of preference prediction. In fact, to the best of our knowledge, this is the first study that examines in detail the effect of preference indicators in enhancing the performance of classification algorithms. Furthermore, we analyzed the performance of deep learning with other conventional classification algorithms, such as k-nearest neighbor (KNN), random forest (RF), and support vector machine (SVM).

The remainder of this paper is arranged as follows: section 2 introduces the main concepts of this study with background details; section 3 presents the related works; section 4 describes the research methodology, i.e., the experiments with EEG data; section 5 discusses the evaluation results; and, finally, section 6 presents the conclusion.

2. BACKGROUND

In this section, we provide an overview of BCI-based preference detection and examine EEG-based preference indices.

2.1. BCI-Based Preference Detection

This section explains the design process of neuromarketing experiments for anticipating customer preferences and choices. First, a customer places a BCI device on his/her head. Then, the customer looks at the products while EEG data are recorded at the same time on the BCI. Next, the customer rates his/her preference on each product using a nine-point subjective ranking scale. After viewing all products, the subjective ranks need to be manually labeled as "preferred" or "unpreferred." Next, the recorded EEG signals go through preprocessing and feature extraction. The training and prediction of the classifier are based on the consumer's choice (subjective ranks). The proposed BCI system for preference detection is shown in **Figure 1**. This system has three fundamental modules: signal preprocessing, feature extraction, and classification modules.

2.2. EEG-Based Preference Indices

This section explains the preference indicators based on EEG signals. Based on our literature review (Aldayel et al., 2020), we defined the following four EEG indices to measure people's responses to marketing stimuli: the approach-withdrawal (AW) index, valence, choice index, and effort index. Such indices help marketers in realizing the reactions of consumers to products (Cartocci et al., 2017; Cherubino, 2018).

2.2.1. AW Index

The AW index measures the frontal alpha asymmetry reflected the difference between the left and right hemispheres; that is, the percentage of participation of the left hemisphere compared to the right one in the frontal alpha band (Cartocci et al., 2017; Touchette and Lee, 2017; Cherubino, 2018; Ramsøy et al., 2018). Several studies have shown the efficacy and precision of frontal alpha asymmetry as an essential determinant in emotion and neuromarketing research (Cartocci et al., 2017; Touchette and Lee, 2017; Al-Nafjan et al., 2017b; Cherubino, 2018; Modica et al., 2018; Ramsøy et al., 2018).

2.2.2. Effort Index

This measure is described as the activity level of the frontal theta in the prefrontal cortex. Higher theta activity has been associated with higher levels of task difficulty and complexity in the frontal area. It is an indication of cognitive processing arising from mental exhaustion (Modica et al., 2018) and has been frequently studied in neuromarketing research (Vecchiato et al., 2010, 2011; Boksem and Smidts, 2015; Telpaz et al., 2015; Modica et al., 2018). This reveals the significance of handling emotional changes for

FIGURE 1 | EEG-based BCI for preference recognition.

TABLE 1 | Classification algorithms employed for preferences detection in neuromarketing.

References	Classification Algorithm	Class	Best accuracy (%)
Chew et al., 2016	SVM	1. Liked 2. Disliked	75
	KNN		80
Kim et al., 2015	SVM	1. Preferred image 2. Unnoticed image	83.64
Hadjidimitriou and Hadjileontiadis, 2012, 2013	KNN	1. Liked 2. Disliked	91.02
Pan et al., 2013	SVM	1. Liked 2. Disliked	74.77
Moon et al., 2013	Quadratic discriminant analysis	1. Most preferred 2. Preferred 3. Less preferred 4. Least preferred	97.39
	KNN		97.99
Teo et al., 2017, 2018a,b	DNN	1. Liked 2. Disliked	74.38
	SVM		60.19
Hakim et al., 2018	Logistic Regression	1. Most favored 2. Least favored	67.32
	SVM		68.50
	KNN		59.98
	Decision trees		63.34
Yadava et al., 2017	DNN	1. Liked 2. Disliked	60.10
	SVM		62.85
	RF		68.41
	HMM		70.33

the formation of sustainable memory in commercials (Cartocci et al., 2017).

2.2.3. Choice Index

The choice index measures the frontal irregular fluctuations in beta and gamma, frequently associated with the actual stage of decision-making. It has been the most associated marker of willingness to pay for assessing customer desire and preferences, particularly in the gamma band. Higher gamma and beta implied greater neural activity of the left frontal area, while smaller amounts are associated with greater neural activity of the right area (Ramsøy et al., 2018).

2.2.4. Valence

Asymmetrical activation of the frontal hemisphere was correlated to preferences interpreted as valence, that is, the orientation of affective status of a consumer). Activation of the right and left prefrontal area is related to negative and positive values of valence, respectively. A large number of studies support the theory that frontal EEG asymmetry can be a measure of valence (Al-Nafjan et al., 2017b).

3. RELATED WORK

EEG-based preference classification normally includes the spectral conversion of waveforms into features exploited by

data-mining algorithms, which are trained on labeled data to forecast whether preferences are presently being detected. The preference classification of EEG varies from binary labels such as ("like" vs. "dislike") and (most favored vs. least favored) to multiple ordinal labels in the form of ranks, such as the nine-scale rank or five-scale rank. Several preference studies have used more than two algorithms of classification to find tuned classifiers for a set of features (Hwang et al., 2013). Chew et al. (Ramadan et al., 2015) evaluated user preferences of aesthetics displayed as virtual three-dimensional objects. The frequency bands were used as features for the EEG classification into two classes—"like" and "dislike"—using SVM and KNN and achieving an accuracy of 75 and 80%, respectively. These results, however, are not considered credible since the authors used a relatively low dataset of five subjects. In their extended research (Teo et al., 2017, 2018b), the authors raised the number of subjects to 16 but did not obtain better results.

By integrating EEG measures with questionnaire measures, Hakim et al. (2018) obtained an accuracy of 68.5% using the SVM to determine the most and least preferred items. Combining classifiers, such as boosting, voting, or stacking, can be used to gather multiple classification algorithms by integrating their outcomes and/or training them to complement each other and improve their performance (Lotte et al., 2018). The choice of classifiers in a BCI system is mainly dependent on both the type of mental signals acquired and the setting in which the application is used. LDA and SVM, however, are the most widely used classification algorithms and were used in over half of the EEG-based BCI experiments. Some works employed graph-based deep learning to study attention behavior (Zhang et al., 2019, 2020). **Table 1** summarizes several studies in neuromarketing in which various classifiers were used to achieve the most accurate accuracy in predicting customer preferences.

Our review in Aldayel et al. (2020) highlighted the need to use further features and fusion of classifiers to boost the accuracy of the prediction. In this study, we used a publicly available neuromarketing dataset (Yadava et al., 2017) that was previously used (Yadava et al., 2017) in building a predictive model for consumer product choice from EEG data. By using a passive BCI, researchers studied the influences of gender and age on consumer preferences in terms of like/dislike. However, all indices of EEG-based preference recognition have not been combined in any study. To the best of our knowledge, this is the first in-depth investigation of the effect of preference indicators in enhancing the performance of classification algorithms.

TABLE 2 | Affective dataset description.

Preference model	Binary (like-dislike)
Stimul	Visual-based stimuli (4 s per product picture)
Participants	Twenty-five participants, aged 18–38
Trials	1,050 trials (42 trials for each subject)
EEG device	Emotiv EPOC+ device includes 14 channels
Experimental method	Each user viewed and evaluated his or her preferences toward 42 pictures of products in form of either like or dislike.

4. MATERIALS AND METHODS

The outcome of preference detection is dependent on the choices of algorithms for feature extraction and classification. In this study, we examined the probability that two affective levels, namely, "like" and "dislike," could be identified employing different feature combinations of EEG indices as well as different approaches of feature extraction and classification algorithms.

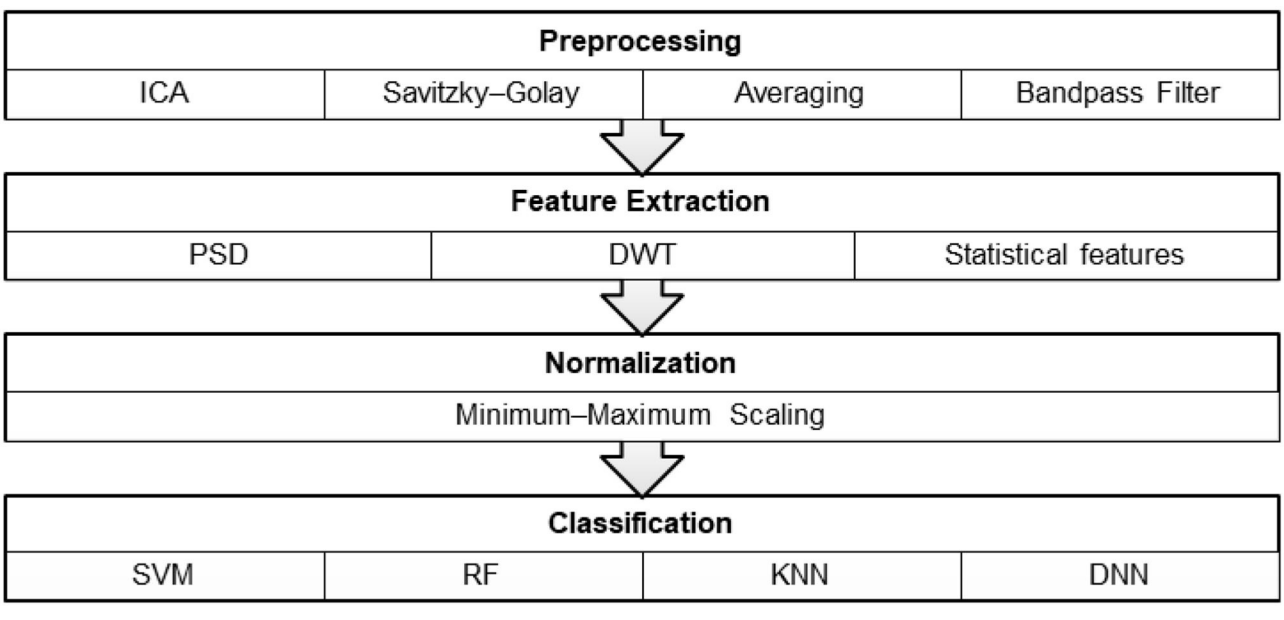

FIGURE 2 | Architecture of the consumer preference prediction system.

We chose these EEG indices based on an analysis of neural correlations of the preference that was explained in our previous research (Aldayel et al., 2020). For EEG feature extraction, we used DWT and PSD. Then, the PSD features were used to calculate the EEG-based preference indices. We applied deep learning classification to identify approaches of using intelligent computational modeling in the form of classification algorithms as these approaches can effectively reflect the subjects' preferred states. Moreover, we compared the efficiency of deep learning with other traditional classifiers, such as SVM, RF, and KNN. We developed our model in Python programming language using the Scikit-Learn, SciPy, and MNE and Keras packages for machine learning, EEG preprocessing and filtration, and signal processing and deep learning, respectively.

In this section, we present our methods and describe the architecture of the proposed EEG-based preference recognition. First, we examine the neuromarketing benchmark dataset and labeling of preferences states. Then, we illustrate how to extract features from EEG signals. Lastly, we explain the DNN classifier for preference detection. **Figure 2** presents the methods used in the consumer preference prediction system.

4.1. Dataset

This section describes a publicly available EEG dataset (**Table 2**) that has been used (Yadava et al., 2017) in neuromarketing experiments. The Emotiv EPOC+ headset was used to record

TABLE 3 | Frequency bands correlated to decomposed coefficients.

Decomposed coefficient	Frequency bands (Hz)	Decomposition level
D1	32–64	Gamma
D2	16–32	Beta
D3	8–16	Alpha
D4	4–8	Theta
A4	0–4	Delta

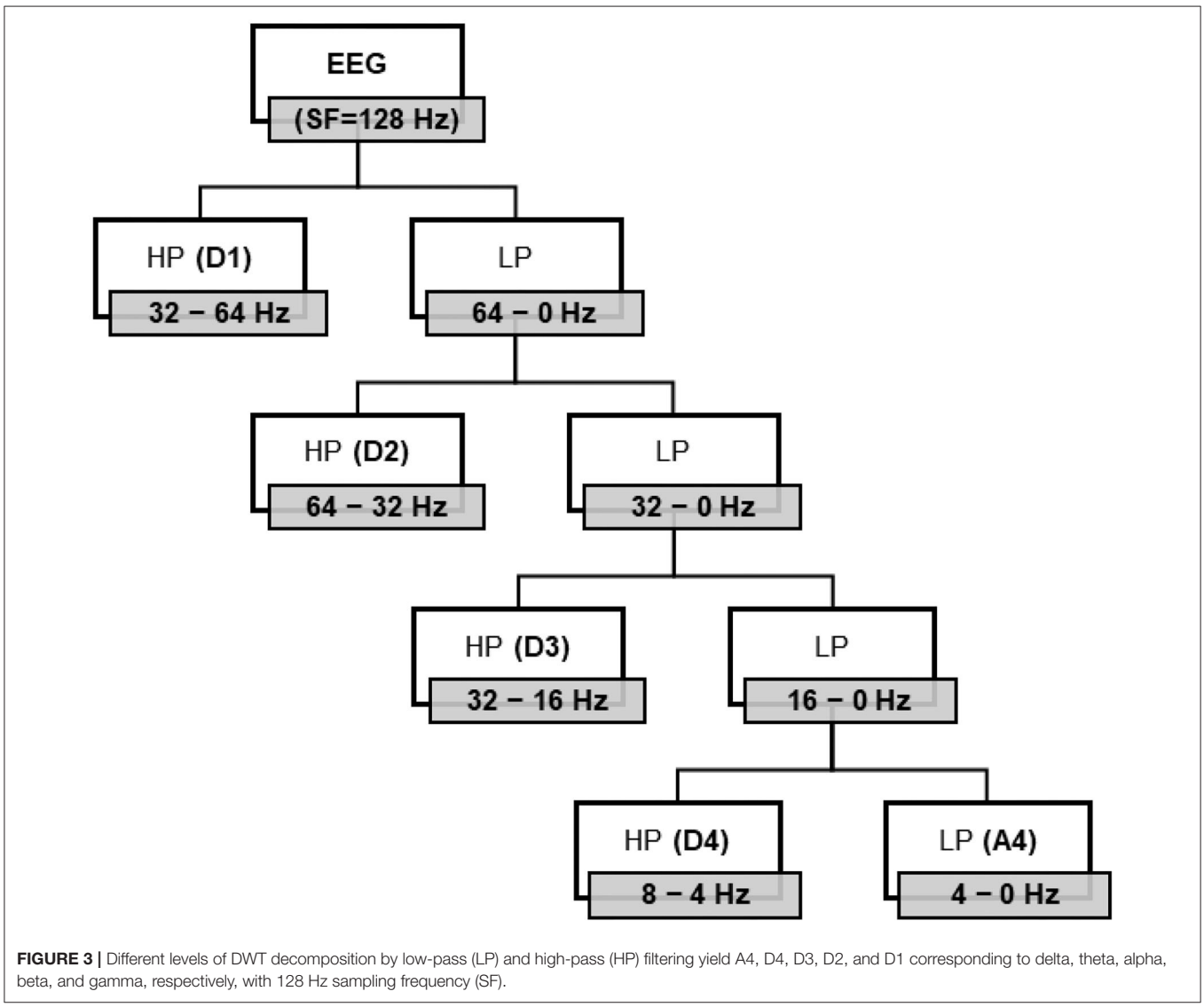

FIGURE 3 | Different levels of DWT decomposition by low-pass (LP) and high-pass (HP) filtering yield A4, D4, D3, D2, and D1 corresponding to delta, theta, alpha, beta, and gamma, respectively, with 128 Hz sampling frequency (SF).

EEG data. Twenty-five users participated, and their EEG data were recorded while they watched products on a computer screen. The age of the users ranged from 18 to 38 years. A set of 14 diverse products, each with three variations, were selected. A total of 42 ($= 14 \times 3$) diverse product pictures were then generated, and 1050 ($= 42 \times 25$) EEG data were therefore logged for all users. The EEG data were downsampled to 128 Hz and preprocessed to 14 channels, resulting in 25 documents or one document per user. The EEG features were collected from 14 channels placed at AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4 locations. Responses in the form of either "like" or "dislike" were collected from the users for each product. Each product was presented for 4 s, and EEG data were logged simultaneously. After each image was presented, the preferred choice of the user was collected.

Since consumers may not be able to express their preferences when asked to clearly articulate them, their subjective labeling is not sufficient. We extracted true hidden preferences (i.e., the ground truth table) from EEG signals. We used two methods to identify preference labels ("like" or "dislike"): (1) subjective self-assessment labels collected during the experiment; and (2) valence-based labels to identify the objective preference states. In this experiment, we used different types of preference labeling to obtain more accurate results. We used the valence index as the determinant of preference to match the target preference state—"like" or "dislike." Valence rates were categorized to lower rates if values ranged from one to five and higher rates if values ranged from six to nine. A lower valence rate is an indicator of a "dislike" preference state, while a higher valence rate is an indicator of a "like" preference state.

We used Cohen's kappa to test the agreement level between two types of labeling, namely, subjective self-assessment and valence-based labels determined from EEG. The kappa score was 0.03, which can be interpreted as a slight agreement between these labels. We also noticed there were differences in 513 of the 1050 trials, which in line with the main goal of this neuromarketing research: real and more accurate identification of preferences using EEG signals.

4.2. Signal Preprocessing

We first averaged the EEG signals and then resampled the frequency to 128 Hz per channel. From prior knowledge of EEG, the correlated signal frequency ranges produced by the brain during preferences states are mainly concentrated below 45 Hz. The useful frequency band in EEG signal data is therefore between 4 and 45 Hz. We used a bandpass filter ranging from 4.0 to 45.0 Hz. Subsequently, we used ICA and Savitzky–Golay filters to remove artifacts. We considered only the following electrodes in the preference calculation: AF3, F3, AF4, and F4.

4.3. Feature Extraction

Feature extraction aims to find important and relevant information from EEG signals. We extracted EEG frequency bands using two approaches: DWT and a PSD method named

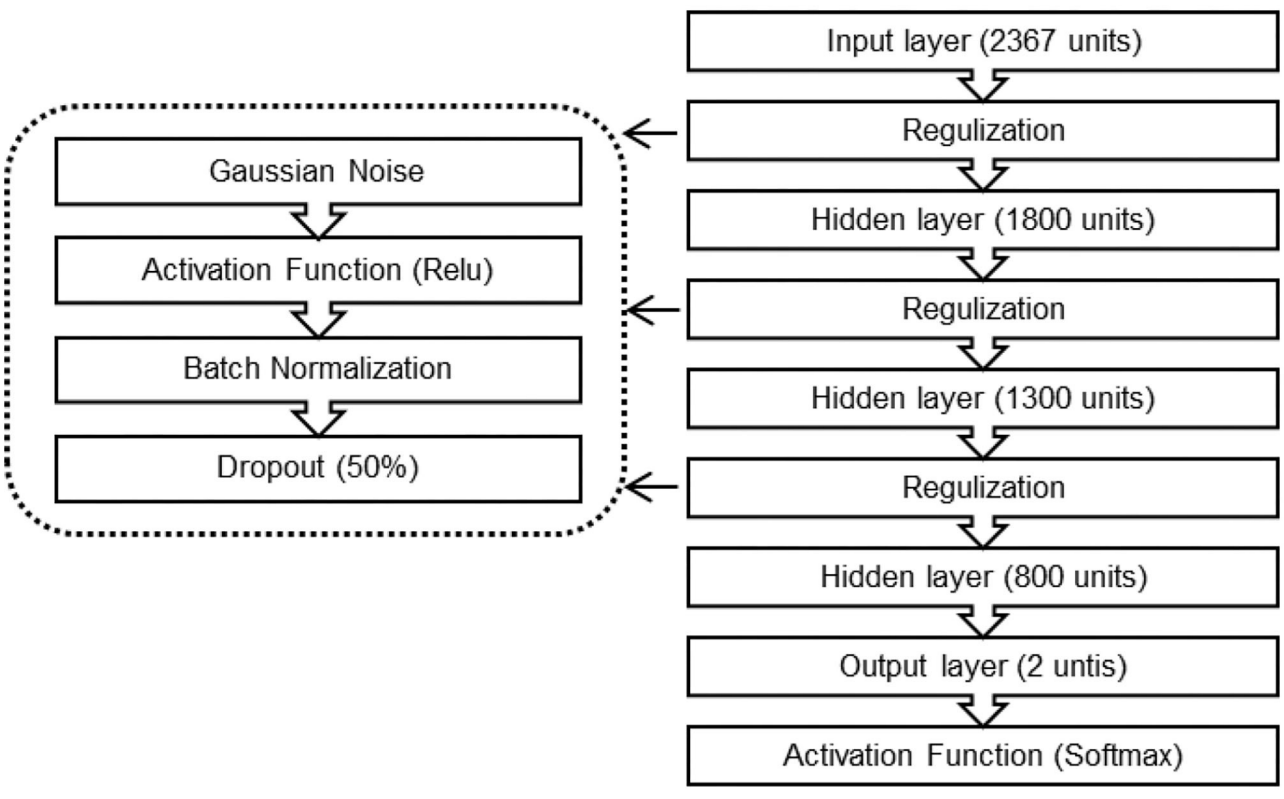

FIGURE 4 | Block diagram of the DNN classifier.

TABLE 4 | Classification results of PSD-based feature extraction with/without preference indices and the valence index (V) and different classifiers: KNN, RF, SVM, and DNN using various activation functions in the DNN: hinge, and cross-entropy (categorical and binary) functions.

Classifiers	DNN									SVM			RF			KNN		
	Hinge cross			Binary cross			Categorical cross											
Preference indices	No (%)	V (%)	All (%)	No (%)	V (%)	All (%)	No (%)	V (%)	All (%)	No (%)	V (%)	All (%)	No (%)	V (%)	All (%)	No (%)	V (%)	All (%)
Accuracy	72	92	93	77	92	93	72	92	92	71	87	86	83	94	93	72	80	78
Recall	72	92	93	77	92	93	72	92	92	71	87	86	83	94	93	72	80	78
Precision	73	92	93	79	92	93	73	92	92	72	88	87	84	94	93	73	80	79

TABLE 5 | Classification results of DWT-based feature extraction with/without preference indices and the valence index (V) and different classifiers: KNN, RF, SVM, and DNN using various activation functions in the DNN: hinge, and cross-entropy (categorical and binary) functions.

Classifiers	DNN									SVM			RF			KNN		
	Hinge cross			Binary cross			Categorical cross											
Preference indices	No	V	All	No	V	All	No	V	All	No	V	All	No	V	All	No	V	All
Accuracy	77	82	83	76	75	80	72	79	80	76	76	81	78	79	87	70	73	73
Recall	77	82	83	76	75	80	72	79	80	76	76	81	78	79	87	70	73	73
Precision	77	82	83	76	75	80	73	79	80	76	76	81	78	79	87	71	73	73

Welch. Then, we used the resulting frequency bands to calculate the preference indices. The first approach extracts a set of statistics-based features for each frequency band (details [D2-D5] and approximation [A5]) computed by DWT. The second approach stacks the features computed by PSD into a single array over the raw EEG of the channels.

4.3.1. Discrete Wavelet Transform

The DWT is a time-frequency domain analysis method that decomposes signals into different coefficients. It can be defined as multi-resolution or multi-scale analysis, where each coefficient is a unique representation of mind signals. The convolution operation is a two-function multiplication process (Chen et al., 2015; Vega-Escobar et al., 2016). Each inner product results in a wavelet coefficient. Therefore, the DWT can be expressed using the following Equation (1):

$$W(j,k) = \sum_{N=0}^{M-1} f(n) \cdot \psi_{j,k}^{*}(n) \tag{1}$$

where $f(n)$ is a signal (sequence) of length n, and $\psi_{j,k}^{*}(n)$ is scaling wavelet function. DWT decomposition can be implemented as a group of high- and low-pass filters in a filter bank. The outputs of the low-pass filters are called approximation coefficients, and those of the high-pass filters are called wavelet detail coefficients. After the filtering, the signal is down-sampled by a factor of two based on the Nyquist Theorem, resulting in a frequency band ranging between $f_n/2$ and f_n. Assuming f_s sampling frequency and L decomposition level, every detail coefficient frequency is related to the sampling frequency rate f_s of the raw signals, given by $f_n = f_s/2L + 1$. The number of wavelet decomposition levels and the selection of a proper wavelet technique are critical to achieving DWT analysis accuracy (Chen et al., 2015; Vega-Escobar et al., 2016; Yadava et al., 2017).

Since the sampling frequency in the present study was 128 Hz, we used four levels of Daubechies (db4) wavelets to decompose EEG signals into five coefficients, namely, A4, D4, D3, D2, and D1. Each coefficient is approximately correlated to the basic frequency bands, namely, (1–4 Hz) delta, (4–8 Hz) theta, (8–13 Hz) alpha, (13–22 Hz) beta, and (22–64 Hz) gamma. The decomposed details D1-D4 and approximation A4 for each of the 14 channels are shown in **Figure 3**, and their correlated ranges of frequency are listed in **Table 3**.

Moreover, we computed the (Shannon) entropy values as measures of signal complexity and extracted the statistical features that are most commonly used for signals, such as variance, standard deviation, mean, median, 25th and 75th percentile values, root mean square of the average amplitude values, zero and mean crossing rates, and the mean of the signal derivatives. These 10 statistical features and entropy and coefficient values were calculated for the five coefficients for the 14 channels. Thus, the number of DWT features was $12 \times 5 \times 14 = 840$.

4.3.2. Power Spectral Density

The PSD is an indicator of power in a certain signal in terms of frequency (Xie and Oniga, 2020). PSD is one of the most common feature extraction approaches in neuromarketing research, based on frequency domain analysis. Previous studies (Ohme et al., 2009, 2010; Khushaba et al., 2013) have demonstrated that the PSD obtained from EEG signals is suitable for determining consumer preferences. The PSD approach transforms the data from the time domain to the frequency domain, and vice versa. This conversation is focused on the fast transformation of Fourier, measuring the discrete transformation of Fourier and its opposite. In addition to DWT, we applied the PSD technique in this study to divide each EEG signal into four different frequency bands: theta θ (4–8 Hz), alpha α (8–13 Hz), beta β (13–30 Hz), and gamma γ (30–40 Hz). The MNE package for signal processing was employed for computing PSD and the average power across the frequency ranges.

4.4. Calculation of Preference Indices

We implemented various equations to measure the following EEG-based preferences indices (Section 2.2): the AW index, effort index, choice index, and valence. The AW index (frontal alpha asymmetry), measures motivation and desire as higher activation of alpha in the left frontal cortex. We used (Equation 2) stated by Touchette and Lee (2017) to measure the AW scores using electrodes F4 and F3 to find the difference between the right and left PSD divided by their amounts.

$$\text{AW index} = \frac{\alpha(F4) - \alpha(F3)}{\alpha(F4) + \alpha(F3)} \tag{2}$$

The effort index measures effort and cognitive processing as higher theta activation in the prefrontal cortex. We used the following equation to calculate the effort index:

$$\text{Effort Index} = \frac{\theta(F4) - \theta(F3)}{\theta(F4) + \theta(F3)} \tag{3}$$

The choice index measures choice possibility in decision making as higher gamma and beta activation in the frontal cortex (Ramsøy et al., 2018). We used Equation (4) reported by Ramsoy et al. to calculate the choice index for each band individually (gamma and beta) using electrodes AF3 and AF4:

$$\text{Choice index} = \frac{log(AF3) - log(AF4)}{log(AF3) + log(AF4)} \tag{4}$$

The valence measures positive emotion as left frontal activation in alpha and beta bands. We applied different valence equations and investigated the relationships between the self-assessment and different valence measurements. We computed the values of valence using Equations (5), (6), (7), and (8), which are well-explained in literature (Al-Nafjan et al., 2017b).

$$\text{Valence} = \frac{\beta(AF3, F3)}{\alpha(AF3, F3)} - \frac{\beta(AF4, F4)}{\alpha(AF4, F4)} \tag{5}$$

$$\text{Valence} = ln[\alpha(Fz, AF3, F3)] - ln[\alpha(Fz, AF4, F4)] \tag{6}$$

$$\text{Valence} = \alpha(F4) - \beta(F3) \tag{7}$$

$$\text{Valence} = \frac{\alpha(F4)}{\beta(F4)} - \frac{\alpha(F3)}{\beta(F3)} \quad (8)$$

For all preference indices, we used the PSD to extract frequency band powers from the neuromarketing data because the PSD is based on frequency analysis, unlike DWT, which is based on time and frequency analysis.

4.5. Preference Classification Algorithms

In our study, two preference states ("like" or "dislike") were detected from EEG neuromarketing data. Mainly, we proposed a DNN classifier and compared its performance with those of the KNN, RF, and SVM classifiers. We applied four classifiers, namely, DNN, KNN, RF, and SVM, to discover the optimal preference index and a well-matched classifier with the best accuracy.

RF is an ensemble learning used for classification and regression problems. It consists of a combination of several decision trees where the final outcome class is the mode of all outcome classes of individual trees. Such advantage resulted in low error rates and robustness against over-fitting while preserving computational efficiency (Al-Nafjan et al., 2017b; Teo et al., 2018a). We used the default hyper-parameters of RF in an sklearn package and adjusted the number of trees in the foreset to 500, which all processed in parallel.

4.5.1. DNN Classification

There is an explosive growth of deep learning in machine learning due to its capacity to learn good feature representations from the raw input. DL was able to provide optimal solutions to many problems in natural language processing, image, and speech. With EEG-based BCI, DL has been proven an effective tool to analyze EEG signals (Roy et al., 2019). We aim to investigate the possibility to detect two preference states in the EEG data. We proposed a DNN classifier and compared its performance with the performances of KNN and RF classifiers. The proposed DNN classifier block diagram is shown in **Figure 4**. The extracted features were first normalized using minimum-maximum normalization (Equation 9) and then fed into the DNN classifier.

$$x\_\text{scaled} = (x - \min)/(\max - \min) \quad (9)$$

In our work, we experimented various techniques and architectures. The optimal DNN architecture and properties are as follows:

- Fully connected feed-forward neural network comprised of three hidden layers.
- The input layer consisted of 2,367 units, and each hidden layer consisted of 50% units from its predecessor layer.
- Rectified Linear Unit (ReLu) as activation functions.
- Cross entropy (cost function) to compute the output of the softmax layer.

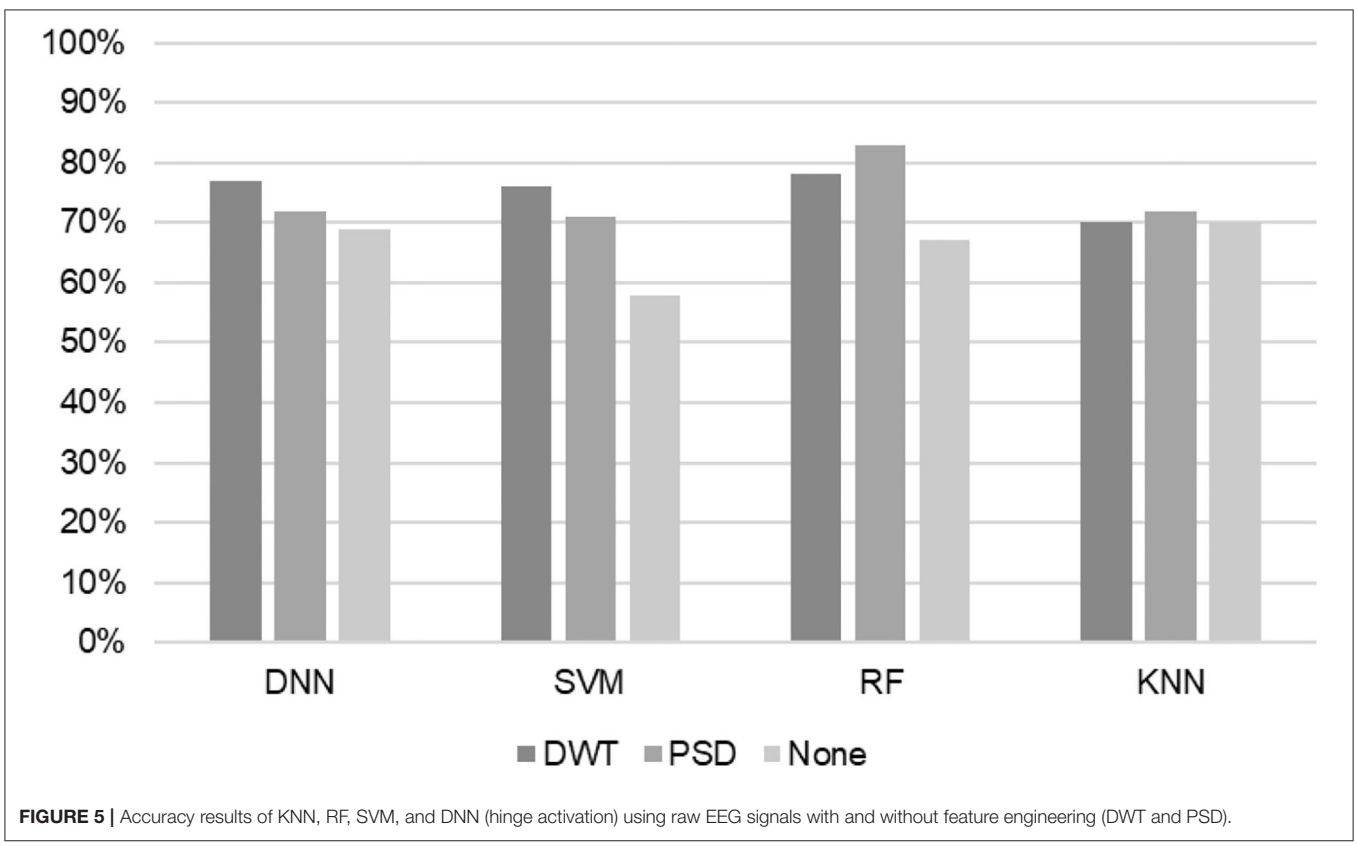

FIGURE 5 | Accuracy results of KNN, RF, SVM, and DNN (hinge activation) using raw EEG signals with and without feature engineering (DWT and PSD).

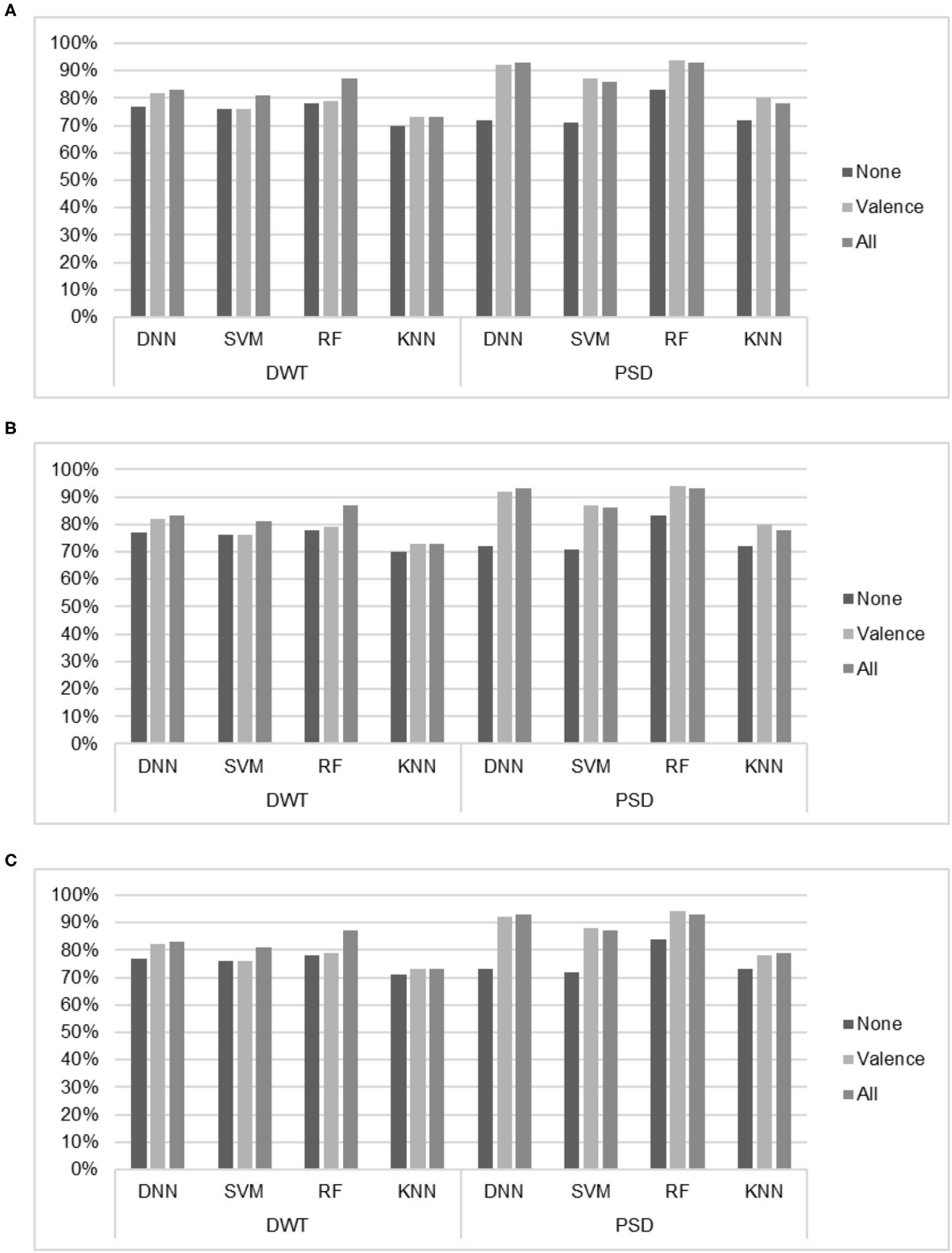

FIGURE 6 | Classification results of the SVM, RF, KNN, and DNN (hinge function) of different combination of preference indices. **(A)** Accuracy. **(B)** Recall. **(C)** Precision.

- The dimension(s) of the output layer was related to the number of target preferences state (2) units.

We used the Adam gradient descent with three objective loss functions: the binary cross-entropy, categorical cross-entropy, and hinge cross functions for training the DNN classifier with the following properties:

- Learning rate was set to 0.001.
- Dropout rate for the input and hidden layers was set to 0.5.
- Stopping criterion, to prevent over-fitting, was determined according to the model performance on a testing set.

Then, we tested our classifier on a test set, which contained approximately 20% of the data samples in the dataset. In our work, we used different approaches to prevent over-fitting including regularization (such as L1 regularization, L2 regularization, and Gaussian noise), early stopping, and dropout. Adding noise to the DNN model in a relatively small dataset can improve its robustness with regularizing effect and decrease over-fitting.

5. RESULTS AND DISCUSSION

We detected the preference states ("like" or "dislike") of the subjects using two different feature extraction methods (PSD and DWT) and four classifiers: DNN, KNN, RF, and SVM. For validation and evaluation, we used various measurements, namely precision, recall, and accuracy. The precision was the percentage of the prediction of "like" states, which was correct. The recall was the percentage of actually expected "like" states. To evaluate the efficiency of the classification algorithms, we split the data into train and test sets with holdout cross-validation.

The proposed DNN classifier was compared with three traditional classifiers for EEG signals: KNN, RF, and SVM using PSD and DWT feature extraction methods as well as various preference indices. **Tables 4**, **5** list the results of recall, accuracy, and precision results of the KNN, RF, SVM, and DNN algorithms using various activation functions in the DNN: hinge and cross-entropy (categorical and binary) functions. To show the importance of the feature extraction (DWT and PSD), **Figure 5** presents the accuracy results of KNN, RF, SVM, and DNN (hinge activation) using raw EEG signals with and without feature engineering (DWT and PSD).

When using PSD-based features, the KNN and SVM classifiers yielded enhanced accuracies of 80 and 87% with the valence index, whereas RF and DNN (binary cross-entropy function) achieved the highest accuracy of 93% with all preference indices. Similar results were achieved with the valence index. Using DWT-based features, the best results were achieved with all preference indices for all classifiers. The KNN and SVM classifiers led to enhanced accuracies of 73 and 81%, respectively. The highest accuracy was 87% with RF and the second-highest accuracy was 83% with DNN and the hinge loss function.

Figure 6 analyzes the results from the viewpoint of preference indices. We consider the DNN results with hinge loss function as it achieved the best accuracy result compared with other loss functions. About EEG features that exclude preference indices, the best accuracy results reached 83% with RF and DWT-based features.

6. CONCLUSIONS

A DNN model is proposed for detecting subject preferences from EEG signals using the benchmark neuromarketing dataset. Two kinds of features—PSD and DWT—have been generated from the EEG to obtain a set of 2367 interesting attributes, which demonstrate the EEG task in each experiment. We used various evaluation measures (recall, accuracy, and precision) to test the performance of the classifiers. We built four classifiers, namely, DNN, KNN, RF, and SVM.

The results demonstrated that RF reached the best results in PSD-based and DWT-based features with either valence or all preference indices, however, RF obtained comparable outcomes to DNN. PSD-based features achieved better results in preference detection than DWT-based features. Moreover, combining preference indices leads to better results with either PSD or DWT-based features. This is perhaps the first study that examines in detail the effect of preference indicators in enhancing the performance of classification algorithms.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://link.springer.com/article/10.1007/s11042-017-4580-6.

AUTHOR CONTRIBUTIONS

MA conceived, designed, and performed the experiment, analyzed and interpreted the data, and drafted the manuscript. MY supervised the analysis and reviewed the manuscript. AA-N co-supervised this study, and contributed to the discussion. All authors have read and approved the submitted version of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the deanship of scientific research for funding and supporting this research through the initiative of DSR Graduate Students Research Support (GSR) at King Saud University.

REFERENCES

Ait Hammou, K., Galib, M. H., and Melloul, J. (2013). The contributions of neuromarketing in marketing research. *J. Manage. Res.* 5:20. doi: 10.5296/jmr.v5i4.4023

Aldayel, M., Ykhlef, M., and Al-Nafjan, A. (2020). Deep learning for EEG-based preference classification in neuromarketing. *Appl. Sci.* 10, 1–23. doi: 10.3390/app10041525

Al-Nafjan, A., Hosny, M., Al-Ohali, Y., and Al-Wabil, A. (2017a). Review and classification of emotion recognition based on EEG brain-computer interface

system research: a systematic review. *Appl. Sci.* 7:1239. doi: 10.3390/app7121239

Al-Nafjan, A., Hosny, M., Al-Wabil, A., and Al-Ohali, Y. (2017b). Classification of human emotions from electroencephalogram (EEG) signal using deep neural network. *Int. J. Adv. Comput. Sci. Appl.* 8, 419–425. doi: 10.14569/ijacsa.2017.080955

Boksem, M. A. S., and Smidts, A. (2015). Brain responses to movie trailers predict individual preferences for movies and their population-wide commercial success. *J. Market. Res.* 52, 482–492. doi: 10.1509/jmr.13.0572

Cartocci, G., Caratu, M., Modica, E., Maglione, A. G., Rossi, D., Cherubino, P., et al. (2017). Electroencephalographic, heart rate, and galvanic skin response assessment for an advertising perception study: application to antismoking Public Service Announcements. *J. Visual. Exp.* 126:e55872. doi: 10.3791/55872

Chen, L. L., Zhao, Y., Zhang, J., and Zou, J. Z. (2015). Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning. *Expert Syst. Appl.* 42, 7344–7355. doi: 10.1016/j.eswa.2015.05.028

Cherubino, P. (2018). "Application of Neuro- Marketing techniques to the wine tasting experience," in *11th Annual Conference of the EuroMed Academy of Business* (Malta), 290–298.

Chew, L. H., Teo, J., and Mountstephens, J. (2016). Aesthetic preference recognition of 3D shapes using EEG. *Cogn. Neurodyn.* 10, 165–173. doi: 10.1007/s11571-015-9363-z

Hadjidimitriou, S. K., and Hadjileontiadis, L. J. (2012). Toward an EEG-based recognition of music liking using time-frequency analysis. *IEEE Trans. Biomed. Eng.* 59, 3498–3510. doi: 10.1109/TBME.2012.2217495

Hadjidimitriou, S. K., and Hadjileontiadis, L. J. (2013). EEG-Based classification of music appraisal responses using time-frequency analysis and familiarity ratings. *IEEE Trans. Affect. Comput.* 4, 161–172. doi: 10.1109/T-AFFC.2013.6

Hakim, A., Klorfeld, S., Sela, T., Friedman, D., Shabat-Simon, M., and Levy, D. J. (2018). Pathways to consumers minds: using machine learning and multiple EEG metrics to increase preference prediction above and beyond traditional measurements. *bioRxiv*. Available online at: https://www.biorxiv.org/content/10.1101/317073v2

Hwang, H.-J., Kim, S., Choi, S., and Im, C.-H. (2013). EEG-based brain-computer interfaces: a thorough literature survey. *Int. J. Hum. Comput. Interact.* 29, 814–826. doi: 10.1080/10447318.2013.780869

Khushaba, R. N., Wise, C., Kodagoda, S., Louviere, J., Kahn, B. E., and Townsend, C. (2013). Consumer neuroscience: assessing the brain response to marketing stimuli using electroencephalogram (EEG) and eye tracking. *Expert Syst. Appl.* 40, 3803–3812. doi: 10.1016/j.eswa.2012.12.095

Kim, Y., Kang, K., Lee, H., and Bae, C. (2015). "Preference measurement using user response electroencephalogram," in Computer Science and Its Applications, eds J. Park, I. Stojmenovic, H. Jeong, and G. Yi (Berlin; Heidelberg: Springer) 1315–1324. doi: 10.1007/978-3-662-45402-2_183

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., et al. (2018). A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update. *J. Neural Eng.* 15:aab2f2. doi: 10.1088/1741-2552/aab2f2

Modica, E., Cartocci, G., Rossi, D., Martinez Levy, A. C., Cherubino, P., Maglione, A. G., et al. (2018). Neurophysiological responses to different product experiences. *Comput. Intell. Neurosci.* 2018,1—10. doi: 10.1155/2018/9616301

Moon, J., Kim, Y., Lee, H., Bae, C., and Yoon, W. C. (2013). Extraction of user preference for video stimuli using eeg-based user responses. *ETRI J.* 35, 1105–1114. doi: 10.4218/etrij.13.0113.0194

Morin, C. (2011). Neuromarketing: the new science of consumer behavior. *Society* 48, 131–135. doi: 10.1007/s12115-010-9408-1

Ohme, R., Reykowska, D., Wiener, D., and Choromanska, A. (2009). Analysis of neurophysiological reactions to advertising stimuli by means of EEG and galvanic skin response measures. *J. Neurosci. Psychol. Econ.* 2, 21–31. doi: 10.1037/a0015462

Ohme, R., Reykowska, D., Wiener, D., and Choromanska, A. (2010). Application of frontal EEG asymmetry to advertising research. *J. Econ. Psychol.* 31, 785–793. doi: 10.1016/j.joep.2010.03.008

Pan, Y., Guan, C., Yu, J., Ang, K. K., and Chan, T. E. (2013). "Common frequency pattern for music preference identification using frontal EEG," in *International IEEE/EMBS Conference on Neural Engineering, NER* (San Diego, CA), 505–508. doi: 10.1109/NER.2013.6695982

Ramadan, R. A., Refat, S., Elshahed, M. A., and Ali, R. A. (2015). *Brain-Computer Interfaces*. Cham: Springer International Publishing. doi: 10.1007/978-3-319-10978-7

Ramsøy, T. Z., Skov, M., Christensen, M. K., and Stahlhut, C. (2018). Frontal brain asymmetry and willingness to pay. *Front. Neurosci.* 12:138. doi: 10.3389/fnins.2018.00138

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng.* 16:051001. doi: 10.1088/1741-2552/ab260c

Telpaz, A., Webb, R., and Levy, D. J. (2015). Using EEG to predict consumers' future choices. *J. Market. Res.* 52, 511–529. doi: 10.1509/jmr.13.0564

Teo, J., Chew, L. H., Chia, J. T., and Mountstephens, J. (2018a). Classification of affective states via EEG and deep learning. *Int. J. Adv. Comput. Sci. Appl.* 9, 132–142. doi: 10.14569/IJACSA.2018.090517

Teo, J., Hou, C. L., and Mountstephens, J. (2017). Deep learning for EEG-based preference classification. *AIP Conf. Proc.* 1891, 020141. doi: 10.1063/1.5005474

Teo, J., Hou, C. L., and Mountstephens, J. (2018b). Preference classification using Electroencephalography (EEG) and deep learning. *J. Telecommun. Electron. Comput. Eng.* 10, 87–91.

Touchette, B., and Lee, S. E. (2017). Measuring neural responses to apparel product attractiveness: an application of frontal asymmetry theory. *Cloth. Tex. Res. J.* 35, 3–15. doi: 10.1177/0887302X16673157

van Erp, J., Lotte, F., and Tangermann, M. (2012). Brain-computer interfaces: beyond medical applications. *Computer* 45, 26–34. doi: 10.1109/MC.2012.107

Vecchiato, G., Astolfi, L., Fallani, F. D. V., Cincotti, F., Mattia, D., Salinari, S., et al. (2010). Changes in brain activity during the observation of TV commercials by using EEG, GSR and HR measurements. *Brain Topogr.* 23, 165–179. doi: 10.1007/s10548-009-0127-0

Vecchiato, G., Toppi, J., Astolfi, L., Fallani, F. D. V., Cincotti, F., Mattia, D., et al. (2011). Spectral EEG frontal asymmetries correlate with the experienced pleasantness of TV commercial advertisements. *Med. Biol. Eng. Comput.* 49, 579–583. doi: 10.1007/s11517-011-0747-x

Vega-Escobar, L., Castro-Ospina, A., and Duque-Munoz, L. (2016). "DWT-based feature extraction for motor imagery classification," in *6th Latin-American Conference on Networked and Electronic Media (LACNEM 2015)* (Medellin). doi: 10.1049/ic.2015.0309

Xie, Y., and Oniga, S. (2020). A review of processing methods and classification algorithm for EEG signal. *Carpathian J. Electron. Comput. Eng.* 13, 23–29. doi: 10.2478/cjece-2020-0004

Yadava, M., Kumar, P., Saini, R., Roy, P. P., and Dogra, D. P. (2017). Analysis of EEG signals and its application to neuromarketing. *Multimedia Tools Applicat.* 76, 19087–19111. doi: 10.1007/s11042-017-4580-6

Zhang, D., Chen, K., Jian, D., and Yao, L. (2020). Motor imagery classification via temporal attention cues of graph embedded EEG signals. *IEEE J. Biomed. Health Inform.* 24, 2570–2579. doi: 10.1109/JBHI.2020.2967128

Zhang, D., Yao, L., Chen, K., Wang, S., Haghighi, P. D., and Sullivan, C. (2019). A graph-based hierarchical attention model for movement intention detection from EEG signals. *IEEE Trans. Neural Syst. Rehabil. Eng.* 27, 2247–2253. doi: 10.1109/TNSRE.2019.2943362

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Human Neuroscience

METHODS
published: 20 January 2021
doi: 10.3389/fnhum.2020.605246

Two-Level Domain Adaptation Neural Network for EEG-Based Emotion Recognition

Guangcheng Bao[1], Ning Zhuang[1], Li Tong[1], Bin Yan[1], Jun Shu[1], Linyuan Wang[1], Ying Zeng[1,2] and Zhichong Shen[1]*

[1] Henan Key Laboratory of Imaging and Intelligent Processing, PLA Strategic Support Force Information Engineering University, Zhengzhou, China, [2] Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China

Emotion recognition plays an important part in human-computer interaction (HCI). Currently, the main challenge in electroencephalogram (EEG)-based emotion recognition is the non-stationarity of EEG signals, which causes performance of the trained model decreasing over time. In this paper, we propose a two-level domain adaptation neural network (TDANN) to construct a transfer model for EEG-based emotion recognition. Specifically, deep features from the topological graph, which preserve topological information from EEG signals, are extracted using a deep neural network. These features are then passed through TDANN for two-level domain confusion. The first level uses the maximum mean discrepancy (MMD) to reduce the distribution discrepancy of deep features between source domain and target domain, and the second uses the domain adversarial neural network (DANN) to force the deep features closer to their corresponding class centers. We evaluated the domain-transfer performance of the model on both our self-built data set and the public data set SEED. In the cross-day transfer experiment, the ability to accurately discriminate joy from other emotions was high: sadness (84%), anger (87.04%), and fear (85.32%) on the self-built data set. The accuracy reached 74.93% on the SEED data set. In the cross-subject transfer experiment, the ability to accurately discriminate joy from other emotions was equally high: sadness (83.79%), anger (84.13%), and fear (81.72%) on the self-built data set. The average accuracy reached 87.9% on the SEED data set, which was higher than WGAN-DA. The experimental results demonstrate that the proposed TDANN can effectively handle the domain transfer problem in EEG-based emotion recognition.

Keywords: EEG, emotion recognition, topological graph feature, maximum mean discrepancy, domain adversarial network

OPEN ACCESS

Edited by:
Hong Gi Yeom,
Chosun University, South Korea

Reviewed by:
Xiangmin Xu,
South China University of Technology, China
Nattapong Thammasan,
University of Twente, Netherlands

***Correspondence:**
Ying Zeng
yingzeng@uestc.edu.cn

Specialty section:
This article was submitted to Brain-Computer Interfaces, a section of the journal Frontiers in Human Neuroscience

Received: *11 September 2020*
Accepted: *22 December 2020*
Published: *20 January 2021*

Citation:
Bao G, Zhuang N, Tong L, Yan B, Shu J, Wang L, Zeng Y and Shen Z (2021) Two-Level Domain Adaptation Neural Network for EEG-Based Emotion Recognition. Front. Hum. Neurosci. 14:605246. doi: 10.3389/fnhum.2020.605246

INTRODUCTION

Emotion recognition plays an important role in the human-computer interaction system (Walter et al., 2014). In addition, accurately identifying the patient's emotions helps improve the quality of medical care (Acharya et al., 2015). Currently, popular emotion detection can be divided into two categories. One is based on non-physiological signals such as facial expressions (Gur et al., 1992). The other is based on physiological signals such as electroencephalogram (EEG) signals (Sourina et al., 2012). Facial expressions are prone to misinterpretation (Saxen et al., 2017), but

EEG signals are directly extracted from the cerebral cortex without damage, accurately reflecting the physiological state of the human brain. Therefore, emotion recognition technology based on EEG signals has received more extensive research interest.

At present, researchers use a variety of traditional machine learning methods to identify emotions via EEG, including support vector machines (SVM) (Alarcao and Fonseca, 1949), linear discriminant analysis (LDA) (Zong et al., 2016), *K*-nearest neighbor (KNN) (Mehmood and Lee, 2015), and more. Although these methods have achieved good performance in EEG emotion recognition, there are still limitations. Due to the individual differences and non-stationarity of EEG signals, traditional machine learning methods have high requirements for extracted features. However, most of the current methods for extracting features from EEG signals are manual, and the results are often not satisfactory.

Researchers have proposed a variety of shallow unsupervised domain adaptation methods to solve the cross-subject classification problem. The main idea of this shallow unsupervised domain adaptation method is to learn shared features by minimizing the distance of the distribution difference between features from different domains. Algorithms for measuring the distance between two distributions usually include KL divergence, Wasserstein distance, Shannon entropy distance, and maximum mean discrepancy (MMD) (Chai et al., 2016). In recent years, the multiple kernel maximum mean discrepancy (MK-MMD) (Hang et al., 2019) has shown a greater advantage in domain adaptation. Pan et al. (2011) proposed a domain adaptation method called Transfer Component Analysis (TCA). The principle was to map two differently distributed data points to a high-dimensional regenerative kernel Hilbert space (RKHS) by learning a set of universal transfer mappings between the source and target domains, and then minimize the MMD in the RKHS to minimize the distribution distance between the source and target domains. The Transformation Parameter Transfer (TPT) method proposed by Sangineto et al. (2014) first trained the classifier of each source domain, then trained a regression function to learn the relationship between the data distribution and the classifier parameters, and finally used the target domain distribution and classifier mapping to obtain the target classifier, thereby realizing distribution transfer. The shallow domain adaptation method has achieved remarkable results in cross-subject classification, but its performance depends in large part on the quality of the features and the classification performance of the classifier. However, it is well-known that it is very difficult to design a general classifier. If the extracted features are inaccurate, the resulting model may lead to reduced classification performance, that is, negative transfer.

Therefore, researchers are more interested in deep domain adaptation methods. Studies have found that deep neural networks can learn more transferable features for domain adaptation (Donahue et al., 2013; Yosinski et al., 2014). Ganin et al. (2016) proposed a domain-adversarial training of neural networks (DANN), an approach composed of two main parts. First, the source and target domains were mapped to a common subspace through shared parameters for alignment, and then the source domain classification loss was minimized. Domain classification loss of the source and target domains was maximized to achieve domain confusion. The deep adaptation network (DAN) (Hang et al., 2019) proposed by Long et al. relied on multi-kernel MMD (MK-MMD) to adapt the source domain and target domain after multiple fully connected layers in the deep layer. In addition, Luo et al. (2018) proposed a domain adaptation framework based on WGAN. There were two main steps; the first was to pre-train the source domain, and then the Wasserstein algorithm was used for adversarial training to adapt the target domain to the source domain. Similar to the WGAN framework, Jimenez-Guarneros and Gomez-Gil (2020) proposed a custom domain adaptive method (CDA). This method used adaptive batch normalization (AdaBN) (Li et al., 2018) and MMD in two independent networks to reduce the marginal and conditional distribution of the source and target domains. Ma et al. (2019) proposed an adversarial domain generalization framework called DResNet, which learned specific biased weights for each source domain and unbiased weights shared by all domains. Unlike the other methods mentioned above, this method did not use any information about the target domain. At present, most of the methods based on deep domain adaptation put the distributed adaptation strategy on the specific task layer of the deep network, which can better reduce the domain difference. However, these deep domain adaptation methods usually only use simple distributed adaptation methods, which cannot confuse the source domain and target domain well. In addition, most of the existing deep domain adaptation methods are based on image classification, and there are few domain adaptation methods based on cross-subject EEG emotion classification. For example, Zheng and Lu (2016) proposed a framework of emotion transfer based on TPT, Luo et al. (2018) proposed a domain adaptation method for EEG emotion based on WGAN, Li Y. et al. (2019) proposed a domain adversarial method for EEG emotion based on Bi-hemisphere, Li J. et al. (2019) proposed a multisource transfer method for EEG emotion, Li et al. (2020) proposed a domain adaptation method for EEG emotion based on latent representation similarity.

Clearly, even if a subject induces the same emotion at different times, some external factors such as temperature and humidity will cause physiological changes (Chueh et al., 2012). This will cause changes in their EEG signals that are called cross-day variability. At present, few researchers analyze and study this problem. Although the tasks of cross-day transfer and cross-subject transfer are the same, they both match the distribution of source domain and target domain to eliminate the distribution difference. But they have different characteristics to learn. The challenge in cross-day transfer is to train a general classification model for the same subject, which must extract the same EEG features for the same emotional states across days. Cross-subject transfer, on the other hand, trains a general classification model for different subjects, and must extract the same EEG features for the same emotional states across subjects. It is very difficult to build a general model and extract high-quality features; a deep neural network is better than traditional methods at learning features.

In this paper, we propose a two-level deep domain adversarial network model based on a deep convolutional neural network to recognize EEG emotion transfer. EEG features are mapped to images, and the spatial topological information of EEG features is simultaneously retained using the method presented by Bashivan et al. (2015) and Hwang et al. (2020). A deep convolutional neural network can learn more transferable features by learning the EEG feature topological map. We use the AdaBN layer to standardize the characteristics of the source and target domains, and then use MMD to reduce the distribution difference between the source and target domains to achieve the domain matching effect. Finally, through the adversarial domain adaptation network, the distribution difference between the source and target domains is further reduced dynamically to achieve complete domain confusion. We verified the cross-day transfer and cross-subject transfer.

The main contributions of this manuscript lie in the following aspects:

1) A two-level domain adaptation neural network (TDANN) was proposed to construct a transfer model for EEG-based emotion recognition. Through the combination of MMD and DANN, the source domain, and the target domain can adapt to each other better.
2) Topology features were used to increase spatial information, which can better describe the state of different emotions. In addition, a convolutional network with adaptive standard layer was proposed to extract effective emotion features from topology graph.
3) A cross-subject and cross-day emotion EEG data set was constructed to study the transfer models for EEG-based emotion recognition. In this data set, each subject participated in six sessions, which is the largest number of sessions in the current public datasets for EEG-based emotion recognition.

EXPERIMENTAL SETUP

Since there is no data set big enough for research on the cross-day transfer model for EEG-based emotion recognition, we designed an experiment to build an EEG data set for emotion recognition. Each subject's EEG signals under different emotion states were collected three times with a 1 week interval, and the sequence was repeated again after 1 month.

Stimuli and Experimental Procedure

Thirty-six video clips of joy, sadness, anger, and fear were chosen for the experiment from the Chinese affective video system (Xu et al., 2010) and from a self-built emotional material library. The self-built library was a standardized multi-sensory emotional stimulation material library built on the basis of psychological methods and composed of various comedy, love, crime, war, documentary, and horror films with a clear picture and good sound. In order to induce a single type of emotion accurately, the length of movie clips was set to 50–335 s and the emotion induced by each video reached the highest intensity at the end.

The experiment was performed in three parts, namely, Experiments A, B, and C. The details of the movie clips used in each part are listed in **Table 1**. See **Figure 1** for an overview of the experimental procedure.

The order of the three parts was random, and the time interval between them was 1 week. In each part, four categories of movie clips (total of 12 movie clips) were randomly presented to the participants in 12 trials, and each trial involved the following steps:

1. 10-s display of the current trial number to inform the participants of their progress
2. 5 s of baseline signal collection (fixation cross)
3. Display of the movie clips
4. 10-s self-assessment for arousal and valence (based on self-assessment manikins)
5. 5 min break between different emotional types of video clips.

EEG Recording and Preprocessing

The Beck Anxiety Inventory (Fydrich et al., 1992), Hamilton Anxiety Rating Scale (Shear et al., 2010), and Hamilton Depression Scale (Hamilton, 2004) were administered to exclude individuals with anxiety, depression, or physical abnormalities and those under sedatives and psychotropic drugs. The participants included 16 college students (eight males and eight females) with an average age of 23.13 years (range = 19–27, SD = r 2.37). All participants were right-handed, with normal or corrected vision and hearing.

EEG signals were recorded with a gtec.HIamp system. The sampling rate was 512 Hz, a band-pass filter in the range of 0.1–100 Hz was utilized to filter EEG signals, and a notch filter with a frequency of 50 Hz was used. The layout of 62 electrodes followed the international 10–20 system. The Fz electrode was used for reference calculation. Thus, the number of effective electrodes was 61.

First, we selected the subjects' EEG data based on their self-evaluated valence. The threshold was set to 5. If a participant's valence for happy videos exceeded five points, and videos with sadness, anger, and fear were <5, we believed that the participant's emotions were accurately induced, and the participant's signal was retained; otherwise the participant's signal was deleted. We also excluded subjects with poor EEG signal quality, for example large EMG artifacts or EEG signal drift. In the end, we eliminated 4 subjects and retained 12 subjects with better signals. Then, we selected the last 50 s of the EEG signal from each video clip for analysis. In the video material, the shortest video length is 50 s. In order to make the sample balanced, we intercepted the data corresponding to all videos in the last 50 s. The EEG signals were passed through a 2-s time window and overlapped by 50%. After segmentation, each video segment had a total of 49 samples, and each participant had a total of 588 samples. There were 3,528 samples over 6 days.

Before extracting features, the data was preprocessed. First, the channels with poor data were recompressed and averaged with the surrounding channels. Next, the blind source analysis algorithm FastICA (Hyvärinen, 1999) was used to remove EOG artifacts. We used FastICA to decompose the original EEG signal into multiple ICs, identifying IC with occasional large amplitude as eye-movement artifact and removed it. Third,

TABLE 1 | Brief description of the movie clips used in the emotion experiment.

No.	Label	Experiment A		Experiment B		Experiment C	
		Movie Name	Length (sec)	Movie Name	Length(sec)	Movie Name	Length (sec)
1	Joy	More Haste Less Speed	109	The Eagle Shooting Heroes (1)	228	Lost on Journey	281
2	Joy	A Big Potato	142	A World Without Thieves	191	Home with Kids	187
3	Joy	Flirting Scholar	112	Chaplin Comedy	244	The Eagle Shooting Heroes (2)	53
4	Sadness	My Brothers and Sisters	146	Dearest (1)	182	Man Phoning in the Snow	142
5	Sadness	Mother Love Me Once Again	137	Tangshan Earthquake;	335	Echoes of the Rainbow	241
6	Sadness	Warm Spring	102	Dearest (2)	120	ROB-B-HOOD	234
7	Anger	Fist of Fury (2)	66	YiP Man II	172	Japanese Aggression	96
8	Anger	Kangxi Dynasty	94	Don't Talk to Strangers	205	Blind Mountain	275
9	Anger	Conman in Tokyo	107	Fist of Fury (1)	258	Poaching Wild Animals	148
10	Fear	Help Me	50	Lights Out	134	A Man Lying in Bed	162
11	Fear	The Game of Killing (1)	159	Man Lying on the Ground	291	The Grudge	167
12	Fear	Inner Senses	247	Snake Eating People	158	A Woman Taking a Gun	190

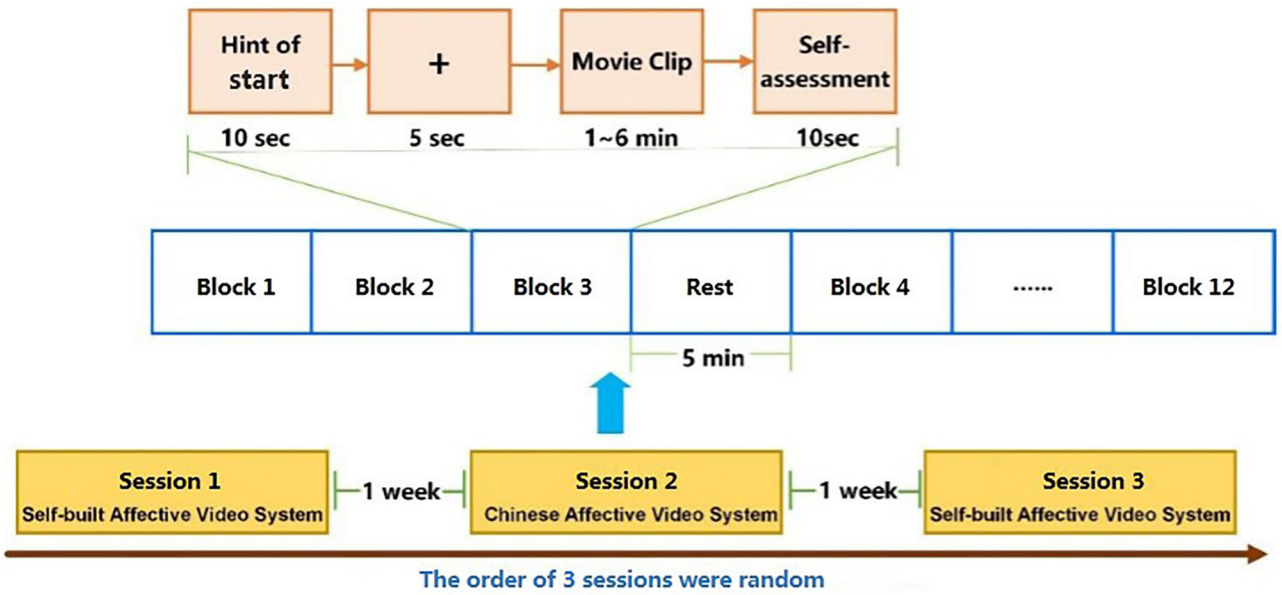

FIGURE 1 | Experimental procedure. The experiment was performed in three parts: Experiments A, B, and C. The order of the three parts was random and the time interval was 1 week. In each part, 12 movie clips with four discrete categories of emotion (joy, sadness, anger, and fear) were presented in 12 trials. Each subject participated in two complete experiments.

we used a band-pass filter of 0.1–64 Hz to filter out high-frequency interference in EEG signals. Then, we used the reference electrode standardization technology (REST) to re-reference the data (Yao, 2001; Yao et al., 2019), and finally, we removed the 5 s of the baseline before the task from the EEG signal.

TWO-LEVEL DOMAIN ADAPTATION NEURAL NETWORK

The two-level deep domain adaptation framework for EEG-based emotion recognition is shown in **Figure 2**. The framework was mainly composed of three parts, namely a feature generator, a domain discriminator, and a classifier. The main task of the generator was to further learn the stable features related to the emotional state in the EEG image and to align the source and target domains in the subspace. The domain discriminator further reduced the distribution distance between the source and target domains.

Feature Generator Based on CNN

Feature extraction is a very critical step in the research of EEG emotion recognition. Features based on EEG emotion

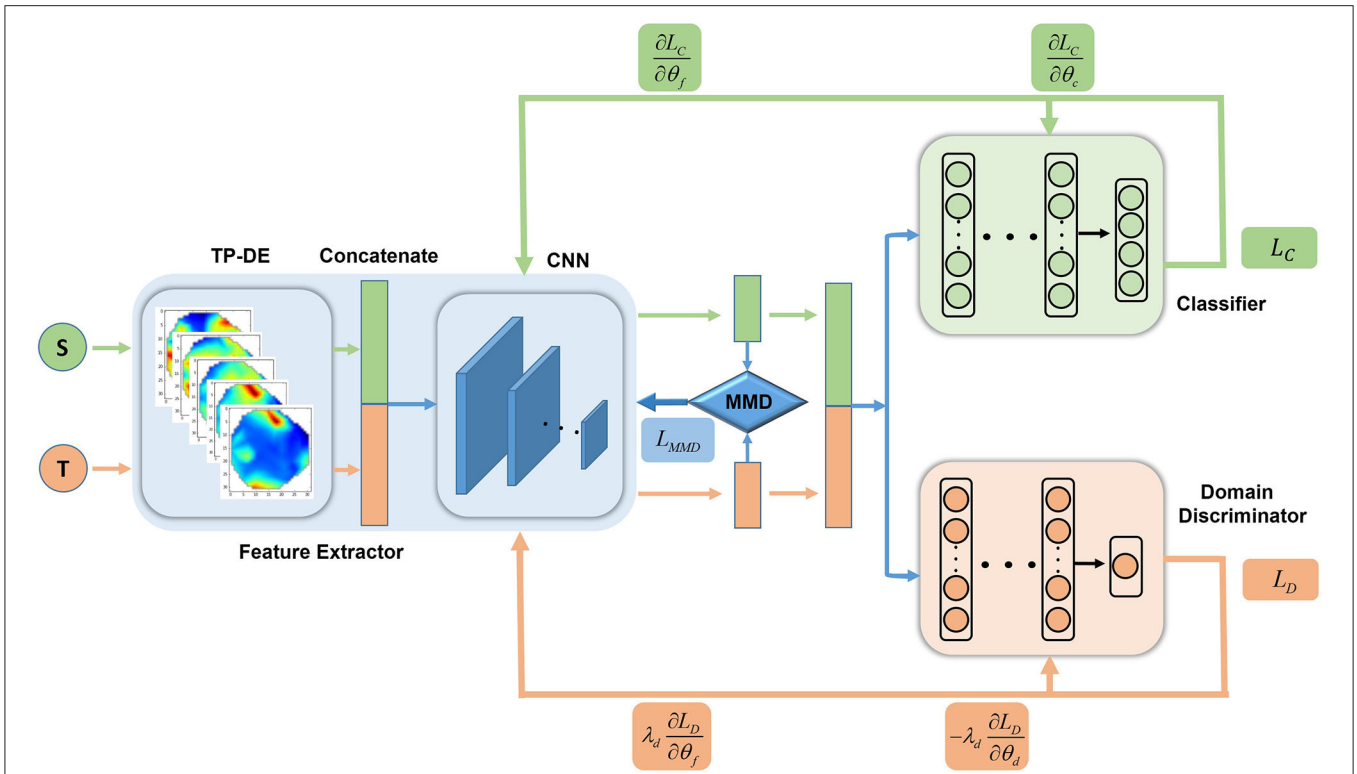

FIGURE 2 | Flowchart of the two-level depth domain adaptation framework, with feature extractors, classifiers, and domain discriminators. The DE feature was converted into a topological map feature as the input of the feature extractor. After processing by the two-level domain adaptation network, the source and target domains were distributed similarly while ensuring classification performance. The first-level domain adaptation network was mainly composed of feature extractors and a traditional algorithm MMD; the second-level domain adaptation network was composed of domain adversarial networks with feature extractors and domain discriminators.

recognition are mainly divided into three categories: time-domain features, frequency-domain features, and time-frequency features (Jenke et al., 2014). Time domain features include energy, average, standard deviation, first-order variance, standard first-order variance, second-order variance, and standard second-order variance. Hjorth (1970) proposed more complex temporal characteristics: Activity, Mobility, and Complexity. There is also the fractal dimension (FD) (Sourina and Liu, 2011), in addition to the high-order cross (HOC) (Petrantonakis and Hadjileontiadis, 2010) feature extraction method, which represents the oscillation mode of the signal and has high stability. The frequency domain features are mainly extracted on five frequency bands, Delta band (1–3 Hz), Theta band (4–7 Hz), Alpha band (8–13 Hz), Beta band (14–30 Hz), and Gamma band (31–50 Hz). Commonly used frequency domain features include energy and power spectral density (PSD) (Jenke et al., 2014). Moreover, time-frequency domain features include differential entropy (DE) (Duan et al., 2013), differential asymmetry (DASM) feature, rational asymmetry (RASM) feature, and differential causality (DCAU) feature (Zheng et al., 2019). Time-frequency domain features is usually extracted by short-time Fourier transform (STFT) (Koenig, 1946), Hilbert-Huang Spectrum (HHS) (Hadjidimitriou and Hadjileontiadis, 2012), discrete wavelet transform (DWT) (Mallat, 2009) and other time-frequency transformation methods. Murugappan et al. (2010) used DWT to extract the energy and entropy of five frequency bands of EEG signal, including root mean square (RMS), and recursive energy efficiency (REE). Alazrai et al. (2018) proposed a quadratic time-frequency distribution (QTFD) to extract time-frequency feature. Most of the current researches extract the DE features of five frequency bands for emotion recognition. Since the EEG signal is non-stationary, it can be approximated that the EEG signals follow the Gaussian distribution $N(\mu, \sigma^2)$, DE can be simply expressed by the following (Duan et al., 2013):

$$h(X) = -\int_{\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \ell^{-\frac{(x-\mu)^2}{2\sigma^2}} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \ell^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx$$

$$= \frac{1}{2} \log\left(2\pi\ell\sigma^2\right) \quad (1)$$

Where X submits the Gaussian distribution$N(\mu, \sigma^2)$, $\frac{1}{\sqrt{2\pi\sigma^2}} \ell^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is the probability density function of X, x is a variable, π and ℓ are constants.

The extracted DE features only consider the temporal information and ignore spatial information. Therefore, we

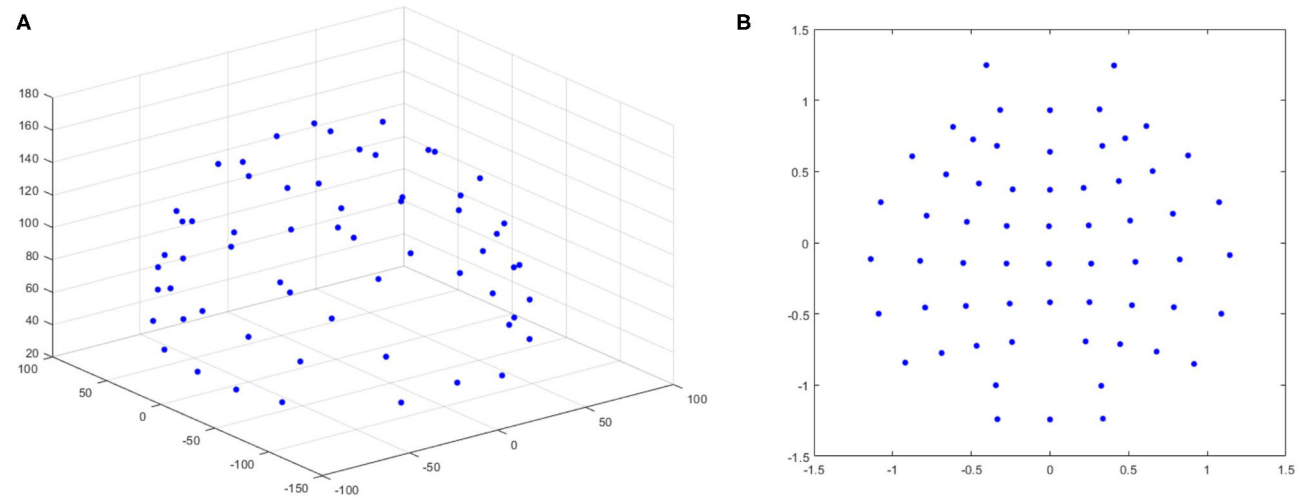

FIGURE 3 | (A) 3-dimensional position of the EEG electrode with 61 channels. **(B)** 2-dimensional position of the electrode using the polar coordinate projection method.

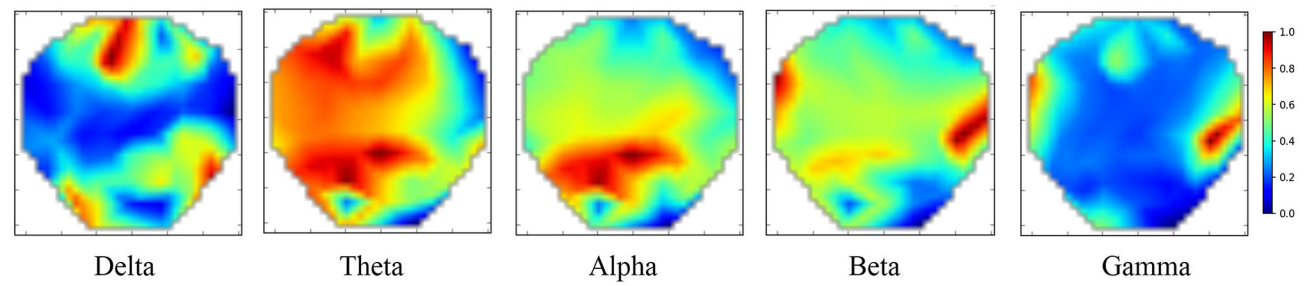

FIGURE 4 | TP-DE images of five frequency bands for a certain participant. Length and width are 32; channel is 5.

adopted a previously tested method using polar coordinate projection to maintain the spatial topology (Bashivan et al., 2015; Hwang et al., 2020). We projected the three-dimensional electrode position onto a two-dimensional plane, as shown in **Figure 3**. We used the Clough–Tocher scheme interpolation method to insert the differential entropy feature on each electrode and to estimate the value between the electrodes to obtain a 32 × 32 × 5 EEG image. **Figure 4** shows the topology-preserving DE (TP-DE) characteristics of five frequency bands of a certain subject after using maximum and minimum standardization.

In the deep CNN, we used a multi-layer convolutional layer and two maximum pooling layers. **Table 2** shows the CNN model structure for cross-day transfer research. We added an AdaBN layer after each set of convolutional layer and fully connected layer. The AdaBN standardized the distribution between the source and target domains in each batch of samples, so that the source and target domains were better matched in the subspace. Each fully connected layer used a dropout layer, with dropout rate = 0.5.

TABLE 2 | CNN model structure for cross-day transfer research.

Layer	Input dimension	Output dimension	Kernel size	Stride size
Conv1	5	6	3 × 3	1 × 1
Maxpool1	6	6	2 × 2	2 × 2
Conv2	6	64	3 × 3	1 × 1
Maxpool2	64	64	2 × 2	2 × 2
FC1	7 × 7 × 64	512		
FC2	512	256		

In the mean, bolding means the accuracy mean is the largest; In the Std, the bold is the smallest.

Two-Level Domain Adaptation Method

In order to understand the deep domain adversarial method more clearly, we first introduce the symbols that will be used here. We assume $X_S \sim x_i^S, i = 1 \cdots n_S$ is the data sample of the source domainD_s, $Y_S \sim y_i^S, i = 1 \cdots n_S$ is the label corresponding to the source domain data sample, and $X_T \sim x_i^T, i = 1 \cdots n_T$ is the data sample of the target domain D_T.

Feature generator $G_f$maps the source domain data $X_S$and the target domain data X_T to the same space:

$$X'_S = G_f(X_S), X'_T = G_f(X_T) \tag{2}$$

The generator G_f shares parameters in the source domains X_S and target domains X_T, so the feature dimensions of X'_S and X'_T are the same.

The function of the domain discriminator G_d is to distinguish the source domain and the target domain. It takes X'_S and X'_T as the input, and outputs the prediction of domain, respectively Y_S^D and Y_T^D:

$$Y_S^D = G_d(X'_S), Y_T^D = G_d(X'_T) \tag{3}$$

The role of the classifier G_c is to classify EEG emotions. It takes X'_S and X'_T as inputs and outputs predictive labels, which Y_S are and Y_T:

$$Y_S = G_c(X'_S), Y_T = G_c(X'_D) \tag{4}$$

We parameterize the generator G_f, domain discriminator G_d, and classifier G_c; their parameters are θ_f, θ_d and θ_c respectively.

First, we optimize the parameters and minimize the cross-entropy:

$$\min_{\theta_f,\theta c} L_C(X_S, X_T) = -\mathbb{E}_{(x_S,y_S)-(X_S,Y_S)}\left[\sum_{c=1}^{M} y_c^S \log G_c(G_f(x_S))\right] \tag{5}$$

Here, M represents the emotion class.

Then, introducing the domain adaptation algorithm, we propose a two-level domain adaptation algorithm based on a deep neural network. In the first-level domain adaptation, we use the MMD algorithm, combined with the AdaBN layer in the feature extractor, to align the class distribution of the source and target domains. Under the premise of ensuring the classification performance, the source and target domains are initially confused, and the MMD distance is minimized by optimizing the parameter θ_f:

$$\min_{\theta_f} L_{MMD}(X_S, X_T) = \mathcal{L}_{MMD}\mathbb{E}_{X_S,X_T}(X_S, X_T) \tag{6}$$

Where $\mathcal{L}_{MMD}$ represents the MMD distance. MMD distance can effectively measure the distance between distributions, and can be expressed by:

$$\mathcal{L}_{MMD}(X_S, X_T) = \frac{1}{n_S^2}\sum_{i,j=0}^{n_S} \kappa(X_S^{(i)}, X_S^{(j)}) - \frac{1}{n_S n_T}\sum_{i,j=0}^{n_s,n_T} \kappa(X_S^{(i)}, X_T^{(j)}) + \frac{1}{n_T^2}\sum_{i,j=0}^{n_T} \kappa(X_T^{(i)}, X_T^{(j)}) \tag{7}$$

Where n_S, n_T represent the number of samples in the source and target domains, respectively, and $\kappa(\cdot,\cdot)$ is a linear combination of multiple radial basis function (RBF) kernels, defined as:

$$\kappa(X_S^{(i)}, X_T^{(j)}) = \sum_n \eta_n \exp\left\{-\frac{1}{2\sigma_n}\left\|X_S^{(i)} - X_T^{(j)}\right\|^2\right\} \tag{8}$$

Where σ_n is the standard deviation of the n^{th} RBF kernel and η_n corresponds to its associated weight.

Using the MMD algorithm alone for domain adaptation is not sufficient for multi-source domain matching. Therefore, the second-level domain adaptation–domain adversarial method is introduced. We use the second-level domain adaptation network to reduce the distribution distance between the source and target domains. The principle of the domain discriminator is to maximize the cross entropy by optimizing the parameters θ_f and θ_d:

$$\max_{\theta_d}, \min_{\theta_f} L_D(X_S, X_T) = -\mathbb{E}_{(x_S,x_T)-(X_S,X_T)}\left[\sum_{d=1}^{N} y_d \log G_d(G_f(x_S, x_T))\right] \tag{9}$$

Where N is the numbers of domains.

Finally, we add gradient penalty to the domain loss to realize the Lipschitz constraint, so that the domain loss function can be more stable and converge faster in training. We also add an extra $L2$ norm regular term:

$$\min_{\theta_f,\theta_c,\theta_d} \mathcal{L}_G = L_C + \lambda_d L_D + \lambda_m L_{MMD} + \lambda_z \|W\|_2 \tag{10}$$

$$\max_{\theta_d} \mathcal{L}_D = -L_D + \lambda_L(\left\|\nabla_x G_d(x)\right\|_2 - 1)^2 \tag{11}$$

Where λ_d, λ_m, λ_z, and λ_L are hyper-parameters, and is the transformation matrix.

RESULTS

Cross-Day Transfer Research

We used a self-built data set for cross-day transfer research. In this data set, each participant had 6 days of data and each participant iterated six times. We used the leave-one-out method for cross-validation, that is, for each subject, 1 day was randomly selected as the test set, and the remaining days as the training set. In the deep network, 15% of the data was randomly selected from the training set every day as the validation set. In the parameter settings of the network model, the batch size was 160, the source and target domains were each 80, and the number of neurons in the fully connected layer was 512 and 256, respectively. The hyperparameters were λ_d (0.1), λ_m (0.1), λ_z (0.01), and λ_L (10). An Adam optimizer was used, and the learning rate was 0.0005. All the methods in this paper were implemented in Python, and the deep neural network was implemented in Tensorflow. The workstation operating system was Windows 7, using Inter(R)Xeon(R) E3-1230v3 CPU, NVIDIA TITAN V GPU, and 16G of RAM.

We studied the characteristics of the CNN learning EEG topological map. We extracted the output of the EEG topological map through the last layer of the convolutional network, and after superimposing and averaging the samples of the source and target domains, we selected nine channels with clear features and drawn feature maps after using maximum and minimum standardization, as shown in **Figure 5**. The first two

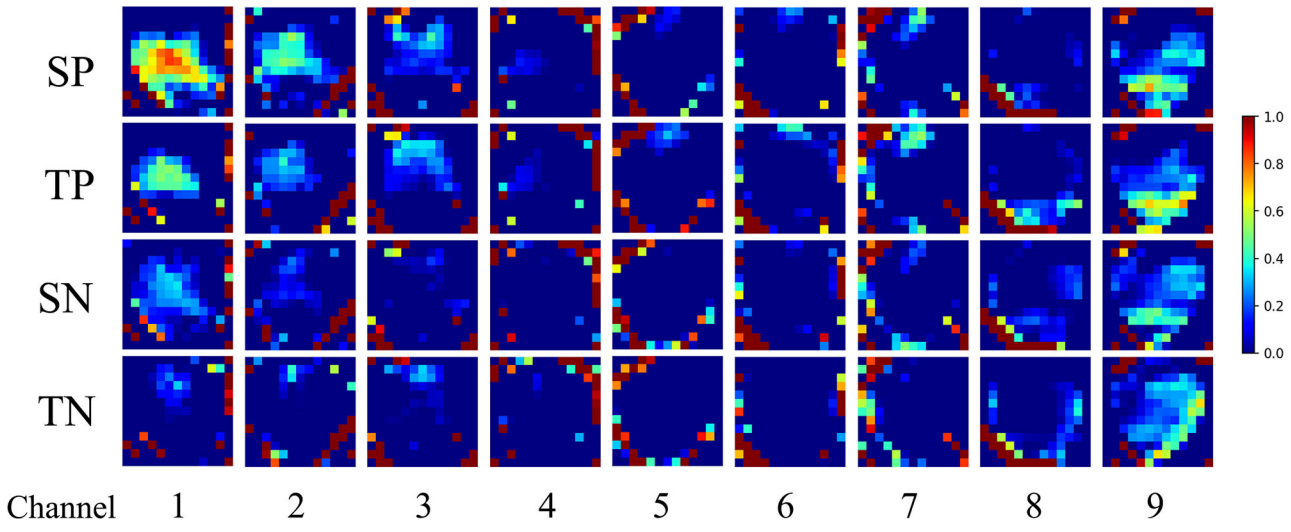

FIGURE 5 | Feature visualization based on cross-day transfer model learning in the second convolutional layer. (SP, source positive; TP, target positive; SN, source negative; TN, target negative).

rows represent the positive characteristics of the source domain and target domain learned by the convolutional network, and the last two rows represent the negative characteristics of the source domain and target domain learned by the network. From channels 1, 2, 3, and 4, we can see that there are differences between positive and negative emotions in the central area of the graph; in channels 5, 6, and 7, there are differences at the top of the graph; in channels 8, There are differences on both sides of the graph; channels 9 are differences at the bottom of the graph. There were obvious differences between positive and negative emotions in the parietal, frontal, and temporal lobes. This result was consistent with that of Zhuang et al. (2018). In addition, the positive and negative emotions of the source and target domains were similar, which proved that the network proposed in this paper can effectively solve the problem of cross-day transfer.

Next, we used the traditional support vector machine (SVM) classification method as the baseline, the RBF kernel is used, and compared the superior traditional transfer method, transfer component analysis (TCA), and the depth domain adaptation network DANN. First, we verified the EEG data set we collected using the leave-one-out method, and the results are shown in **Table 3**. In the self-built database, due to the difference in the data distribution of the training set and the test set, the baseline SVM classification performance was poor. In the second classification, for Joy-Sadness, Joy-Anger, and Joy-Fear, the accuracy rates were 70.02%, 71.16%, and 69.01%, and the accuracy rate for the four categories was 40.29%.Compared with the SVM method, the classification accuracy was slightly improved with the traditional TCA transfer method, but the improvement was not obvious. Using the DANN, the classification accuracy was significantly improved. The accuracy of the two classifications was 80.84%, 81.27%, and 80.20%, and the accuracy of the four classifications was 49.67%. Compared with the baseline SVM classifier, the accuracy of the classification was improved by 10%, 10%, 11%, and 9%. This showed that deep neural networks can effectively learn more transferable features for domain adaptation. The accuracy of the method proposed in this paper reached 84.0%, 87.04%, and 85.32% in the second classification. The accuracy of the four classifications reached 56.88%. Compared with the DANN network, it increased by 4%, 6%, 5%, and 7% respectively.

Moreover, we used SEED data set for cross-day transfer research. The SEED data set was proposed by Zheng and Lu (2017). They used scores (1–5) and keywords to evaluate subjects' emotions (positive, neutral, and negative) when watching video clips. There were 15 movie clips (5 positive, 5 neutral, and 5

TABLE 3 | Performance of adaptive methods in different domains for self- built EEG data set (cross-day).

Methods	Two classification						Four classification	
	Joy-Sadness		Joy-Anger		Joy-Fear			
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
SVM	0.7002	0.159	0.7160	0.162	0.6901	0.137	0.4029	0.102
TCA	0.7429	0.172	0.7343	0.149	0.7256	0.131	0.4373	0.108
DANN	0.8084	**0.123**	0.8127	0.128	0.8020	**0.117**	0.4967	**0.083**
MMD	0.7997	0.153	0.8094	0.136	0.8038	0.118	0.4298	0.105
TDANN	**0.8400**	0.149	**0.8704**	**0.119**	**0.8532**	0.120	**0.5688**	0.097

TABLE 4 | Performance of SEED adaptation methods in different domains for the public data set (cross-day).

	SVM	TCA	DANN	MMD	TDANN
Mean	0.5884	0.6827	0.6972	0.6817	**0.7493**
Std.	0.1142	0.1670	**0.0900**	0.1350	0.0927

negative) and each movie clip lasted about 4 minutes. Fifteen healthy subjects (8 females, 7 males, MEAN: 23.27, SD: 2.37) were selected and scanned using the ESI NeuroScan System. The distribution of 62 electrodes conformed to the international 10–20 standard and the sampling rate was 1000 Hz. The EEG signal was down-sampled to 200 Hz, the signals that were heavily polluted by EOG and EMG were screened, and the screened signals were then passed through a 0.3–50 Hz bandpass filter. Then the EEG signal was divided into 1s-long data segments without overlap. Thus, there were 3,394 samples for each subject, and the sample sizes of the three emotions were basically the same. Each subject had three experiments. We used the leave-one-out method for cross-validation. The results are shown in the **Table 4**. Compared with SVM, TCA, DANN, and MMD, the accuracy of TDANN is improved by 16, 6, 5, and 6% respectively.

In order to show the transfer process of feature distribution, we selected one subject's EEG data in our self-built data set to visualize by t-SNE (Donahue et al., 2013) in different domain adaptation algorithms in the leave-one-out method verification (see **Figure 6**). **Figure 6A** shows the original distribution of the source and target domains of the subject. It can be seen that the distribution of EEG features in the source and target domains was different, which was confusing and resulted in a very poor classification effect using the SVM classifier directly. **Figure 6B** shows the feature distribution map after feature mapping by the TCA method. It can be seen that mapping the feature to the feature subspace effectively distinguished the source domain from the target domain, but for multi-source domains transfer it was not enough; the feature distribution of the source domain was still very scattered. **Figure 6C** shows the feature distribution map learned by the DANN network. Still, some of the features of the source and target domains were confused, and the features of the source and target domains were relatively scattered and not clustered together. **Figure 6D** shows the distribution of features learned by the MMD. It can reduce the intra class distance, but can't widen the class spacing. **Figure 6E** shows the distribution of features learned by our method. It is evident that the features learned by our method are easier to distinguish than those learned by the DANN. Moreover, the class spacing became larger and the class inner distance became smaller.

Cross-Subject Transfer Research

Currently, the most used data set for cross-subject transfer research is SEED, so we first chose to use SEED for this as well. When using the SEED data set to verify the cross-subject transfer research, we also used the leave-one-out method for cross-validation, that is, one subject was randomly selected as the test set, and the rest were the training set, so 15 iterations were required. Compared with the cross-day transfer study, the tasks were different, and the selected data and sample sizes were also different. The number of samples in the cross-day transfer study was small, while the number in the cross-subject transfer study was large. Therefore, the CNN in the cross-subject transfer study had a deeper network structure than in the cross-day transfer study. The CNN structure is shown in **Table 5**. Similarly, we added an AdaBN layer after each convolutional layer and fully connected layer. The AdaBN standardized the distribution between the source and target domains in each batch of samples, making the source and target domains better in the subspace matched by one (Donahue et al., 2013). In addition, each fully connected layer used a dropout layer, with a dropout rate of 0.5.

We then conducted cross-subject transfer research on the SEED data set. When using the SEED data set to verify the cross-subject transfer research, we also used the leave-one-out method

TABLE 5 | CNN model structure for cross-subject transfer research.

Layer	Input dimension	Output dimension	Kernel size	Stride size
Conv1	5	32	3 × 3	1 × 1
Conv2	32	32	3 × 3	1 × 1
Maxpool1	32	32	2 × 2	2 × 2
Conv3	32	64	3 × 3	1 × 1
Conv4	64	64	3 × 3	1 × 1
Conv5	64	128	3 × 3	1 × 1
Conv6	128	128	3 × 3	1 × 1
Maxpool2	128	128	2 × 2	2 × 2
FC1	6 × 6 × 128	1,024		
FC2	1,024	512		
FC3	512	256		

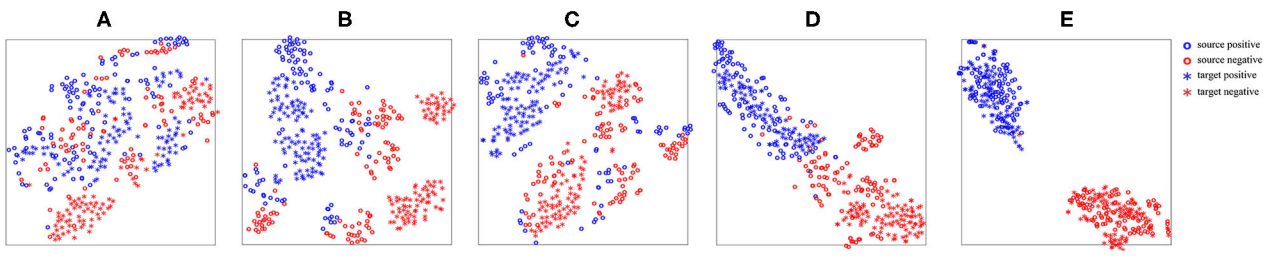

FIGURE 6 | Feature visualization diagram. **(A)** original distribution of the features of the source and target domains; **(B)** distribution of the features after being mapped by the TCA algorithm; **(C)** distribution of the features learned by the DANN algorithm; **(D)** distribution of the features learned by the MMD algorithm; **(E)** feature distribution of TDANN learning.

TABLE 6 | Performance of SEED adaptation methods in different domains for the public data set (cross-subject).

	SVM	TCA	TPT	DANN	MMD	DAN	DResNet	WGAN-DA	TDANN
Mean	0.5818	0.6400	0.7517	0.7919	0.6655	0.8381	0.8530	0.8707	**0.8790**
Std.	0.1385	0.1466	0.1283	0.1314	**0.0483**	0.0856	0.0832	0.0714	0.0613

In the mean, bolding means the accuracy mean is the largest; In the Std, the bold is the smallest.

for cross-validation, that is, we randomly selected one subject as the test set, and the rest as the training set. Therefore, 15 iterations were required. The batch size was 224, the source domain and target domain were each 112, and the number of neurons in the fully connected layer was 1,024, 512, and 256, respectively. The hyperparameters were λ_d (0.1), λ_m (0.1), λ_z (0.01), and λ_L (0.1). An Adam optimizer was used, and the learning rate was 0.0005.

We simultaneously compared the current best-performing algorithms in the cross-subject transfer of EEG emotions, including shallow algorithms such as TCA and TPT, and deep algorithms such as DANN, DResNet, and WGAN-DA. We continued to use the SVM classifier as the baseline. **Table 6** shows the average and variance obtained with different algorithms. Among the shallow transfer algorithms, TPT had the best effect, with an accuracy rate of 75.17%. Among the deep transfer algorithms, WGAN-DA had the best classification performance, with an accuracy rate of 87.07%. Although the accuracy of DResNet was not as high as that of WGAN-DA, DResNet did not use any information about the target domain data. TDANN's recognition accuracy rate was 87.9%, the highest recognition rate achieved by any of the algorithms, and it was more stable than WGAN-DA.

Then, we used a self-built data set for cross-subject transfer research. Twelve subjects' EEG data collected for the first time were used in this cross-subject transfer experiment. We used the leave-one-out method for cross-validation, and compared with TCA, DANN, and MMD algorithms. The results are shown in the **Table 7**. The accuracy of the method TDANN reached 83.79, 84.13, and 81.72% in the second classification. The accuracy of the four classifications reached 47.28%. Compared with the MMD, it increased by 5, 5, 6, and 4%, respectively. However, in the cross-subject transfer experiment of self-built data set, the overall accuracy is lower than that of cross day transfer experiment. The reason for this may be that there exists intrinsic differences among subjects, and more data collected from different subjects are needed to remove this intrinsic differences among subjects.

TABLE 7 | Performance of adaptive methods in different domains for self-built EEG data set (cross-subject).

Methods	Two classification						Four classification	
	Joy-Sadness		Joy-Anger		Joy-Fear			
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
SVM	0.6726	0.147	0.6995	0.1474	0.6565	0.120	0.3411	0.089
TCA	0.7505	**0.040**	0.7544	0.049	0.7327	0.459	0.4202	**0.025**
DANN	0.7299	0.046	0.7168	**0.023**	0.6624	**0.025**	0.4120	0.043
MMD	0.7837	0.151	0.7993	0.154	0.7568	0.146	0.4341	0.100
TDANN	**0.8379**	0.155	**0.8413**	0.137	**0.8172**	0.130	**0.4728**	0.079

In the mean, bolding means the accuracy mean is the largest; In the Std, the bold is the smallest.

CONCLUSIONS

Emotion recognition is the most important part of human-computer interaction. EEG emotion recognition research has been developed for decades, and many impressive results have been obtained. However, there are still quite a few problems, among which the most important are cross-day transfer and cross-subject transfer. Because EEG signals are non-stationary, the signal distribution of each subject is different. Even for the same subject, there are differences in the EEG signals collected at different times.

In this paper, we propose a domain adaptation framework using deep neural networks for EEG emotion recognition. We have verified the performance of the framework on two data sets: our self-built data set, and the public data set SEED. In the cross-day transfer evaluation, we compared the currently favored transfer algorithms TCA and DANN. In the self-built data set, the accuracy rates of Joy-Sadness, Joy-Anger, and Joy-Fear were 84.0, 87.04, and 85.32%, respectively, and the accuracy rate of the four categories was 56.88%. In the SEED data set, the accuracy of three classification reached 74.93%. For the cross-subject transfer evaluation, the algorithm we proposed achieved an average accuracy rate of 87.9% in SEED data set. In the self-built data set, the accuracy rates of Joy-Sadness, Joy-Anger, and Joy-Fear were 83.79, 84.13, and 81.72%, respectively, and the accuracy rate of the four categories was 47.28%. Visualizing the features learned by the feature extractor, it can be clearly seen that different brain regions are activated by different emotions. The energy of positive emotions in the parietal, and frontal lobes is significantly higher than that of negative emotions.

In our cross-day transfer research, although we established a data set with the largest amount of data available at present for deep neural network training, the amount of data is still far from enough. The labor and funds required to build a sufficiently large data set are beyond the scope of most research institutions. Some studies have found that sample generation through a generative adversarial network (GAN) can effectively increase sample size and improve the training performance of a neural network to a certain extent. In follow-up research, we will study data enhancement based on a GAN to further address the problem of EEG emotion transfer.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

GB is mainly responsible for research design, data analysis, and manuscript writing of this study. NZ is mainly responsible for data collection and data analysis. LT is mainly responsible for research design and data analysis. JS is mainly responsible for data collection and production of charts. LW is mainly responsible for data analysis and document retrieval. BY is mainly responsible for research design and manuscript writing. YZ is mainly responsible for data collection and manuscript writing. ZS is mainly responsible for data collection. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by the National Key Research and Development Plan of China under Grant 2017YFB1002502, in part by the National Natural Science Foundation of China under Grant 61701089, and in part by the Natural Science Foundation of Henan Province of China under Grant 162300410333.

ACKNOWLEDGMENTS

The authors would like to thank all the subjects who participated in the experiment.

REFERENCES

Acharya, U. R., Sudarshan, V. K., Adeli, H., Santhosh, J., Koh, J. E. W., Puthankatti, S. D., et al. (2015). A novel depression diagnosis index using nonlinear features in EEG signals. *Eur. Neurol.* 74, 79–83. doi: 10.1159/000438457

Alarcao, S. M., and Fonseca, M. J. (1949). Emotions recognition using EEG signals: a survey. *IEEE Trans. Affect. Comput.* 1. doi: 10.1109/TAFFC.2017.2714671

Alazrai, R., Homoud, R., Alwanni, H., and Daoud, M. (2018). EEG-based emotion recognition using quadratic time-frequency distribution. *Sensors* 18, 2739. doi: 10.3390/s18082739

Bashivan, P., Rish, I., Yeasin, M., and Codella, N. (2015). Learning representations from EEG with deep recurrent-convolutional neural networks. *Comput. Sci. arXiv:1511.06448.* doi: 10.1109/CVPR.2016.522

Chai, X., Wang, Q., Zhao, Y., Liu, X., Bai, O., and Li, Y. (2016). Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition. *Comput. Biol. Med.* 79, 205–214. doi: 10.1016/j.compbiomed.2016.10.019

Chueh, T. H., Chen, T. B., L.u, H. H.-S., J.u, S.-S., Tao, T. H., and Shaw, J. H. (2012). Statistical Prediction of emotional states by physiological signals with manova and machine learning. *Int. J. Pattern Recogn. Artif. Intell.* 26, 1250008-1–1250008-18. doi: 10.1142/S0218001412500085

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., and Darrell, T. (2013). "DeCAF: a deep convolutional activation feature for generic visual recognition," in *International Conference on Machine Learning* (Tianjin).

Duan, R. N., Zhu, J. Y., and Lu, B. L. (2013). "Differential entropy feature for EEG-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)* (San Diego, CA).

Fydrich, T., Dowdall, D., and Chambless, D. L. (1992). Reliability and validity of the Beck Anxiety Inventory. *J. Anxiety Disord.* 6, 55–61. doi: 10.1016/0887-6185(92)90026-4

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *arXiv:1505.07818.* Available online at: http://arxiv.org/abs/1505.07818 (accessed August 29, 2020).

Gur, R. C., Erwin, R. J., Gur, R. E., Zwil, A. S., and Kraemer, H. C. (1992). Facial emotion discrimination: II. Behavioral findings in depression. *Psychiatry Res.* 42, 241–251. doi: 10.1016/0165-1781(92)90116-K

Hadjidimitriou, S. K., and Hadjileontiadis, L. J. (2012). Toward an EEG-based recognition of music liking using time-frequency analysis. *IEEE Trans. Biomed. Eng.* 59, 3498–3510. doi: 10.1109/TBME.2012.2217495

Hamilton, M. (2004). *Hamilton Depression Scale*. Group.

Hang, W., Feng, W., Du, R., Liang, S., Chen, Y., Wang, Q., et al. (2019). Cross-subject EEG signal recognition using deep domain adaptation network. *IEEE Access.* 7, 128273–128282. doi: 10.1109/ACCESS.2019.2939288

Hjorth, B. (1970). EEG analysis based on time domain properties. *Electroencephalogr. Clin. Neurophysiol.* 29, 306–310. doi: 10.1016/0013-4694(70)90143-4

Hwang, S., Hong, K., Son, G., and Byun, H. (2020). Learning CNN features from DE features for EEG-based emotion recognition. *Pattern Anal. Applic.* 23, 1323–1335. doi: 10.1007/s10044-019-00860-w

Hyvärinen, A. (1999). The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Process. Lett.* 10, 1–5. doi: 10.1023/A:1018647011077

Jenke, R., Peer, A., and Buss, M. (2014). Feature extraction and selection for emotion recognition from EEG. *IEEE Trans. Affective Comput.* 5, 327–339. doi: 10.1109/TAFFC.2014.2339834

Jimenez-Guarneros, M., and Gomez-Gil, P. (2020). Custom domain adaptation: a new method for cross-subject, EEG-based cognitive load recognition. *IEEE Signal Process. Lett.* 27, 750–754. doi: 10.1109/LSP.2020.2989663

Koenig, W. (1946). The sound spectrograph. *J. Acoust. Soc. Am.* 18:19. doi: 10.1121/1.1902419

Li, J., Qiu, S., Du, C., Wang, Y., and He, H. (2020). Domain adaptation for EEG emotion recognition based on latent representation similarity. *IEEE Trans. Cogn. Dev. Syst.* 12, 344–353. doi: 10.1109/TCDS.2019.2949306

Li, J., Qiu, S., Shen, Y.-Y., Liu, C.-L., and He, H. (2019). Multisource transfer learning for cross-subject EEG emotion recognition. *IEEE Trans. Cybern.* 99, 1–13. doi: 10.1109/TCYB.2019.2904052

Li, Y., Wang, N., and Shi, J., Hou, H., Liu, J. (2018). Adaptive Batch Normalization for practical domain adaptation. *Pattern Recogn.* 80, 109–117. doi: 10.1016/j.patcog.2018.03.005

Li, Y., Zheng, W., Zong, Y., Cui, Z., Zhang, T., and Zhou, X. (2019). A bi-hemisphere domain adversarial neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comput.*, 2018, 1. doi: 10.1109/TAFFC.2018.2885474

Luo, Y., Zhang, S.-Y., Zheng, W.-L., and Lu, B.-L. (2018). "WGAN domain adaptation for EEG-based emotion recognition," in *Neural Information Processing Lecture Notes in Computer Science*, eds L. Cheng, A. C. S. Leung, and S. Ozawa (Cham: Springer International Publishing), 275–286.

Ma, B.-Q., Li, H., Zheng, W.-L., and Lu, B.-L. (2019). "Reducing the subject variability of EEG signals with adversarial domain generalization," in *Neural Information Processing*, eds T. Gedeon, K. W. Wong, and M. Lee (Cham: Springer International Publishing), 30–42.

Mallat, S. (2009). *A Wavelet Tour of Signal Processing* China Machine Press.

Mehmood, R. M., and Lee, H. J. (2015). "Emotion classification of EEG brain signal using SVM and KNN," in *2015 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. (Turin).

Murugappan, M., Ramachandran, N., and Sazali, Y. (2010). Classification of human emotion from EEG using discrete wavelet transform. *J. Biomed. Sci. Eng.* 03, 390–396. doi: 10.4236/jbise.2010.34054

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2011). Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* 22, 199–210. doi: 10.1109/TNN.2010.2091281

Petrantonakis, P. C., and Hadjileontiadis, L. J. (2010). Emotion recognition from EEG using higher order crossings. *IEEE Trans. Inf. Technol. Biomed.* 14, 186. doi: 10.1109/TITB.2009.2034649

Sangineto, E., Zen, G., Ricci, E., and Sebe, N. (2014). "We are not all equal: personalizing models for facial expression analysis with transductive parameter transfer," in *Proceedings of the ACM International Conference on Multimedia - MM '14* (Orlando, FL: ACM Press), 357–366.

Saxen, F., Werner, P., and Al-Hamadi, A. (2017). "Real vs. fake emotion challenge: learning to rank authenticity from facial activity descriptors," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* (Venice: IEEE), 3073–3078.

Shear, M. K., Bilt, J. V., Rucci, P., Endicott, J., and Frank, D. M. (2010). Reliability and validity of a structured interview guide for the Hamilton Anxiety Rating Scale (SIGH-A). *Depress. Anxiety* 13, 166–178. doi: 10.1002/da.1033

Sourina, O., and Liu, Y. (2011). "A fractal-based algorithm of emotion recognition from EEG using Arousal-Valence model," in *Biosignals - International Conference on Bio-inspired Systems and Signal Processing* (Rome).

Sourina, O., Liu, Y., and Nguyen, M. K. (2012). Real-time EEG-based emotion recognition for music therapy. *J. Multimodal User Interf.* 5, 27–35. doi: 10.1007/s12193-011-0080-6

Walter, S., Wendt, C., Boehnke, J., Crawcour, S., Tan, J. W., Chan, A., et al. (2014). Similarities and differences of emotions in human-machine and human-human interactions: what kind of emotions are relevant for future companion systems? *Ergonomics* 57, 374–386. doi: 10.1080/00140139.2013.822566

Xu, P., Huang, Y., and, Luo, Y. (2010). Establishment and assessment of native Chinese affective video system. *Chin. Mental Health J. 24*, 551–561.

Yao, D. (2001). A method to standardize a reference of scalp EEG recordings to a point at infinity. *Physiol. Meas.* 22, 693–711. doi: 10.1088/0967-3334/22/4/305

Yao, D., Qin, Y., Hu, S., Dong, L, Bringas Vega, M. L., Valdés Sosa, P. A. (2019). Which reference should we use for EEG and ERP practice? *Brain Topogr.* 32, 530–549. doi: 10.1007/s10548-019-00707-x

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). "How transferable are features in deep neural networks?," in *International Conference on Neural Information Processing Systems*.

Zheng, W.-L., and Lu, B.-L. (2016). "Personalizing EEG-Based Affective Models with Transfer Learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)* (New York, NY).

Zheng, W.-L., Zhu, J.-Y., and Lu, B.-L. (2019). Identifying stable patterns over time for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* 10, 417–429. doi: 10.1109/TAFFC.2017.2712143

Zheng, W. L., and Lu, B. L. (2017). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Autonomous Mental Dev.* 7, 162–175. doi: 10.1109/TAMD.2015.2431497

Zhuang, N., Zeng, Y., Yang, K., Zhang, C., Tong, L., and Yan, B. (2018). Investigating patterns for self-induced emotion recognition from EEG signals. *Sensors* 18:841. doi: 10.3390/s18030841

Zong, Y., Zheng, W., Huang, X., Yan, K., Yan, J., and Zhang, T. (2016). Emotion recognition in the wild via sparse transductive transfer linear discriminant analysis. *J. Multimodal User Interf.* 10, 163–172. doi: 10.1007/s12193-015-0210-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 11 March 2021
doi: 10.3389/fnhum.2021.645952

Data Augmentation: Using Channel-Level Recombination to Improve Classification Performance for Motor Imagery EEG

Yu Pei[1,2], *Zhiguo Luo*[2,3], *Ye Yan*[2,3], *Huijiong Yan*[2,3], *Jing Jiang*[4], *Weiguo Li*[1], *Liang Xie*[2,3]* *and Erwei Yin*[2,3]

[1] School of Software, Beihang University, Beijing, China, [2] Tianjin Artificial Intelligence Innovation Center (TAIIC), Tianjin, China, [3] Unmanned Systems Research Center, National Innovation Institute of Defense Technology, Academy of Military Sciences China, Beijing, China, [4] National Key Laboratory of Human Factors Engineering, China Astronaut Research and Training Center, Beijing, China

The quality and quantity of training data are crucial to the performance of a deep-learning-based brain-computer interface (BCI) system. However, it is not practical to record EEG data over several long calibration sessions. A promising time- and cost-efficient solution is artificial data generation or data augmentation (DA). Here, we proposed a DA method for the motor imagery (MI) EEG signal called brain-area-recombination (BAR). For the BAR, each sample was first separated into two ones (named half-sample) by left/right brain channels, and the artificial samples were generated by recombining the half-samples. We then designed two schemas (intra- and adaptive-subject schema) corresponding to the single- and multi-subject scenarios. Extensive experiments using the classifier of EEGnet were conducted on two public datasets under various training set sizes. In both schemas, the BAR method can make the EEGnet have a better performance of classification ($p < 0.01$). To make a comparative investigation, we selected two common DA methods (noise-added and flipping), and the BAR method beat them ($p < 0.05$). Further, using the proposed BAR for augmentation, EEGnet achieved up to 8.3% improvement than a typical decoding algorithm CSP-SVM ($p < 0.01$), note that both the models were trained on the augmented dataset. This study shows that BAR usage can significantly improve the classification ability of deep learning to MI-EEG signals. To a certain extent, it may promote the development of deep learning technology in the field of BCI.

Keywords: brain-computer interface, electroencephalogram, motor imagery, deep learning, inter-subject transfer learning, pre-training, data augmentation

OPEN ACCESS

Edited by:
Minkyu Ahn,
Handong Global University,
South Korea

Reviewed by:
Minpeng Xu,
Tianjin University, China
Jing Jin,
East China University of Science and Technology, China

***Correspondence:**
Liang Xie
xielnudt@gmail.com

Specialty section:
This article was submitted to Brain-Computer Interfaces, a section of the journal Frontiers in Human Neuroscience

Received: *24 December 2020*
Accepted: *17 February 2021*
Published: *11 March 2021*

Citation:
Pei Y, Luo Z, Yan Y, Yan H, Jiang J, Li W, Xie L and Yin E (2021) Data Augmentation: Using Channel-Level Recombination to Improve Classification Performance for Motor Imagery EEG. Front. Hum. Neurosci. 15:645952. doi: 10.3389/fnhum.2021.645952

1. INTRODUCTION

The brain-computer interface (BCI) is a communication control system directly established between the brain and external devices (computers or other electronic devices), using signals generated during brain activity (Wolpaw et al., 2000). Instead of relying on the muscles and organs, the system directly builds communication between the brain and the machine. Electroencephalogram (EEG) is one of the most common signals used for building a BCI

system because of its cost-effectiveness, non-invasive implementation, and portability. BCIs have shown potentials in applying various fields such as communication, control, and rehabilitation (Abdulkader et al., 2015).

Recent years have witnessed intense researches into different types of BCI systems. According to the signal acquisition method, BCI technology can be divided into three types: non-implantable system, semi-implantable system, an implantable system (Wolpaw et al., 2000). Non-implantable BCI systems mainly use EEG to recognize human's intention. According to the signal generation mechanism, BCI systems can be divided into induced BCI systems and spontaneous BCI systems. The induced BCI systems are: steady-state visual evoked potentials (SSVEP) (Friman et al., 2007; Ko et al., 2020), slow cortical potentials (Beuchat et al., 2013), and the P300 (Yin et al., 2016; Yu et al., 2017; Chikara and Ko, 2019), and the spontaneous BCI systems are: motor imagery (MI) (Choi and Cichocki, 2008; Belkacem et al., 2018; Chen et al., 2019; Wang et al., 2020).

The motor imagery (MI) BCI system's framework is based on the fact that the brain's activity in a specific area will be changed when the patients (or subjects) imagine moving any part of their bodies. For example, when a person imagines moving his/her right arm, there is a desynchronization of neural activity in the primary motor cortex on the left side of the brain. This desynchronization is called event-related desynchronization (ERD), which can be observed in the EEG signals transitioning from resting-state energy level to a lower energy level. The spatial location, temporal onset, amount of decrease, and ERD's stability are all subject-dependent factors (McFarland et al., 2000; Lotze and Halsband, 2006), bringing challenges for detecting changes in MI's neural activity.

In recent years, based on the considerable amount of data and sophisticated model structure, deep learning has been proved its strong learning ability to classify linguistic features, images, and sounds (Zhong et al., 2015; Song et al., 2017; Alom et al., 2018; Cooney et al., 2019). However, it is difficult to collect sufficient data in practice due to the limited available subjects, experimental time, and operation complexity in BCI. This problem is pronounced in MI-based BCI. The performance of deep neural networks (DNNs) is susceptible to the number of samples. A small scale dataset often leads to poor generalization during model training, reducing the decoding accuracy (LeCun et al., 2015).

A feasible approach to improve deep networks' performance and to avoid the overfitting caused by lack of training data is data augmentation (DA) methods (Salamon and Bello, 2017). These methods augment training data by artificially generating new samples based on existing data (Roy et al., 2019). Yin and Zhang (2017) added Gaussian white noise to the EEG feature vector to improve their deep learning model's accuracy on the classification task of Mental Workload (MW). Sakai et al. (2017) shifted EEG trials in time axis and amplified the amplitude to generate artificial EEG signals for augmentation. The results showed that their augmentation method improved the classification performance when the training set's size was 20, but this method has no significant effect on the more extensive training set. In another work, artificial EEG trials were generated by segmentation and recombination in time and frequency domains (Lotte, 2015), and the results were more convincing. Other studies have used more advanced techniques such as variational auto-encoders (VAE) (Aznan et al., 2019) and generative adversarial networks (GANs) (Goodfellow et al., 2014). However, tens of thousands of parameters in these methods need to be trained using the original data, which creates a certain degree of demand for the original data scale. It is a conflict with our goal of data augmentation on a tiny training set. Besides, the huge consumption of computing resources and the difficulty of being reproduced are also their shortcomings, although they have achieved a certain degree of success in some aspects (Karras et al., 2017; Kodali et al., 2017).

We proposed a new motor imagery EEG DA method, called Brain Area Recombination (BAR), which first decomposes the training dataset from the left and right brains and reassembles them into a new training dataset. Pre-training on the datasets of other subjects is also a common way to solve the insufficient training of deep neural networks (Fahimi et al., 2020). There are two types of pre-training, one is to use the source subjects' data in the same dataset as the pre-training training dataset, and the other is to use another dataset as the pre-training training dataset (Xu et al., 2020). The first type of pre-training is used in our study. Fortunately, experiments show that our method can be well-embedded with the pre-training framework to improve the deep learning network's classification performance.

Compared with the methods above, the proposed BAR has the following advantages:

(1) Low computational complexity;
(2) High and fixed expansion ratio;
(3) Great quality of new artificial samples.

This paper's remainder is organized as follows: section 2 introduced the public dataset used in the study. Section 3 proposed a preliminary experiment and our method's hypothesis in detail and then introduced the method's mathematical definition. We used two common schemas to evaluate our method and implemented the other two common DA methods as a comparison. Experiment results were showed in section 4, which demonstrated that the proposed BAR had achieved significant results. Section 5 presented the discussion. Section 6 concluded the study.

2. MATERIALS AND METHODS

2.1. Materials

Dataset 1: The first dataset was from BCI-Competition-III-IVa and was collected in a cue-based setting. Only cues for the classes "right" and "foot" are provided. This dataset was recorded from five healthy subjects (aa, al, av, aw, ay) at 100 Hz. The subjects sat in a comfortable chair with arms resting on armrests. The timeline of the dataset was shown in **Figure 1A**. The raw data were continuous signals of 118 EEG channels and markers that indicate the time points of 280 cues. Each sample was segmented from [0, 2.5] s by marks, then passed a band-pass filter (5-order Butterworth digital filter with cut-off frequencies at [8, 30] Hz) to remove muscle artifacts, line-noise contamination,

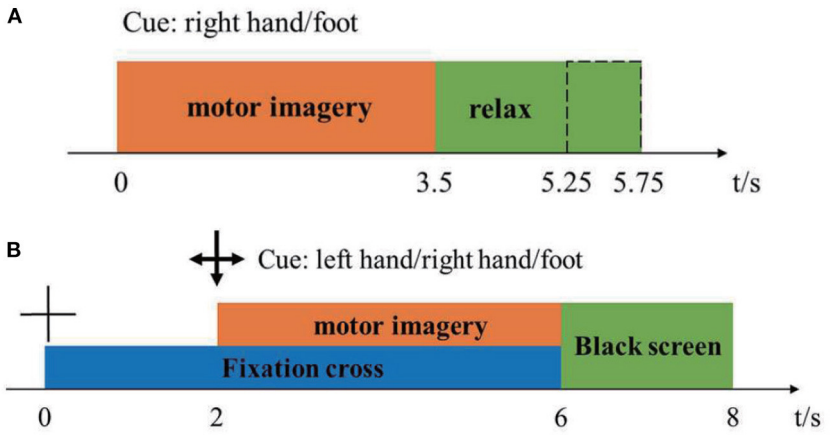

FIGURE 1 | Timeline of one trial in the dataset 1 **(A)** and dataset 2 **(B)**.

and DC drifts. Under the condition that the positive sample and the negative sample were balanced, 100 samples were randomly selected as the training pool. The remaining samples were used as the test samples. The details of the competition, including ethical approval, and the raw data can be download from http://www.bbci.de/competition/iii/.

Dataset 2: The second dataset was from BCI-Competition-IV-1. The dataset was recorded from seven healthy subjects (a, b, c, d, e, f, g), including four healthy individuals (named "a," "b," "f," "g") and three artificially generated "participants" (named "c," "d," "e"). 59-channel EEG signals were recorded at 100 Hz. Two motor imagery classes were selected for each subject from the three classes: left hand, right hand, and foot. The timeline of the dataset was shown in **Figure 1B**. There were two subjects (a, f) whose motor imaging tasks were different from the others, so they were eliminated. Here we only used the calibration data because of the complete marker information. Each sample was segmented from [0, 2.5] s by marks, then passed a band-pass filter (5-order Butterworth digital filter with cut-off frequencies at [8, 30] Hz) to remove muscle artifacts, line-noise contamination, and DC drifts. After preprocessing, we obtained 200 samples for each subject. We randomly selected 100 samples as a training pool and the rest as test samples, like dataset 1. The details of the competition, including ethical approval, and the raw data can be download from http://www.bbci.de/competition/iv/.

2.2. Methods

2.2.1. Core Assumption

Consider that we select two samples from the original samples randomly, and take out the left brain part of the first sample and the right brain part of the second sample, and recombine these two parts together to form an artificial sample. This artificial sample is still a normal EEG sample (1).

$$\textbf{If } x_i, x_j \sim P_{MI-EEG}, \textbf{ Then } \hat{x} = \begin{bmatrix} x_i^{(R)} \\ x_j^{(L)} \end{bmatrix} \sim P_{MI-EEG} \quad (1)$$

where $x_i, x_j \in \mathbb{R}^{C\times T}$, C is the number of electrodes, T is the sample-points, P_{MI-EEG} is the distribution of MI-EEG data. $x^{(R)}, x^{(L)} \in \mathbb{R}^{\frac{C}{2}\times T}$ represent samples containing only the right brain channels and the left brain channels, respectively.

2.2.2. Brain Area Recombination

Based on the assumption described in (1), we propose two similar DA methods for single-subject and multi-subject scenes for EEG of motor imagination. The whole framework of our proposed method was shown in **Figure 2**.

For the single-subject scene, the data augmentation method is:

$$E_{single}^{(s)} = \bigcup_{c=1}^{N_c} \{x^{(R)} | x^{(R)} \in D_s^{(R)}, y(x^{(R)}) = c\} \times \{x^{(L)} | x^{(L)} \in D_s^{(L)}, y(x^{(L)}) = c\} \quad (2)$$

For the multi-subject scene, the data augmentation method is:

$$E_{multi}^{(s)} = \bigcup_{i,j\neq s}^{N_s} \bigcup_{c=1}^{N_c} \{x^{(R)} | x^{(R)} \in D_i^{(R)}, y(x^{(R)}) = c\} \times \{x^{(L)} | x^{(L)} \in D_j^{(L)}, y(x^{(L)}) = c\} \quad (3)$$

where N_s and N_c represent the number of subjects and the number of classification tasks, respectively. $y(\cdot)$ is a mapping to get the label. s is the index of the subject. $D_i^{(R)}$ and $D_i^{(L)}$ represent the training dataset from the right brain and the left brain for the i-th subject, respectively. "$\times$" was defined as the cartesian-like product operator. For example, there are two matrix sets $A = \{a_1, a_2\}, B = \{b_1, b_2\}$ and their product $C = A \times B = \{\begin{bmatrix} a_1 \\ b_1 \end{bmatrix}, \begin{bmatrix} a_1 \\ b_2 \end{bmatrix}, \begin{bmatrix} a_2 \\ b_1 \end{bmatrix}, \begin{bmatrix} a_2 \\ b_2 \end{bmatrix}\}$. Note that $E_{multi}^{(s)}$ does not contain any sample from the s-th subject. In other words, it is a cross-subject training dataset. **Figure 2A** represents the meaning of Equation (2) and **Figure 2B** represents the meaning of Equation (3). The zero midline electrodes are Fpz, Fz, FCz, Cz, CPz, Pz, POz, and

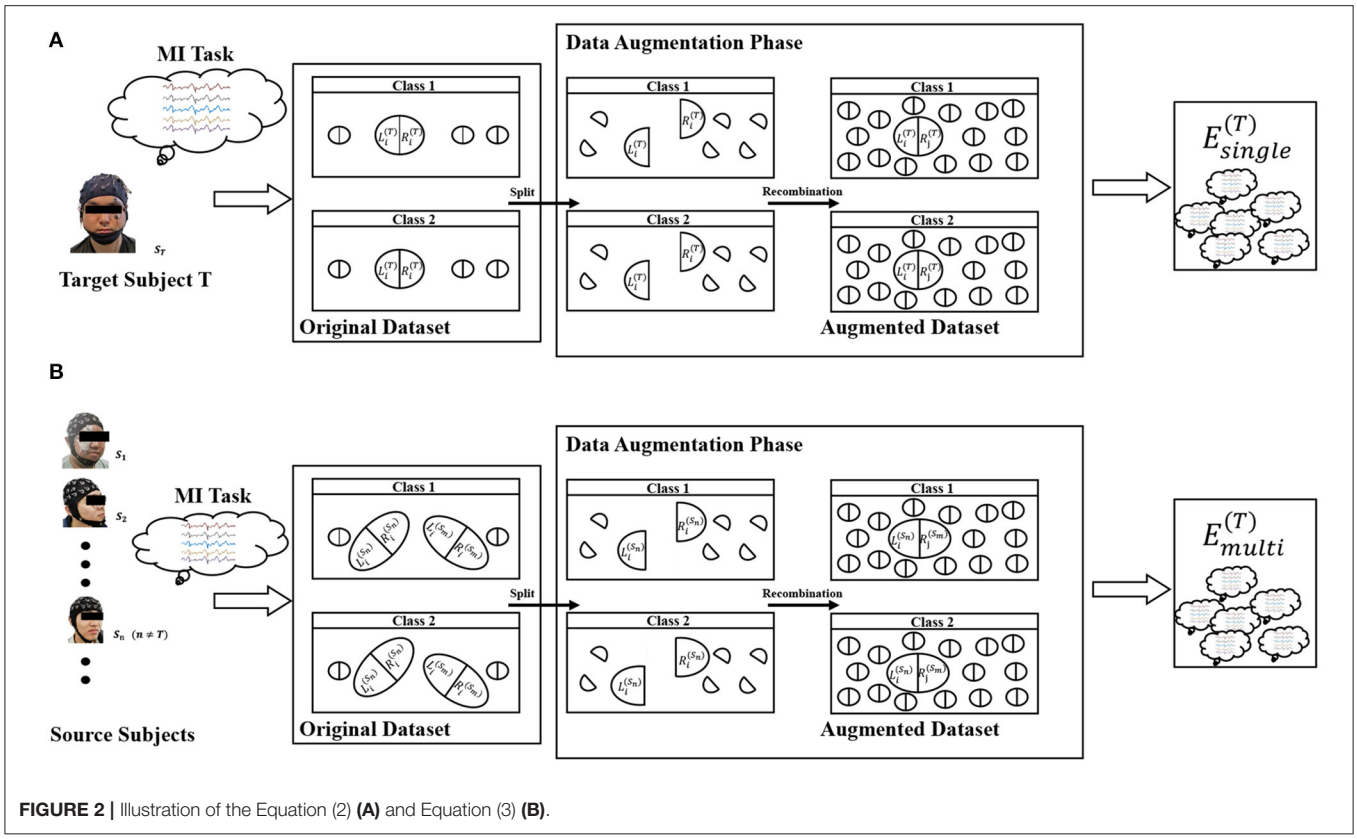

FIGURE 2 | Illustration of the Equation (2) **(A)** and Equation (3) **(B)**.

Oz in dataset 1. We alternately divide them into two sets in turn. Specifically, Fpz, FCz, CPz, and Pz are selected to be the brain's left part, and Fz, Cz, Pz, Oz were are selected to be the brain's right part. In dataset 2, the zero midline electrodes are Fz, FCz, Cz, CPz, and Pz. Fz, Cz, and Pz are selected to be the brain's left part. FCz and CPz were are selected to be the brain's right part.

2.2.3. Noise-Added and Flipping

The problem we want to solve is data augmentation on a tiny training dataset. Methods like GANs and VAEs having massive parameters to be learned are not suitable for this situation. Moreover, considering that the no-parameter (or few parameters) methods will often achieve better results for this situation, we selected noise-added and flipping methods for comparative investigation (Lashgari et al., 2020). Concerning the noise-added method, a Gaussian noise matrix with SNR (signal to noise ratio) of 5 is calculated, and then this noise matrix is added to the original sample. The rule of the flipping DA method is to reverse each real sample in the time axis. Because the noise-added DA method does not have a fixed expansion ratio constant, we have implemented two versions of the noise-added DA method for a more objective comparison. One implementation (version 1) makes the noise-added DA method have the same expansion ratio as the flipping DA method. The other implementation (version 2) makes the noise-added DA method and the proposed BAR DA method have the same expansion ratio constant.

2.2.4. The EEGnet

We use the end-to-end deep learning model, named EEGnet (Lawhern et al., 2018). The EEGnet takes the EEG segments as the input, passes them through three convolution layers for feature extraction, and uses a fully connected layer to classify. The first layer is a temporal convolution to learn frequency filters. The second layer is a depthwise convolution layer. This layer connects to each feature map individually and learns frequency-specific spatial filters. The third layer is a separable convolution layer. The separable convolution is a combination of depthwise convolution, which learns a temporal summary for each feature map individually, followed by a pointwise convolution, which learns how to mix the feature maps optimally. All feature maps are flattened and are fed into a fully connected layer. Full details about the network architecture can be found in the open-source project: https://github.com/vlawhern/arl-eegmodels. In this study, we use the default hyperparameters provided by the open-source project.

2.2.5. The Common Spatial Patterns Extraction

Due to the strong spatial distribution characteristics of motor imagery EEG signals, a feature extraction method called Common Spatial Patterns (CSP) is designed (Koles et al., 1990). The CSP aims to construct spatial filters which can maximize the variance of band-pass filtered EEG signals from one class and minimize the variance of EEG signals from the other class (Lotte et al., 2018). Formally, CSP uses the spatial filters w to assign a

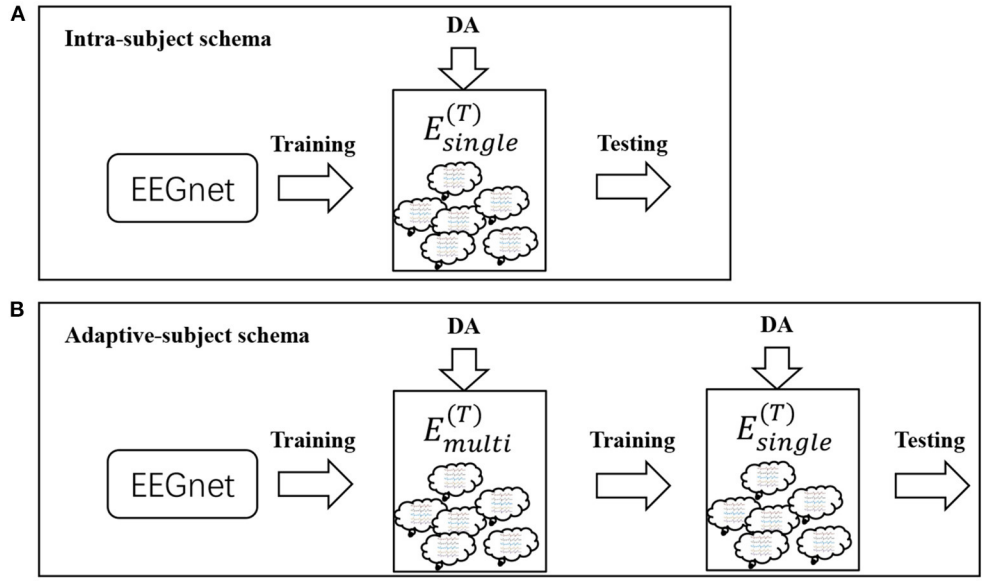

FIGURE 3 | Workflows for IS-schema and AS-schema. Panel **(A)** is the workflow of intra-subject-schema and panel **(B)** is the workflow of adaptive-subject-schema.

weight to each EEG sample channel. The w can be calculated by extremizing the following function:

$$J(w) = \frac{w'X_1'X_1w}{w'X_2'X_2w} = \frac{w'C_1w}{w'C_2w} \tag{4}$$

By maximizing Equation (4), we can calculate the spatial filter focusing on class 1. Indeed, $J(k \times w) = J(w)$, with k a real constant, means that the rescaling of the w is arbitrary. To calculate the only maximizer, we need a condition that $w'C_2w = 1$. Using the Lagrange multiplier method, the constrained optimization problem is equivalent to maximizing the following function:

$$L(\lambda, w) = w'C_1w - \lambda(w'C_2w - 1) \tag{5}$$

The filters w maximizing L can be calculated by setting the derivative of L concerning w to 0:

$$\begin{aligned} \frac{\partial L}{\partial w} &= 2w'C_1 - 2\lambda w'C_2 = 0 \\ &\Leftrightarrow C_1w = \lambda C_2w \\ &\Leftrightarrow C_2^{-1}C_1w = \lambda w \end{aligned} \tag{6}$$

We obtain an eigenvalue problem in Equation (6). Therefore, the spatial filters maximizing Equation (4) are the eigenvectors of $M = C_2^{-1}C_1$ which correspond to its largest eigenvalue. Empirically, we select the eigenvectors corresponding to the top $k(k = 3)$ eigenvalues and concatenate them into a matrix $W^{(1)}_{N_c \times k}$. This matrix can capture the three components that most relate to class 1. For the class 2, we can swap the numerator and denominator in the Equation (4) and repeat the above process. Finally, we concatenate the two matrices together to obtain the CSP-feature extraction matrix $W = \left[W^{(1)}_{N_c \times k}, W^{(2)}_{N_c \times k}\right]$.

3. RESULTS

We designed two schemas to evaluate these DA methods. Both schemas used the same deep learning model called EEGnet (Lawhern et al., 2018). The setting of the dataset was different between both schemas. In the first schema, referred to as intra-subject(IS)-schema, each subject's EEGnet was only trained on the subject's own dataset. In the second schema, referred to as adaptive-subject(AS)-schema, each subject's EEGnet should be trained in two stages. In the first stage, named the pre-training stage, the EEGnet should be trained on other subjects' datasets. In the second stage, named the adaptive-training stage, the EEGnet should be trained on the target subject's dataset. **Figure 3** showed the two schemas' workflow. We only used one DA method in each experiment instead of experimenting with multiple DA methods' additive effects. Two DA methods (noise-added and flipping) were implemented as reference methods (Lashgari et al., 2020). The flipping DA method flipped each sample along the time axis to generate a new sample. So this DA method can only get a double-sized dataset. The noise-added DA method added noise to each sample to generate new samples. Since the noise-added DA method had no fixed expansion factor, we implemented two versions with different expansion factors for the more objective comparison experiments. We repeated 10 times of experiments in each training set size for a specific subject.

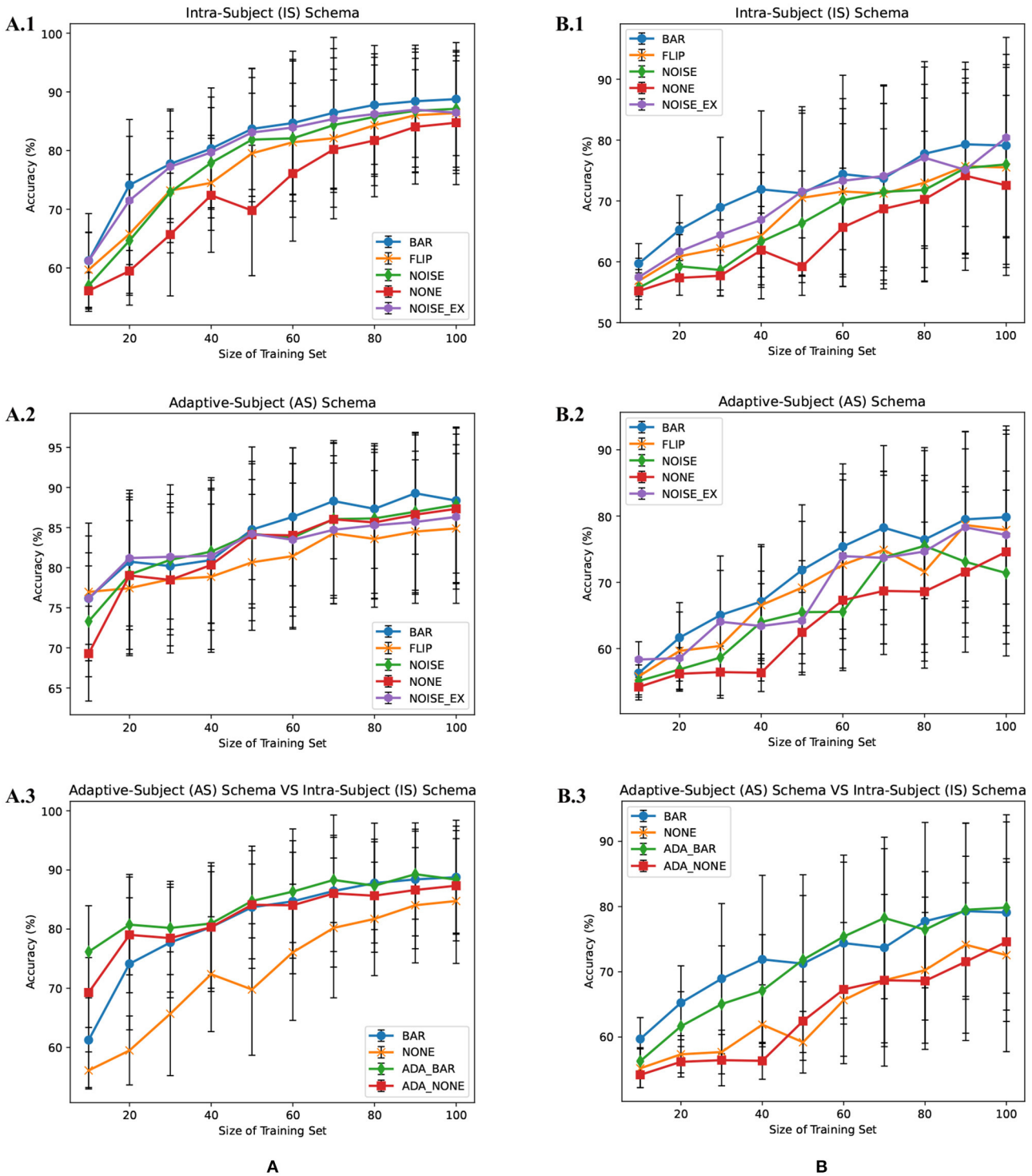

FIGURE 4 | Average subjects performance on different size of training dataset in the IS-schema and AS-schema from dataset 1 **(A)** and dataset 2 **(B)** was plotted. The red line with "NONE" in **(A.1,A.2,B.1,B.2)** indicates the case that the training set was not augmentated. The lines with "NOISE" and "NOISE_EX" represent version 1 of noise-added method and version 2 of the noise-added method, respectively. The green line and red line with the prefix "ADA_" in **(A.3,B.3)** represent AS-schema's accuracies. The blue line and yellow line in **(A.3,B.3)** represent IS-schema's accuracies.

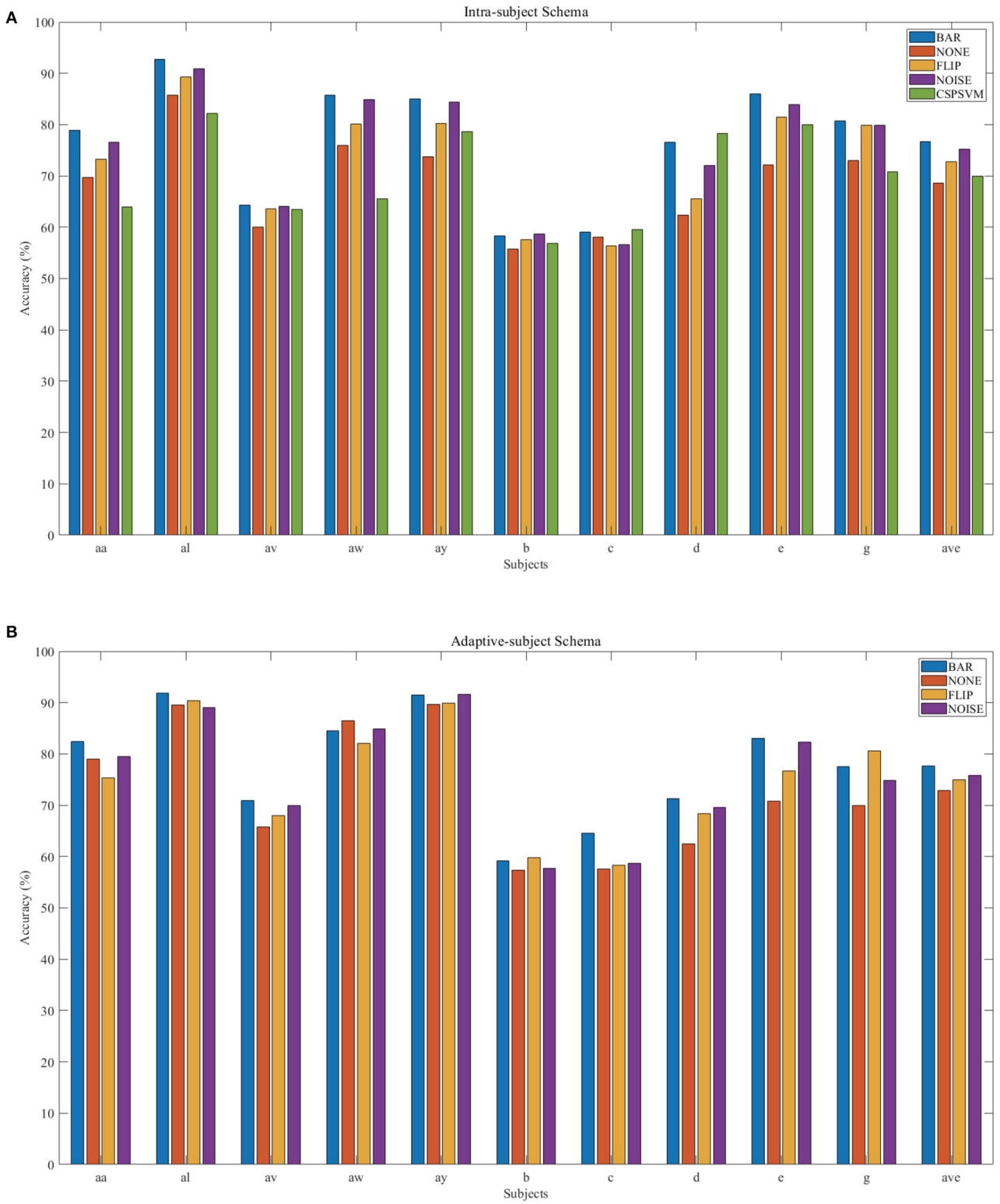

FIGURE 5 | Individual result in IS-schema **(A)** and AS-schema **(B)**. Note that the green bar ("CSPSVM") in **(A)** is the traditional classifier's result using the BAR method. The bar named "NOISE" represents the version 2 of noise-added method.

TABLE 1 | Paired-sample *t*-test result on the test accuracy in **Figures 4A.1,B.1**.

	Dataset 1			
	FLIP	**NOISE**	**NONE**	**NOISE_EX**
BAR	1×10^{-4}	2×10^{-3}	1×10^{-5}	2×10^{-3}
	Dataset 2			
	FLIP	**NOISE**	**NONE**	**NOISE_EX**
BAR	1×10^{-4}	8×10^{-6}	6×10^{-7}	2.6×10^{-3}

TABLE 2 | Paired-sample *t*-test result on the test accuracy in **Figures 4A.2,B.2**.

	Dataset 1			
	FLIP	**NOISE**	**NONE**	**NOISE_EX**
BAR	2×10^{-4}	2×10^{-3}	4×10^{-3}	5×10^{-2}
	Dataset 2			
	FLIP	**NOISE**	**NONE**	**NOISE_EX**
BAR	9×10^{-5}	3×10^{-4}	6×10^{-7}	1.2×10^{-3}

TABLE 3 | Paired-sample *t*-test result on the test accuracy in **Figure 5**.

	IS-schema			
	FLIP	**NOISE**	**NONE**	**CSPSVM**
BAR	3×10^{-3}	7.2×10^{-3}	3×10^{-4}	1.5×10^{-2}
	AS-schema			
	FLIP	**NOISE**		**NONE**
BAR	2.3×10^{-2}	9.9×10^{-4}		5.2×10^{-3}

3.1. IS-Schema Performance

There were multiple subjects in each dataset. For the s_{th} subject, we first randomly selected some samples to construct the original-training dataset D_s. The augmentation methods were applied to the D_s in turn to construct the corresponding augmented training datasets. The EEGnet was trained on these training datasets separately and was tested on the testing dataset. To test the proposed BAR's sensitivity to the training dataset's size, we conducted extensive experiments on different training dataset sizes. For a specific subject, we repeated the experiment 10 times on a specific training set size from 10 to 100 and then took the averaged accuracy as the final one. The results of those experiments were plotted in **Figures 4A.1,B.1**, **5A**. **Figures 4A.1,B.1** shown the performance of each DA method under different training set sizes. This performance was the average result of all subjects. For dataset 1, as the size of the training set increases, all DA methods' performance is improving, but our BAR method is always ahead of other methods by about 3%, except for the case where the training set size is 10. **Figure 5A** shown the performance of each DA method on different subjects. These accuracies come from the averaged accuracy of all experiments. For a specific subject, we run many experiments on different training set sizes from 10 to 100, and all accuracies are averaged to be the final one. A paired-sample *t*-test was used to measure the significance of our proposed BAR, and the results were shown in **Tables 1**, **3**. The test results were $p < 0.005$ between the referenced methods and the proposed method.

3.2. AS-Schema Performance

In the AS-schema: For each subject, we selected the same samples randomly from other subjects to construct the training dataset. In this schema, two training sets were needed, and the EEGnet was trained on them as **Figure 3** shows. The first dataset was a cross-subject dataset which is constructed by (3) for the s_{th} subject. The second dataset was constructed by (2), which was the same as the training dataset in the intra-subject schema. For the flipping method, we first mix the source subjects' data and then flip each sample in this mixed dataset to obtain a new artificial sample. For the noise-added method, we have two versions of strategies. Version 1: We first mix the source subjects' data and then add gaussian noise to each sample in this mixed dataset to obtain a new artificial sample. Version 2: We randomly select an original sample with replacement from the source subjects' mixed data and add Gaussian noise to it to obtain a new artificial sample. Repeat the process until the original samples, and the artificial samples are equals to the augmented dataset by the BAR. To investigate our proposed BAR's performance in different sizes of selected samples for each subject, we run the adaptive-subject experiment many times in each size of selected samples. The result was demonstrated in **Figures 4A.2,B.2**, **5B**. **Figures 4A.2,B.2** shown the performance of each DA method under different training set sizes. This performance was the average result of all subjects. For dataset 1 and dataset 2, as the training set size increases, all DA methods' performance increases, but our BAR method has always been ahead of other methods except for a few cases. **Figure 5B** shown the performance of each DA method on different subjects. These accuracies come from the averaged accuracy of all experiments. For a specific subject, we run a lot of experiments on different training set sizes from 10 to 100, and all accuracies are averaged to be the final one. A Paired-sample *t*-test was used to measure our proposed BAR's significance, and the result was shown in **Tables 2**, **3**. The test results were $p < 0.05$ between the referenced methods and the proposed method.

3.3. EEGnet vs. CSP-SVM

The CSP-SVM, a traditional classifier for MI-EEG signals, does not support the AS-schema. Therefore, we can only compare the performance of EEGnet and CSP-SVM in IS-schema. The results were plotted in **Figure 6**. The training set and testing set were the same as those used by EEGnet. In **Figure 6A**, our BAR method enabled EEGnet to obtain a huge improvement in classification performance compared to CSP-SVM. In **Figure 6B**, the improvement of classification performance was not obvious, but it still exceeded the performance of traditional CSP-SVM. This showed that our method can enable deep learning models

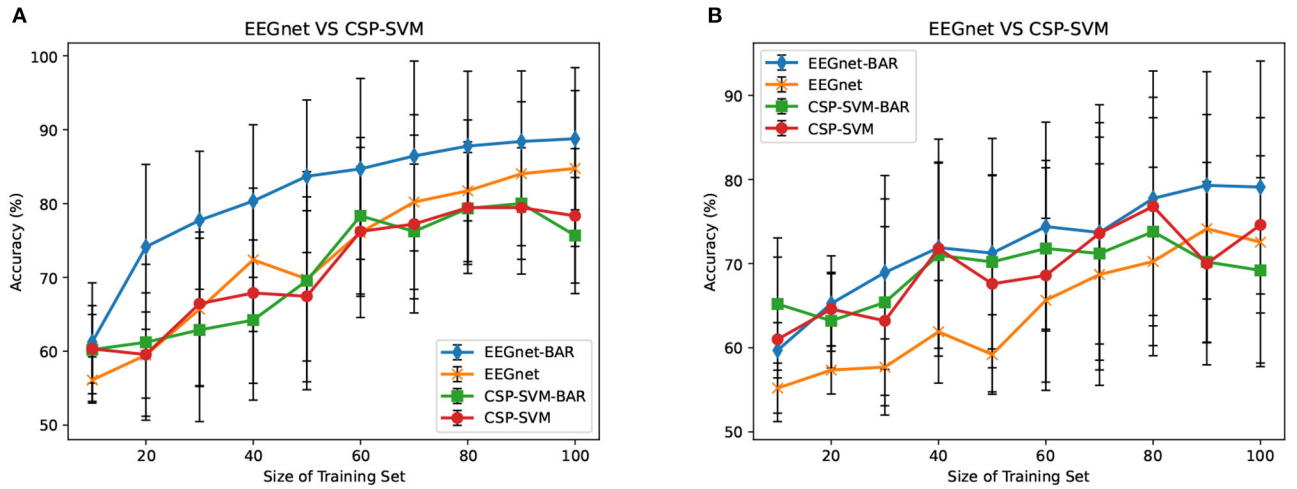

FIGURE 6 | EEGnet vs. CSP-SVM on dataset 1 **(A)** and dataset 2 **(B)**.

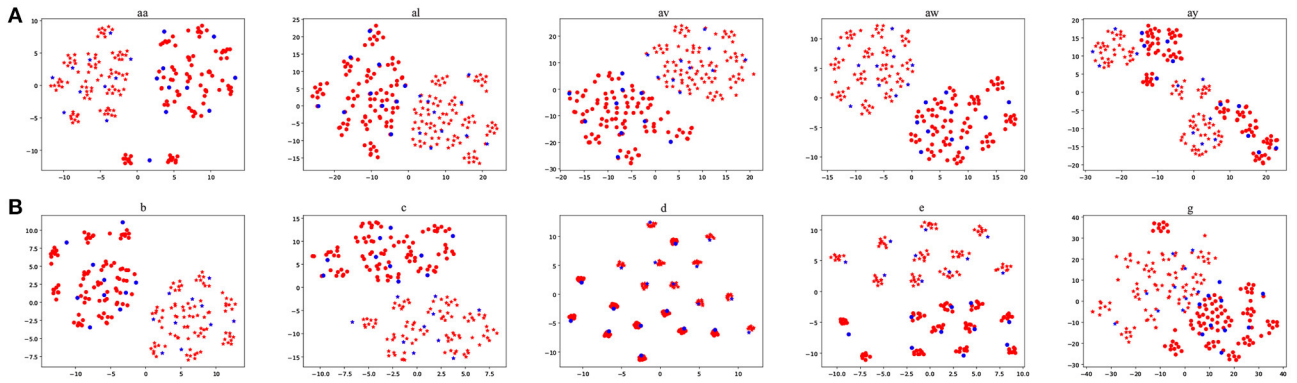

FIGURE 7 | The red points were artificial samples by the proposed BAR and the blue ones were sampled from the recording process. The star points ("*") and circle points ("o") represent the first class and the second class samples, respectively.

to be more fully trained, whether the quality of the dataset was poor or better, and the classification performance exceeded the traditional method CSP-SVM's.

3.4. Data Visualization

It is interesting to visualize the locations of the EEG trials generated by our proposed BAR. We used t-Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008), a non-linear dimensionality reduction technique that embeds high-dimensional data in a two-or-three-dimensional space, to show and compare the original EEG trials and generated EEG trials in the intra-subject schema. **Figure 7** shows the result of t-SNE on augmented training dataset from each subject, where the size of the training set we used is 20. An overall characteristic can be found that BAR's generated EEG trials may not be scattered far away from the original EEG trials. Note that the subject d and subject e are artificially generated "participants." So that the artificial samples closely surround real samples, which is slightly different from other subjects'.

3.5. Experimental Setup

We used ubuntu 18.04.5 LTS with a GPU TITAN V as the experiment platform. We chose 60 epochs determined by our iterated experiments for early stopping. We set the batch size to 16. Adam optimizer was used in all experiments with $lr = 0.001, \beta_1 = 0.9, \beta_2 = 0.999$.

4. DISCUSSION

This article developed a new DA method to generate artificial EEG data from the recorded samples. These artificial data can be used to supplement the training set, which can improve the EEGnet's decoding accuracy. The human brain is composed of two parts, the left hemisphere and the right hemisphere. Many studies have shown cooperative relationships between brain areas under specific tasks (Rubinov and Sporns, 2010). We believe that the motor imagery induced EEG patterns contain three parts: the left hemibrain independent component, the right hemibrain independent component, and the left and right brain cooperative

components. Moreover, our data augmentation method may strengthen the left and right brain collaboration components through channel-level reorganization and constructs its more robust training samples. We designed two schemas toward two application scenarios: single-subject scenario and multi-subjects scenario. We have demonstrated that the augmented datasets significantly improved the performance of detection in deep-learning-based MI-BCI systems. Furthermore, we reimplemented the noise-added method and the flipping method, known as common DA methods for time series data (Wen et al., 2020). The results showed that the proposed BAR significantly outperformed them in the MI-EEG classification task. Although the final binary classification accuracy has not improved much, it is a reliable improvement because it passed the *t*-test.

Merely expanding the size of the training dataset can improve the classification performance of the deep learning network. To get a more objective conclusion, the expansion ratio of the noise-added DA method was set to be the same as our proposed BAR's. Moreover, the results were plotted in **Figure 4**, which illustrated that with the help of our proposed BAR method, the EEGnet had been more fully trained to achieve the best classification performance. We found that as the size of the training set increases, the classification accuracy of deep learning increases very quickly at the beginning. After a certain threshold, the increase rate will slow down. In the **Figures 4A.1,A.2** the threshold is 20. In the **Figures 4B.1** the threshold is 40. In the **Figure 4B.2** the threshold is 70.

The flipping method destroyed the time-domain characteristics of EEG signals. Comparing **Figures 4A.1,A.2**, we found that the performance of the flipping method had dropped. In these two schemas, the only difference was that EEGnet was pre-trained in the mixed dataset of multiple subjects in the second schema. The spatial distribution characteristics of datasets mixed by multiple subjects would be reduced by the differences between subjects (Ang et al., 2008). Therefore, training EEGnet on a multi-subject mixed data would force the model to pay more attention to temporal features. However, the temporal features had been destroyed by the flipping method's operation in the time axis. EEGnet would perform worse in the AS-schema if the flipping DA method was used. This result was also consistent with prior knowledge that the two important dimensions of motor imaging EEG signal characteristics were space and time (Sakhavi et al., 2018).

The noise-added method was difficult to tune. We implemented two versions for the comparative experiments. One version had the same expansion ratio as the flip method, and the other version had the same expansion ratio as our proposed BAR method. In our experiment, the noise-added method was not adjusted to the optimal state. The tuning process of the noise-added method was complicated and required massive experiments. There were too many factors affecting the noise-added method's performance, such as the type of noise distribution, the signal-to-noise ratio, and the ratio of the generated data volume, original data volume, etc.

The BAR may promote the application of advanced DA methods such as GANs in the BCI field. After long-term development, the Generative Adversarial Network had evolved various variants, improving the training process's stability and the diversity of the generated samples (Goodfellow et al., 2014; Radford et al., 2015; Isola et al., 2017). Nevertheless, its essence was still a deep generative model that contained two deep modules (generator and discriminator), which included massive parameters to be learned (Gui et al., 2020). To obtain a generator with superior performance, a certain amount of data was needed to support generator and discriminator adversarial training. Still, the motivation for data augmentation in the BCI field was that we did not have enough real training data. This was a conflict. Therefore, when the original dataset's size was very small, using advanced methods such as GAN for data augmentation was not a good choice. However, our method was parameterless, and experiments proved that it can still enhance the deep-learning-based classifier under a small training set size. So, we want to say that our method may help GANs to improve their performance. In other words, the BAR may be a parameterless DA method that can assist the parameterized DA method.

Although AS-schema was dependent on the dataset, the DA method we proposed can improve the deep learning model's classification performance. The viewpoint that AS-schema was dependent on the dataset can be understanding by comparing the red line and the orange line in **Figures 4A.3,B.3**. As can be seen from **Figure 4**, the data quality of dataset 1 was better than that of dataset 2. The improvement effect of AS-schema on the dataset recorded from the high-quality subjects was more obvious. With the augmentation of our method, the deep learning model can improve the dataset with many poor subjects, which can be seen in **Figure 4B.3**. On dataset 1 with the better overall quality, our method can further improve the classification performance of deep learning models. The green line beat the others in **Figure 4A.3**.

DA method was not effective for CSP-SVM (traditional methods), but it was effective for EEGnet (deep learning methods). Two facts may explain the phenomenon. The perspective of features: In the traditional CSP-SVM framework, the features are extracted by the CSP algorithm. Although These features are highly explainable, they are too simple to reflect the data's original appearance. However, the deep-learning-based classifier is an end-to-end method, and the features are automatically learned from the massive training samples. These features, which are learned by many samples, often reflect more information of the original data. The perspective of non-linear fitting ability: The non-linear fitting ability of the SVM comes from the kernel function. Choosing a suitable kernel function is very dependent on experience (Cortes and Vapnik, 1995). However, the non-linear fitting ability of the deep-learning-based classifier is automatically learned from the massive training samples. Many facts have proved that data-driven non-linear expression capabilities are often better than that of manually selected kernel functions in recent years.

An interesting phenomenon discovered by comparing **Tables 1**, **2** was that the significance of the proposed BAR was reduced in dataset 1. The reason for the result may be that the size of the training set was enough to train a good neural network in the AS-schema. In the pre-training stage, our neural network was first trained on data from other subjects. The amount of data

in this stage was large enough to make the neural network to converge to a not bad point. So the effect of our proposed BAR will be reduced in the pre-training schema.

This article has several limitations that call for future investigation. (1) For multi subjects, we use the pre-training pipeline, which is an approach relying on experience stem from natural language processing (NLP) (Xipeng et al., 2020). Experiments show that this pipeline is not completely suitable for the BCI field. It is worth seeking the best way to transfer knowledge from other subjects to the target subject. (2) Influenced by the phenomenon of ERD and ERS, we choose the left brain part and the right brain part as the region to be divided and be recombined. Although we demonstrate this divided approach's effectiveness by extensive experiments, the optimal dividend approach is a crucial problem for the channel-wise-recombined DA method. (3) Through a large number of artificial samples obtained by the BAR in a short time, these samples have a large number of redundant samples. They contain countless repetitive information, which will significantly reduce the training speed of the model. Selecting high-quality samples from artificially generated samples has become a problem, which would become a potential application scenario for active learning (Settles, 2009). These questions will guide our next research direction.

5. CONCLUSION

In this study, a data augmentation method (denoted as BAR) based channel-level recombination was proposed for MI-BCI systems. In our method, to obtain an augmented training set, we divided each sample into two samples according to the brain region to which the channel belongs and then regroup them in the same category. After that, the EEGnet was trained on the augmented training set. Two common DA methods were implemented as comparisons in two training schemas to verify the proposed BAR method. All comparative experimental results passed the paired-sample *t*-test, which fully demonstrated our proposed BAR's effectiveness. At the same time, we found that AS-schema was dependent on the dataset. It performed well on dataset 1 but badly on dataset 2. One possible reason was that dataset 2 was of poor quality, and AS-schema did not apply. How to match the AS-schema with a poor quality dataset will be our next research direction. In bad situations, our method can still improve the decoding performance of deep learning models. The proposed BAR may promote the application of deep learning technology in BCI systems.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

YP and LX: conceptualization. YP: methodology, data curation, and writing—original draft preparation. HY: software, formal analysis, and visualization. YP, ZL, and YY: validation. YP and JJ: investigation. EY: resources and funding acquisition. YP, ZL, and LX: writing—review and editing. YY and WL: supervision. YP and LX: project administration. All authors: contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant no. 61901505).

REFERENCES

Abdulkader, S. N., Atia, A., and Mostafa, M.-S. M. (2015). Brain computer interfacing: applications and challenges. *Egypt. Inform. J.* 16, 213–230. doi: 10.1016/j.eij.2015.06.002

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., et al. (2018). The history began from alexnet: a comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*.

Ang, K. K., Chin, Z. Y., Zhang, H., and Guan, C. (2008). "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (Singapore: IEEE), 2390–2397.

Aznan, N. K. N., Atapour-Abarghouei, A., Bonner, S., Connolly, J. D., Al Moubayed, N., and Breckon, T. P. (2019). "Simulating brain signals: Creating synthetic EEG data via neural-based generative models for improved SSVEP classification," in *2019 International Joint Conference on Neural Networks (IJCNN)* (Durham: IEEE), 1–8. doi: 10.1109/IJCNN.2019.8852227

Belkacem, A. N., Nishio, S., Suzuki, T., Ishiguro, H., and Hirata, M. (2018). Neuromagnetic decoding of simultaneous bilateral hand movements for multidimensional brain-machine interfaces. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 1301–1310. doi: 10.1109/TNSRE.2018.2837003

Beuchat, N. J., Chavarriaga, R., Degallier, S., and Millán, J. D. R. (2013). "Offline decoding of upper limb muscle synergies from EEG slow cortical potentials," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Lausanne: Swiss Federal Institute of Technology), 3594–3597. doi: 10.1109/EMBC.2013.6610320

Chen, C., Zhang, J., Belkacem, A. N., Zhang, S., Xu, R., Hao, B., et al. (2019). G-causality brain connectivity differences of finger movements between motor execution and motor imagery. *J. Healthcare Eng.* 2019. doi: 10.1155/2019/5068283 Available online at: https://www.hindawi.com/journals/jhe/2019/5068283/

Chikara, R. K., and Ko, L.-W. (2019). Neural activities classification of human inhibitory control using hierarchical model. *Sensors* 19:3791. doi: 10.3390/s19173791

Choi, K., and Cichocki, A. (2008). "Control of a wheelchair by motor imagery in real time," in *International Conference on Intelligent Data Engineering and Automated Learning* (Springer), 330–337. doi: 10.1007/978-3-540-88906-9_42

Cooney, C., Folli, R., and Coyle, D. (2019). "Optimizing layers improves cnn generalization and transfer learning for imagined speech decoding from EEG," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (Derry: IEEE), 1311–1316. doi: 10.1109/SMC.2019.8914246

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn*. 20, 273–297. doi: 10.1007/BF00994018

Fahimi, F., Dosen, S., Ang, K. K., Mrachacz-Kersting, N., and Guan, C. (2020). Generative adversarial networks-based data augmentation for brain-computer interface. *IEEE Trans. Neural Netw. Learn. Syst*. doi: 10.1109/TNNLS.2020.3016666

Friman, O., Volosyak, I., and Graser, A. (2007). Multiple channel detection of steady-state visual evoked potentials for brain-computer interfaces. *IEEE Trans. Biomed. Eng*. 54, 742–750. doi: 10.1109/TBME.2006.889160

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems* (Montreal, QC: NIPS 2014), 2672–2680.

Gui, J., Sun, Z., Wen, Y., Tao, D., and Ye, J. (2020). A review on generative adversarial networks: algorithms, theory, and applications. *arXiv preprint arXiv:2001.06937*.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Los Angeles, CA: IEEE), 1125–1134. doi: 10.1109/CVPR.2017.632

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Ko, L.-W., Chikara, R. K., Lee, Y.-C., and Lin, W.-C. (2020). Exploration of user's mental state changes during performing brain-computer interface. *Sensors* 20:3169. doi: 10.3390/s20113169

Kodali, N., Abernethy, J., Hays, J., and Kira, Z. (2017). On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*.

Koles, Z. J., Lazar, M. S., and Zhou, S. Z. (1990). Spatial patterns underlying population differences in the background EEG. *Brain Topogr*. 2, 275–284. doi: 10.1007/BF01129656

Lashgari, E., Liang, D., and Maoz, U. (2020). Data augmentation for deep-learning-based electroencephalography. *J. Neurosci. Methods* 346:108885. doi: 10.1016/j.jneumeth.2020.108885

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). EEGnet: a compact convolutional neural network for EEG-based brain-computer interfaces. *J. Neural Eng*. 15:056013. doi: 10.1088/1741-2552/aace8c

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lotte, F. (2015). Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain-computer interfaces. *Proc. IEEE* 103, 871–890. doi: 10.1109/JPROC.2015.2404941

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., et al. (2018). A review of classification algorithms for eeg-based brain-computer interfaces: a 10 year update. *J. Neural Eng*. 15:031005. doi: 10.1088/1741-2552/aab2f2

Lotze, M., and Halsband, U. (2006). Motor imagery. *J. Physiol*. 99, 386–395. doi: 10.1016/j.jphysparis.2006.03.012

Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res*. 9, 2579–2605.

McFarland, D. J., Miner, L. A., Vaughan, T. M., and Wolpaw, J. R. (2000). Mu and beta rhythm topographies during motor imagery and actual movements. *Brain Topogr*. 12, 177–186. doi: 10.1023/A:1023437823106

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng*. 16:051001. doi: 10.1088/1741-2552/ab260c

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003

Sakai, A., Minoda, Y., and Morikawa, K. (2017). "Data augmentation methods for machine-learning-based classification of bio-signals," in *2017 10th Biomedical Engineering International Conference (BMEiCON)* (Tokyo), 1–4. doi: 10.1109/BMEiCON.2017.8229109

Sakhavi, S., Guan, C., and Yan, S. (2018). Learning temporal information for brain-computer interface using convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst*. 29, 5619–5629. doi: 10.1109/TNNLS.2018.2789927

Salamon, J., and Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process. Lett*. 24, 279–283. doi: 10.1109/LSP.2017.2657381

Settles, B. (2009). *Active Learning Literature Survey*. Technical report, Department of Computer Sciences, University of Wisconsin-Madison.

Song, X., Shibasaki, R., Yuan, N. J., Xie, X., Li, T., and Adachi, R. (2017). DeepMOB: learning deep knowledge of human emergency behavior and mobility from big and heterogeneous data. *ACM Trans. Inform. Syst*. 35, 1–19. doi: 10.1145/3057280

Wang, K., Xu, M., Wang, Y., Zhang, S., Chen, L., and Ming, D. (2020). Enhance decoding of pre-movement eeg patterns for brain-computer interfaces. *J. Neural Eng*. 17:016033. doi: 10.1088/1741-2552/ab598f

Wen, Q., Sun, L., Song, X., Gao, J., Wang, X., and Xu, H. (2020). Time series data augmentation for deep learning: a survey. *arXiv preprint arXiv:2002.12478*.

Wolpaw, J. R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., et al. (2000). Brain-computer interface technology: a review of the first international meeting. *IEEE Trans. Rehabil. Eng*. 8, 164–173. doi: 10.1109/TRE.2000.847807

Xipeng, Q., TianXiang, S., Yige, X., Yunfan, S., Ning, D., and Xuanjing, H. (2020). *Pre-trained Models for Natural Language Processing: A Survey*. Science China Technological Sciences.

Xu, L., Xu, M., Ke, Y., An, X., Liu, S., and Ming, D. (2020). Cross-dataset variability problem in eeg decoding with deep learning. *Front. Hum. Neurosci*. 14:103. doi: 10.3389/fnhum.2020.00103

Yin, E., Zeyl, T., Saab, R., Hu, D., Zhou, Z., and Chau, T. (2016). An auditory-tactile visual saccade-independent p300 brain-computer interface. *Int. J. Neural Syst*. 26:1650001. doi: 10.1142/S0129065716500015

Yin, Z., and Zhang, J. (2017). Cross-subject recognition of operator functional states via EEG and switching deep belief networks with adaptive weights. *Neurocomputing* 260, 349–366. doi: 10.1016/j.neucom.2017.05.002

Yu, Y., Zhou, Z., Liu, Y., Jiang, J., Yin, E., Zhang, N., et al. (2017). Self-paced operation of a wheelchair based on a hybrid brain-computer interface combining motor imagery and p300 potential. *IEEE Trans. Neural Syst. Rehabil. Eng*. 25, 2516–2526. doi: 10.1109/TNSRE.2017.2766365

Zhong, Z., Jin, L., and Xie, Z. (2015). "High performance offline handwritten Chinese character recognition using googlenet and directional feature maps," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (Shanghai: IEEE), 846–850. doi: 10.1109/ICDAR.2015.7333881

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 12 March 2021
doi: 10.3389/fnhum.2021.646915

Subject-Independent Functional Near-Infrared Spectroscopy-Based Brain–Computer Interfaces Based on Convolutional Neural Networks

Jinuk Kwon[1,2] *and Chang-Hwan Im*[1,2]*

[1] Department of Biomedical Engineering, Hanyang University, Seoul, South Korea, [2] Department of Electronic Engineering, Hanyang University, Seoul, South Korea

Functional near-infrared spectroscopy (fNIRS) has attracted increasing attention in the field of brain–computer interfaces (BCIs) owing to their advantages such as non-invasiveness, user safety, affordability, and portability. However, fNIRS signals are highly subject-specific and have low test-retest reliability. Therefore, individual calibration sessions need to be employed before each use of fNIRS-based BCI to achieve a sufficiently high performance for practical BCI applications. In this study, we propose a novel deep convolutional neural network (CNN)-based approach for implementing a subject-independent fNIRS-based BCI. A total of 18 participants performed the fNIRS-based BCI experiments, where the main goal of the experiments was to distinguish a mental arithmetic task from an idle state task. Leave-one-subject-out cross-validation was employed to evaluate the average classification accuracy of the proposed subject-independent fNIRS-based BCI. As a result, the average classification accuracy of the proposed method was reported to be 71.20 ± 8.74%, which was higher than the threshold accuracy for effective BCI communication (70%) as well as that obtained using conventional shrinkage linear discriminant analysis (65.74 ± 7.68%). To achieve a classification accuracy comparable to that of the proposed subject-independent fNIRS-based BCI, 24 training trials (of approximately 12 min) were necessary for the traditional subject-dependent fNIRS-based BCI. It is expected that our CNN-based approach would reduce the necessity of long-term individual calibration sessions, thereby enhancing the practicality of fNIRS-based BCIs significantly.

Keywords: brain–computer interface, functional near-infrared spectroscopy, deep learning, convolutional neural network, binary communication

OPEN ACCESS

Edited by:
Sung Chan Jun,
Gwangju Institute of Science and Technology, South Korea

Reviewed by:
Dalin Zhang,
Aalborg University, Denmark
Jinung An,
Daegu Gyeongbuk Institute of Science and Technology (DGIST), South Korea
Sangtae Ahn,
Kyungpook National University, South Korea

***Correspondence:**
Chang-Hwan Im
ich@hanyang.ac.kr

Specialty section:
This article was submitted to Brain-Computer Interfaces, a section of the journal Frontiers in Human Neuroscience

Received: *28 December 2020*
Accepted: *19 February 2021*
Published: *12 March 2021*

Citation:
Kwon J and Im C-H (2021) Subject-Independent Functional Near-Infrared Spectroscopy-Based Brain–Computer Interfaces Based on Convolutional Neural Networks. Front. Hum. Neurosci. 15:646915. doi: 10.3389/fnhum.2021.646915

INTRODUCTION

Brain–computer interfaces (BCIs) have been developed to decode a user's intention from their neural signals with the ultimate goal of providing non-muscular communication channels to those who experience difficulties communicating with the external environment (Wolpaw et al., 2002; Daly and Wolpaw, 2008). Various neuroimaging modalities such as electroencephalography (EEG), magnetoencephalography, and functional magnetic resonance imaging have been employed to implement BCIs (Mellinger et al., 2007; Sitaram et al., 2007; Hwang et al., 2013).

Recently, functional near-infrared spectroscopy (fNIRS), which is also one of the representative brain-imaging modalities, has attracted increasing attention owing to its advantages, including non-invasiveness, affordability, low susceptibility to noise, and portability (Naseer and Hong, 2015; Shin et al., 2017a). fNIRS is an optical brain-imaging technology used to record hemodynamic responses of the brain using near-infrared-range light of wavelength 600–1,000 nm. fNIRS can measure oxy- and deoxy-hemoglobin concentration changes (ΔHbO and ΔHbR) while an individual performs specific mental tasks such as mental arithmetic (MA), motor imagery (MI), mental singing, and imagining of object rotation. During these mental tasks, increased cerebral blood flow caused by neural activities leads to an increase and decrease in ΔHbO and ΔHbR, respectively, which have been utilized to implement fNIRS-based BCIs (Ferrari and Quaresima, 2012; Schudlo and Chau, 2015). Previous studies (Coyle et al., 2007; Naseer and Hong, 2013; Hong et al., 2020) have reported that the performance of fNIRS-based BCI is high enough to be applied to practical binary communication systems that require a threshold classification accuracy of at least 70% (Vidaurre and Blankertz, 2010).

Recently, many researchers have proposed new approaches to improve the performance of fNIRS-based BCIs. For example, recent studies have reported significant improvements in the classification accuracy of fNIRS-based BCIs by employing high-density multi-distance fNIRS devices (Shin et al., 2017a) and using ensemble classifiers based on bootstrap aggregation Shin and Im (2020). von Lühmann et al. (2020) proposed a general linear model-based preprocessing method to improve the classification accuracy of fNIRS-based BCI. The combination of fNIRS with other brain-imaging modalities also demonstrated a potential to improve the classification accuracy of the BCI system (Fazli et al., 2012; Shin et al., 2018b). Recently, Kwon and Im (2020) demonstrated that photobiomodulation before a BCI experiment could enhance the overall classification accuracy of fNIRS-based BCIs. Besides, a number of studies have attempted to improve the information transfer rate (ITR) of fNIRS-based BCI by increasing the number of commands (i.e., mental tasks) (Khan et al., 2014; Hong and Khan, 2017; Shin et al., 2018a). In addition, researchers have also been interested in implementing portable BCI systems with a small number of sensors while preserving the overall BCI performance to elevate their practical applicability (Kazuki and Tsunashima, 2014; Shin et al., 2017b; Kwon et al., 2020a).

Although fNIRS-based BCI technology has advanced considerably, it is still challenging to use fNIRS-based BCIs in real-world applications because neural signals generally exhibit high inter-subject variability and non-stationarity. Moreover, because fNIRS signals are readily affected by a user's mental state, such as cognitive load and fatigue, they can change during the course of same-day experiments (Holper et al., 2012; Hu et al., 2013). Therefore, individual training sessions need to be performed before each usage of the BCI system to acquire high-performance BCI systems. However, such relatively long calibration sessions to obtain enough training data degrade their practicality and sometimes cause user fatigue even before using the BCI system. Various strategies have been proposed to reduce the necessity of such long-term calibration sessions in the field of EEG-based BCIs (Fazli et al., 2009; Wang et al., 2015; Yuan et al., 2015; Jayaram et al., 2016; Waytowich et al., 2016; Joadder et al., 2019; Xu et al., 2020). Recently, Kwon et al. (2020b) proposed a subject-independent EEG-based BCI framework based on deep convolutional neural networks (CNNs), which does not require any calibration sessions, with a fairly high classification accuracy. However, to the best of our knowledge, no previous study has successfully implemented a deep CNN-based subject-independent fNIRS-based BCI that outperforms conventional machine-learning-based subject-independent fNIRS-based BCIs.

In this study, we proposed a novel CNN-based deep-learning approach for subject-independent fNIRS-based BCIs. fNIRS signals were recorded using a portable fNIRS recording system that covers the prefrontal cortex while the participants were performing MA and idle state (IS) tasks. The leave-one-subject-out cross-validation (LOSO-CV) strategy was employed to evaluate the performance of the proposed method. The resultant classification accuracy was then compared with the threshold accuracy for effective binary BCIs (70%) and the classification accuracy was achieved using the conventional machine learning method, which has been widely employed for fNIRS-based BCIs. To the best of our knowledge, this is the first study that has applied a deep learning approach to subject-independent fNIRS-based mental imagery BCIs.

MATERIALS AND METHODS

Dataset

In this study, a part of an fNIRS dataset collected in our previous study (Shin et al., 2018b) was used to evaluate the proposed method. The original dataset consisted of 21-channel EEG data and 16-channel fNIRS data, which were recorded from 18 healthy adult participants (10 males and 8 females, 23.8 $\pm$ 2.5 years). From the original dataset, only the fNIRS data measured during the MA and IS tasks at all 16 prefrontal NIRS channels were selectively used in this study. A commercial NIRS recording system (LIGHTNIRS; Shimadzu Corp.; Kyoto, Japan) was used to record fNIRS signals at a sampling rate of 13.3 Hz. The arrangement of the fNIRS channels is shown in **Figure 1**.

Experiment Paradigm

The timing sequence of a single trial is shown in **Figure 2**. Each task trial consisted of an instruction (2 s), task (10 s), and inter-trial rest (a randomized interval of 16–18 s). During the instruction period, a specific task to be performed during the task period was displayed at the center of the monitor. The participants were provided with either a mathematical expression showing a "random three-digit number minus a one-digit number between 6 and 9 (e.g., 123–9)" for the MA task or a fixation cross for the IS task. During the task period, the participants were asked to perform either MA or IS tasks as instructed. During the MA task, the participants had to repetitively subtract the designated one-digit number from the result of the former calculation as quickly as possible (e.g., 123–9 = 114, 114–9 = 105, 105–9 = 96, …), until the stop sign was presented. During the IS task, the participants stayed relaxed

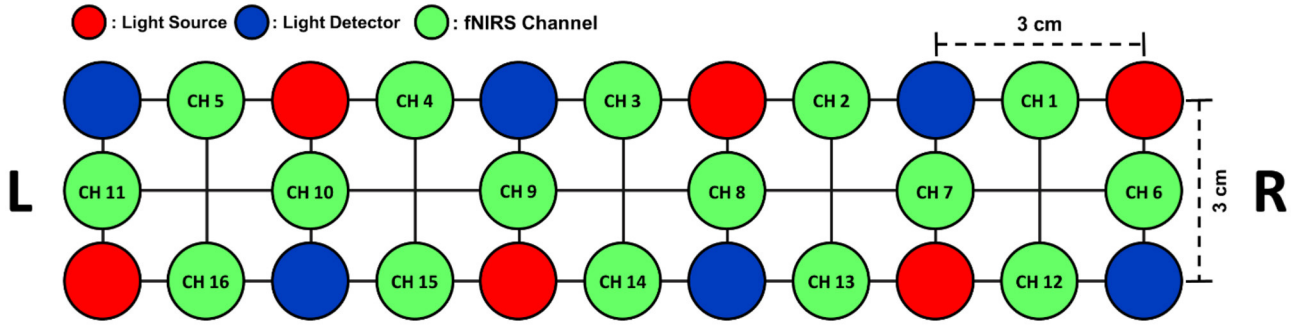

FIGURE 1 | The arrangement of six light emitters (red) and six light detectors (blue) on the forehead over the prefrontal area. A total of 16 fNIRS channels (green) were formed by the pairs of neighboring light emitters and detectors with a distance of 3 cm between them.

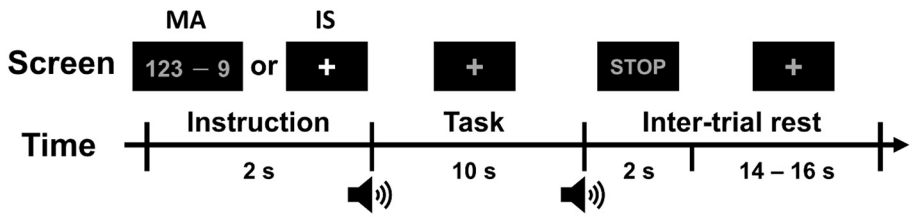

FIGURE 2 | Timing sequence of a single trial. Each trial consisted of an introduction period of 2 s, a task period of 10 s, and an inter-trial rest period of 16–18 s. During the introduction period, the task to be performed was displayed at the center of the monitor. After a short beep, the participants were asked to perform the designated task while looking at a fixation cross. When a STOP sign was displayed with a second short beep, the participants stopped performing the task and relaxed during the inter-trial rest period.

without performing any mental imagery task. The MA and IS tasks were performed 30 times each.

Preprocessing

MATLAB 2018b (MathWorks; Natick, MA, USA) was used to analyze the recorded fNIRS data, when functions implemented in the BBCI toolbox[1] were employed. The raw optical densities (ODs) were converted to ΔHbR and ΔHbO using the following formula (Matcher et al., 1995):

$$\begin{pmatrix}\Delta\text{HbR}\\ \Delta\text{HbO}\end{pmatrix} = \begin{pmatrix}1.8545 & -0.2394 & -1.0947\\ -1.4887 & 0.5970 & 1.4847\end{pmatrix}\begin{pmatrix}\Delta\text{OD}_{780}\\ \Delta\text{OD}_{805}\\ \Delta OD_{830}\end{pmatrix}\ (\text{mM}\cdot\text{cm}),$$

where ΔOD represents the optical density changes at wavelengths of 780, 805, and 830 nm. The converted ΔHbR and ΔHbO values were band-pass filtered at 0.01–0.09 Hz using a 6th-order Butterworth zero-phase filter to remove physiological noise. fNIRS data were then segmented into epochs from 0 to 15 s considering the hemodynamic delay of the order of several seconds (Naseer and Hong, 2013). Baseline correction was performed by subtracting the temporal mean value within the (−1 s, 0 s) interval from each fNIRS epoch.

[1]https://github.com/bbci/bbci_public

Performance Evaluation

Shrinkage Linear Discriminant Analysis

A shrinkage linear discriminant analysis (sLDA), which is a combination of linear discriminant analysis (LDA) and a shrinkage tool, was employed as the representative conventional classification method as it has been widely employed in recent fNIRS-based BCI studies owing to its high classification performance (Shin et al., 2017a, 2018b). This method is known to be particularly useful for improving the estimation of covariance matrices in situations where the number of training samples is small compared to the number of features. The feature vectors to train the sLDA were constructed using the temporal mean amplitudes of fNIRS data within multiple windows of 0–5, 5–10, and 10–15 s for each epoch. As a result, the dimension of fNIRS feature vectors was 96 (= 16 channels × 2 fNIRS chromophores × 3 intervals).

Proposed Deep Learning Approach

We proposed a one-dimensional CNN-based deep-learning approach for subject-independent fNIRS-based BCI. The detailed network architecture is listed in **Table 1**. The proposed model consisted of an input layer, two 1-dimensional convolutional layers, and a single fully connected layer. The input layer had a dimension of 201 (time samples) × 32 (= 16 channels × 2 chromophores), followed by two convolutional layers with 32 filters. The kernel sizes of the two layers were set to 13 and 6, and the stride sizes of the two layers were set to 9 and 4. The

TABLE 1 | The architecture of the deep-learning model based on 1-dimensional CNN.

Layer	Number of filters	Kernel size	Normalization, dropout, activation layer	Output shape	Options
Input			EvoNorm Dropout ($p = 0.5$)	(201, 32)	
1D Conv	32	13	EvoNorm Dropout ($p = 0.5$)	(21, 32)	Stride = 9 Padding = Valid
1D Conv	32	6	EvoNorm Dropout ($p = 0.5$)	(4, 32)	Stride = 4 Padding = Valid
Flatten				(128)	
Dense	128 × 2		Softmax	(2)	

TABLE 2 | The architecture of EEGNet.

Layer	Number of filters	Kernel size	Normalization, Dropout, activation layer	Output shape	Options
Input				(32, 201, 1)	
2D Conv	F_1	(1, 6)	BatchNorm	(32, 201, F_1)	Padding = same
2D Depthwise Conv	$D \times F_1$	(32, 1)	BatchNorm ELU	(1, 201, $D \times F_1$)	Padding = valid Depth = D Max norm = 1
2D Average Pooling		(1, 4)	Dropout ($p = 0.25$)	(1, 50, $D \times F_1$)	
2D Separable Conv	F_2	(1, 2)	BatchNorm ELU	(1, 50, F_2)	Padding = same
2D Average Pooling		(1, 8)	Dropout ($p = 0.25$)	(1, 6, F_2)	
Flatten				$1 \times 6 \times F_2$	
Dense	$(6 \times F_2) \times 2$		Softmax	2	

F_1, F_2, and D were set to 8, 16, and 2, respectively.

flattened output of the last convolutional layer, which had the dimension of 128, was fed into the fully connected layer, followed by the Softmax activation function. Consequently, the output of the proposed method had a dimension of two, corresponding to the number of tasks to be classified. The normalization and dropout layers were added after the input layer and the two convolutional layers to improve the generalization performance and training speed of the networks (Ravi et al., 2020). An evolving normalization-activation layer (EvoNorm) (Liu et al., 2020) was employed as the normalization layer, and the dropout probability was set to 0.5. The weights of the layers were initialized using a He-Normal initializer.

Ensemble of Regularized LDA

Recently, Shin and Im (2020) demonstrated that ensemble of weak classifiers resulted in a better classification accuracy than that of a single strong classifier. Based on this work, the ensemble of regularized LDA based on bootstrap aggregating (Bagging) algorithm was employed to validate the performance of subject-independent fNIRS-based BCI. The Bagging algorithm creates multiple training sets by sampling with replacement, then builds weak classifiers using each training set. The final classification result is decided by a majority vote of results from weak classifiers. In this study, the ensemble classifier was implemented using the MATLAB "fitcensemble" function. According to the previous study (Shin and Im, 2020), the number of weak classifiers, fraction of training set to resample, and gamma value for regularized LDA were set to 50, 100%, and 0.1, respectively. The feature vectors of training sets were set to be the same as those used to train sLDA.

EEGNet

Lawhern et al. (2018) introduced a compact CNN-based deep-learning architecture (EEGNet) that contains a small number of training parameters but showed robust classification performance in various EEG-based BCI paradigms such as P300, error-related negativity, movement-related cortical potential, and sensory-motor rhythm during MI. In this study EEGNet was employed as a conventional CNN-based classification method to verify the performance of the proposed method. EEGNet consists of an input layer, three 2-dimensional convolutional layers of temporal, spatial, and separable layers, and a single fully connected layer as listed in **Table 2**. The input layer had a dimension of 32 (= 16 channels × 2 chromophores) × 201 (time samples) × 1, followed by a 2-dimenssional temporal convolutional layer with F_1 filters. The kernel size of the temporal convolutional layer was set to (1, 6), chosen to be half the sampling rate of the data. The spatial convolutional layer had $D \times F_1$ filters with the kernel size of (32, 1), and the separable convolutional layer had F_2 filters with the kernel size of (2, 1). Each convolutional layer was followed by a Batch Normalization layer (BatchNorm) and a linear or exponential linear unit activation layer (ELU). Two average pooling layers were located after spatial and separable layers to reduce the size of feature maps, with the kernel sizes of (1, 4) and (1, 8), respectively. In this study, all the hyper parameters were determined based on the

previous studies (Lawhern et al., 2018). F_1, F_1, and D were set to 8, 16, and 2, respectively, and the kernel sizes of each convolutional layer were set considering the sampling rate of the fNIRS device.

Training Details

All the training and simulation processes were run on a desktop computer with a 12-core Ryzen 9 3900x processor, 64 GB memory, and an NVIDIA RTX 2080Ti GPU, using Keras (https://keras.io) with a Tensorflow backend, which is an open-source library for deep learning. Ten percent of the training data was split as the validation set, and an early stopping technique with a patience of 20 was used to avoid over-fitting with a batch size of 100. The hyper-parameters were empirically determined, and the random seed was set to 0. The pre-processed fNIRS data were fed into the proposed network after z-score normalization over the time axis to compensate for intrinsic amplitude differences among participants (Erkan and Akbaba, 2018). The network was trained to minimize the categorical cross-entropy loss function using the Adamax optimizer (Kingma and Ba, 2014; Vani and Rao, 2019) with a learning rate of 0.0005, decay of 5×10^{-8}.

Leave-One-Subject-Out Cross-Validation

A leave-one-subject-out cross-validation (LOSO-CV) strategy was employed to evaluate the performance of subject-independent fNIRS-based BCIs. In LOSO-CV, all the datasets except for a test participant—that is, the dataset of 1,020 samples (= 17 participants × 30 trials × 2 classes)—were used to train the classifier, and then data from the test participant (30 trials × 2 classes = 60 samples) were classified to evaluate the performance of the trained classifier. For example, when participant #1 was a test participant, the classification model for the participant #1 was trained using the data of the other 17 participants (participants #2 to #18). Then, the accuracy of the trained model was evaluated by applying the participant #1's data that were not used for the training to the trained model. This process was repeated until all participants' data were tested.

Pseudo-Online Simulation of Subject-Dependent fNIRS-Based BCI

A pseudo-online simulation of subject-dependent fNIRS-based BCI was performed to investigate how many training trials were required to achieve a classification accuracy higher than that of subject-independent fNIRS-based BCI. The dataset of each participant was split into training data and test data. For each task, the first N trials and the remaining (30 - N) trials were used as the training and test datasets, respectively. sLDA was employed as the classifier (Shin et al., 2017a) for this subject-dependent fNIRS-based BCI, and the classification accuracy was evaluated for different sizes (N) of training datasets to investigate how many training trials each participant should undergo before using the fNIRS-based BCI. It should be noted that data from other participants were not utilized to train the classifier.

RESULTS

The binary classification accuracies of individual participants are shown in **Figure 3**. The white and gray bars represent the classification accuracies of subject-independent fNIRS-based BCIs implemented using sLDA and the proposed CNN-based methods, respectively. The error bars represent the standard errors. The red dotted horizontal line denotes the threshold accuracy for the effective binary BCI (70%). The average classification accuracy of the proposed method was reported to be 71.20 ± 8.74% (mean ± standard deviation), which was higher than that obtained using the conventional sLDA (65.74 ± 7.68%) as well as the threshold accuracy for effective binary BCI communications (70%). The Wilcoxon signed rank sum test was conducted to statistically compare the difference in the classification accuracies, and statistically significant improvement of classification accuracy was observed for the proposed method ($p < 0.05$).

Figure 4 shows the results of the pseudo-online simulation of subject-dependent fNIRS-based BCI (denoted by "sLDA-Dependent" in the figure) with respect to different numbers of training data per class. The two horizontal lines denoted by "sLDA-independent" and "CNN-independent" represent the average accuracies of subject-independent fNIRS-based BCIs achieved using sLDA (65.74%) and CNN (71.20%), respectively. The black dotted line represents the threshold accuracy for an effective binary BCI (70%). It can be seen from the figure that the overall classification accuracy of the subject-dependent BCI increased as the number of training data increased. Notably, at least 12 training data per class were required to realize a subject-dependent fNIRS-based BCI with better performance than the subject-independent fNIRS-based BCI implemented using the proposed CNN-based method. This implies that an approximately 12 m-long training session may not be necessary before using the fNIRS-based BCI if the proposed subject-independent fNIRS-based BCI is employed.

Figure 5 illustrates the average classification accuracies of subject-independent fNIRS-based BCI evaluated using different classification methods. The red dotted horizontal line denotes the threshold accuracy for the effective binary BCI (70%) and the error bars represent the standard errors. The average classification accuracies evaluated using sLDA, ensemble of regularized LDA (denoted by "Bagging" in the figure), EEGNet, and proposed CNN-based methods were reported to be 65.74 ± 7.68%, 66.39 ± 7.44%, 67.96 ± 9.35%, and 71.20 ± 8.74%. Among all classification methods, only the proposed CNN-based method achieved higher classification accuracy than the threshold accuracy for effective binary BCI communications.

DISCUSSION

In this study, we investigated the feasibility of implementing a subject-independent fNIRS-based BCI using a deep learning-based approach. We proposed a novel deep-learning-based model architecture based on a CNN to effectively differentiate the two mental tasks, MA and IS. fNIRS signals were recorded from 16 sites covering the prefrontal cortex while participants performed either MA or IS task. The classification accuracy obtained using the proposed CNN-based method was reported

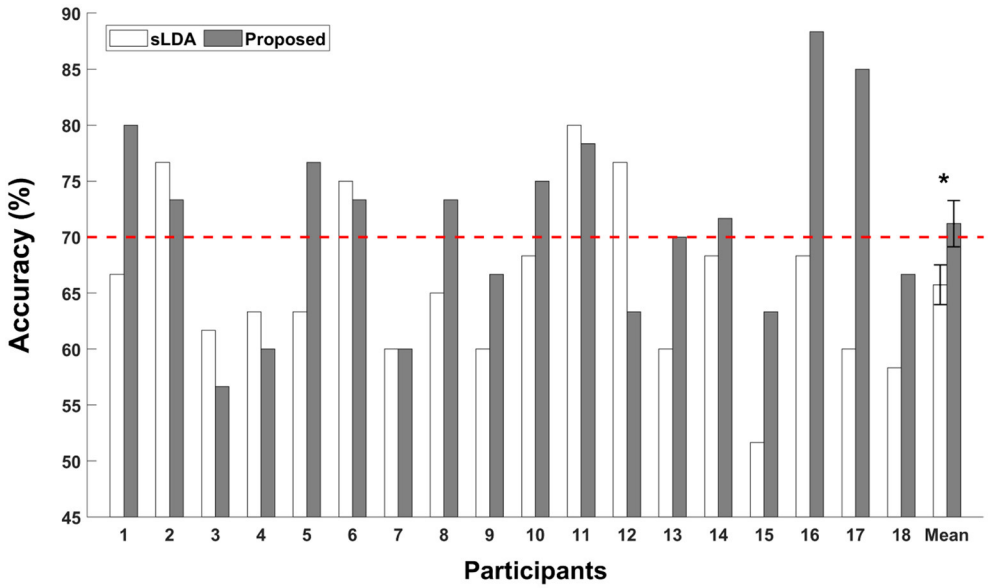

FIGURE 3 | Individual classification accuracies of the subject-independent fNIRS-based BCI. White and gray bars indicate the classification accuracies obtained using the shrinkage linear discriminant analysis (sLDA) classifier and the proposed method. The red horizontal dashed line indicates the effective BCI threshold level (70.0%). Error bars represent the standard errors. The grand average classification accuracies were 65.74 ± 7.68% and 71.20 ± 8.74% (mean ± standard deviation) for the sLDA and the proposed method, respectively. The asterisk (*) represents $p < 0.05$ (Wilcoxon signed rank test).

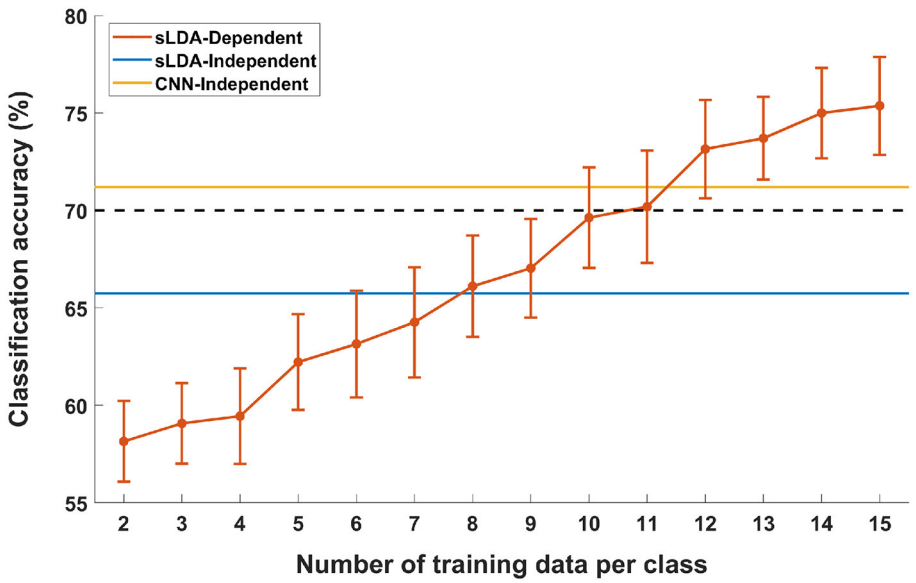

FIGURE 4 | Comparison of MA vs. IS classification accuracies of subject-independent (sLDA-independent and CNN-independent) and subject-dependent (sLDA-dependent) scenarios as a function of the number of individual training data. Vertical lines indicate the standard errors. The black horizontal dashed line represents the threshold accuracy of the effective BCI application (70.0%).

to be 71.20 ± 8.74%, which was not only higher than the threshold accuracy for effective BCI communication, but also higher than that obtained using the conventional sLDA method. Our experimental results demonstrated that our deep-learning-based approach has great potential to be adopted to establish a zero-training fNIRS-based BCI that could significantly enhance the practicality of fNIRS-based BCIs.

We believe that the improvement in the overall BCI performance stemmed from the synergetic effect of three factors employed to construct the proposed CNN-based

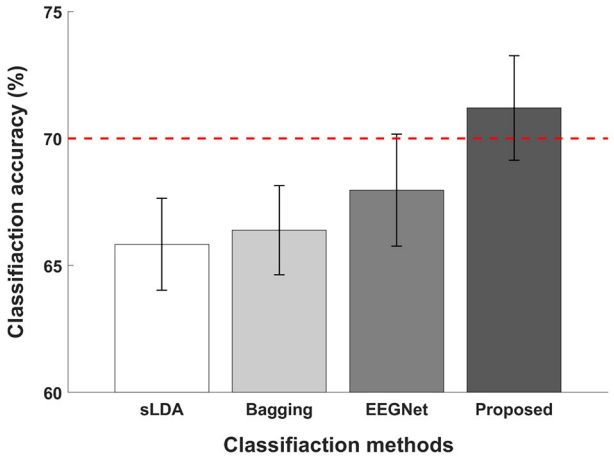

FIGURE 5 | Comparison of average classification accuracies of sLDA, ensemble of regularized LDA (denoted by "Bagging"), EEGNet, and the proposed method. The red horizontal dashed line indicates the effective BCI threshold level (70.0%) and error bars represent the standard errors. The average classification accuracies were 65.74 ± 7.68, 66.39 ± 7.44, 67.96 ± 9.35, and 71.20 ± 8.74% for sLDA, Bagging, EEGNet, and the proposed method, respectively.

model architecture. First of all, the CNN layer had high automatic feature extraction ability compared to that of the conventional feature extraction method (Shaheen et al., 2016). Additionally, construction of an appropriate structure of fully-connected layers is also an important factor. The performances of subject-independent fNIRS-based BCIs using various fully-connected layers with different structures are listed in **Supplementary Table 1**. Finally, to improve the generalization performance, we adopted EvoNorm, a recently introduced normalization-activation layer (Liu et al., 2020), instead of a batch normalization layer followed by the ReLU activation layer, which is a widely-used approach in deep learning. The classification accuracy evaluated using the EvoNorm (71.20%) was significantly higher than that obtained using the batch normalization and the ReLU activation layers (68.43%, $p < 0.05$, Wilcoxon signed rank test).

A previous study on the implementation of a subject-independent EEG-based BCI (Kwon et al., 2020b) reported the average classification accuracy of 74.15% in the two-class MI task classification problem. Since the modalities and paradigms of the previous study and this study are quite different with each other, direct comparison of BCI performance may not be meaningful; however, some important clues that can be employed in our future studies could be found in the previous study. In Kwon et al.'s study, EEG data were recorded from a total of 54 participants, which was almost three times more than the number of participants participated in our experiments. The authors of the previous study (Kwon et al., 2020b) demonstrated that a deep neural network model trained with a larger number of training data could result in a better classification accuracy and reduce the differences in BCI performance among participants. Thus, it may be a promising topic to investigate whether the performance of subject-independent fNIRS-based BCI based on our proposed CNN model could be further enhanced by increasing the size of the fNIRS dataset through additional experiments with a larger number of participants. The application of data augmentation techniques (Luo and Lu, 2018) or the employment of open-access datasets (Shin et al., 2018c) could also be promising options to increase the training data without additional experiments. After increasing the number of training data large enough to improve the overall BCI performance and investigating more appropriate deep learning structures, we will implement a real-time fNIRS-based BCI communication system that does not require any training session.

Current trends in BCI research are moving toward a hybrid BCI approach that combines more than two neuroimaging modalities to improve BCI performance. Among the various possible hybrid BCIs, a hybrid fNIRS-EEG BCI has been widely studied and has demonstrated the potential to increase the overall performance of BCIs—particularly compared to that of unimodal BCIs in terms of both classification accuracy and ITR (Hong and Khan, 2017; Shin et al., 2018b). Because Kwon et al. (2020b) recently demonstrated the feasibility of implementing a subject-independent EEG-based BCI using CNN, it is expected that a subject-independent hybrid fNIRS-EEG BCI could also be implemented by incorporating our proposed CNN model for fNIRS-based BCI with Kwon et al.'s CNN model for EEG-based BCI.

In this study, the proposed CNN-based model was trained using the data from different participants, excluding the data from the test participant. Although this study focused only on the feasibility of implementing subject-independent BCIs, the classification accuracy could be further improved by adopting a fine-tuning technique (Bengio, 2012; Anderson et al., 2016) with a small portion of the test subject's data. The fine-tuning technique has shown promising results, particularly when a deep learning model needs to be trained using only a small number of datasets. If this "few-training" approach could dramatically increase the classification accuracy of the fNIRS-based BCI, then just a few minute training sessions before the use of the BCI system would be manageable. This would be one of the promising areas we would like to investigate in our future studies.

In this study, the proposed CNN-based approach has demonstrated its potential to be used to implement a practical subject-independent fNIRS-based BCI; however, we believe that there is still room for improvement in future studies. First, the proposed deep learning approach is based on CNNs, but there are other promising neural network models—such as long short-term memory (LSTM)—which are known to be particularly effective for dealing with time-series data. Asgher et al. (2020) reported that the deep learning framework based on LSTM outperformed conventional machine learning and CNN-based algorithms in the assessment of cognitive and mental workload using fNIRS. Therefore, it would be worthwhile to compare the performance of various deep learning approaches in the implementation of subject-independent fNIRS-based BCI. In addition, we used raw fNIRS data without any particular feature extraction method except for band-pass filtering and Z-score normalization as the input tensor of the CNN model.

Furthermore, investigating the feasibility of new forms of input tensors (e.g., adjacency matrix of functional connectivity network) to implement a subject-independent fNIRS-based BCI would be an interesting research topic.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://doi.org/10.6084/m9.figshare.9198932.v1.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board Committee of Hanyang University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JK planned the study and analyzed the data. C-HI supervised the study. Both authors wrote and reviewed the manuscript.

FUNDING

This work was supported by the Institute for Information & Communications Technology Promotion (IITP) funded by the Korean Government, Ministry of Science and ICT (MSIT), under Grant 2017-0-00432 and Grant 2020-0-01373.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2021.646915/full#supplementary-material

REFERENCES

Anderson, A., Shaffer, K., Yankov, A., Corley, C. D., and Hodas, N. O. (2016). Beyond fine tuning: a modular approach to learning on small data. *arXiv preprint arXiv:*1611.01714.

Asgher, U., Khalil, K., Khan, M. J., Ahmad, R., Butt, S. I., Ayaz, Y., et al. (2020). Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain–computer interface. *Front. Neurosci.* 14:584. doi: 10.3389/fnins.2020.00584

Bengio, Y. (2012). "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning* (Edinburgh), 17–36.

Coyle, S. M., Ward, T. E., and Markham, C. M. (2007). Brain–computer interface using a simplified functional near-infrared spectroscopy system. *J. Neural Eng.* 4:219. doi: 10.1088/1741-2560/4/3/007

Daly, J. J., and Wolpaw, J. R. (2008). Brain–computer interfaces in neurological rehabilitation. *Lancet Neurol.* 7, 1032–1043. doi: 10.1016/S1474-4422(08)70223-0

Erkan, E., and Akbaba, M. (2018). A study on performance increasing in SSVEP based BCI application. *Eng. Sci. Technol.* 21, 421–427. doi: 10.1016/j.jestch.2018.04.002

Fazli, S., Grozea, C., Danóczy, M., Blankertz, B., Popescu, F., and Müller, K.-R. (2009). "Subject independent EEG-based BCI decoding," in *Advances in Neural Information Processing Systems* (Vancouver, BC), 513–521.

Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K.-R., et al. (2012). Enhanced performance by a hybrid NIRS–EEG brain computer interface. *Neuroimage* 59, 519–529. doi: 10.1016/j.neuroimage.2011.07.084

Ferrari, M., and Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *Neuroimage* 63, 921–935. doi: 10.1016/j.neuroimage.2012.03.049

Holper, L., Kobashi, N., Kiper, D., Scholkmann, F., Wolf, M., and Eng, K. (2012). Trial-to-trial variability differentiates motor imagery during observation between low vs. high responders: a functional near-infrared spectroscopy study. *Behav. Brain Res.* 229, 29–40. doi: 10.1016/j.bbr.2011.12.038

Hong, K.-S., Ghafoor, U., and Khan, M. J. (2020). Brain–machine interfaces using functional near-infrared spectroscopy: a review. *Artif. Life Robot.* 25, 204–218. doi: 10.1007/s10015-020-00592-9

Hong, K.-S., and Khan, M. J. (2017). Hybrid brain–computer interface techniques for improved classification accuracy and increased number of commands: a review. *Front. Neurorobot.* 11:35. doi: 10.3389/fnbot.2017.00035

Hu, X.-S., Hong, K.-S., and Ge, S. S. (2013). Reduction of trial-to-trial variability in functional near-infrared spectroscopy signals by accounting for resting-state functional connectivity. *J. Biomed. Opt.* 18:017003. doi: 10.1117/1.JBO.18.1.017003

Hwang, H.-J., Kim, S., Choi, S., and Im, C.-H. (2013). EEG-based brain-computer interfaces: a thorough literature survey. *Int. J. Hum. Comput. Interact.* 29, 814–826. doi: 10.1080/10447318.2013.780869

Jayaram, V., Alamgir, M., Altun, Y., Scholkopf, B., and Grosse-Wentrup, M. (2016). Transfer learning in brain-computer interfaces. *IEEE Comput. Intell. Mag.* 11, 20–31. doi: 10.1109/MCI.2015.2501545

Joadder, M. A., Siuly, S., Kabir, E., Wang, H., and Zhang, Y. (2019). A new design of mental state classification for subject independent BCI systems. *IRBM* 40, 297–305. doi: 10.1016/j.irbm.2019.05.004

Kazuki, Y., and Tsunashima, H. (2014). "Development of portable brain-computer interface using NIRS," in *2014 UKACC International Conference on Control (CONTROL)* (Loughborough), 702–707. doi: 10.1109/CONTROL.2014.6915225

Khan, M. J., Hong, M. J., and Hong, K.-S. (2014). Decoding of four movement directions using hybrid NIRS-EEG brain-computer interface. *Front. Hum. Neurosci.* 8:244. doi: 10.3389/fnhum.2014.00244

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint] arXiv:*1412.6980.

Kwon, J., and Im, C.-H. (2020). performance improvement of near-infrared spectroscopy-based brain-computer interfaces using transcranial near-infrared photobiomodulation with the same device. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 28, 2608–2614. doi: 10.1109/TNSRE.2020.3030639

Kwon, J., Shin, J., and Im, C.-H. (2020a). Toward a compact hybrid brain-computer interface (BCI): performance evaluation of multi-class hybrid EEG-fNIRS BCIs with limited number of channels. *PLoS ONE* 15:e0230491. doi: 10.1371/journal.pone.0230491

Kwon, O.-Y., Lee, M.-H., Guan, C., and Lee, S.-W. (2020b). Subject-independent brain-computer interfaces based on deep convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 31, 3839–3852. doi: 10.1109/TNNLS.2019.2946869

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural. Eng.* 15:056013. doi: 10.1088/1741-2552/aace8c

Liu, H., Brock, A., Simonyan, K., and Le, Q. V. (2020). Evolving normalization-activation layers. *arXiv [Preprint] arXiv:*2004.02967.

Luo, Y., and Lu, B.-L. (2018). "EEG data augmentation for emotion recognition using a conditional wasserstein GAN," in *Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Honolulu, Hi), 2535–2538. doi: 10.1109/EMBC.2018.8512865

Matcher, S. J., Elwell, C. E., Cooper, C. E., Cope, M., and Delpy, D. T. (1995). Performance comparison of several published tissue near-infrared spectroscopy algorithms. *Anal. Biochem.* 227, 54–68. doi: 10.1006/abio.1995.1252

Mellinger, J., Schalk, G., Braun, C., Preissl, H., Rosenstiel, W., Birbaumer, N., et al. (2007). An MEG-based brain–computer interface (BCI). *Neuroimage* 36, 581–593. doi: 10.1016/j.neuroimage.2007.03.019

Naseer, N., and Hong, K.-S. (2013). Classification of functional near-infrared spectroscopy signals corresponding to the right-and left-wrist motor imagery for development of a brain–computer interface. *Neurosci. Lett.* 553, 84–89. doi: 10.1016/j.neulet.2013.08.021

Naseer, N., and Hong, K.-S. (2015). fNIRS-based brain-computer interfaces: a review. *Front. Hum. Neurosci.* 9:3. doi: 10.3389/fnhum.2015.00003

Ravi, A., Beni, N. H., Manuel, J., and Jiang, N. (2020). Comparing user-dependent and user-independent training of CNN for SSVEP BCI. *J. Neural Eng.* 17:026028. doi: 10.1088/1741-2552/ab6a67

Schudlo, L. C., and Chau, T. (2015). Towards a ternary NIRS-BCI: single-trial classification of verbal fluency task, Stroop task and unconstrained rest. *J. Neural Eng.* 12:066008. doi: 10.1088/1741-2560/12/6/066008

Shaheen, F., Verma, B., and Asafuddoula, M. (2016). "Impact of automatic feature extraction in deep learning architecture," in *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (Gold Coast, QLD), 1–8. doi: 10.1109/DICTA.2016.7797053

Shin, J., and Im, C.-H. (2020). performance improvement of near-infrared spectroscopy-based brain-computer interface using regularized linear discriminant analysis ensemble classifier based on bootstrap aggregating. *Front. Neurosci.* 14:168. doi: 10.3389/fnins.2020.00168

Shin, J., Kwon, J., Choi, J., and Im, C.-H. (2017a). Performance enhancement of a brain-computer interface using high-density multi-distance NIRS. *Sci. Rep.* 7:16545. doi: 10.1038/s41598-017-16639-0

Shin, J., Kwon, J., Choi, J., and Im, C.-H. (2018a). Ternary near-infrared spectroscopy brain-computer interface with increased information transfer rate using prefrontal hemodynamic changes during mental arithmetic, breath-holding, and idle state. *IEEE Access* 6, 19491–19498. doi: 10.1109/ACCESS.2018.2822238

Shin, J., Kwon, J., and Im, C.-H. (2018b). A ternary hybrid EEG-NIRS brain-computer interface for the classification of brain activation patterns during mental arithmetic, motor imagery, and idle state. *Front. Neuroinform.* 12:5. doi: 10.3389/fninf.2018.00005

Shin, J., Müller, K.-R., Schmitz, C. H., Kim, D.-W., and Hwang, H.-J. (2017b). Evaluation of a compact hybrid brain-computer interface system. *Biomed. Res. Int.* 2017:6820482. doi: 10.1155/2017/6820482

Shin, J., von Lühmann, A., Blankertz, B., Kim, D.-W., Mehnert, J., Jeong, J., et al. (2018c). "Open access repository for hybrid EEG-NIRS data," in *2018 6th International Conference on Brain-Computer Interface (BCI)* (GangWon), 1–4. doi: 10.1109/IWW-BCI.2018.8311523

Sitaram, R., Caria, A., Veit, R., Gaber, T., Rota, G., Kuebler, A., et al. (2007). fMRI brain-computer interface: a tool for neuroscientific research and treatment. *Comput. Intell. Neurosci.* 2007:025487. doi: 10.1155/2007/25487

Vani, S., and Rao, T. M. (2019). "An experimental approach towards the performance assessment of various optimizers on convolutional neural network," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (Tirunelveli), 331–336. doi: 10.1109/ICOEI.2019.8862686

Vidaurre, C., and Blankertz, B. (2010). Towards a cure for BCI illiteracy. *Brain Topogr.* 23, 194–198. doi: 10.1007/s10548-009-0121-6

von Lühmann, A., Ortega-Martinez, A., Boas, D. A., and Yücel, M. A. (2020). Using the general linear model to improve performance in fNIRS single trial analysis and classification: a perspective. *Front. Hum. Neurosci.* 14:30. doi: 10.3389/fnhum.2020.00030

Wang, P., Lu, J., Zhang, B., and Tang, Z. (2015). "A review on transfer learning for brain-computer interface classification," in *2015 5th International Conference on Information Science and Technology (ICIST)* (Changsha), 315–322. doi: 10.1109/ICIST.2015.7288989

Waytowich, N. R., Faller, J., Garcia, J. O., Vettel, J. M., and Sajda, P. (2016). "Unsupervised adaptive transfer learning for steady-state visual evoked potential brain-computer interfaces," in *Proceedings of IEEE International Conference on Systems, Man and Cybernetics (SMC)* (Budapest), 004135–004140. doi: 10.1109/SMC.2016.7844880

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clin. Neurophysiol.* 113, 767–791. doi: 10.1016/S1388-2457(02)00057-3

Xu, L., Xu, M., Ke, Y., An, X., Liu, S., and Ming, D. (2020). Cross-dataset variability problem in EEG decoding with deep learning. *Front. Hum. Neurosci.* 14:103. doi: 10.3389/fnhum.2020.00103

Yuan, P., Chen, X., Wang, Y., Gao, X., and Gao, S. (2015). Enhancing performances of SSVEP-based brain–computer interfaces via exploiting inter-subject information. *J. Neural Eng.* 12:046006. doi: 10.1088/1741-2560/12/4/046006

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Human Neuroscience

REVIEW
published: 28 May 2021
doi: 10.3389/fnhum.2021.643386

A Survey on Deep Learning-Based Short/Zero-Calibration Approaches for EEG-Based Brain–Computer Interfaces

Wonjun Ko[1†], *Eunjin Jeon*[1†], *Seungwoo Jeong*[2], *Jaeun Phyo*[1] *and Heung-Il Suk*[1,2*]

[1] *Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea,* [2] *Department of Artificial Intelligence, Korea University, Seoul, South Korea*

OPEN ACCESS

Edited by:
Hong Gi Yeom,
Chosun University, South Korea

Reviewed by:
Carmen Vidaurre,
Public University of Navarre, Spain
Dalin Zhang,
Aalborg University, Denmark

***Correspondence:**
Heung-Il Suk
hisuk@korea.ac.kr

[†] *These authors have contributed equally to this work and share first authorship*

Specialty section:
This article was submitted to Brain-Computer Interfaces, a section of the journal Frontiers in Human Neuroscience

Received: *18 December 2020*
Accepted: *27 April 2021*
Published: *28 May 2021*

Citation:
Ko W, Jeon E, Jeong S, Phyo J and Suk H-I (2021) A Survey on Deep Learning-Based Short/Zero-Calibration Approaches for EEG-Based Brain–Computer Interfaces. Front. Hum. Neurosci. 15:643386. doi: 10.3389/fnhum.2021.643386

Brain–computer interfaces (BCIs) utilizing machine learning techniques are an emerging technology that enables a communication pathway between a user and an external system, such as a computer. Owing to its practicality, electroencephalography (EEG) is one of the most widely used measurements for BCI. However, EEG has complex patterns and EEG-based BCIs mostly involve a cost/time-consuming calibration phase; thus, acquiring sufficient EEG data is rarely possible. Recently, deep learning (DL) has had a theoretical/practical impact on BCI research because of its use in learning representations of complex patterns inherent in EEG. Moreover, algorithmic advances in DL facilitate short/zero-calibration in BCI, thereby suppressing the data acquisition phase. Those advancements include data augmentation (DA), increasing the number of training samples without acquiring additional data, and transfer learning (TL), taking advantage of representative knowledge obtained from one dataset to address the so-called data insufficiency problem in other datasets. In this study, we review DL-based short/zero-calibration methods for BCI. Further, we elaborate methodological/algorithmic trends, highlight intriguing approaches in the literature, and discuss directions for further research. In particular, we search for *generative model*-based and *geometric manipulation*-based DA methods. Additionally, we categorize TL techniques in DL-based BCIs into *explicit* and *implicit* methods. Our systematization reveals advances in the DA and TL methods. Among the studies reviewed herein, ∼45% of DA studies used generative model-based techniques, whereas ∼45% of TL studies used explicit knowledge transferring strategy. Moreover, based on our literature review, we recommend an appropriate DA strategy for DL-based BCIs and discuss trends of TLs used in DL-based BCIs.

Keywords: brain–computer interface, electroencephalography, deep learning, data augmentation, transfer learning

1. INTRODUCTION

1.1. Overview

Brain–computer interfaces (BCIs) (Dornhege et al., 2007; Lotte et al., 2018; Roy et al., 2019) provide communication pathways between a user and an external device (e.g., robotic arm, speller, seizure alarm system, etc.) by measuring and analyzing brain signals. Owing to its practicality, non-invasive BCIs based on electroencephalography (EEG) are commonly exploited (Suk and Lee, 2012; Roy et al., 2019). The *real-world* impact of BCIs is promising because they can identify intention-reflected brain activities. In the past decade, human-centered BCIs, such as those in mental fatigue detection tasks (Binias et al., 2020; Ko et al., 2020b), emotion recognition (Qing et al., 2019), and controlling exoskeletons (Lee et al., 2017) have shed light on the success of improving human ability. An *active* BCI (Fahimi et al., 2020) recognizes complex patterns from EEG spontaneously caused by a user's intention independent of external stimuli, and a *reactive* BCI (Won et al., 2019) identifies brain activities in reaction to external events. A *Passive* BCI (Ko et al., 2020b) is exploited to acquire implicit information of a user's cognitive status without any voluntary control.

EEG-based BCIs generally benefit from machine learning techniques (Lotte et al., 2018). Specifically, EEG features of various paradigms are crafted using machine learning algorithms, such as *common spatial pattern* (CSP) (Ramoser et al., 2000) and *canonical correlation analysis* (Lin et al., 2006), including preprocessing techniques. Further, the extracted EEG features are discriminated by successful machine learning algorithms used in classification tasks, e.g., *support vector machines* (Bishop, 2006). These feature extraction and classification algorithms have shown their ability in EEG-based BCIs but have also been limited because of the lack of representation power for complex EEG patterns (Schirrmeister et al., 2017). In addition, since feature extractions using these machine learning methods are widely performed in a *hand-crafted manner* (Lawhern et al., 2018), it is difficult for *unskilled personnel* to develop a novel BCI framework.

Deep learning (DL) methodologies (Schirrmeister et al., 2017; Sakhavi et al., 2018; Zhang et al., 2019c; Ko et al., 2020a) have become the core of BCI research owing to their representational power for complex patterns in EEG. Specifically, DL significantly simplifies the EEG analysis pipeline (Lawhern et al., 2018) by learning preprocessing, feature representation, and decision-making in an *end-to-end* manner. Furthermore, architectural developments in DL have been very successful in representing complicated patterns. DL learns the hierarchical representations of input data through stacked non-linear transformations (LeCun et al., 2015). In DL, stacked layers apply a linear transformation to the input, and the transformation is fed through non-linear activation. The parameters of these stacked layers are automatically learned by exploiting an *objective* function. In the machine learning field, various DL architectures have been developed. Examples include convolutional neural networks (CNNs), which have been well-suited for *structural* pattern representation and are thus widely used to learn *spatio-spectral-temporal* patterns of EEG (Schirrmeister et al., 2017; Ko et al., 2020a). Additionally, owing to the ability of sequential data modeling, recurrent neural networks and their variants, e.g., long short-term memory (LSTM) networks, have achieved considerable success in the temporal embedding of EEG (Zhang et al., 2019c; Freer and Yang, 2020). Moreover, recent research has shown interest in hybrid forms of recurrent layers and convolutional layers (Ko et al., 2018; Zhang et al., 2019a).

Although DL has been demonstrated to be a powerful tool in EEG analysis, there are some limitations. First, typically available EEG datasets contain substantially fewer training samples than do other datasets that are commonly used in DL-based *computer vision* or *natural language processing* task development. However, EEG acquisition is an expensive and time-consuming task. Further, data accessibility is often hindered because of privacy concerns, especially in the clinical domain. Thus, collecting large amounts of training EEG samples for DL training is rarely possible. Owing to the nature of EEG properties, such as low signal-to-noise ratio and inter/intra-variability (Jayaram et al., 2016), DL-based BCIs are rarely trained only with a different user's or even multiple users' training EEG samples.

To address the aforementioned problems, recent research has focused on *data augmentation* (DA) (Luo and Lu, 2018; Zhang et al., 2019d; Fahimi et al., 2020) and *transfer learning* (TL) (Jayaram et al., 2016; Kwon et al., 2019; Jeon et al., 2020). The use of DL has shown the possibility of synthesizing high-dimensional image data (Goodfellow et al., 2014), audio data (Donahue et al., 2019), and EEG data (Hartmann et al., 2018). Further, traditional DA techniques used in DL fields, such as image rotation have demonstrated their own efficiency and effectiveness (Simonyan and Zisserman, 2014). By exploiting these DA techniques, DL-based BCIs have improved the performance with a short-calibration phase producing little data (Fahimi et al., 2020; Zhang et al., 2020b). In terms of TL, DL has also been widely used to suppress the training EEG data acquisition phase (Chai et al., 2016; Jeon et al., 2020; Tang and Zhang, 2020). In particular, DL-based BCIs can be designed in a short/zero-calibration manner by appropriately conducting 2-fold TL strategies, i.e., explicit TL and implicit TL.

Overall, several DL methods have been proven to improve existing EEG processing techniques. The end-to-end strategy allows DL to simply learn existing EEG analysis pipelines, reducing paradigm-specific processing and feature extraction. Objective function-based automatic learning requires only raw or minimally preprocessed EEG data. The feature representation of DL can also be more effective and richer than features engineered by humans. Moreover, DL can pave the way for methodological advances in EEG analysis, such as generative modeling (Goodfellow et al., 2014) and knowledge transfer (Jayaram et al., 2016) to handle the lack of EEG data problems and the data variability issue.

1.2. Our Contributions

In this study, we review DL-based BCI studies that mostly focused on suppressing the EEG calibration phase. Unlike recent survey papers for EEG-based BCIs that are mostly focused on introducing machine learning/DL algorithms for

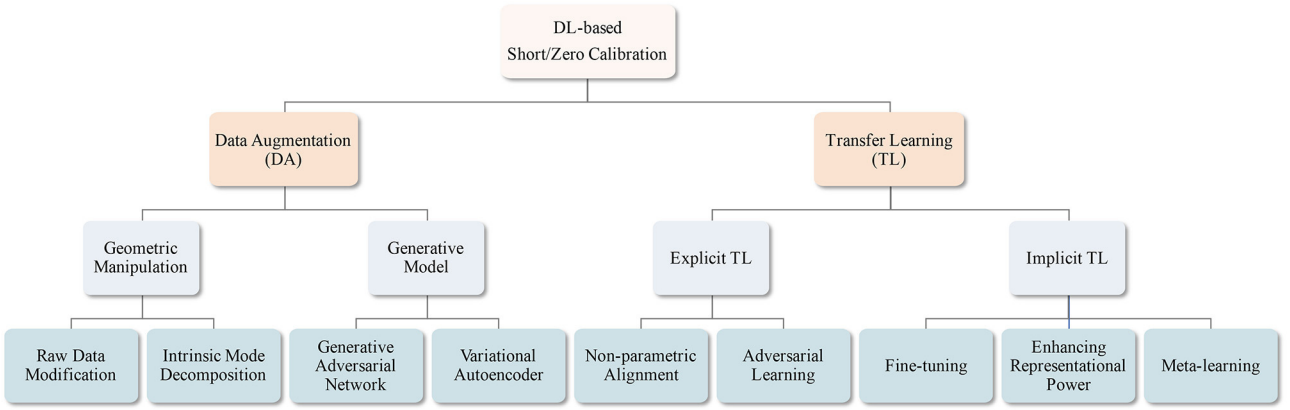

FIGURE 1 | Overview of DL-based short/zero calibration approaches.

BCIs (Lotte et al., 2018; Craik et al., 2019; Zhang et al., 2020d), summarizing EEG analysis studies (Roy et al., 2019), providing comprehensive information on EEG-based BCIs, including sensing technology and healthcare systems (Gu et al., 2020), and surveying application of machine learning/DL-based TLs (Zhang et al., 2020c), our review aims to address short-/zero-calibration techniques for EEG-based BCIs. In detail, we categorize these studies into two different groups, based on the manner of increasing the number of training samples: (i) manipulating the given training data without using an additional one and (ii) exploiting other subjects/sessions' EEG samples. Specifically, (i) is further categorized into generative model-based and geometric manipulation-based methods, and (ii) is classified into explicit and implicit knowledge transfer. In the case of (i), 45% of the studies proposed generative model (Goodfellow et al., 2014; Kingma and Welling, 2014)-based DA methodologies, whereas 45% of the case of (ii) developed explicit knowledge transfer strategies. Further, we recommend a training technique for DL-based BCI models with a generative model-based DA based on our literature review and discuss trends of recent knowledge transfer methods. We summarize the taxonomy of our review in **Figure 1**.

The remainder of this paper is organized as follows. In section 2, we describe DL methods to augment training samples and review the methods proposed in various BCI studies. In section 3, we discuss and review DL methods for transferring knowledge of other subjects/sessions' samples in BCIs. For both sections 2 and 3, we summarize our review in **Tables 1**–**4**. Section 4 presents our discussion and recommendations for DA-based short-calibration techniques to develop a new DL-based BCI system. Further, section 4 details trends of recent knowledge transfer methods in DL research. Finally, section 5 provides concluding statements.

2. ADVANCES IN DATA AUGMENTATION

2.1. What Is Data Augmentation?

Recently, DL-based BCIs have shown promising results in both active and passive BCI applications. However, a sufficient number of training EEG samples are required to train DL-based BCIs to avoid *overfitting* problems. DA is one way to address the data insufficiency problem. Specifically, DA increases the amount of data by synthesizing samples from the existing training data. Thus, DL models cannot overfit all samples and are forced to generalize well. Commonly, in the DL-based computer vision field (Simonyan and Zisserman, 2014; He et al., 2016), image samples are rotated/shifted/rescaled/flipped/sheared/stretched to be augmented. Further, generating extra samples from the existing ones by exploiting DL-based generative models is one of the most important strategies in DA. Because DA techniques help reduce the necessity of acquiring new EEG data, which is hindered by its cost-/time-consuming properties (Hartmann et al., 2018; Freer and Yang, 2020), they have gained significant attention in the BCI field. Here, we review the DA methodologies used for improving the performance of DL-based BCIs.

2.2. Challenges in Data Augmentation

A major difference between EEG data and image data is *translational invariance*, a property that an output value is invariant with respect to positional transformations of an input. Common computer vision tasks have to solve the problems of viewpoint, lightness, background, scale, etc. Therefore, in the computer vision field, widely used DA techniques, such as translation and rotation, are designed to improve the translational invariance of the training dataset. Further, those computer vision methods mostly use CNNs that exploit two-dimensional (height × width) and/or three-dimensional (height × width × depth) convolutional kernels. A CNN learns local features by sharing kernel weights, thus translational invariance is naturally followed. In other words, it represents patterns regardless of the position of the object in an input image. In contrast, for raw EEG analysis, DL-based BCIs (Schirrmeister et al., 2017; Lawhern et al., 2018; Ko et al., 2020a) are widely designed to extract features of EEG by using one-dimensional (temporal or spatial) convolution kernels. Furthermore, retraining the *spatio-spectral-temporal* information of raw EEG is also important for these DL-based BCIs. Hence, commonly used DA methods in computer vision tasks, e.g., rotating, cropping, scaling, are rarely applicable to DL-based

BCIs, because those methods may harm the spectro-spatio-temporal information in EEG signals. In other words, we cannot augment raw EEG signals using simple techniques. Moreover, labeling augmented EEG samples via geometric manipulation is also difficult. In this regard, many DL methods for DL-based BCIs apply geometric manipulation to *spectrogram* images estimated from raw EEGs (Shovon et al., 2019; Zhang et al., 2020b), or cropped EEGs using a sliding window (Schirrmeister et al., 2017; Ko et al., 2018; Majidov and Whangbo, 2019). Meanwhile, other DA methods for DL-based BCIs (Hartmann et al., 2018; Luo and Lu, 2018; Hwang et al., 2019) have focused on synthesizing EEG signals from existing ones. These works generally introduce DL-based generative models (Goodfellow et al., 2014; Kingma and Welling, 2014)-based augmenting methods. However, as synthesized signals are not sufficiently realistic to be used as training samples, many studies have tried to improve the generation ability, i.e., the quality of augmented samples by regularizing their generative models (Arjovsky et al., 2017).

2.3. Approaches in Data Augmentation

DA methods in BCI can be categorized into two groups—geometric manipulation-based and deep generative model-based methods—depending on modifying existing samples and synthesizing novel training samples with an additional deep generative model, respectively. First, as the direct application of data modification used in computer vision to DL-based BCIs is somewhat difficult, Lotte et al. (2018) showed that geometric manipulation-based EEG DA can improve the BCI performance of linear machine learning models. Inspired by these intriguing results, in case of the geometric manipulation-based group, it was hypothesized that traditional DA techniques used in computer vision can be extended to DL-based BCIs. Further, some pioneering studies (Liu et al., 2016; Zhang et al., 2019d) have attempted to learn the intrinsic mode, i.e., subspaces of the training data, and controlled them to generate new data. Second, generative model-based approaches have gained attention from the BCI society with algorithmic advancements of generative models. DL-based generative model *explicitly*, e.g., *variational autoencoder* (VAE) (Kingma and Welling, 2014), or *implicitly*, e.g., *generative adversarial network* (GAN) (Goodfellow et al., 2014), learn the distribution of input data as well as output result. Generation of synthetic data in the input data space is possible by sampling from the learned distribution. The size of the training dataset can be considerably expanded by adopting deep generative model for BCI methods, using a limited number of samples, i.e., less than hundreds (Hartmann et al., 2018; Roy et al., 2020). In addition, some studies (Ko et al., 2019; Panwar et al., 2019a) use *min-max game*-based training algorithms, a core of GAN for DL-based BCI model training, thereby improving the BCI performance even with fewer training samples.

2.3.1. Geometric Manipulation-Based Data Augmentation Methods

Geometric manipulation is one of the most simple and efficient DA ways. It modifies data without additional learning, hence is applicable directly and intuitively. Geometric manipulation-based DA methods show promising results for performance improvements in several computer vision tasks (Simonyan and Zisserman, 2014; He et al., 2016); thus, many attempts have been made to apply similar approaches to EEG data. In this section, we review many interesting DL-based BCI methods that take traditional DA strategies developed in computer vision tasks, such as geometric transformation (Schirrmeister et al., 2017), noise addition (Parvan et al., 2019), and mixup (Kostas and Rudzicz, 2020). Some studies used the segmentation and recombination approach for DA (Freer and Yang, 2020), whereas other studies learned the intrinsic modes of EEG data and generated novel samples by modifying the learned modes (Liu et al., 2016).

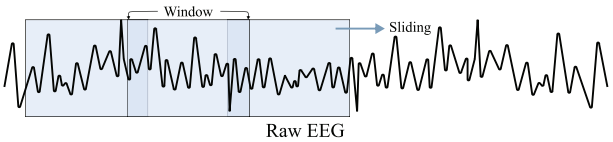

FIGURE 2 | Cropping strategy using a sliding window (Schirrmeister et al., 2017; Ko et al., 2018). For a raw EEG signal, a sliding window with a length shorter than that of the signal moves on EEG with a predefined stride. Subsequently, the window crops a part of signal for augmentation.

2.3.1.1. Raw Data Modification

A straightforward means of raw data modification is *geometric transformation*, which includes rotating, shifting, flipping, lightening, zooming, and cropping. As geometric transformation is easily applicable, many DL-based BCI methods use it as DA, based on Lotte et al. (2018)'s pioneering approaches, e.g., segmentation and recombination of EEG signals. For instance, Zhang et al. (2020b) performed three different geometric transformation-based DAs. First, Zhang et al. rotated spectrogram images of EEG signals estimated by using *short-time Fourier transform* (STFT). Further, they shifted the spectrogram and filled the remaining space with random noise and finally, perturbed the RGB values of the STFT image in the color space. Shovon et al. (2019) also performed DA by rotating, flipping, zooming, and brightening spectrogram images of motor imagery EEG signals. Moreover, as depicted in **Figure 2**, Schirrmeister et al. (2017), Ko et al. (2018), and Majidov and Whangbo (2019) used similar approaches to augment raw motor imagery EEG samples; they cropped EEG signals from an EEG epoch by using a sliding window having a shorter time length than that of the epoch. Freer and Yang (2020) performed flipping raw motor imagery samples to augment their training data. Furthermore, Mousavi et al. (2019) conducted a sliding window-based DA technique to increase the number of training EEG samples for sleep stage recognition. Supratak and Guo (2020) also focused on the sleep stage classification task but augmented the training dataset using the shifting technique. Finally, Sakai et al. (2017) used shifting to augment their cognition classification task, classifying EEG signals acquired at *motivated* status and *unmotivated* statuses.

Similar to the geometric transformation method, a *noise addition*-based DA technique has also been widely used in many successful DL-based computer vision studies (Simonyan and Zisserman, 2014; He et al., 2016). The noise addition

facilitates DA by adding randomly sampled noise values to the original samples. In terms of DA for EEG, Zhang et al. (2020b) augmented spectrogram images of motor imagery EEG by adding Gaussian noise. Similarly, Parvan et al. (2019) and Freer and Yang (2020) performed noise addition using uniform distribution and Gaussian distribution to augment raw motor imagery EEG samples, respectively. Finally, Wang F. et al. (2018) added Gaussian noise to differential entropy values estimated from emotion EEG signals for the DA. Interestingly, all DL-based BCIs that exploit the noise addition method use Gaussian distribution to sample noise, with a mean value of 0 and a small standard deviation value, e.g., 0.01 or 0.001.

Another intuitive geometric manipulation is segmenting and recombining the EEG samples (Lotte et al., 2018). There are two methods for the segmentation and recombination methods. First, let us denote the ith epoch of EEG samples as $\mathbf{x}^i$. Then, with the predefined segmentation hyperparameter, T, the given trial is segmented to $\mathbf{x}_1^i, \mathbf{x}_2^i, ..., \mathbf{x}_T^i$. Finally, these segments are recombined with other segments from the other EEG epoch, i.e., $\mathbf{x}^j, \forall i \neq j$. Thus, the augmented new sample, $\mathbf{x}_{\text{aug}}$, can be made as, for instance, $\mathbf{x}_{\text{aug}} = \text{Concat}(\mathbf{x}_1^1, \mathbf{x}_2^6, ..., \mathbf{x}_T^4)$, where Concat denotes a concatenation operation. Refer to **Figure 3** for the concept of temporal signal segmentation and recombination. The other method includes spectral transformation, such as STFT. In this case, EEG samples are mapped into the spectro-temporal domain by a transformation method, segmented, and recombined. Subsequently, the augmented combinations of spectrogram segments are mapped into the temporal domain using an inverse transformation method. Recently, Cho et al. (2020), Dai et al. (2020), Freer and Yang (2020), and Huang et al. (2020) used segmentation and recombination in a temporal manner, i.e., without STFT, to augment their raw motor imagery EEG. Additionally, Huang et al. performed the same augmentation method in a spectro-temporal manner. Specifically, Huang et al. swapped entire segments in a specific frequency band of two randomly sampled EEG signals. Further, Fahimi et al. (2020) performed both segmentation and recombination methods, i.e., both temporal and spectral methods, to augment the motor execution EEG samples. Zhao X. et al. (2020) also effectively acquired artificial ictal EEG samples with a *discrete cosine transform* (DCT)-based spectral transformation. Finally, Fan et al. (2020) and Supratak and Guo (2020) performed the temporal segmentation and recombination-based DA technique to increase the training data for the sleep stage classification.

The *synthetic minority oversampling technique* (SMOTE) (Chawla et al., 2002) is one of the most widely used oversampling techniques to address the class imbalance problem in machine learning fields. Let us assume that A is a minority class set and its elements are $\mathbf{x}_i \in A$. Subsequently, for each sample $\mathbf{x}_i$, we obtain its k-nearest neighbors, $\mathbf{x}_i^{(k)}$, with some distance metrics, for example, Euclidean distance. Then, a new augmented sample is acquired by using $\mathbf{x}_{i,\text{aug}} = \mathbf{x}_i + \epsilon|\mathbf{x}_i - \mathbf{x}_i^{(k)}|$ for $\forall k$, where $\epsilon \sim \text{Uniform}(0, 1)$ denotes a random number drawn from a uniform distribution. Owing to its simplicity and power, some DL-based BCI studies have used SMOTE to augment the imbalanced training data. Lee T. et al. (2020) oversampled raw *target* class EEG samples that generally belong to the minority class in the *event-related potential* (ERP) paradigm. Similarly, Romaissa et al. (2019) used SMOTE (Chawla et al., 2002) to oversample ictal EEG signals. Interestingly, Romaissa et al. first extracted the spectral features of EEG signals and performed SMOTE on the spectral domain. Sun et al. (2019) also oversampled minor epochs in the sleep stage classification by conducting SMOTE on hand-crafted features.

In addition, some studies amplified given EEG samples to augment them. Amplification-based DA can be performed by using $\mathbf{x}_{\text{aug}} = (1 \pm C)\mathbf{x}$, where $C \in \mathbb{R}$ is a predefined amplification-control hyperparameter. Freer and Yang (2020) amplified raw motor imagery samples with $C = 0.02, 0.05, 0.1$, and 0.2. Furthermore, Sakai et al. (2017) amplified EEG signals with $C = 0.1$. Sakai et al. established a 2-fold strategy of amplifying (i) all-time data and (ii) near-peak data. In the second strategy, Sakai et al. only multiplied $(1 \pm C)$ to near-peak data.

Mixup (Zhang et al., 2018b) is a recently proposed DA technique for computer vision tasks. For two given training samples $\mathbf{x}_i$ and $\mathbf{x}_j$, $\forall i \neq j$ with labels $\mathbf{y}_i$ and $\mathbf{y}_j$, respectively, an augmented sample is then estimated by using $\mathbf{x}_{\text{aug}} = \lambda\mathbf{x}_i + (1 - \lambda)\mathbf{x}_j$, and its label is defined as $\mathbf{y}_{\text{aug}} = \lambda\mathbf{y}_i + (1 - \lambda)\mathbf{y}_j$, where $\lambda \in [0, 1]$ is a random number. In case of DL-based BCI, Kostas and Rudzicz (2020) used mixup to augment raw motor imagery/ERP/*rapid serial visual presentation* (RSVP) EEG samples and improved the BCI performance.

Raw data modification-based DA methods are easily applicable and do not require any further networks optimization. Meanwhile, because of the EEG data's spectro-spatio-temporal properties, these methods barely improve the performance and make model interpretation complicated.

2.3.1.2. Intrinsic Mode Decomposition

As EEG is a very non-stationary and non-linear time-series data, *empirical mode decomposition* (EMD) (Flandrin et al., 2004), illustrated in **Figure 4**, can be an appropriate spatio-temporal analysis method. To be specific, EEG is partitioned into *modes* called *intrinsic mode functions* (IMFs) without leaving the time domain by EMD. Similar to the segmentation and recombination, EMD-based DA first estimates IMFs of EEG signals, and IMFs are then recombined to create artificial EEG samples. Importantly, the mode of each IMF used in the DA does not overlap. Dinarès-Ferran et al. (2018) and Zhang et al. (2019d) performed EMD to acquire IMFs of motor imagery EEG samples and generated artificial samples by recombining IMFs. Kalaganis et al. (2020) created spatio-temporal graphs by using EEG signals acquired from cognitive tasks and estimated graph IMFs using EMD. Subsequently, Kalaganis et al. recombined these graph IMFs to augment the training data.

Another way to learn the intrinsic modes of the data is the *self-organizing map* (SOM) (Kohonen, 1990), which discretizes the training samples to a *map*. SOM training utilizes competitive learning. For a given training sample fed into a neural network, the Euclidean distance between each weight vector and the input data is estimated. Then, a neuron having the shortest distance is called the *best matching unit* (BMU). The weights

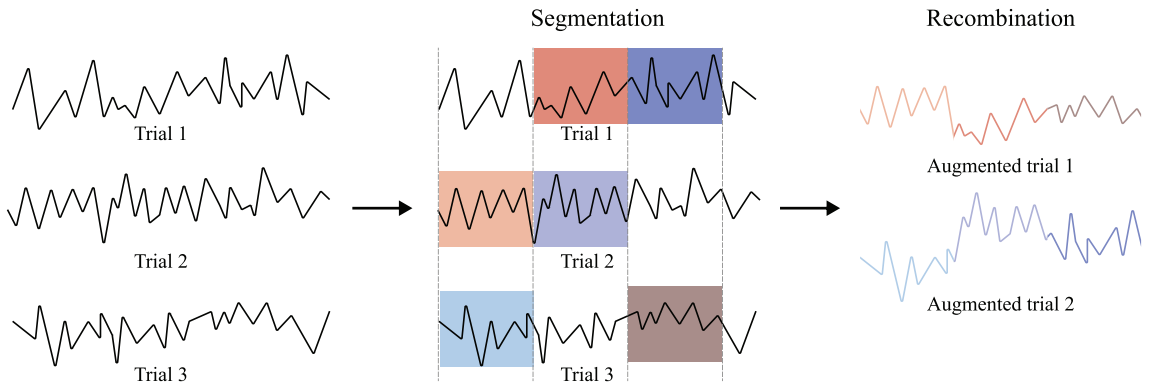

FIGURE 3 | EEG segmentation and recombination method (referred to Dai et al., 2020). EEG samples are segmented into constant lengths. The divided pieces are then randomly recombined to generate new signals.

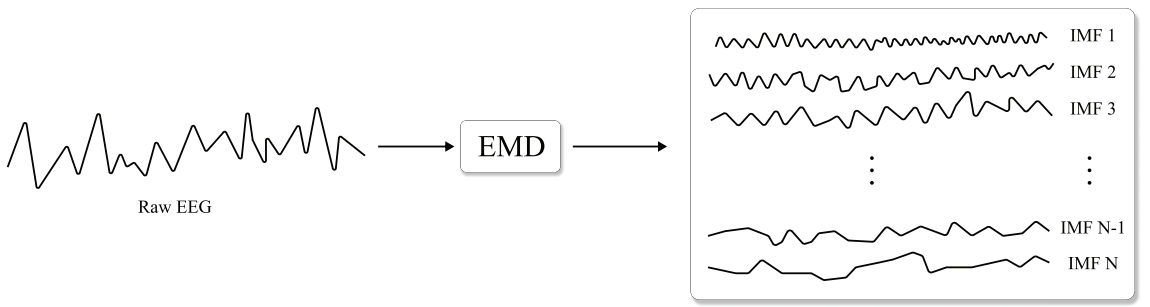

FIGURE 4 | Illustration of empirical mode decomposition (EMD) (Flandrin et al., 2004) which decomposes modes, i.e., intrinsic mode function (IMF) of the input signal (re-illustrated from Dinarès-Ferran et al., 2018).

of the BMU and neurons that are close to it in the SOM grid are adjusted to the input data. When adjusting, the magnitude of the change decreases with time and the grid-distance from the BMU. In this regard, Liu et al. (2016) applied a variant of SOM, named adaptive subspace SOM (ASSOM), trained it with predefined numbers, N, of quadratic modules and achieved N subspace representations of data $\mathbf{x}$. Finally, N numbers of synthetic samples could be obtained by inversely transforming the representations. Even though intrinsic mode decomposition-based DAs effectively learn internal modes of EEG data, they still show limitations. For instance, they introduce additional hyperparameters to be found, e.g., the number of IMFs and BMUs, thus require extra tuning phase. We summarize our review of the geometric manipulation-based DA methods in **Table 1**.

2.3.2. Generative Model-Based Data Augmentation Methods

A characteristic of generative model-based DA methods is exploiting additional DL for synthesizing training samples. Among recent successes of deep generative models, GAN (Goodfellow et al., 2014) and VAE (Kingma and Welling, 2014) demonstrate their caliber by showing practical use with sound theoretical foundations. We herein review the advances in GAN-based DA methods for BCIs (Hartmann et al., 2018; Hwang et al., 2019; Ko et al., 2019; Luo et al., 2020). These methods exploit GAN and its variants (Radford et al., 2015; Arjovsky et al., 2017; Mao et al., 2017) to learn the distribution of training samples. Those GAN-based DA methods can effectively generate artificial samples and stabilize DL-based BCI training. The autoencoder (AE) (Ballard, 1987) and VAE are also used for learning the *latent space* of the training dataset. Subsequently, some DL-based BCIs (Fahimi et al., 2020; Zhang et al., 2020b) are employed to generate artificial samples from the learned latent space, thereby augmenting the data.

2.3.2.1. Generative Adversarial Network

Recently, Goodfellow et al. (2014) proposed a DL-based generative model named GAN to learn deep representations of data distribution without extensively annotated training data. As depicted in **Figure 5**, GAN comprises two networks: a *generator* and a *discriminator*. In GAN, generator $\mathcal{G}$ tries to generate a *realistic* sample, $\mathcal{G}(\mathbf{z})$, from a latent code vector, $\mathbf{z}$. Discriminator $\mathcal{D}$ tries to discriminate the real sample, $\mathbf{x}$, from the generated one and outputs a probability of whether the input is real. To simultaneously train those two networks, i.e., the generator and

TABLE 1 | Geometric manipulation data augmentation methods.

	Approach	References	Paradigm	Summary
Raw data modification	Geometric transformation	Zhang et al., 2020b	Motor imagery	Rotated (180°), shifted, and changed RGB values of STFT images estimated from raw EEGs
		Shovon et al., 2019		Rotated (5°), flipped, zoomed, brightened (±30%) STFT images estimated from raw EEGs
		Schirrmeister et al., 2017		Cropped raw EEG using a sliding window
		Ko et al., 2018		Cropped raw EEG using a sliding window
		Majidov and Whangbo, 2019		Cropped raw EEG using a sliding window
		Freer and Yang, 2020		Flipped raw EEG
		Mousavi et al., 2019	Sleep	Cropped raw EEG using a sliding window
		Supratak and Guo, 2020		Shifted raw EEG
		Sakai et al., 2017	Cognition	Shifted raw EEG
	Noise addition	Zhang et al., 2020b	Motor imagery	Added Gaussian noise (std of 0.1)
		Freer and Yang, 2020		Used uniform noise ([−0.5, 0.5])
		Wang F. et al., 2018	Emotion	Added Gaussian noise (std of 0.001 ∼ 0.5)
	Recombination	Freer and Yang, 2020	Motor imagery	Segmented and recombined raw EEGs
		Cho et al., 2020		Segmented and recombined raw EEGs
		Dai et al., 2020		Segmented and recombined raw EEGs
		Huang et al., 2020		Segmented and recombined STFT images
		Fahimi et al., 2020	Motor	Segmented and recombined both raw EEGs and STFT images
		Zhao X. et al., 2020	Seizure	Segmented and recombined DCT images
		Fan et al., 2020	Sleep	Segmented and recombined raw EEGs; compared synthesizing qualities to other DA methods
		Supratak and Guo, 2020		Segmented and recombined raw EEGs
	SMOTE (Chawla et al., 2002)	Lee T. et al., 2020	ERP	Used borderline-SMOTE algorithm to raw EEGs
		Sun et al., 2019	Sleep	Used SMOTE algorithm to hand-crafted features
	Amplifying	Freer and Yang, 2020	Motor imagery	Amplified raw EEG ±2 ∼ 20%
		Sakai et al., 2017	Cognition	Amplified raw EEG ±10%
	Mixup (Zhang et al., 2018b)	Kostas and Rudzicz, 2020	Multi	Conducted mixup algorithm to raw EEGs; experimented TL experiments
Intrinsic mode decomposition	EMD (Flandrin et al., 2004)	Zhang et al., 2019d	Motor imagery	Estimated and recombined IMFs of raw EEGs
		Dinarès-Ferran et al., 2018		Estimated and recombined IMFs of raw EEGs
		Kalaganis et al., 2020	Cognition	Estimated and recombined IMFs of graphs estimated by raw EEGs
	SOM (Kohonen, 1990)	Liu et al., 2016	Drowsy	Conducted ASSOM algorithm

the discriminator, GAN uses a min-max objective function:

$$\max_{\mathcal{D}} \mathbb{E}_{p_{\mathbf{x}}}[\log \mathcal{D}(\mathbf{x})] + \mathbb{E}_{p_{\mathbf{z}}}[\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z})))] \quad \text{and} \quad \min_{\mathcal{G}} \mathbb{E}_{p_{\mathbf{z}}}[\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z})))], \tag{1}$$

where $p_{\mathbf{x}}$ and $p_{\mathbf{z}}$ denote the distribution of real samples and latent code, respectively. In Equation (1), the Jensen-Shannon distance (JSD) is used for estimating the distance between the real sample distribution and the generated sample distribution. Here, $\mathcal{G}$ is minimized when $\mathcal{D}(\mathcal{G}(\mathbf{z})) \rightarrow 1$, i.e., the generator tries to make realistic samples, and $\mathcal{D}$ is maximized when $\mathcal{D}(\mathbf{x}) \rightarrow 1$ and $\mathcal{D}(\mathcal{G}(\mathbf{z})) \rightarrow 0$; thus, $\mathcal{D}$ determines the real and fake samples correctly.

Based on the use of GAN (Goodfellow et al., 2014), some DL-based BCIs use GAN as the DA method. Roy et al. (2020) proposed a GAN-based motor imagery EEG augmentation method, named *MIEEG-GAN*. Roy et al. developed an LSTM-based generator and an LSTM-based discriminator to augment both raw motor imagery EEG signals and spectrum images generated by STFT. Further, Roy et al. analyzed generated samples both qualitatively and quantitatively. Similarly, Krishna et al. (2020) constructed a *gated recurrent unit* (GRU) (Chung et al., 2014)-based generator and a GRU-based discriminator with the GAN loss function, i.e., Equation (1). Thus, Krishna

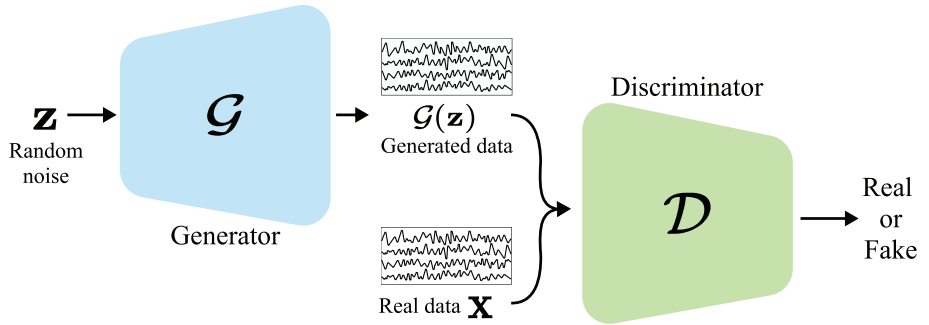

FIGURE 5 | Illustration of generative adversarial network (GAN) (Goodfellow et al., 2014). Generator $\mathcal{G}$ outputs generated data $\mathcal{G}(\mathbf{z})$ using a random noise vector $\mathbf{z}$. Then, discriminator $\mathcal{D}$ distinguishes generated data $\mathcal{G}(\mathbf{z})$ from real data $\mathbf{x}$.

et al. augmented EEG data for speech recognition and achieved performance improvement. Although these studies showed promising results for GAN-based DA, there is still room for improvement with a minor modification of the GAN loss function (Arjovsky et al., 2017); thus, many DL-based BCIs that use GAN for the DA exploited variants of GAN.

In this regard, Mao et al. (2017) proposed a modified version of the GAN loss function. They minimized the Pearson-χ^2 distance between the real distribution and the generated data distribution instead of the JSD used for the original GAN loss function (Goodfellow et al., 2014). Thus, Mao et al. modified the loss to:

$$\min_{\mathcal{D}} \frac{1}{2}\mathbb{E}_{p_{\mathbf{x}}}[\log(\mathcal{D}(\mathbf{x})-1)^2] + \frac{1}{2}\mathbb{E}_{p_{\mathbf{z}}}[\log(\mathcal{D}(\mathcal{G}(\mathbf{z}))-0)^2]$$
$$\text{and } \min_{\mathcal{G}} \frac{1}{2}\mathbb{E}_{p_{\mathbf{z}}}[\log(\mathcal{D}(\mathcal{G}(\mathbf{z}))-1)^2], \quad (2)$$

and named their method least-squares GAN (LSGAN). This LSGAN objective function gives a larger gradient to fake samples farther from the real samples decision boundary, thereby suppressing the gradient vanishing phenomenon. In case of DA for BCI, Pascual et al. (2019) adopted LSGAN to epileptic EEG DA. Specifically, Pascual et al. used a conditional vector (Mirza and Osindero, 2014) in their model to generate ictal EEG samples from given inter-ictal EEG samples. They also exploited U-Net (Ronneberger et al., 2015) for both the generator and the discriminator. By doing so, Pascual et al. synthesized numerous ictal samples and improved the performance with the generated samples.

Meanwhile, Radford et al. (2015) focused on solving the min-max objective of GAN (Goodfellow et al., 2014) as inherently unstable. With exhaustive attempts to design a stable CNN-based GAN from scratch, Radford et al. showed that the generator of a deconvolutional network without fully-connected layers and pooling layers and the discriminator of a convolutional network without pooling layers makes GAN robust. Their successful achievement is commonly called deep convolutional GAN (DCGAN). In a BCI society, DCGAN is also widely used for DA. For instance, Zhang et al. (2020b) augmented spectrograms of motor imagery EEG estimated by applying STFT using DCGAN. Zhang and Liu (2018) also showed improved motor imagery-based BCI performance by DA using DCGAN. Fahimi et al. (2020) generated raw EEG signals using DCGAN and analyzed the generated signals using t-stochastic neighbor embedding (Maaten and Hinton, 2008) and STFT. Additionally, Lee Y. E. et al. (2020) reconstructed ERP signals using DCGAN for mobile BCI. They also showed the performance of reconstructed ERP signals and visualized the generated samples. Truong et al. (2019a,b) applied DA to STFT transforms of epileptic EEG signals using DCGAN. Finally, Fan et al. (2020) performed the DA using DCGAN to tackle a class imbalance problem in sleep staging tasks and demonstrated the validity of GAN-based DA.

Similar to LSGAN (Mao et al., 2017), Arjovsky et al. (2017) focused on changing the JSD to the Wasserstein distance. Arjovsky et al. showed that the Wasserstein distance can be applied to GAN in a theoretically rigorous manner and proposed a modified version of the objective function:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}_{p_{\mathbf{x}}}[\mathcal{D}(\mathbf{x})] - \mathbb{E}_{p_{\mathbf{z}}}[\mathcal{D}(\mathcal{G}(\mathbf{z}))]. \quad (3)$$

To satisfy a constraint, i.e., to restrict the discriminator to the Lipschitz function, Arjovsky et al. used weight clipping on discriminator $\mathcal{D}$. However, Gulrajani et al. (2017) removed the weight clipping by adding a gradient penalty regularization to the objective function and made the training stable. These methods are widely known as Wasserstein GAN (WGAN). Several researchers of DL-based BCIs showed interest in a WGAN-based DA method. Ko et al. (2019) exploited WGAN with a gradient penalty to improve the BCI performance in motor imagery. They used WGAN, rather than the DA method, for DL-based BCI model training, and improved performance even with fewer training datasets. In addition, Hartmann et al. (2018) proposed *EEG-GAN* which is a modified version of WGAN to generate artificial raw EEG data. Aznan et al. (2019) also used WGAN to augment *steady-state visual evoked potential* (SSVEP) and improved the BCI performance. Panwar et al. (2019a,b) exploited WGAN with the gradient penalty to generate raw EEG data of RSVP and drowsiness and significantly improved the BCI performance. Luo and Lu (2018) and Luo et al. (2020) modified WGAN and synthesized *differential entropy* values calculated from emotion EEG signals. As the aforementioned methods

require a calibration phase, Hwang et al. (2019) tried to introduce zero-calibration. They used WGAN to generate raw EEG data acquired from a protocol of watching natural objects, such as a pizza and a banana. GAN-based DA methods synthesize realistic EEG samples by learning the data distribution implicitly, thereby showing great opportunity for DA. Nevertheless, these methods need (relatively) large amounts of data to train to network modules, i.e., the generator and the discriminator.

2.3.2.2. Variational Autoencoder

As GAN (Goodfellow et al., 2014) and its variants (Radford et al., 2015; Arjovsky et al., 2017; Mao et al., 2017) demonstrated their ability in DA, some studies focused on learning a latent representation of EEG data distribution in an explicit manner. AE (Ballard, 1987) is a neural network trained to replicate the input and the output data. AE has an encoder and a decoder; the encoder describes a *code* that is used for representing the input data, and the decoder reconstructs the input data from the code. Modern AE models have tried to generalize the encoder and the decoder functions to learn the distribution of the input data and the code. In particular, as depicted in **Figure 6**, VAE, which is a type of AE, learns encoder $\mathcal{Q}$ and decoder $\mathcal{P}$ through *variational inference*. The VAE (Kingma and Welling, 2014) is trained by the objective function:

$$\min_{\mathcal{P},\mathcal{Q}} -\mathbb{E}_{\mathcal{Q}}[\log(\mathcal{P}(\mathbf{x}|\mathbf{z}))] + \mathrm{KLD}(\mathcal{Q}(\mathbf{z}|\mathbf{x})||\mathcal{P}(\mathbf{z})). \quad (4)$$

where KLD denotes the *Kullback-Leibler divergence* (KLD). In Equation (4), the first term represents a *negative log-likelihood* of the latent code, $\mathbf{z}$, and can be considered as a reconstruction error. The second term is a regularization term to constrain the variational distribution, $\mathcal{Q}(\mathbf{z}|\mathbf{x})$, to be similar to $\mathcal{P}(\mathbf{z})$. Based on the objective function in Equation (4), the VAE effectively represents the latent space of the data distribution and can generate novel samples from the learned latent distribution.

In this regard, some DL-based BCIs use AEs (Ballard, 1987) and VAEs (Kingma and Welling, 2014) for DA. For example, Zhang et al. (2020b) transformed EEG signals into spectrograms using STFT and reconstructed them using both an AE and a VAE. By reconstructing STFT images from the learned code, Zhang et al. could effectively acquire novel training samples. Fahimi et al. (2020) exploited a VAE to synthesize artificial motor EEG signals. Furthermore, Aznan et al. (2019) performed DA of SSVEP EEG signals using a VAE. Finally, to augment the raw emotion EEG signals, Luo et al. (2020) learned the latent space of the data distribution and generated artificial samples using a VAE. Even though VAE-based DAs learn the training data distribution and generate augmentation samples, the synthesizing quality still lacks. We summarize our review of both the GAN and VAE-based DA methods in **Table 2**.

3. ADVANCES IN TRANSFER LEARNING

3.1. What Is Transfer Learning?

In recent years, efforts have been made to take advantage of other real EEG samples (i.e., from a session or a subject) to train deep neural networks that decode EEG samples, thereby mitigating the data insufficiency problem (Chai et al., 2016; Andreotti et al., 2018; Fahimi et al., 2019; Özdenizci et al., 2020). These studies known as TL have focused on transferring knowledge from one dataset to another one. Generally, the TL methods aim to learn well-generalized representation among different tasks (e.g., classification, regression, clustering, etc.) or multiple datasets following different but similar distributions (i.e., domains) in other fields. Meanwhile, various TL-based BCIs have leveraged other subjects' or sessions' data to solve the same task. The representation trained from those TL methods can be applied to the seen domains (e.g., domain adaptation) or an unseen domain (e.g., domain generalization) in a short/zero-calibration manner. Hence, we mainly focus on domain adaptation/generalization-based TL approaches in this study.

3.2. Challenges in Transfer Learning

When designing transfer methods in BCI, there are two major concerns: (i) intra- and inter-subject variabilities and (ii) negative transfer. First, as brain signals contain their inherent background activities and vary according to their conditions, e.g., fatigue, drowsiness, excitation, and agitation, high variabilities have been observed for different subjects and even for sessions of the same subjects (Jayaram et al., 2016), which are regarded as non-stationary EEG characteristics (Chai et al., 2016; Raza and Samothrakis, 2019). In this respect, when training a DL-based BCI method with samples of one subject or session, the trained DL method cannot be deployed to another subject or session directly, because unseen data (from new subject or session) can be misaligned with the training data in the trained feature space, referred to as a *domain shift* (Ganin et al., 2016). In other words, owing to the large discrepancy between training and unseen data, the trained DL-based BCI can be degraded drastically in testing unseen data. Domain adaptation (Wang and Deng, 2018) is proposed to diminish the domain shift in other fields, such as computer vision. Owing to its goal, domain adaptation-based approaches have been widely used in DL-based BCIs (Jeon et al., 2019; Özdenizci et al., 2020; Wei et al., 2020a; Zhao H. et al., 2020). Each subject or session is regarded as one domain in most studies. Recent studies have introduced a question: what should be transferred between various domains? Although the domain-invariant features can be obtained through TL, mainly via domain adaptation techniques, it can also induce degradation of unseen data because all information is not equally transferable (Lin and Jung, 2017; Wang and Deng, 2018; Peng et al., 2019; Jeon et al., 2020), which is denoted as a *negative transfer*.

3.3. Approaches in Transfer Learning

TL methods in BCI can be categorized into two approaches—explicit TL and implicit TL—depending on whether to explicitly use a discrepancy between two domains in the objective function. Explicit TL-based approaches have commonly focused on minimizing a divergence between multiple domains during the training process. These methods have been fundamentally devised according to *domain theory* (Ben-David et al., 2010). In domain theory, when training a model with a labeled source domain and an unlabeled target domain, the expected error of the target domain is upper bounded as the sum of the error of the labeled source domain and the discrepancy

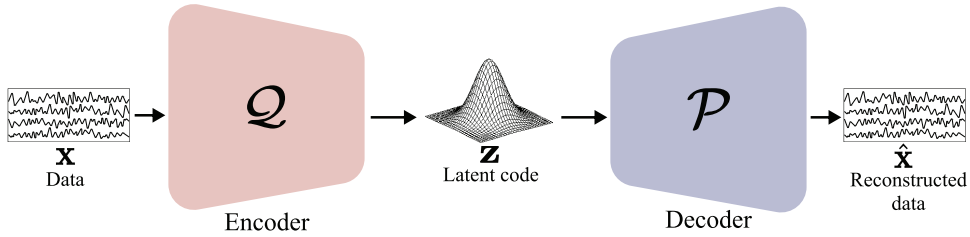

FIGURE 6 | Illustration of variational autoencoder (VAE) (Kingma and Welling, 2014). Encoder $\mathcal{Q}$ learns latent space of input data $\mathbf{x}$. From the learned latent space, latent code $\mathbf{z}$ is sampled and input to decoder $\mathcal{P}$. Finally, decoder $\mathcal{P}$ reconstructs input data $\hat{\mathbf{x}}$.

TABLE 2 | Deep generative data augmentation methods.

	Approach	References	Paradigm	Summary
GAN	GAN (Goodfellow et al., 2014)	Roy et al., 2020	Motor imagery	Devised LSTM-based generator and discriminator; qualitatively analyzed generated signals
		Krishna et al., 2020	Speech	Devised GRU-based generator and discriminator
	LSGAN (Mao et al., 2017)	Pascual et al., 2019	Seizure	Devised U-Net-based generator and discriminator; used conditional GAN concept
	DCGAN (Radford et al., 2015)	Zhang et al., 2020b	Motor imagery	Generated STFT images estimated from raw EEGs; compared synthesizing quality to other DA methods
		Zhang and Liu, 2018		Compared classification accuracy of testing dataset for different ratio of raw data and artificial data; used conditional GAN concept
		Fahimi et al., 2020	Motor	Used feature vector with the random noise for the generator input
		Lee Y. E. et al., 2020	ERP	Used features of EEG signals during walking as the generator input to reconstruct EEG signals similar to ones during standing
		Truong et al., 2019a	Seizure	Generated STFT images estimated from raw EEGs
		Truong et al., 2019b		Generated STFT images estimated from raw EEGs
		Fan et al., 2020	Sleep	Compared synthesizing quality to other DA methods
	WGAN (Arjovsky et al., 2017)	Ko et al., 2019	Motor imagery	Conducted gradient penalty rather than weight clipping; used semi-supervised GAN concept
		Hartmann et al., 2018	Motor	Conducted gradient penalty rather than weight clipping
		Aznan et al., 2019	SSVEP	Compared synthesizing quality to VAE-based DA methods; experimented TL setting
		Panwar et al., 2019b	RSVP	Conducted gradient penalty rather than weight clipping; used conditional GAN concept
		Luo et al., 2020	Emotion	Conducted gradient penalty rather than weight clipping; used conditional GAN concept
		Luo and Lu, 2018		Conducted gradient penalty rather than weight clipping; used conditional GAN concept
		Panwar et al., 2019a	Drowsy	Conducted gradient penalty rather than weight clipping
		Hwang et al., 2019	Cognition	Designed zero-calibration experiments
VAE	AE (Ballard, 1987)	Zhang et al., 2020b	Motor imagery	Generated STFT images estimated from raw EEGs; compared synthesizing quality to other DA methods
	VAE (Kingma and Welling, 2014)	Zhang et al., 2020b	Motor imagery	Generated STFT images estimated from raw EEGs; compared synthesizing quality to other DA methods
		Fahimi et al., 2020	Motor	Compared synthesizing quality to other DA methods
		Aznan et al., 2019	SSVEP	Compared synthesizing quality to VAE-based DA methods; experimented TL setting
		Luo et al., 2020	Emotion	Compared synthesizing quality to VAE-based DA methods

between the source and target domains. In other words, minimizing the divergence between multiple domains is key regardless of the labels in the target domain. The question here is why TL can be considered as an effort to reduce cost/time-consuming calibration. Most studies assumed that the subject-invariant feature space can be directly applied with zero or short-calibrations for new subjects' EEG data (Jeon et al., 2020; Özdenizci et al., 2020).

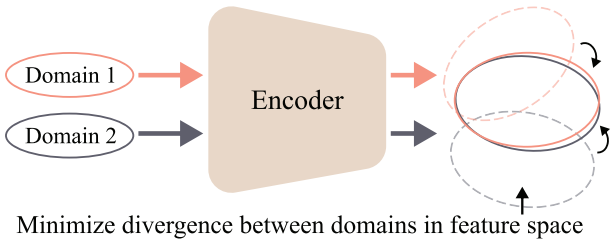

FIGURE 7 | Concept of explicit TL-based methods. The alignment can be achieved by minimizing a divergence between different domains.

Contrary to explicit TL-based methods, implicit TL-based approaches follow the hypothesis that their method can train domain-invariant feature spaces on the basis of only their internal architectures without explicitly minimizing the discrepancy. For instance, they merely perform fine-tuning with a new dataset (Andreotti et al., 2018; Fahimi et al., 2019; Zhang et al., 2021) or applied meta-learning framework (An et al., 2020; Duan et al., 2020). Furthermore, well-trained feature representation capturing multi-scale discriminative EEG patterns or focusing more discriminative temporal regions can be employed to evaluate new datasets (Kwon et al., 2019; Zhang et al., 2019a, 2020a; Ko et al., 2020a). We describe deep TL methods for zero/short-calibrations in more detail.

3.3.1. Explicit Transfer Learning Methods

Explicit TL-based methods define the distributional discrepancy between multiple domains, i.e., subjects or sessions, and then minimize the discrepancy during the training by appropriately designing their objective functions, thereby achieving an alignment in the feature space. We have witnessed the success of TLs that exploit subspace alignment methods in DL-based BCIs (Chai et al., 2016; Zhang et al., 2017; Özdenizci et al., 2020; Wei et al., 2020b; Wang et al., 2021). These methods can require additional DLs (adversarial learning) or not (non-parametric). Non-parametric alignment-based methods define a distributional discrepancy between different domains at various distances (Gretton et al., 2012; He and Wu, 2019) and then minimize it during optimization. Therefore, this minimization term is considered to be a regularization on a latent feature space. In contrast, adversarial learning-based methods require at least one neural network. Subsequently, the additional network identifies the domain from which the input data is sampled and denotes it as a domain discriminator. Through the min-max game between the domain discriminator and a feature extractor, adversarial learning induces domain-invariant features (Ganin et al., 2016). The conceptual schematization of the explicit TL is shown in **Figure 7**.

3.3.1.1. Non-parametric Alignment

To align features between different domains, three divergences are mainly introduced in DL-based BCIs: (i) *maximum mean discrepancy* (MMD) (Chai et al., 2016; Hang et al., 2019), (ii) KLD (Zhang et al., 2017), and (iii) *Euclidean distance* (Kostas and Rudzicz, 2020). First, MMD is the distance between two distributions S and T in a kernel embedding space and is defined as follows:

$$\mathrm{MMD}(S,T) = \left\| \frac{1}{n_S}\sum_{i=1}^{n_S}\phi(\mathbf{x}_i) - \frac{1}{n_T}\sum_{j=1}^{n_T}\phi(\mathbf{x}_j) \right\|_{\mathcal{H}}^2, \ \mathbf{x}_i \sim S, \ \mathbf{x}_j \sim T \tag{5}$$

where ϕ denotes a mapping function for *reproducing kernel Hilbert space* (RKHS) and $||\cdot||_{\mathcal{H}}$ is the RKHS norm (Gretton et al., 2006). Here, n_S and n_T denote the number of samples drawn from the S and T distributions, respectively. In terms of TL for DL-based BCIs, Hang et al. utilized MMD to minimize the distance between the source and target domains in features extracted from fully-connected layers after convolutional layers. They deployed another loss function named the *center-based discriminative feature learning* (CDFL) method. CDFL is referred to as a regularization technique, that compels the distance between each sample feature and the corresponding class center point to become less than thresholds for better separability between different classes. As a result, Hang et al. acquired a domain-invariant feature of motor imagery EEG signals at the class level by minimizing MMD as well as CDFL. Chai et al. also minimized MMD in a hidden feature space among source and target samples during training an AE and obtained a domain-invariant subspace for the emotion recognition task. However, the classifier was not jointly trained with the AE.

Similar to Chai et al. (2016)'s work, Zhang et al. (2017) constrained a hidden space in their AE to train a subject-invariant feature of the sleep EEG. However, according to the existing AE-based TL method (Zhuang et al., 2015), they only reduced a symmetric KLD between the source and target features by using an identity function as ϕ in Equation (5). In other words, they did not transform their features to another space during training. Although they trained all parameters of the AE and the classifier in an end-to-end manner, their method diminished only the marginal distribution difference, disregarding the conditional distributions of the two domains in classification (Ding et al., 2018).

Kostas and Rudzicz (2020) performed raw EEG data alignment from many subjects at the preprocessing step by applying the *Euclidean alignment* (EA) method (He and Wu, 2019). As raw EEG signals can be transformed into covariance matrices, i.e., *symmetric positive definite*, they can be operated on a Riemannian manifold (Wang et al., 2021). However, He and Wu demonstrated that covariance matrix alignment on the Riemannian space for TL required high computational costs and showed unstable operations compared with the Euclidean space. For this reason, Kostas and Rudzicz constrained the mean covariance matrix to become an identity matrix according to the EA method and then used the aligned samples as the input of their TL for the DL-based BCI method. Thus, Kostas and Rudzicz developed the TL method for motor imagery, ERP, and RSVP.

These non-parametric alignment-based methods do not require additional trainable parameters whereas they can be employed between only two domains. Accordingly, they selected two subjects (i.e., source and target subject) in their dataset (Chai

et al., 2016; Hang et al., 2019) or considered the remaining subjects except for a target subject as one source subject (Zhang et al., 2017; Kostas and Rudzicz, 2020). Consequently, we cannot easily utilize their methods in order for a zero-calibration BCI.

3.3.1.2. Adversarial Learning

In the BCI field, many TL methods have applied an adversarial learning (Goodfellow et al., 2014) concept. Among them, the *adversarial conditional VAE* (A-cVAE) (Wang Y. et al., 2018) and *domain adversarial neural network* (DANN) (Ganin et al., 2016) have shown their potential in training domain-invariant features from cross-subjects or cross-sessions. Özdenizci et al. (2019) proposed an adversarial neural network to learn subject-invariant latent representations by using an A-cVAE. They combined a *conditional VAE* (cVAE) (Sohn et al., 2015) and an adversarial network. To be specific, in their network, an encoder and a decoder were trained to learn latent EEG representations from multiple subjects under the subjects' ID, and an adversary was trained for subject identification. These two steps are conducted alternatively so that they can learn subject-invariant EEG representations. Subsequently, the output of the frozen encoder for the same training samples was fed into a new classifier for classification. Hence, there still exists a limitation that both subject-invariant learning class-discriminative learning did not train in an end-to-end manner.

Most adversarial learning-based methods adopt DANN (Ganin et al., 2016) for designing their TL frameworks. DANN comprises three components a feature extractor $\mathcal{F}$, domain discriminator $\mathcal{D}$, and classifier $\mathcal{C}$, as shown in **Figure 8**. The domain discriminator and the classifier identify the domains or classes to which the incoming features belong, whereas the feature extractor is trained to minimize the classification loss and maximize the domain loss through a *gradient reversal layer* (GRL) where gradients are multiplied by a negative value during the back-propagation process. The objective function of the DANN is defined as follows:

$$\min_{\mathcal{F},\mathcal{C}} \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_s(\mathbf{x},\mathbf{y})} \mathrm{CCE}(\mathcal{C}(\mathcal{F}(\mathbf{x})), \mathbf{y}) \tag{6}$$

$$\max_{\mathcal{F}} \min_{\mathcal{D}} -\mathbb{E}_{\mathbf{x}\sim p_s(\mathbf{x})}[\log \mathcal{D}(\mathcal{F}(\mathbf{x}))] - \mathbb{E}_{\mathbf{x}\sim p_t(\mathbf{x})}[\log(1 - \mathcal{D}(\mathcal{F}(\mathbf{x})))] \tag{7}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote the input and corresponding labels, respectively. Here, p_s and p_t indicate distributions from the source and target domains, respectively, and CCE is the categorical cross-entropy loss that is widely used for classification tasks. Thus, Equation (6) is used to train the feature extractor $\mathcal{F}$ and classifier $\mathcal{C}$ to represent the input data and discriminate it correctly and is considered as the classification loss. In addition, in Equation (7), similar to the GAN objective function, i.e., Equation (1), feature extractor $\mathcal{F}$ tries to extract domain-indiscriminative features, whereas the domain discriminator $\mathcal{D}$ focuses on classifying the domains. In this regard, Equation (7) is commonly referred to as the domain loss. Therefore, the feature extractor output can be class-discriminative and domain-invariant by optimizing Equations (6) and (7).

Based on DANN, Özdenizci et al. (2020) introduced an adversarial learning-based TL network where the domain discriminator identifies whether features belong to which subjects, similar to the previous study of using an A-cVAE (Özdenizci et al., 2019). Özdenizci et al. demonstrated that any decoding models for EEG can be applied to their DANN-based methods by considering various CNN-based architectures (Schirrmeister et al., 2017; Lawhern et al., 2018). In this study, Özdenizci et al. effectively represented the domain-invariant features of multiple subjects' motor imagery signals.

Recently, several methods have shown that the use of only DANN (Ganin et al., 2016) has some limitations and challenges (Ma et al., 2019; Nasiri and Clifford, 2020; Tang and Zhang, 2020; Zhao H. et al., 2020). First, Zhao et al. considered a single subject as a target and the remaining subjects of datasets as source sets; therefore, the domain discriminator was trained to distinguish between the target and the sources. Furthermore, Zhao et al. exploited a classification loss and a center loss (Wen et al., 2016) for the target subject to strengthen class-discriminative power by minimizing intra-class compactness and maximizing inter-class separability. In addition, Tang and Zhang addressed that DANN cannot capture complex multimodal structures because even a perfectly trained domain discriminator cannot ensure perfect alignment between different domains. In this regard, Tang and Zhang performed an outer product between the output of the feature extractor and the output of the classifier (class probabilities) and then fed it into the domain discriminator for better alignment between the two domains according to the conditional GAN (Mirza and Osindero, 2014). Additionally, Ma et al. introduced a domain residual connection for domain generalization. They assumed that domain-invariant features and domain-specific features can be separately trained by using additional parameters in the feature extractor. In detail, the domain-invariant (denoted as common in Ma et al.'s work) parameters are shared among all source domains and the additional parameters are used only for the corresponding domain samples per domain. Subsequently, the sum between the domain-invariant outputs and the domain-specific outputs is taken as inputs of the domain discriminator and classifier. Here, the common parameters of the feature extractor and the classifier are activated on testing the unseen target's data. However, as there are no decomposition strategies, it does not ensure that the subject-specific parameters capture the real subject-specific information regardless of the subject-invariant information.

Further, to mitigate negative transfer, two approaches have been proposed: (i) source selection (Jeon et al., 2019; Wei et al., 2020b; Wang et al., 2021) and (ii) transferable attention (Nasiri and Clifford, 2020). Regarding the source selection methods, they introduced the need to obtain the most similar subjects due to the high variability between subjects. Specifically, Jeon et al. assumed that before adapting other subjects' samples, they first must select a source subject whose properties were similar to those of a target subject by performing hierarchical clustering based on resting-state EEG signal candidates in the source pool. Although their feature extractor embeds both the source and target's EEG samples to the subject-invariant representations in accordance with DANN (Ganin et al., 2016), each classifier was separately trained between source and target subjects to capture the subject-specific characteristics. Similar to Jeon et al.'s work, Wei et al. selected source subjects based on the target

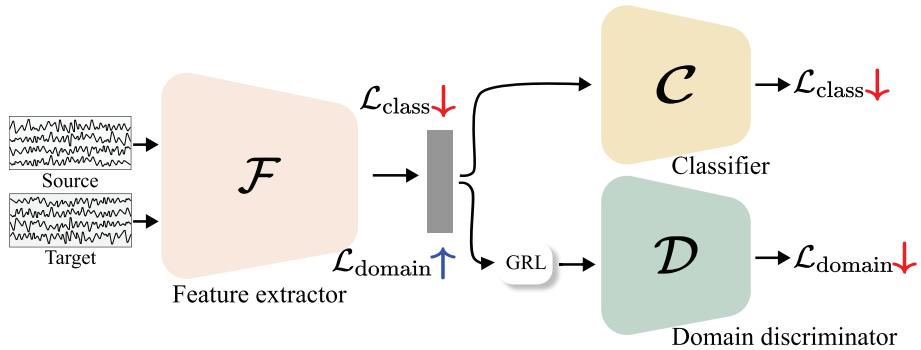

FIGURE 8 | Illustration of domain adversarial neural network (DANN) (Ganin et al., 2016). $\mathcal{L}_{class}$ and $\mathcal{L}_{domain}$ denote a classification loss and domain loss, respectively. Through a GRL where gradients of a domain loss are reversed by multiplying a negative value, a domain loss is minimized in a domain discriminator and maximized in a feature extractor.

subject's classification performance among the source subject-specific classifiers. In detail, they first trained different classifiers for each subject and then evaluated all trained classifiers with a target subject to rank them with respect to the target subject. After ranking the performances, they selected the top K subjects and then used them as a source domain set. Subsequently, the classification outputs were also regarded as inputs of the domain discriminator with features in the same manner (Mirza and Osindero, 2014; Tang and Zhang, 2020). Following Wei et al.'s source selection strategy, Wang et al. trained their network with the selected sources' samples and the target samples. In Wang et al.'s work, domain adaptation was achieved by using both adversarial loss and centroid alignment loss. They considered the geometric means of each class as each class-prototype and then minimized the discrepancy between the same class-prototypes among different domains in the Riemannian space.

In the meantime, Nasiri and Clifford (2020) also described that all features can contain considerably dissimilar information among various subjects so that they are not necessarily transferable. To focus on more important or class-relevant local parts of data, Nasiri and Clifford added channel-wise domain discriminators and then used their output to generate attention maps which can be a criterion for transferability by transforming entropy.

To sum up, these adversarial learning-based methods assumed that the well-trained feature representation can be validated for unseen domains, thus, they can accomplish the zero-calibration BCI. However, in the adversarial learning-based methods, additional trainable parameters are demanded to align distributions between two or more domains. Moreover, they can cause any distortion of feature representations on account of disregarding class-related information between domains (Liu et al., 2019; Jeon et al., 2020). We summarize our review of both non-parametric alignment/adversarial learning-based TL methods in **Table 3**.

3.3.2. Implicit Transfer Learning Methods

In this section, we describe the implicit TL approaches in DL-based BCIs. Implicit knowledge transferring methods do not explicitly minimize the discrepancy objective functions but only depend on their network (i.e., architecture). Most existing implicit TL methods have been used in the *leave-one subject-out* (LOO) scenario to fine-tune the trained parameters totally or partially using new target data (Andreotti et al., 2018; Fahimi et al., 2019; Shovon et al., 2019; Phan et al., 2020; Raghu et al., 2020; Zhang et al., 2021). Furthermore, various studies have only focused on enhancing the representational power of EEG features with only their well-designed architectures (Kwon et al., 2019; Zhang et al., 2019a; Jeon et al., 2020; Ko et al., 2020a). The remaining methods of implicit TLs (An et al., 2020; Duan et al., 2020) are based on meta-learning, which has drawn increasing attention for few-shot tasks in machine learning fields (Hospedales et al., 2020).

3.3.2.1. Fine-Tuning

Fine-tuning is a retraining strategy to initialize parameters of a network as learned parameters of another identical network trained with diverse source datasets to adapt them to the target dataset. Fine-tuning can be regarded as the most naive approach to transfer knowledge. In this respect, many studies have taken advantage of fine-tuning for TL (Andreotti et al., 2018; Fahimi et al., 2019; Zhao et al., 2019; Raghu et al., 2020; Zhang et al., 2021). Deep networks have been pre-trained with multiple subjects' samples in a large source pool dataset, and entire parameters or parts of parameters have been fine-tuned to capture more target-related information. For example, Shovon et al. (2019) fine-tuned the parameters of the entire network for transferring knowledge of natural image classification tasks to motor imagery EEG classification. Specifically, they trained the pre-trained network with natural images by using STFT from motor imagery EEGs. Raghu et al. fine-tuned the last layers that were learned using the source subjects for the seizure classification task. Aznan et al. (2019) first trained a network using synthetic SSVEP samples and then fine-tuned the pre-trained network with real SSVEP samples, which leads to carrying information of synthetic SSVEP to a real SSVEP classification. In addition, Vilamala et al. (2017), Phan et al. (2020), and Andreotti et al. fine-tuned the entire network for sleep stage classification.

On the contrary to those methods, the existing works (Zhao et al., 2019; Olesen et al., 2020; Zhang et al., 2021) fine-tuned parts

TABLE 3 | Explicit transfer learning methods.

	Approach	References	Paradigm	Summary
Non-parametric alignment	MMD	Hang et al., 2019	Motor imagery	Minimized MMD in a feature level and introduced CDFL
		Chai et al., 2016	Emotion	Minimized MMD in a feature level and trained AE and classifier separately
	KLD	Zhang et al., 2017	Sleep	Minimized KLD in a feature level and trained with classifier in an end-to-end manner
	EA	Kostas and Rudzicz, 2020	Multi	Constrained that the mean covariance matrix becomes an identity matrix in a raw data level
Adversarial learning	A-cVAE (Wang Y. et al., 2018)	Özdenizci et al., 2019	Motor imagery	Added an adversarial network to cVAE, and trained cVAE and classifier separately
	DANN (Ganin et al., 2016)	Özdenizci et al., 2020	Motor imagery	Devised DANN by exploiting various CNN-based architectures as their feature extractor
		Zhao H. et al., 2020		Added center loss for target to minimize intra-class compactness and maximize inter-class separability
		Tang and Zhang, 2020		Fed output of a classifier into a domain discriminator
		Jeon et al., 2019		Selected source based on resting-state EEG signals
		Wei et al., 2020b	RSVP	Selected sources based on a ranking of performances in subject-specific classifiers
		Wang et al., 2021	Emotion	Selected sources based on a ranking of performances in subject-specific classifiers and devised centroid alignment loss
		Nasiri and Clifford, 2020	Sleep	Estimated attention maps using channel-wise domain discriminators
		Ma et al., 2019	Drowsy	Trained additional parameters capturing subject-specific features

of the pre-trained network to transfer knowledge of EEG. For a new subject, Zhang et al. fine-tuned only the parameters of fully-connected layers while freezing the previous layers. Especially, Zhao et al. conducted ablation studies to identify which layers of their network should be transferred to the target. Whereas, those methods performed with motor imagery EEGs, Olesen et al. fine-tuned the last few layers with different samples for sleep stage classification. However, even though fine-tuning can be easily implemented, it is not performed within one process and cannot achieve zero-calibration efficiently. In addition, fine-tuning can cause over-fitting because of the small amount of target data (Kostas and Rudzicz, 2020).

3.3.2.2. Enhancing Representational Power

Several studies have focused on learning better EEG representations to concentrate on more discriminative temporal slices (Zhang et al., 2018a, 2019a,b, 2020a) or capture multi-scale spatio-temporal characteristics (Kwon et al., 2019; Ko et al., 2020a) and to separate class-relevant information (Jeon et al., 2020) among diverse subjects. First, Zhang et al. investigated the temporal dynamics of EEG signals based on the attention mechanism that emphasizes on more informative region on the basis of self-relationships. In their work, raw EEG signals were first divided into various slices by applying a sliding window technique with a window size of a shorter length than the overall length of the time sequence. Next, the segmented EEG slices in the form of raw slices (Zhang et al., 2018a, 2019a) or graphs (Zhang et al., 2019b, 2020a) embedded their features via the encoding module. Subsequently, Zhang et al. used a self-attention module to obtain more class-discriminative segments among those features and then aggregated all slices by means of a weighted sum with the attention maps (Zhang et al., 2018a). Further, in order for the attentive temporal dynamics between those features, Zhang et al. (2019a, 2020a) employed a recurrent self-attention module (e.g., LSTM). Additionally, Zhang et al. (2019b) discovered more discriminative EEG channels by introducing another attention module.

Meanwhile, Kwon et al. (2019) applied band-pass filtering for various predefined frequency bands to raw EEG samples from source subjects. Subsequently, by employing a CSP (Ramoser et al., 2000), they extracted spatio-spectral features for all frequency bands. They calculated mutual information between the spatio-spectral features and class labels and then sorted mutual information of all frequency bands in the descending order. They selected the top K frequency bands in the list and then used them as their CNN input. Ko et al. (2020a) also demonstrated that it is of substantial importance to discover multi-scale features in terms of frequency/time ranges, considering spatial patterns. Unlike Kwon et al.'s work, Ko et al.'s network is composed of only convolutional layers; thus, it can be trained with raw EEGs in an end-to-end manner. Specifically, they first extracted spatio-temporal features in multi-scale by gathering intermediate representations of three convolutional layers and applying different spatial convolutional layers to them.

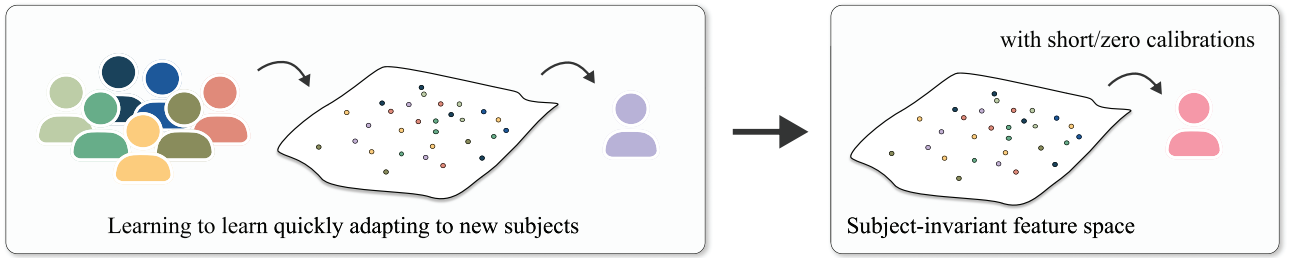

FIGURE 9 | Conceptual illustration of meta-learning in BCI. Meta-learning aims to learn how to quickly adapt to new subjects by updating parameters based on a variety of tasks acquired from multiple subjects. Subsequently, the trained feature space can be considered as subject-invariant space that can be efficiently applied to new subjects with short/zero calibrations.

After concatenating the multi-scale features, Ko et al. applied global average pooling (Lin et al., 2013) to them and fed the results to a fully-connected layer.

Jeon et al. (2020) proposed an information-theoretic method that decomposes an intermediate feature of the existing CNN models (Schirrmeister et al., 2017; Lawhern et al., 2018) into class-relevant and class-irrelevant features by estimating mutual information between them to mitigate a negative transfer. Furthermore, to enrich the representational power of their features, they maximized mutual information between class-relevant features and global features, i.e., an output of the last convolutional layer by regarding it as a more high-level representation, utilizing two mutual information neural estimators (MINEs) (Belghazi et al., 2018) from the local and global viewpoints, inspired by Hjelm et al. (2019). In detail, they exploited three MINEs (Belghazi et al., 2018); one to ensure good decomposition between class-relevant features and class-irrelevant features and the other two to make the global features contain more class-relevant information.

These methods (Kwon et al., 2019; Zhang et al., 2019a; Jeon et al., 2020; Ko et al., 2020a) have great significance in the sense that they showed the importance of exploring better EEG representation and enabled zero calibration in terms of TL. However, most of the methods for better EEG representation, except for (Ko et al., 2020a), focused on the motor imagery EEG and used the characteristics of it, which can be a limitation to apply them for other paradigms of EEG.

3.3.2.3. Meta-Learning

Meta-learning is known as *learning to learn*, which allows a model to learn a method that enables fast adaptation to a new task or environment for a few-shot learning task (Hospedales et al., 2020). After the successful application of meta-learning in machine learning fields, the meta-learning framework has recently been applied to DL-based BCIs (An et al., 2020; Duan et al., 2020). **Figure 9** represents a basic concept of meta-learning with respect to TL in BCIs. As shown in **Figure 9**, some researchers assumed that learning to learn a task (e.g., classification, regression, etc.) among multiple subjects can result in a subject-invariant feature space that can be quickly applied to the target subject. Specifically, Duan et al. deployed a model-agnostic meta learning (MAML) (Finn et al., 2017) to obtain optimal parameters that can be rapidly adapted to target data through gradient-based optimization across multiple subjects. After dividing various source subjects' EEG data into many small groups, they updated the parameters of their network based on their gradients in two phases, meta-training and meta-test phase, and then fine-tuned the trained parameters with a small amount of target data. However, MAML easily induces over-fitting (Zintgraf et al., 2019), therefore, Duan et al. designed shallow convolutional layers for feature extraction. For this reason, their method cannot learn sufficient representation to capture class-discriminative information, which can be one of the limitations in applying their method. Another meta-learning example in BCI is the work of An et al. (2020). An et al. adopted a metric-based meta-learning framework, relation network (Sung et al., 2018), to efficiently learn class-representative features among multiple subjects. An et al. introduced three components: (i) an embedding module that extracts multi-scale features for support (labeled samples) and query (unlabeled samples) sets from source subjects, (ii) an attention module that generates a class-representative vector considering class-related importance among support sets, and (iii) a relation module to estimate the relation score between each class-representative vector and the query samples. An et al. optimized all these components by simply minimizing a cross-entropy loss, i.e., classification loss, and evaluated their network in 5-, 10-, and 20-shot settings, i.e., 5, 10, and 20 EEG samples per class. Their relational learning with attention improved the performances of all scenarios compared with a case with only relation network. However, since this metric-based meta-learning required a pair-wise input during training and evaluation, it can show difference performances depending on the support sets. We summarize our review of the fine-tuning/enhancing representational power/meta-learning-based TL methods in **Table 4**. Furthermore, all acronyms are listed in **Appendix**: List of Acronyms.

4. DISCUSSION

In section 2, we review many DA methods for DL-based BCIs. From now on, we directly compare generative model-based DA methods and geometric manipulation-based DA

TABLE 4 | Implicit transfer learning methods.

	Approach	References	Paradigm	Summary
Fine-tuning	Whole	Shovon et al., 2019	Motor imagery	Pre-trained with natural images
		Aznan et al., 2019	SSVEP	Pre-trained with synthetic SSVEP samples
		Andreotti et al., 2018	Sleep	Trained their network with source subjects and fine-tuned it with target subject (LOO)
		Phan et al., 2020		Pre-trained network with different dataset
		Vilamala et al., 2017		Pre-trained network with natural images
		Fahimi et al., 2019	Cognition	Trained their network with source subjects and fine-tuned it with target subject (LOO)
	Partial	Zhang et al., 2021	Motor imagery	Fine-tuned only fully-connected layers
		Zhao et al., 2019		Conducted ablation studies to identify which layer should be transferred target
		Raghu et al., 2020	Seizure	Fine-tuned the last some layers of pre-trained network
		Olesen et al., 2020	Sleep	Fine-tuned parts of parameters
Enhancing representational power	Attention	Zhang et al., 2018a	Motor imagery	Designed a self-attention module to find more class-discriminative segments
		Zhang et al., 2019a		Designed a recurrent self-attention module
		Zhang et al., 2020a		Presented raw EEG to a spatial graph and designed a recurrent self-attention module
		Zhang et al., 2019b		Presented raw EEG to a spatial graph and designed two attention modules; one for attentive temporal dynamics and the other for attentive channels
	Multi-scale features	Kwon et al., 2019		Extracted spatio-spectral features in multi-frequency bands using CSP and selected top bands to use them as inputs
		Ko et al., 2020a	Multi	Extracted multi-scale features including spatio-temporal-spectral patterns
	Maximize mutual information	Jeon et al., 2020	Motor imagery	Decomposed an intermediate feature into a class-relevant and class-irrelevant feature and maximized mutual information between low-level and high-level representations
Meta-learning	MAML (Finn et al., 2017)	Duan et al., 2020	Multi	Trained optimal parameters through gradient-based optimization and conducted fine-tuning with a small amount of target data
	Relation (Sung et al., 2018)	An et al., 2020	Motor imagery	Estimated relation scores between support and query sets among source subjects in few-shot scenarios

methods and recommend a DA method for DL-based BCIs. Approximately 45% of generative model-based DA methods are reviewed, whereas ∼55% of geometric manipulation-based methods are reviewed. Interestingly, Zhang et al. (2020b) and Fahimi et al. (2020) used both generative model and geometric manipulation-based DA methods. Specifically, Zhang et al. used geometric transformation, noise addition, AE (Ballard, 1987), VAE (Kingma and Welling, 2014), and DCGAN (Radford et al., 2015) to augment motor imagery data. Zhang et al. conducted classification experiments using a CNN with various real data to generate data ratio values of 1:1, 1:3, 1:5, 1:7, and 1:9. Regardless of the ratio, DCGAN-based DA achieved a high degree of consistency for the average classification accuracy whereas geometric transformation and noise addition-based methods mostly underperform with the baseline, i.e., CNN without any DA method. In addition, Fahimi et al. conducted motor execution EEG classification experiments with various augmentation methods, segmentation and recombination, VAE, and DCGAN. Similar to Zhang et al.'s work, Fahimi et al. also achieved the best performance improvement with DCGAN whereas segmentation and recombination-based augmentation did not achieve significant improvement. Based on these two results, even geometric manipulation techniques have room for improvement, and we recommend a generative model-based DA method for DL-based BCI research. Furthermore, the Wasserstein distance can be directly adapted to DCGAN, and it is expected that the BCI will have performance improvements with DCGAN trained on the Wasserstine distance (Arjovsky et al., 2017). As some pioneering studies (Hartmann et al., 2018; Hwang et al., 2019; Ko et al., 2019) have demonstrated the validity of WGAN, we anticipate that the WGAN-based DA method with careful structural design and training can improve many DL-based BCI methods.

In section 3, we summarize various TL approaches for DL-based BCIs. To achieve a short/zero calibration task, many studies performed TL across different subjects/sessions in a single dataset (Fahimi et al., 2019; Kwon et al., 2019; Özdenizci et al., 2020), inter-dataset (Phan et al., 2020), and even different data paradigms (Vilamala et al., 2017; Aznan et al., 2019). In our review, explicit TL-based methods account for nearly 45% and the remaining works are categorized as implicit TL-based methods. With regard to explicit TL-based methods, there exist two approaches, non-parametric and parametric (i.e., adversarial learning) alignment methods, for a feature space among multiple domains (subjects or sessions) (Jeon et al., 2019; Nasiri and Clifford, 2020; Özdenizci et al., 2020; Zhao H. et al., 2020; Wang et al., 2021). In **Table 3**, we observe that most of the existing adversarial methods employ DANN (Ganin et al., 2016). Further, modified adversarial objective functions, such as WGAN (Arjovsky et al., 2017; Gulrajani et al., 2017) and LSGAN (Mao et al., 2017), have been employed to stabilize the training process in adversarial learning-based TL approaches (Wei et al., 2020b; Zhao H. et al., 2020). In this regard, we expect that numerous variants of DANN (Tzeng et al., 2017; Xu et al., 2018; Zhang et al., 2018c; Peng et al., 2019; Wang et al., 2019) can be applied to DL-based BCI tasks. Although most implicit TL-based methods fine-tune their pre-trained network using the new target's data, there are still few limitations: (i) fine-tuning cannot reach zero-calibration and (ii) fine-tuning may lead to an overfitting problem with a small amount of target data (Kostas and Rudzicz, 2020). An et al. (2020) and Duan et al. (2020) showed successful applications of common meta-learning methods (Finn et al., 2017; Sung et al., 2018) for DL-based BCIs. However, there still remain concerns: (i) a constraint in architectures of the feature extractor (Duan et al., 2020) and (ii) variations of performances depending on varying support samples (An et al., 2020). Meanwhile, a few methods (Zhang et al., 2018a, 2019a,b, 2020a; Kwon et al., 2019; Ko et al., 2020a) demonstrated that their intrinsic architectures are sufficient to cover the new target's characteristics even in the zero-calibration scenario. Most of these methods highly rely on EEG paradigm. In this respect, despite the success of the implicit TL-based methods, there are still several points to be considered for practical applications. Hence, when first trying the short/zero-calibration BCI, we recommend the explicit TL-based methods.

Based on our survey about many pioneering DA and TL approaches for BCIs, we conclude that both strategies can be beneficial to the short- and/or zero-calibration BCIs. Especially, it can be an interesting future research direction to combine both DA and TL approaches. For instance, before performing any TL strategies, a series of DAs would augment the number of samples, thereby improving the zero-calibration BCIs. Moreover, let us assume that there exist a large amount of source data samples and a few target samples. Then, it can be considered, inter alia, some strategic TL methods to build a good starting *backbone* network. Then, DA methods are applied to the target samples to augment them. Finally, these augmented target samples can fine-tune the backbone network to improve the short-calibration BCIs.

5. CONCLUSION

In this study, we surveyed recent advances in the field of DL-based BCIs, especially for short/zero-calibration techniques. We focused on several important aspects of the short/zero-calibration techniques. Various generative model-based and geometric manipulation-based DA methods have demonstrated their promising potential in the short-calibration technique. Moreover, we summarized recent trends in TL used in DL-based BCIs. Overall, explicit TL-based and implicit TL-based TL strategies significantly improve the zero-calibration BCIs.

Presently, increasing interests in DL have considerably increased the use of BCI technologies in the *real world*. Moreover, advancements in other fields, such as computer vision will benefit from more practical and powerful DL-based BCIs. We hope that this review contributes to the BCI field as a good summary of short/zero-calibration techniques for the design of DL-based BCI studies.

AUTHOR CONTRIBUTIONS

WK, EJ, SJ, and JP performed the literature search. WK and EJ discussed and clustered the results and drafted the manuscript together. H-IS organized the overall structure of the manuscript and revised the manuscript, approved the final version, and agreed to be accountable for all aspects of the work. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No. 2017-0-00451; Development of BCI based Brain and Cognitive Computing Technology for Recognizing User's Intentions using Deep Learning and No. 2019-0-00079; Department of Artificial Intelligence, Korea University).

REFERENCES

An, S., Kim, S., Chikontwe, P., and Park, S. H. (2020). "Few-shot relation learning with attention for EEG-based motor imagery classification," in *IEEE/RSH International Conference on Intelligent Robots and Systems (IROS)* (Las Vegas, NV). doi: 10.1109/IROS45743.2020.9340933

Andreotti, F., Phan, H., Cooray, N., Lo, C., Hu, M. T., and De Vos, M. (2018). "Multichannel sleep stage classification and transfer learning using convolutional neural networks," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Honolulu, HI: IEEE), 171–174. doi: 10.1109/EMBC.2018.8512214

Arjovsky, M., Chintala, S., and Bottou, L. (2017). "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)* (Sydney), 214–223.

Aznan, N. K. N., Atapour-Abarghouei, A., Bonner, S., Connolly, J. D., Al Moubayed, N., and Breckon, T. P. (2019). "Simulating brain signals: creating synthetic EEG data via neural-based generative models for improved SSVEP

classification," in *2019 International Joint Conference on Neural Networks (IJCNN)* (Budapest: IEEE), 1–8.

Ballard, D. H. (1987). "Modular learning in neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (Seattle, WA), 279–284.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., et al. (2018). "MINE: mutual information neural estimation," in *Proceedings of the 35th International Conference on Machine Learning (ICML)* (Stockholm).

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Mach. Learn*. 79, 151–175. doi: 10.1007/s10994-009-5152-4

Binias, B., Myszor, D., Palus, H., and Cyran, K. A. (2020). Prediction of pilot's reaction time based on EEG signals. *Front. Neuroinform*. 14:6. doi: 10.3389/fninf.2020.00006

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Chai, X., Wang, Q., Zhao, Y., Liu, X., Bai, O., and Li, Y. (2016). Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition. *Comput. Biol. Med*. 79, 205–214. doi: 10.1016/j.compbiomed.2016.10.019

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res*. 16, 321–357. doi: 10.1613/jair.953

Cho, J. H., Jeong, J. H., and Lee, S. W. (2020). Decoding of grasp motions from EEG signals based on a novel data augmentation strategy. *arXiv* 2005.04881. doi: 10.1109/EMBC44109.2020.9175784

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* 14 12.3555.

Craik, A., He, Y., and Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. *J. Neural Eng*. 16:031001. doi: 10.1088/1741-2552/ab0ab5

Dai, G., Zhou, J., Huang, J., and Wang, N. (2020). HS-CNN: a CNN with hybrid convolution scale for EEG motor imagery classification. *J. Neural Eng*. 17:016025. doi: 10.1088/1741-2552/ab405f

Dinarès-Ferran, J., Ortner, R., Guger, C., and Solé-Casals, J. (2018). A new method to generate artificial frames using the empirical mode decomposition for an EEG-based motor imagery BCI. *Front. Neurosci*. 12:308. doi: 10.3389/fnins.2018.00308

Ding, Z., Nasrabadi, N. M., and Fu, Y. (2018). Semi-supervised deep domain adaptation via coupled neural networks. *IEEE Trans. Image Process*. 27, 5214–5224. doi: 10.1109/TIP.2018.2851067

Donahue, C., McAuley, J., and Puckette, M. (2019). "Adversarial audio synthesis," in *International Conference on Learning Representations (ICLR)* (New Orleans, LA).

Dornhege, G., del R. Millán, J., Hinterberger, T., McFarland, D. J., and Müller, K. R. (eds.). (2007). "An introduction to brain-computer interfacing," in *Toward Brain-Computer Interfacing* (MIT Press), 1–25. doi: 10.7551/mitpress/7493.003.0003

Duan, T., Chauhan, M., Shaikh, M. A., and Srihari, S. (2020). Ultra efficient transfer learning with meta update for cross subject EEG classification. *arXiv* 2003.06113.

Fahimi, F., Dosen, S., Ang, K. K., Mrachacz-Kersting, N., and Guan, C. (2020). Generative adversarial networks-based data augmentation for brain-computer interface. *IEEE Trans. Neural Netw. Learn. Syst*. doi: 10.1109/TNNLS.2020.3016666. [Epub ahead of print].

Fahimi, F., Zhang, Z., Goh, W. B., Lee, T. S., Ang, K. K., and Guan, C. (2019). Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI. *J. Neural Eng*. 16:026007. doi: 10.1088/1741-2552/aaf3f6

Fan, J., Sun, C., Chen, C., Jiang, X., Liu, X., Zhao, X., et al. (2020). EEG data augmentation: towards class imbalance problem in sleep staging tasks. *J. Neural Eng*. 17:056017. doi: 10.1088/1741-2552/abb5be

Finn, C., Abbeel, P., and Levine, S. (2017). "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)* (Sydney).

Flandrin, P., Rilling, G., and Goncalves, P. (2004). Empirical mode decomposition as a filter bank. *IEEE Signal Process. Lett*. 11, 112–114. doi: 10.1109/LSP.2003.821662

Freer, D., and Yang, G. Z. (2020). Data augmentation for self-paced motor imagery classification with C-LSTM. *J. Neural Eng*. 17:016041. doi: 10.1088/1741-2552/ab57c0

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res*. 17, 2096–2030. doi: 10.1007/978-3-319-58347-1_10

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, Vol.27, 2672–2680.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2006). A kernel method for the two-sample-problem. *In Advances in Neural Information Processing Systems (NIPS)*, *volume* 19, 513–520. doi: 10.5555/2188385.2188410

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res*. 13, 723–773.

Gu, X., Cao, Z., Jolfaei, A., Xu, P., Wu, D., Jung, T. P., et al. (2020). EEG-based brain-computer interfaces (BCIs): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications. *arXiv* 2001.11337.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). "Improved training of Wasserstein GANs," in *Advances in Neural Information Processing Systems (NIPS)*, Vol. 30, 5767–5777.

Hang, W., Feng, W., Du, R., Liang, S., Chen, Y., Wang, Q., et al. (2019). Cross-subject EEG signal recognition using deep domain adaptation network. *IEEE Access* 7, 128273–128282. doi: 10.1109/ACCESS.2019.2939288

Hartmann, K. G., Schirrmeister, R. T., and Ball, T. (2018). EEG-GAN: generative adversarial networks for electroencephalograhic (EEG) brain signals. *arXiv* 1806.01875.

He, H., and Wu, D. (2019). Transfer learning for brain-computer interfaces: a Euclidean space data alignment approach. *IEEE. Trans. Biomed. Eng*. 67, 399–410. doi: 10.1109/TBME.2019.2913914

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., et al. (2019). "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations (ICLR)* (New Orleans, LA).

Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. (2020). Meta-learning in neural networks: a survey. *arXiv* 2004.05439.

Huang, W., Wang, L., Yan, Z., and Liu, Y. (2020). "Classify motor imagery by a novel CNN with data augmentation," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Montreal, QC: IEEE), 192–195. doi: 10.1109/EMBC44109.2020.9176361

Hwang, S., Hong, K., Son, G., and Byun, H. (2019). "EZSL-GAN: EEG-based zero-shot learning approach using a generative adversarial network," in *2019 7th International Winter Conference on Brain-Computer Interface (BCI)* (Jeongseon: IEEE), 1–4. doi: 10.1109/IWW-BCI.2019.8737322

Jayaram, V., Alamgir, M., Altun, Y., Scholkopf, B., and Grosse-Wentrup, M. (2016). Transfer learning in brain-computer interfaces. *IEEE Comput. Intell. Mag*. 11, 20–31. doi: 10.1109/MCI.2015.2501545

Jeon, E., Ko, W., and Suk, H. I. (2019). "Domain adaptation with source selection for motor-imagery based BCI," in *2019 7th International Winter Conference on Brain-Computer Interface (BCI)* (Jeongseon: IEEE), 1–4. doi: 10.1109/IWW-BCI.2019.8737340

Jeon, E., Ko, W., Yoon, J. S., and Suk, H. (2020). Mutual information-driven subject invariant and class relevant deep representation learning in BCI. *CoRR* abs/1910.07747.

Kalaganis, F. P., Laskaris, N. A., Chatzilari, E., Nikolopoulos, S., and Kompatsiaris, I. (2020). A data augmentation scheme for geometric deep learning in personalized brain-computer interfaces. *IEEE Access* 8, 162218–162229. doi: 10.1109/ACCESS.2020.3021580

Kingma, D. P., and Welling, M. (2014). "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)* (Banff, AB).

Ko, W., Jeon, E., Jeong, S., and Suk, H. I. (2020a). Multi-scale neural network for EEG representation learning in BCI. *IEEE Comput. Intell. Mag.* 16, 31–45. doi: 10.1109/MCI.2021.3061875

Ko, W., Jeon, E., Lee, J., and Suk, H. I. (2019). "Semi-supervised deep adversarial learning for brain-computer interface," in *2019 7th International Winter Conference on Brain-Computer Interface (BCI)* (Jeongseon: IEEE), 1–4. doi: 10.1109/IWW-BCI.2019.8737345

Ko, W., Oh, K., Jeon, E., and Suk, H. I. (2020b). "VIGNet: a deep convolutional neural network for EEG-based driver vigilance estimation," in *2020 8th International Winter Conference on Brain-Computer Interface (BCI)* (Jeongseon: IEEE), 1–3. doi: 10.1109/BCI48061.2020.9061668

Ko, W., Yoon, J., Kang, E., Jun, E., Choi, J. S., and Suk, H. I. (2018). "Deep recurrent spatio-temporal neural network for motor imagery based BCI," in *2018 6th International Conference on Brain-Computer Interface (BCI)* (Jeongseon: IEEE), 1–3. doi: 10.1109/IWW-BCI.2018.8311535

Kohonen, T. (1990). The self-organizing map. *Proc. IEEE* 78, 1464–1480. doi: 10.1109/5.58325

Kostas, D., and Rudzicz, F. (2020). Thinker invariance: enabling deep neural networks for BCI across more people. *J. Neural Eng.* 17:056008. doi: 10.1088/1741-2552/abb7a7

Krishna, G., Tran, C., Carnahan, M., Han, Y., and Tewfik, A. H. (2020). Generating EEG features from acoustic features. *arXiv* 2003.00007.

Kwon, O. Y., Lee, M. H., Guan, C., and Lee, S. W. (2019). Subject-independent brain-computer interfaces based on deep convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 31, 3839–3852. doi: 10.1109/TNNLS.2019.2946869

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *J. Neural Eng.* 15:056013. doi: 10.1088/1741-2552/aace8c

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, K., Liu, D., Perroud, L., Chavarriaga, R., and Millán, J. d. R. (2017). A brain-controlled exoskeleton with cascaded event-related desynchronization classifiers. *Robot. Auton. Syst.* 90, 15–23. doi: 10.1016/j.robot.2016.10.005

Lee, T., Kim, M., and Kim, S. P. (2020). "Data augmentation effects using borderline-SMOTE on classification of a P300-based BCI," in *2020 8th International Winter Conference on Brain-Computer Interface (BCI)* (Jeongseon: IEEE), 1–4. doi: 10.1109/BCI48061.2020.9061656

Lee, Y. E., Lee, M., and Lee, S. W. (2020). Reconstructing ERP signals using generative adversarial networks for mobile brain-machine interface. *arXiv* 2005.08430.

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv* 1312.4400.

Lin, Y. P., and Jung, T. P. (2017). Improving EEG-based emotion classification using conditional transfer learning. *Front. Hum. Neurosci* 11:334. doi: 10.3389/fnhum.2017.00334

Lin, Z., Zhang, C., Wu, W., and Gao, X. (2006). Frequency recognition based on canonical correlation analysis for ssvep-based bcis. *IEEE. Trans. Biomed. Eng.* 53, 2610–2614. doi: 10.1109/TBME.2006.886577

Liu, H., Long, M., Wang, J., and Jordan, M. (2019). "Transferable adversarial training: a general approach to adapting deep classifiers," in *International Conference on Machine Learning (ICML)* (Long Beach, CA: PMLR), 4013–4022.

Liu, Y. T., Pal, N. R., Wu, S. L., Hsieh, T. Y., and Lin, C. T. (2016). "Adaptive subspace sampling for class imbalance processing," in *2016 International Conference on Fuzzy Theory and Its Applications (iFuzzy)* (Taichung: IEEE), 1–5. doi: 10.1109/iFUZZY.2016.8004947

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., et al. (2018). A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update. *J. Neural Eng.* 15:031005. doi: 10.1088/1741-2552/aab2f2

Luo, Y., and Lu, B. L. (2018). "EEG data augmentation for emotion recognition using a conditional Wasserstein GAN," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Honolulu, HI: IEEE), 2535–2538. doi: 10.1109/EMBC.2018.8512865

Luo, Y., Zhu, L. Z., Wan, Z. Y., and Lu, B. L. (2020). Data augmentation for enhancing EEG-based emotion recognition with deep generative models. *arXiv* 2006.05331. doi: 10.1088/1741-2552/abb580

Ma, B. Q., Li, H., Luo, Y., and Lu, B. L. (2019). "Depersonalized cross-subject vigilance estimation with adversarial domain generalization," in *2019 International Joint Conference on Neural Networks (IJCNN)* (Budapest: IEEE), 1–8. doi: 10.1109/IJCNN.2019.8852347

Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* 9, 2579–2605. doi: 10.1145/3021604

Majidov, I., and Whangbo, T. (2019). Efficient classification of motor imagery electroencephalography signals using deep learning methods. *Sensors* 19:1736. doi: 10.3390/s19071736

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. (2017). "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Venice), 2794–2802. doi: 10.1109/ICCV.2017.304

Mirza, M., and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv* 1411.1784.

Mousavi, Z., Rezaii, T. Y., Sheykhivand, S., Farzamnia, A., and Razavi, S. (2019). Deep convolutional neural network for classification of sleep stages from single-channel EEG signals. *J. Neurosci. Methods* 324:108312. doi: 10.1016/j.jneumeth.2019.108312

Nasiri, S., and Clifford, G. D. (2020). "Attentive adversarial network for large-scale sleep staging," in *Proceedings of the 5th Machine Learning for Healthcare Conference (PMLR)*, 457–478.

Olesen, A. N., Jennum, P., Mignot, E., and Sorensen, H. B. D. (2020). Deep transfer learning for improving single-EEG arousal detection. *arXiv* 2004.05111. doi: 10.1109/EMBC44109.2020.9176723

Özdenizci, O., Wang, Y., Koike-Akino, T., and Erdoğmuş, D. (2019). "Transfer learning in brain-computer interfaces with adversarial variational autoencoders," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)* (San Francisco, CA: IEEE), 207–210. doi: 10.1109/NER.2019.8716897

Özdenizci, O., Wang, Y., Koike-Akino, T., and Erdoğmuş, D. (2020). Learning invariant representations from EEG via adversarial inference. *IEEE Access* 8, 27074–27085. doi: 10.1109/ACCESS.2020.2971600

Panwar, S., Rad, P., Quarles, J., Golob, E., and Huang, Y. (2019a). "A semi-supervised Wasserstein generative adversarial network for classifying driving fatigue from EEG signals," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (Bari: IEEE), 3943–3948. doi: 10.1109/SMC.2019.8914286

Panwar, S., Rad, P., Quarles, J., and Huang, Y. (2019b). "Generating EEG signals of an RSVP experiment by a class conditioned Wasserstein generative adversarial network," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (Bari: IEEE), 1304–1310. doi: 10.1109/SMC.2019.8914492

Parvan, M., Ghiasi, A. R., Rezaii, T. Y., and Farzamnia, A. (2019). "Transfer learning based motor imagery classification using convolutional neural networks," in *2019 27th Iranian Conference on Electrical Engineering (ICEE)* (Yazd: IEEE), 1825–1828. doi: 10.1109/IranianCEE.2019.8786636

Pascual, D., Aminifar, A., Atienza, D., Ryvlin, P., and Wattenhofer, R. (2019). Synthetic epileptic brain activities using generative adversarial networks. *arXiv* 1907.10518.

Peng, X., Huang, Z., Sun, X., and Saenko, K. (2019). "Domain agnostic learning with disentangled representations," in *Proceedings of the 36th International Conference on Machine Learning (ICML)* (Long Beach, CA), 5102–5112.

Phan, H., Chén, O. Y., Koch, P., Lu, Z., McLoughlin, I., Mertins, A., et al. (2020). Towards more accurate automatic sleep staging via deep transfer learning. *IEEE. Trans. Biomed. Eng.* doi: 10.1109/TBME.2020.3020381. [Epub ahead of print].

Qing, C., Qiao, R., Xu, X., and Cheng, Y. (2019). Interpretable emotion recognition using EEG signals. *IEEE Access* 7, 94160–94170. doi: 10.1109/ACCESS.2019.2928691

Radford, A., Metz, L., and Chintala, S. (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks," in *International Conference on Learning Representations (ICLR)* (San Diego, CA).

Raghu, S., Sriraam, N., Temel, Y., Rao, S. V., and Kubben, P. L. (2020). EEG based multi-class seizure type classification using convolutional neural network and transfer learning. *Neural Netw.* 124, 202–212. doi: 10.1016/j.neunet.2020.01.017

Ramoser, H., Muller-Gerking, J., and Pfurtscheller, G. (2000). Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans. Neural Syst. Rehabil. Eng.* 8, 441–446. doi: 10.1109/86.895946

Raza, H., and Samothrakis, S. (2019). "Bagging adversarial neural networks for domain adaptation in non-stationary EEG," in *2019 International Joint Conference on Neural Networks (IJCNN)* (Budapest: IEEE), 1–7. doi: 10.1109/IJCNN.2019.8852284

Romaissa, D., El Habib, M., and Chikh, M. A. (2019). "Epileptic seizure detection from imbalanced EEG signal," in *2019 International Conference on Advanced Electrical Engineering (ICAEE)* (Algiers: IEEE), 1–6. doi: 10.1109/ICAEE47123.2019.9015113

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Munich: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Roy, S., Dora, S., McCreadie, K., and Prasad, G. (2020). "MIEEG-GAN: generating artificial motor imagery electroencephalography signals," in *2020 International Joint Conference on Neural Networks (IJCNN)* (Glasgow: IEEE), 1–8. doi: 10.1109/IJCNN48605.2020.9206942

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng.* 16:051001. doi: 10.1088/1741-2552/ab260c

Sakai, A., Minoda, Y., and Morikawa, K. (2017). "Data augmentation methods for machine-learning-based classification of bio-signals," in *2017 10th Biomedical Engineering International Conference (BMEiCON)* (Hokkaido: IEEE), 1–4. doi: 10.1109/BMEiCON.2017.8229109

Sakhavi, S., Guan, C., and Yan, S. (2018). Learning temporal information for brain-computer interface using convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 29, 5619–5629. doi: 10.1109/TNNLS.2018.2789927

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* 38, 5391–5420. doi: 10.1002/hbm.23730

Shovon, T. H., Al Nazi, Z., Dash, S., and Hossain, M. F. (2019). "Classification of motor imagery EEG signals with multi-input convolutional neural network by augmenting STFT," in *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)* (Dhaka: IEEE), 398–403. doi: 10.1109/ICAEE48663.2019.8975578

Simonyan, K., and Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)* (Banff, AB).

Sohn, K., Lee, H., and Yan, X. (2015). "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems (NIPS)*, Vol. 28, 3483–3491.

Suk, H. I., and Lee, S. W. (2012). A novel Bayesian framework for discriminative feature extraction in brain-computer interfaces. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 286–299. doi: 10.1109/TPAMI.2012.69

Sun, C., Fan, J., Chen, C., Li, W., and Chen, W. (2019). A two-stage neural network for sleep stage classification based on feature learning, sequence learning, and data augmentation. *IEEE Access* 7, 109386–109397. doi: 10.1109/ACCESS.2019.2933814

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. (2018). "Learning to compare: relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, UT: IEEE), 1199–1208. doi: 10.1109/CVPR.2018.00131

Supratak, A., and Guo, Y. (2020). "TinySleepNet: an efficient deep learning model for sleep stage scoring based on raw single-channel EEG," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (IEEE), 641–644. doi: 10.1109/EMBC44109.2020.9176741

Tang, X., and Zhang, X. (2020). Conditional adversarial domain adaptation neural network for motor imagery EEG decoding. *Entropy* 22:96. doi: 10.3390/e22010096

Truong, N. D., Kuhlmann, L., Bonyadi, M. R., Querlioz, D., Zhou, L., and Kavehei, O. (2019a). Epileptic seizure forecasting with generative adversarial networks. *IEEE Access* 7, 143999–144009. doi: 10.1109/ACCESS.2019.2944691

Truong, N. D., Zhou, L., and Kavehei, O. (2019b). "Semi-supervised seizure prediction with generative adversarial networks," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Berlin: IEEE), 2369–2372. doi: 10.1109/EMBC.2019.8857755

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI: IEEE), 7167–7176. doi: 10.1109/CVPR.2017.316

Vilamala, A., Madsen, K. H., and Hansen, L. K. (2017). "Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)* (Tokyo: IEEE), 1–6. doi: 10.1109/MLSP.2017.8168133

Wang, F., Zhong, S. h., Peng, J., Jiang, J., and Liu, Y. (2018). "Data augmentation for EEG-based emotion recognition with deep convolutional neural networks," in *International Conference on Multimedia Modeling (ICMM)* (Bangkok: Springer), 82–93. doi: 10.1007/978-3-319-73600-6_8

Wang, M., and Deng, W. (2018). Deep visual domain adaptation: a survey. *Neurocomputing* 312, 135–153. doi: 10.1016/j.neucom.2018.05.083

Wang, X., Li, L., Ye, W., Long, M., and Wang, J. (2019). "Transferable attention for domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (Honolulu, HI), Vol. 33, 5345–5352. doi: 10.1609/aaai.v33i01.33015345

Wang, Y., Koike-Akino, T., and Erdogmus, D. (2018). Invariant representations from adversarially censored autoencoders. *arXiv* 1805.08097.

Wang, Y., Qiu, S., Ma, X., and He, H. (2021). A prototype-based SPD matrix network for domain adaptation EEG emotion recognition. *Pattern Recognit.* 110:107626. doi: 10.1016/j.patcog.2020.107626

Wei, W., Qiu, S., Ma, X., Li, D., Wang, B., and He, H. (2020a). Reducing calibration efforts in RSVP tasks with multisource adversarial domain adaptation. *IEEE Trans. Neural Syst. Rehabil. Eng.* 28, 2344–2355. doi: 10.1109/TNSRE.2020.3023761

Wei, W., Qiu, S., Ma, X., Li, D., Zhang, C., and He, H. (2020b). "A transfer learning framework for RSVP-based brain computer interface," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (IEEE), 2963–2968. doi: 10.1109/EMBC44109.2020.9175581

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision (ECCV)* (Amsterdam: Springer), 499–515. doi: 10.1007/978-3-319-46478-7_31

Won, K., Kwon, M., Jang, S., Ahn, M., and Jun, S. C. (2019). P300 speller performance predictor based on RSVP multi-feature. *Front. Hum. Neurosci.* 13:261. doi: 10.3389/fnhum.2019.00261

Xu, R., Chen, Z., Zuo, W., Yan, J., and Lin, L. (2018). "Deep cocktail network: multi-source unsupervised domain adaptation with category shift," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, UT: IEEE), 3964–3973. doi: 10.1109/CVPR.2018.00417

Zhang, D., Chen, K., Jian, D., and Yao, L. (2020a). Motor imagery classification via temporal attention cues of graph embedded EEG signals. *IEEE J. Biomed. Health Inform.* 24, 2570–2579. doi: 10.1109/JBHI.2020.2967128

Zhang, D., Yao, L., Chen, K., and Monaghan, J. (2019a). A convolutional recurrent attention model for subject-independent EEG signal analysis. *IEEE Signal Process. Lett.* 26, 715–719. doi: 10.1109/LSP.2019.2906824

Zhang, D., Yao, L., Chen, K., and Wang, S. (2018a). "Ready for use: subject-independent movement intention recognition via a convolutional attention model," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)* (Torino), 1763–1766. doi: 10.1145/3269206.3269259

Zhang, D., Yao, L., Chen, K., Wang, S., Haghighi, P. D., and Sullivan, C. (2019b). A graph-based hierarchical attention model for movement intention detection from EEG signals. *IEEE Trans. Neural Syst. Rehabil. Eng.* 27, 2247–2253. doi: 10.1109/TNSRE.2019.2943362

Zhang, G., Davoodnia, V., Sepas-Moghaddam, A., Zhang, Y., and Etemad, A. (2019c). Classification of hand movements from EEG using a deep attention-based LSTM network. *IEEE Sens. J.* 20, 3113–3122. doi: 10.1109/JSEN.2019.2956998

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018b). "Mixup: beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)* (Vancouver, BC).

Zhang, K., Xu, G., Han, Z., Ma, K., Zheng, X., Chen, L., et al. (2020b). Data augmentation for motor imagery signal classification based on a hybrid neural network. *Sensors* 20:4485. doi: 10.3390/s20164485

Zhang, K., Xu, G., Zheng, X., Li, H., Zhang, S., Yu, Y., et al. (2020c). Application of transfer learning in EEG decoding based on brain-computer interfaces: a review. *Sensors* 20:6321. doi: 10.3390/s20216321

Zhang, Q., and Liu, Y. (2018). Improving brain computer interface performance by data augmentation with conditional deep convolutional generative adversarial networks. *arXiv* 1806.07108.

Zhang, R., Zong, Q., Dou, L., Zhao, X., Tang, Y., and Li, Z. (2021). Hybrid deep neural network using transfer learning for EEG motor imagery decoding. *Biomed. Signal Proces.* 63:102144. doi: 10.1016/j.bspc.2020.102144

Zhang, W., Ouyang, W., Li, W., and Xu, D. (2018c). "Collaborative and adversarial network for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, UT: IEEE), 3801–3809. doi: 10.1109/CVPR.2018.00400

Zhang, X., Yao, L., Wang, X., Monaghan, J. J., Mcalpine, D., and Zhang, Y. (2020d). A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers. *J. Neural Eng.* 18:031002l. doi: 10.1088/1741-2552/abc902

Zhang, X. Z., Zheng, W. L., and Lu, B. L. (2017). "EEG-based sleep quality evaluation with deep transfer learning," in *International Conference on Neural Information Processing (ICNIP)* (Guangzhou: Springer), 543–552. doi: 10.1007/978-3-319-70093-9_57

Zhang, Z., Duan, F., Sole-Casals, J., Dinares-Ferran, J., Cichocki, A., Yang, Z., et al. (2019d). A novel deep learning approach with data augmentation to classify motor imagery signals. *IEEE Access* 7, 15945–15954. doi: 10.1109/ACCESS.2019.2895133

Zhao, D., Tang, F., Si, B., and Feng, X. (2019). Learning joint space-time-frequency features for EEG decoding on small labeled data. *Neural Netw.* 114, 67–77. doi: 10.1016/j.neunet.2019.02.009

Zhao, H., Zheng, Q., Ma, K., Li, H., and Zheng, Y. (2020). Deep representation-based domain adaptation for nonstationary EEG classification. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 535–545. doi: 10.1109/TNNLS.2020.3010780

Zhao, X., Solé-Casals, J., Li, B., Huang, Z., Wang, A., Cao, J., et al. (2020). "Classification of epileptic IEEG signals by CNN and data augmentation," in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona: IEEE), 926–930. doi: 10.1109/ICASSP40776.2020.9052948

Zhuang, F., Cheng, X., Luo, P., Pan, S. J., and He, Q. (2015). "Supervised representation learning: transfer learning with deep autoencoders," in *24th International Joint Conference on Artificial Intelligence (IJCAI)* (Buenos Aires), 4119–4125.

Zintgraf, L., Shiarli, K., Kurin, V., Hofmann, K., and Whiteson, S. (2019). "Fast context adaptation via meta-learning," in *International Conference on Machine Learning (ICML)* (Long Beach, CA: PMLR), 7693–7702.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX: LIST OF ACRONYMS

A-cVAE	Adversarial conditional variational autoencoder
AE	Autoencoder
ASSOM	Adaptive subspace self-organizing map
BCI	Brain–computer interface
BMU	Best matching unit
BN	Batch normalization
CCE	Categorical cross-entropy
CDFL	Center-based discriminative feature learning
CNN	Convolutional neural network
CSP	Common spatial pattern
cVAE	Conditional variational autoencoder
DA	Data augmentation
DANN	Domain adversarial neural network
DCGAN	Deep convolutional generative adversarial network
DCT	Discrete cosine transform
DL	Deep learning
EA	Euclidean alignment
EEG	Electroencephalography
EMD	Empirical mode decomposition
ERP	Event-related potential
GAN	Generative adversarial network
GRL	Gradient reversal layer
GRU	Gated recurrent unit
IMF	Intrinsic mode functions
JSD	Jensen-Shannon distance
KLD	Kullback-Leibler divergence
LOO	Leave-one subject-out
LSGAN	Least square generative adversarial network
LSTM	Long-short term memory
MAML	Model-agnostic meta learning
MINE	Mutual information neural estimator
MMD	Maximum mean discrepancy
RKHS	Reproducing kernel Hilbert space
RSVP	Rapid serial visual presentation
SMOTE	Synthetic minority oversampling technique
SOM	Self-organizing map
SPD	Symmetric positive definite
SSVEP	Steady-state visual evoked potential
STFT	Short-time Fourier transform
TL	Transfer learning
VAE	Variational autoencoder
WGAN	Wasserstein generative adversarial network

frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 23 June 2021
doi: 10.3389/fnhum.2021.653659

BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn From Massive Amounts of EEG Data

Demetres Kostas[1,2]*, *Stéphane Aroca-Ouellette*[1,2] *and Frank Rudzicz*[1,2,3]

[1] *Department Computer Science, University of Toronto, Toronto, ON, Canada,* [2] *Vector Institute for Artificial Intelligence, Toronto, ON, Canada,* [3] *Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, ON, Canada*

Deep neural networks (DNNs) used for brain–computer interface (BCI) classification are commonly expected to learn general features when trained across a variety of contexts, such that these features could be fine-tuned to specific contexts. While some success is found in such an approach, we suggest that this interpretation is limited and an alternative would better leverage the newly (publicly) available massive electroencephalography (EEG) datasets. We consider how to adapt techniques and architectures used for language modeling (LM) that appear capable of ingesting awesome amounts of data toward the development of encephalography modeling with DNNs in the same vein. We specifically adapt an approach effectively used for automatic speech recognition, which similarly (to LMs) uses a self-supervised training objective to learn compressed representations of raw data signals. After adaptation to EEG, we find that a single pre-trained model is capable of modeling completely novel raw EEG sequences recorded with differing hardware, and different subjects performing different tasks. Furthermore, both the internal representations of this model and the entire architecture can be fine-tuned to a *variety* of downstream BCI and EEG classification tasks, outperforming prior work in more *task-specific* (sleep stage classification) self-supervision.

Keywords: brain computer interface, deep learning - artificial neural network, transformers, semi-supervised learning, contrastive learning, convolutional neural network, sequence modeling

OPEN ACCESS

Edited by:
Sung Chan Jun, Gwangju Institute of Science and Technology, South Korea

Reviewed by:
Dalin Zhang, Aalborg University, Denmark
Tomasz Maciej Rutkowski, RIKEN Center for Advanced Intelligence Project (AIP), Japan

***Correspondence:**
Demetres Kostas
demetres@cs.toronto.edu

Specialty section:
This article was submitted to Brain-Computer Interfaces, a section of the journal Frontiers in Human Neuroscience

Received: *14 January 2021*
Accepted: *23 April 2021*
Published: *23 June 2021*

Citation:
Kostas D, Aroca-Ouellette S and Rudzicz F (2021) BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn From Massive Amounts of EEG Data. Front. Hum. Neurosci. 15:653659. doi: 10.3389/fnhum.2021.653659

1. INTRODUCTION

To classify raw electroencephalography (EEG) using deep neural network models (DNNs), these models need to both develop useful features from EEG signals and subsequently classify those features. This frames both the promise and the challenge of using DNNs for supervised EEG classification. On the one hand, it promises to almost entirely circumvent the need for feature engineering, but on the other hand, both feature discovery and classification need to be learned from a *limited*[1] supply of (relevant) high-dimensional data. A paradigmatic way in which we observe this challenge is with brain–computer interface (BCI) applications[2] (Lotte et al., 2018; Roy et al., 2019; Kostas and Rudzicz, 2020b). Shallower neural network models have tended to be more

[1] Consider the difficulty of collecting and labeling 100 more BCI trials as compared to the same for 100 more images.
[2] Though we believe that it is likely that a similar tendency to what we characterize herein holds for most applications of DNNs outside of their core artificial intelligence applications.

effective classifiers than their deeper counterparts in BCI (markedly so when trained independently for each user) (Schirrmeister et al., 2017; Lawhern et al., 2018; Lotte et al., 2018; Roy et al., 2019; Kostas and Rudzicz, 2020b). With these shallower networks, the range of *learnable* features is relatively limited. By design, they employ constrained linear operations, and a limited few include non-linear activations between subsequent layers (Kostas and Rudzicz, 2020b), an otherwise crucial feature of DNN complexity. In prior work, we observed that if some inter-personal variability had been adjusted, the performance of shallower models more quickly saturated to lower performance levels as compared to a deeper network alternative (Kostas and Rudzicz, 2020b), suggesting that more complex raw-BCI-trial features *could* be developed using deeper neural networks when using training data that was more consistent. Understood differently, overcoming the limitations of shallower networks in favor of deeper DNNs that *could* surpass feature engineering approaches likely requires addressing the large variability between different contexts.

A natural framework to understand this problem is transfer learning (TL), which is an area of machine learning that aims to leverage knowledge learned from one context such that it may be useful in a different one. Consider a supervised learning problem, which consists of first, a domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$, itself a representation of a feature space $\mathcal{X}$ (e.g., the set of all possible raw EEG recordings of a certain length) and the probability $P(X)$ of observing a particular configuration of features ($x \in X$, e.g., a particular observation of a raw EEG recording). Second, a task $\mathcal{T} = \{\mathcal{Y}, f(x)\}$, a representation of the possible labels for a particular task, and a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ that maps individual instances to the correct labels. TL means to break down a problem into *source* and *target* problems, with $\mathcal{D}_S \neq \mathcal{D}_T$ (and/)or $\mathcal{T}_S \neq \mathcal{T}_T$.

Evidence abounds in BCI and EEG generally that differences in domain are a critical challenge. For example, under the sensory motor rhythm (SMR) BCI paradigm, different subjects exhibit extremely different capacities at performing this task, and even different sessions from the same users can exhibit enough variation that classifiers trained in one session are ill-suited to the next (Vidaurre and Blankertz, 2010; Ahn and Jun, 2015; Sannelli et al., 2019). This indicates that (at least for the feature representations being considered) the domain of each person and even session differs. Beyond these inter- and intra-personal variations, different features are relevant for different BCI tasks. Hand-selected features (sets possibly pruned later on) are also typically distinct under different BCI paradigms, as different features better discriminate different tasks[3] (Lotte et al., 2018), e.g., P300 vs. SMR. Thus, an explicit imposition of difference in domain is imposed between different BCI task paradigms (as their feature spaces are distinct, e.g., $\mathcal{X}_{SMR} \neq \mathcal{X}_{P300}$), which to us implies that it is fair to expect that this is indicative of strong differences in domain (and of course task) when considering *raw* data. In other words, the very effort of selecting different features for different tasks (rather than only changing classifier) is recognition of a difference in domain. Furthermore, we have found in previous work that the different domains represented by particular individuals seem to be *readily*[4] identifiable from arbitrary raw sequences of EEG using DNNs (Kostas and Rudzicz, 2020a). In summary, a DNN trained with a certain set of contexts (e.g., subjects), intent on transferable performance to novel contexts (e.g., an unseen subject), is required to develop some universal features and/or classifier for possible novel target domains from the sources it was prepared with. Some have argued that this universality is achievable through the selection of the right DNN, or DNN layers (Cimtay and Ekmekcioglu, 2020; Zhang et al., 2020a), but through a questioning of the apparent ideal approaches to TL in the wider DNN literature (presented in section 1.1), we argue that the development of such universal features requires developing pre-training procedures that transfer from *general tasks* to *specific* ones instead.

The interest in these universal, or invariant features are not however limited to better classification performance, but may be of wider importance. While it may be difficult to determine within a DNN when "features" start and "classifier" begins, in applications such as computer vision there is a clear understanding that nearly all transferrable DNNs have tended to learn "low-level" features in *earlier layers* (e.g., edge-detector-like primitives) (Krizhevsky et al., 2012; Yosinski et al., 2015; Raghu et al., 2019). The promise of some such transferable early layers or operations that are easily extended to any subject, session, or task may open valuable lines of inquiry, or novel explicit (rather than implicitly learned) methods (say if these early layers do or do not correspond to existing methodologies, respectively) of analysis. Importantly, the determination of which "low-level" features DNNs developed in computer vision was revealed through models that *had* transferable performance from general to specific tasks (Yosinski et al., 2015; Raghu et al., 2019).

In this work, we argue that self-supervised sequence learning is such a general task. It would be an effective approach for developing and deploying more complex and universal DNNs in BCI and in potentially wider EEG-based analysis. We present a methodology that can learn from many people, sessions, and tasks using *unlabeled* data; in other words, it samples the more general distribution of EEG data. Thus, we attempt to learn $\mathcal{D}_{EEG}$ with self-descriptive features, with the goal that they exhibit little variability across typical context boundaries (invariant between expected domains) like dataset and subjects. More specifically, we investigate techniques inspired by language modeling (LM) that have found recent success in self-supervised end-to-end speech recognition and image recognition in an effort to develop *encephalography models* (EM). We first begin by investigating fully supervised transfer learning (which has been frequently looked to as an EEG/BCI TL solution), finding inconsistency in the extension of computer vision-style pre-training to BCI

[3]While this is typical, some procedures, like covariance-based Riemannian classification schemes, do not necessarily need different features for different tasks (Lotte et al., 2018; Zanini et al., 2018). These are a very interesting exception to the argument we develop.

[4]With a strong latent representation, a nearest-neighbors labeling is sufficient to be nearly 100% accurate for some datasets, despite being recordings made on different hardware (Kostas and Rudzicz, 2020a).

(and by extension the data domain of EEG). We then evaluate a simple adaptation of previous work in self-supervised speech recognition called `wav2vec 2.0` (Baevski et al., 2020) to EEG. With this framework, arbitrary EEG segments are encoded as a sequence of learned vectors we call BErt-inspired Neural Data Representations (or "BENDR"). We ask whether BENDR are transferable to unseen EEG datasets recorded from unseen subjects, different hardware, and different tasks, and how generally suitable BENDR are (both as-is or fine-tuned) to a battery of downstream EEG classification tasks with respect to the same architecture without first being trained with more general EEG data (i.e., "pre-training").

1.1. Pre-training With DNNs

For inspiration on tackling DNN transfer learning in BCI, one can look to other successful approaches, starting with the modern deep learning (DL) "revolution," which was ushered in on the back of computer vision and image recognition (LeCun et al., 2015; Sejnowski, 2020). The successes of DL in these applications have stemmed from a lineage of massive *labeled* datasets (LeCun et al., 2015), such as the ImageNet dataset (Deng et al., 2009). These datasets were (and are) used to train deep convolutional neural networks, often one of the variants or progeny of ResNet (He et al., 2016) and DenseNet (Huang et al., 2017). Crucially, these are labeled datasets, featuring—especially in the case of ImageNet—an enormous number of unique possible classification labels (or equivalently *targets*, with 1000 being common when using ImageNet[5], but more are possible[6]). Leveraging labeled data (especially for a singular domain such as a single subject, session and task) of a similar scale in BCI is impractical but, despite this, a sizeable amount of prior work tries to fashion a transfer learning strategy after the successes of ImageNet "pre-training." These take the form of transferring knowledge from a network *pre-trained* with more data, typically more subjects, to a target domain with less data, typically a single subject (Lin and Jung, 2017; Dose et al., 2018; Schwemmer et al., 2018; Fahimi et al., 2019; Xu et al., 2019; Cimtay and Ekmekcioglu, 2020; Kostas and Rudzicz, 2020b; Zhang et al., 2020a), with some work transferring between entire datasets of the same paradigm, rather than subjects (Ditthapron et al., 2019). On the surface, these embody a *general-to-specific* supervised transfer learning scheme *reminiscent* of ImageNet pre-training where models trained on an ImageNet problem are adapted to a novel (but related) application. However, these particular framings lack the label *diversity* when pre-training with ImageNet. In other words, a narrow set of labels are used to pre-train a model, and these simply overlap with the target context, i.e., $\mathcal{Y}_S = \mathcal{Y}_T$. This approach is in fact distinct from the approach taken as inspiration where $\mathcal{Y}_S \neq \mathcal{Y}_T$ (or possibly $\mathcal{Y}_S \subset \mathcal{Y}_T$). We remain unaware of any work that pre-trains a DNN using a *wide gamut of BCI-relevant targets* in the services of a *more narrow* target set, as would be more analogous to using ImageNet as pre-training toward more specific computer vision tasks[7]. This is noteworthy, as this is what makes ImageNet a *general task*. Evidence suggests that pre-training label diversity is important for effective ImageNet transfer learning (Huh et al., 2016), though an excess could be detrimental (Huh et al., 2016; Ngiam et al., 2018). Furthermore, this general task appears to be responsible for developing the *transferable* early layers (Raghu et al., 2019; Neyshabur et al., 2020) that would seem to embody the desired goal of overcoming "hand-crafted" or developing "invariant" features, and partially appear to be learning data statistics (Neyshabur et al., 2020) [i.e., $P(X)$; recall this as one aspect of a domain for a supervised learning problem, the other is the feature representation]. More fundamentally, however, this pre-training paradigm has begun to be questioned altogether, with some work finding that it does not necessarily improve downstream performance, where commonly it has been assumed that it should (e.g., in medical images or object localization; though it *speeds up* training considerably) (Ngiam et al., 2018; He et al., 2019; Kornblith et al., 2019; Raghu et al., 2019).

1.2. Are There Alternatives?

What has begun to emerge as a potential alternative in computer vision—and markedly so when there is limited labeled downstream data—is self-supervised learning (Chen et al., 2016; van den Oord et al., 2018; Grill et al., 2020; Hénaff, 2020)[8]. These works are inspired by the recent success in natural language processing (NLP) using LMs, which can be used to greatly affect the transfer learning, but also for few-shot and zero-shot learning (Brown et al., 2020; Raffel et al., 2020). These models are understood to work by making a very general model of language and appear even immediately capable of performing tasks they were not explicitly trained to accomplish. We propose that DNN transfer learning in BCI and neuroimaging analysis generally could follow a similar line, with *encephalography models* (EM) in place of LMs. The important question being *how best to construct such an EM so that it learns features that are general enough, while remaining usable for any analysis task?*

To our knowledge, the most similar prior work to this line of inquiry has been the approaches developed for (EEG) self-supervised sleep stage classification (SSC) through contrastive learning (Banville et al., 2019). Contrastive learning is a more particular, yet generally applicable training process that consists of identifying positive representations from a set that also includes incorrect or negative distractor representations (Arora et al., 2019). Banville et al. proposed two potential contrastive learning tasks—a "relative positioning" task and an extension they termed "temporal shuffling" (Banville et al., 2019). Underlying both tasks is the notion that neighboring

[5]image-net.org/challenges/LSVRC/2012/
[6]https://www.image-net.org/index.php

[7]It is also worth noting that our own prior work does not consider or identify this.
[8]Terminology here can be somewhat fuzzy. What is meant by self-supervision is a supervision-like task that requires domain-relevant understanding in some sense. Sometimes, "semi-supervised" is used instead, as it is often also a semi-supervised procedure (Chen et al., 2016), since the task is learned in an unsupervised fashion first and then classic supervised learning is used with labels. Typically, though, semi-supervision involves inferring labels for unlabeled data during training. Instead, self-supervision is loosely a particular case of representation learning, which is not historically uncommon in BCI (Zhang et al., 2020b). Though this work is different given that typically the loss is domain or data agnostic.

representations share a label. The representations themselves are a learned mapping (in their case, a convolutional neural network, but ostensibly arbitrary) of raw EEG time-windows to a feature vector. This assumption of similar neighboring labels is fair for SSC, where sleep stages change slowly, and is generally reasonable for continuous problems, where some notion of smoothness can be assumed. Their proposed "relative positioning" task is a binary classification problem distinguishing whether a pair of representations are within a local or positive window τ_{pos}, or outside a long-range or negative window τ_{neg} (when $\tau_{neg} > \tau_{pos}$, those falling within τ_{neg} but outside τ_{pos} are ignored). Their alternative "temporal shuffling" method adds a third window or representation with which to contrast that is within τ_{pos} of one (arbitrary) window called the "anchor," and again learns the representations through a binary classification task. In this case, the classification determines whether the three representations are ordered sequentially, or are out of order. *Downstream* (loose terminology used to mean the step after pre-training when a model is leveraged and evaluated for a particular task), both contrastive learning tasks ultimately improved SSC classification performance over the *same* network trained in a fully supervised manner *from scratch* (with randomly initialized weights rather than those that accomplish the self-supervised task) and their results further agree with the common finding that self-supervision appears distinctly better with limited fine-tuning[9] data (Brown et al., 2020; Chen et al., 2020). Furthermore, self-supervised pre-training also outperformed an autoencoder-based pretraining, an alternative and historically common pretraining option where a network is pretrained to reconstruct its original input. "Relative positioning" performed better on average (and no statistical significance expressed) when compared to its counterpart, but a linear classification of simple hand-crafted features was still highest performing overall. These results demonstrate the promise of self-supervised learning with DNNs for EEG over a supervised approach, but contextualize them as early in development. This is perhaps best seen by considering the lengths of the time windows (τ_{pos} and τ_{neg}). The shortest windows employed in this particular investigation were 2 min for τ_{pos} and τ_{neg}, which seems prohibitively long for most immediate applications outside of SSC. As it is assumed that representations within τ_{pos} are similarly labeled, it may be difficult to expand the use of this technique to time scales closer to that of say, a BCI trial (across any paradigm), which tend to be no more than several seconds at most. In this work, we focus our efforts on adapting a relevant strategy from the wider ML literature that could develop features on smaller time scales effective for BCI trials *as well as* time scales appropriate for SSC.

Returning to a consideration of how one might adapt LM pre-training to EM, the *masked* language model (MLM) is a slight variation on the typical LM that has been essential to the success of recent LMs like BERT (Devlin et al., 2019) and its lineage (Raffel et al., 2020) of similar models. Where a LM estimates the probability of encountering a language token (a word or *sub-word* Aroca-Ouellette and Rudzicz, 2020) given previous (or, in some cases, also subsequent) tokens, a MLM scheme instead learns to *reconstruct* language token(s) given surrounding context (fashioned after the Cloze task). This family of models may deploy a variety of auxiliary tasks (Aroca-Ouellette and Rudzicz, 2020) for transfer learning capabilities, but the task currently at the heart of this family is as follows: given a sequence of N tokens $t_1, \ldots t_N$, and a subset of token indexes I_m, for each token index $i \in I_m$, tokens are masked with some mask M so that:

$$q_i = \begin{cases} M; i \in I_m \\ t_i; \text{otherwise} \end{cases}, \forall i \in N \tag{1}$$

A transformer encoder (Vaswani et al., 2017; Devlin et al., 2019) then reconstructs the original sequence of tokens from the *masked* sequence [t_i and $q_i, \forall i \in N$, respectively, in Equation (1)]. M could be a single learned token (Baevski et al., 2020), or in the case of BERT: 80% of the time a fixed `[MASK]` token, 10% a random token or 10% the original token (with 15% of tokens masked within each sequence) (Devlin et al., 2019).

Could an EM be developed in this vein, using say, individual samples rather than tokens (i.e., could a direct application of the above be done with raw EEG)? Unfortunately, the highly correlated nature of neighboring samples in EEG (or most other continuous data for that matter) is not conducive to this approach. The likely result would be that, instead of an EM, a method for interpolation would be learned, the model would simply learn how to average neighboring samples, as has been argued in similar work in self-supervised learning with speech (Jiang et al., 2020). In other words, the smoothness of these data would make it hard to produce general features simply through recovering missing individual samples. Masking a contiguous span of tokens instead, which is beneficial in NLP (Joshi et al., 2020; Raffel et al., 2020), could avoid simply learning to interpolate missing samples, but the *reconstruction* of time-series data is difficult, due to the challenge (among other things) of capturing the degree of error in time (within contiguous sequences) (Rivest and Kohar, 2020). The losses used for such reconstruction, commonly mean squared error (or mean absolute error), erroneously assume independence in the error between elements in the series, causing inappropriate error signals when (among other things) simply shifting a reconstruction in time (Rivest and Kohar, 2020).

Contrastive predictive coding (CPC), is a particular contrastive learning approach that is intended for sequence learning. With CPC, the correct *learned representation* for a particular sequence offset is predicted relative to distractor representations, typically those of other positions in the same sequence (van den Oord et al., 2018). What is notable about this is that it is not as susceptible to degeneration into interpolation, nor is it similarly affected by the issues of time-series reconstruction (van den Oord et al., 2018). This task enables learning both a good feature representation *and* an understanding of the sequence of data by modeling the progression of the representations, learned with a single loss function. Indeed, the RP and TS tasks discussed above for SSC can be understood as special cases of the more general CPC,

[9]As is perhaps obvious in the name, though potentially misleading. Fine-tuning is the process of further training on a reserved portion of a target dataset, unless stated otherwise, this is typically through standard supervised training.

though performance appears largely similar when comparing all three (Banville et al., 2020).

Prior work in self-supervised *speech recognition* has begun to synthesize parts of CPC and MLM to produce methodologies for self-learning with raw waveforms (van den Oord et al., 2018; Baevski and Mohamed, 2020; Baevski et al., 2020; Chung et al., 2020; Jiang et al., 2020). In our work, we adapt one of these approaches called `wav2vec 2.0` (Baevski et al., 2020) (its particular formulation is detailed in section 2.4.1) to EEG. We consider how efficient the approach is at developing representations (BENDR), and how general these and the accompanying sequence model are across multiple task paradigms/datasets (not seen during pre-training) and across the subjects that constitute them. Since interestingly, both the representations alone (Chen et al., 2020), and the addition of the sequence model (Baevski et al., 2020) have proven potentially useful for supervised fine-tuning *after* pre-training, we then characterize a variety of "fine-tuning" approaches "downstream." In other words, finally, we compare which aspect of our overall scheme is best leveraged and how toward classifying a variety of publicly available EEG classification task datasets.

2. MATERIALS AND METHODS

All experiments are implemented using the *deep neural networks for neurophysiology* (DN3) library[10]. The source code and pre-trained BENDR models can be found at https://github.com/SPOClab-ca/BENDR.

2.1. Datasets

2.1.1. Pre-training

We intend to learn our proposed general task across a large number of typically confounding domains, which means the ideal pre-training dataset for our purposes would feature many subjects, each recorded over many sessions. These sessions would also ideally be distributed across large time scales and consist of a variety of performed tasks. In other words, the pre-training dataset should consist of a representative sample of EEG data. This also means that these data should include multiple different recording hardware and configurations. The closest publicly accessible dataset, to our current knowledge, was the Temple University Hospital EEG Corpus (TUEG) (Obeid and Picone, 2016). It consists of clinical recordings using a mostly conventional recording configuration (monopolar electrodes in a 10–20 configuration) of over 10,000 people, some with recording sessions separated by as much as 8 months apart. The subjects were 51% female, and ages range from under 1 years old to over 90 (Obeid and Picone, 2016). We focused specifically on versions 1.1 and 1.2 of this dataset which amounted to approximately 1.5 TB of European-data-format (EDF) EEG recordings *before* preprocessing.

2.1.2. Downstream

To investigate the practical utility of the learned representations, we compiled a non-exhaustive battery of publicly accessible EEG data classification tasks—or *downstream tasks*—summarized in **Table 1**. Most of these were BCI task datasets, which could readily be compared to previous work with DNNs trained without any additional unlabeled data (Lawhern et al., 2018; Kostas and Rudzicz, 2020b). We also included one of the SSC tasks used by Banville et al. (2019) in their work on sleep stage self-supervision described above, for comparison. This particular dataset afforded some further insight into generality, as BCI data are typically classified in the context of particular trials or events, and SSC is a more continuous problem, requiring that large spans of time are labeled with the particular sleep stage a subject is undergoing. These segments are distinctly longer than the BCI trials we considered in the remaining battery (an order of magnitude difference in our case when compared to the largest BCI task sequence length), and are distinctly closer in length to the way the pre-training task is formulated (see section 2.4.1). We specifically segmented these sequences into periods of 30 s to be classified into 5 sleep stages as in prior work (Banville et al., 2019; Mousavi et al., 2019). Another potentially notable difference with the SSC dataset was the scale of available labels, which seems to have enabled prior work to consider deeper and more complex models (Mousavi et al., 2019).

2.2. Preprocessing

The focus of the preprocessing stage was to create a maximally consistent representation of EEG sequences across datasets (which implied differences in hardware), so that a pre-trained network was well suited to the downstream tasks. More or less, this amounted to modifying downstream datasets to match the configuration of the pre-training dataset. The first aspect of this was to remove spurious differences in channel amplitude. Each sequence gathered for training was linearly scaled and shifted (a weight and offset for each sequence adjusts every sample in the sequence) so that the maximum and minimum values within each sequence equal 1 and −1, respectively. To account for the lost relative (to the entire dataset) amplitude information, a single channel was added with the constant value $\frac{max(s_i)-min(s_i)}{max(S_{ds})-min(S_{ds})}$, where S_{ds} is the set of all samples in the dataset and $s_i \subset S_{ds}$ is a particular sub-sequence (i.e., trial). We additionally addressed the differences in sampling frequency and electrode sets of the different dataset. Our solutions to these problems were similarly minimalist and were achieved using standard features in DN3 (Kostas and Rudzicz, 2020a). Specifically, we over- or undersampled (by whole multiples, for lower and higher sampling frequencies, respectfully) to get nearest to the target sampling frequency of 256 H*z*. Then, nearest-neighbor interpolation was used to obtain the precise frequency (as was done in prior work Kostas and Rudzicz, 2020a). Additionally, the P300 dataset was low-pass filtered below 120 H*z* to avoid aliasing due to its higher sampling rate (and associated higher original low-pass filter). Furthermore, the SSC dataset featured two bi-polar electrodes: FPz-Cz and Pz-Oz, which were simply mapped to FPz and Pz, respectively. The TUEG dataset itself featured some higher sampling rate signals; we included those with low-pass filters that did not violate the Nyquist criterion

[10]https://github.com/SPOClab-ca/dn3

TABLE 1 | Summary of downstream dataset battery and number of cross-validation folds used.

Dataset	Paradigm	sfreq. Hz	# Ch.	Subjects	Targets	Folds
MMI Goldberger et al. (2000), Schalk et al. (2004)	MI (L/R)	160	64	105	2	5
BCIC Tangermann et al. (2012)	MI (L/R/F/T)	250	22	9	4	9
ERN Margaux et al. (2012)	Error related negativity	200	56	26 (10)	2	4
P300 Goldberger et al. (2000), Citi et al. (2010, 2014)	Donchin speller	2,048	64	9	2	9
SSC Goldberger et al. (2000), Kemp et al. (2000, 2018)	Sleep Staging	100	2	83	5	10

Cross-validation splits were in a leave-multiple-subjects-out configuration if Folds < Subjects, or leave-one-subject-out if Folds = Subjects (as in prior work Kostas and Rudzicz, 2020b). The ERN dataset was featured in an online competition[11] which featured 10 held-out test subjects (not used during training), which we used as a test dataset for all four validation splits of this dataset.

TABLE 2 | Performances of downstream datasets.

Dataset	Start (s)	Length (s)	Metric	Best	Model config.
MMI	0	6	BAC	86.7	Linear (2.)
BCIC	–2	6	Accuracy	42.6	Linear (2.)
ERN	–0.7	2	AUROC	0.65	Linear (2.)
SSC	0	30	BAC	0.72	Linear (2.)
P300	–0.7	2	AUROC	0.72	BENDR (1.)

Start and length refer to length of trials and start with respect to event markers in seconds. Best performance specifies average performance across all subjects (and therefore folds) for best performing model configuration. BAC: class balanced accuracy; AUROC: area under the receiver operating characteristic curve. Model configurations are numbered in accordance with the list presented in section 2.4.2.

(and subsequently re-sampled them as above), and ignored the rest.

A reduced subset of the `Deep1010` channel mapping from DN3 (Kostas and Rudzicz, 2020a) was used throughout. This ensured that particular channels were mapped to a consistent index for each loaded trial. The original mapping was designed to be more inclusive, and thus assumed up to 77 possible EEG electrodes. In the interest of minimizing unnecessary electrodes for an already high-dimensional problem, we focused on the 19 EEG channels of the *unambiguously illustrated 10/20* channel set (UI 10/20) (Jurcak et al., 2007), as the TUEG dataset recordings were done using a roughly 10/20 channel scheme. We simply ignored reference electrodes, electro-oculograms, and any other auxiliary channels. When also accounting for the additional relative amplitude channel described above, every sequence from every dataset used 20 channels. All surplus channels were ignored, and missing channels set to 0.

During pre-training, we extracted sequences of 60 s (every 60 s) from each usable sequence, which amounted to 15,360 samples per subsequence. We observed in early testing that there was better performance with larger sequences (see **Figure 4** for more). As can be seen in **Table 2**, the downstream datasets all classified sequence lengths shorter than this, but the architecture we employed (see section 2.3) was ostensibly agnostic to sequence length (see section 4 for caveats).

2.3. Model Architecture

The model architecture displayed in **Figure 1** closely follows that of `wav2vec 2.0` (Baevski et al., 2020) and is composed of two stages. A first stage takes raw data and dramatically downsamples it to a new sequence of vectors using a stack of short-receptive field 1D convolutions. The product of this stage is what we call BENDR (specifically in our case, when trained with EEG). A second stage uses a transformer *encoder* (Vaswani et al., 2017) (layered, multi-head self-attention) to map BENDR to some new sequence that embodies the target task.

Raw data are downsampled through the stride (number of skipped samples) of each convolution block in the first stage (rather than pooling, which would require greater memory requirements). Each of our convolution blocks composed of the sequence: 1D convolution, GroupNorm (Wu and He, 2020), and GELU activation (Hendrycks and Gimpel, 2016). Our own encoder features six sequential blocks, each with receptive fields of 2, except for the first block, which has 3. Strides match the length of the receptive field for each block. Thus, the *effective sampling frequency* of BENDR is 96 times smaller ($\approx$ 2.67 Hz) than the original sampling frequency (256 Hz). Each block consists of 512 filters, meaning each resulting vector has a length of 512.

The transformer follows the standard implementation of Vaswani et al. (2017), but with internal batch normalization layers removed and with an accompanying weight initialization scheme known as T-Fixup (Huang et al., 2020). Our particular transformer architecture uses 8 layers, with 8 heads, model dimension of 1536 and an internal feed-forward dimension of 3076. As with `wav2vec 2.0`, we use GELU activations (Hendrycks and Gimpel, 2016) in the transformer, and additionally include LayerDrop (Fan et al., 2019) and Dropout at probabilities 0.01 and 0.15, respectively, during pre-training but neither during fine-tuning. We represent position using an

[11] https://www.kaggle.com/c/inria-bci-challenge

additive (grouped) convolution layer (Mohamed et al., 2019; Baevski et al., 2020) with a receptive field of 25 and 16 groups before the input to the transformer. This allows the entire architecture to be sequence-length independent, although it may come at the expense of not properly understanding position for short sequences.

Originally, the downstream target of the `wav2vec 2.0` process was a speech recognition *sequence* (it was fine-tuned on a sequence of characters or phonemes) (Baevski et al., 2020). Instead, here the entire sequence is classified. To do this using a transformer, we adopt the common practice (Devlin et al., 2019) of feeding a fixed token (*a.k.a.* `[CLS]` in the case of BERT or, in our case, a vector filled with an arbitrary value distinct from the input signal range, in this case: −5) as the first sequence input (prepended to BENDR). The transformer output of this initial position was not modified during pre-training, and only used for downstream tasks.

The most fundamental differences in our work as compared to that of the speech-specific architecture that inspired it are as follows: (1) we do not quantize BENDR for creating pre-training *targets*, and (2) we have *many* incoming channels. In `wav2vec 2.0`, a *single* channel of raw audio was used. While a good deal of evidence (Schirrmeister et al., 2017; Chambon et al., 2018; Lawhern et al., 2018; Lotte et al., 2018; Kostas et al., 2019; Kostas and Rudzicz, 2020b) supports the advantage of temporally focused stages (no EEG channel mixing) separate from a stage (or more) that integrates channels, we elected to preserve the 1D convolutions of the original work to minimize any additional confound and to reduce complexity (compute and memory utilization $\propto N_{filters}$ with 2D rather than $\propto \frac{N_{filters}}{N_{EEG}}$ for 1D convolutions). This seemed fair, as there is also evidence that 1D convolutions are effective feature extractors for EEG, particularly with large amounts of data (Gemein et al., 2020; Kostas and Rudzicz, 2020a). Notably, `wav2vec 2.0` downsampled raw audio signals by a much larger factor (320) than our own scheme, but speech information is localized at much higher frequencies than encephalographic data are expected to be. The new effective sampling rate of BENDR is $\approx$ 2.67 Hz, or a feature-window (no overlap) of $\approx$ 375 *ms*. We selected this downsampling factor as it remained stable (i.e., it did not degenerate to an infinite loss, or simply memorize everything immediately) during training.

2.4. Training and Evaluation

We used the Adam (Kingma and Ba, 2015) optimizer throughout training (during pre-training and fine-tuning with downstream data), with weight decay set to 0.01. We additionally used a cosine learning rate decay with linear warm-up for 5 and 10% of total training steps (batches) for pre-training and fine-tuning, respectively. The peak learning rate itself varied by dataset; this and other variable hyperparameters are further documented in **Appendix A**.

2.4.1. Pre-training

The pre-training procedure largely follows `wav2vec 2.0` but we make some notable hyperparameter changes documented below. The procedure itself is as follows: first, the convolutional stage of the overall architecture develops a sequence of representations (in our case BENDR) that summarizes the original input. An input token is prepended to this sequence (a BENDR-lengthed vector filled with −5), and contiguous spans of the remaining sequence are masked. This modified sequence is provided as input to the transformer stage, which is expected to develop outputs that are *most similar* to the *un-masked* input at a position t. Specifically, we use the self-supervised loss function for a masked token localized at t:

$$\mathcal{L} = -log\frac{exp(cossim(c_t, b_t))/\kappa}{\sum_{b_i \in B_D} exp(cossim(c_t, b_i))/\kappa} \tag{2}$$

where c_t is the output of the transformer at position t, b_i is the (original/un-masked) BENDR vector at some offset i, and B_D is a set of 20 uniformly selected distractors/negatives from the same sequence, plus b_t. We use the cosine similarity $cossim(x, y) = x^T y/(|x||y|)$ function to determine how similar vectors are, and the sensitivity of this is adjusted by a temperature factor κ, set to 0.1. This loss is expected to operate by adjusting the output of the transformer at position t to be *most similar to the encoded representation at t, despite that this input to the transformer is masked*. This means the transformer must learn a general enough model of *BENDR* (not EEG *per se*) such that the entire sequence of BENDR can characterize position t well. We also add the mean squared activation of the BENDR to the loss to keep the activations from growing too large, as was similarly done previously (Baevski et al., 2020), but we set the weight of this additional term to 1 (rather than 10).

Contiguous sequences of 10 BENDR are masked before input to the transformer with probability $p_{mask} = 0.065$, such that, for each sample, the likelihood of being the *beginning* of a contiguous section was p_{mask}, and overlap is allowed. We learn a single mask vector during pre-training of the same length as each BENDR vector, and use this as the transformer input to masked positions; masking is done by replacing a masked BENDR with a learned vector. The number of negatives/distractors was set to 20 and uniformly sampled from the *same* sequence as the masked vector, i.e., negatives do not cross trials or sequences.

To evaluate how generalizable the sequence model and vectors were to unseen data *after* pre-training, we evaluated the contrastive task, expressed as the transformer accuracy in constructing c_t to be most similar to b_t rather than the distractors/negatives, with respect to unseen data (in this case the downstream datasets). Note here that no further training or any evaluation with respect to downstream task labels was performed. This was done to evaluate the variability of the representations after pre-training. During this evaluation step, we masked half the amount expected during training, but did so such that masked spans were evenly spaced through the sequence (so that there were no overlapping sequences, and sufficient context was available). That is, for a sequence length of N_S, we masked $0.5 \times N_S \times p_{mask} = N_m$ contiguous sequences (of 10), and spaced them every $\left\lfloor \frac{N_S}{N_m} \right\rfloor$ steps (starting at the first sample). N_S first remained at 15,360 (60 s as in training, no overlap between subsequent sequence representations) for all datasets except P300, where sessions were too short and instead 5,120 (20s) was used. We then evaluated the change in performance

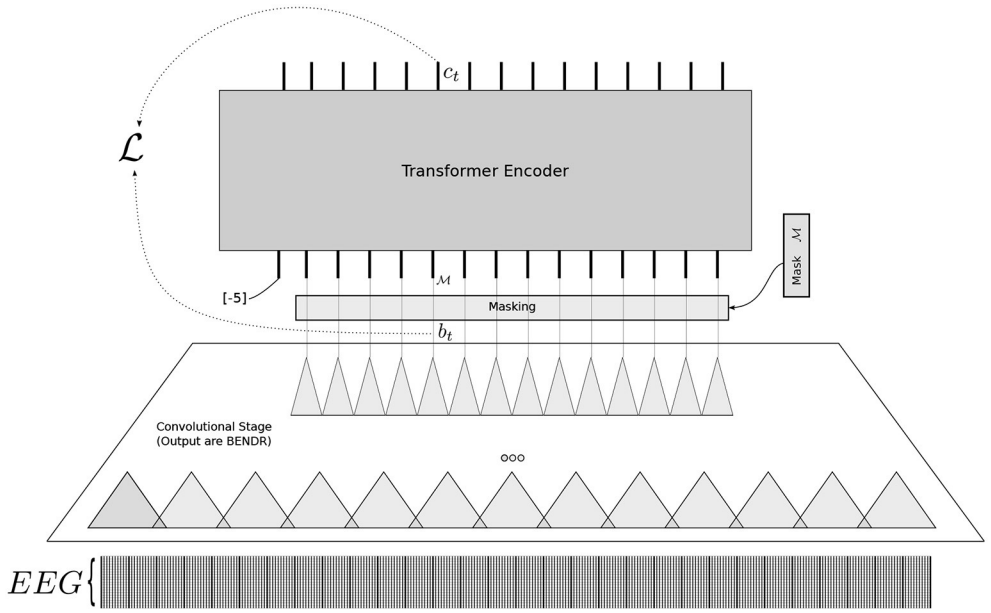

FIGURE 1 | The overall architecture used to construct BENDR. Loss $\mathcal{L}$ is calculated for a masked BErt-inspired Neural Data Representations (BENDR) b_t (after masking, it is replaced by the learned mask $\mathcal{M}$), itself produced from the original raw EEG (bottom) via a progression of convolution stages. The transformer encoder attempts to produce c_t to be more similar to b_t (despite that it is masked) than it is to a random sampling of over BENDR.

across the downstream datasets, excluding P300, as N_S varied from 20 to 60 s.

2.4.2. Downstream Fine-Tuning

Ultimately, our aims for subject-, session-, and dataset-generalizable representations were not simply to accurately select for the correct input (what was evaluated of the pre-training BENDR and sequence model), but with the intent that these representations (BENDR)—and potentially the sequence model itself—could be effectively transferred to specific and arbitrary tasks. We considered six different variations of TL across the battery of downstream EEG classification tasks (classification tasks listed in **Table 1**):

1. Add a new linear layer with softmax activation (classification layer) to the first (recall this position was pre-pended with an *input* value of −5 to the BENDR) output token of the transformer. Then, fine-tune the entire model (continue training the pre-trained model and start training the new layer) to classify the downstream targets using the output of this layer (ignoring the remaining sequence outputs) (shown in **Figure 2.1**).
2. Ignore the pre-trained transformer entirely, and use only the pre-trained convolutional stage (i.e., only use the BENDR). Create a consistent-length representation by dividing the BENDR into four contiguous sub-sequences, average each sub-sequence and concatenate them[12]. Add a new linear layer with softmax activation to classify the downstream targets with respect to this concatenated vector of averaged BENDRs (shown in **Figure 2.2**).
3. The same as **Figure 2.1**, but perform no pre-training; start with a randomly initialized DNN, as shown in **Figure 2.3**.
4. The same as **Figure 2.1**, but keep the BENDR (convolutional stage) fixed and continue training the transformer (and start training the new classification layer) to classify downstream targets, as shown in **Figure 2.4**.
5. The same as **Figure 2.2**, but perform no pre-training; start with randomly initialized convolution stage, as shown in **Figure 2.5**.
6. The same as **Figure 2.2**, but keep the first stage weights fixed and train only the new classification layer, as shown in **Figure 2.6**.

Figure 2 provides some illustration of each variation, where the respective indexed subfigures correspond to the list numbers above. These were considered so that we could speak to the effect each stage had on downstream performance, at least to some degree. We were interested in (1) determining whether the new sequence representation (BENDR) contained valuable features *as-is* (as they appear to be for speech Baevski et al., 2020) or if they required specific adaptation, and (2) whether the sequence model learned characteristics of the BENDR that were informative to the classification task. Finally, ignoring pre-training all-together, of course, was to examine how effective the network would be at learning the task otherwise, without the general pre-training task.

At this stage, we also included the sequence regularization proposed by `wav2vec 2.0` (Baevski et al., 2020), although we adjusted it for our more varied trial lengths. That is, in

[12]The selection of four here was arbitrary.

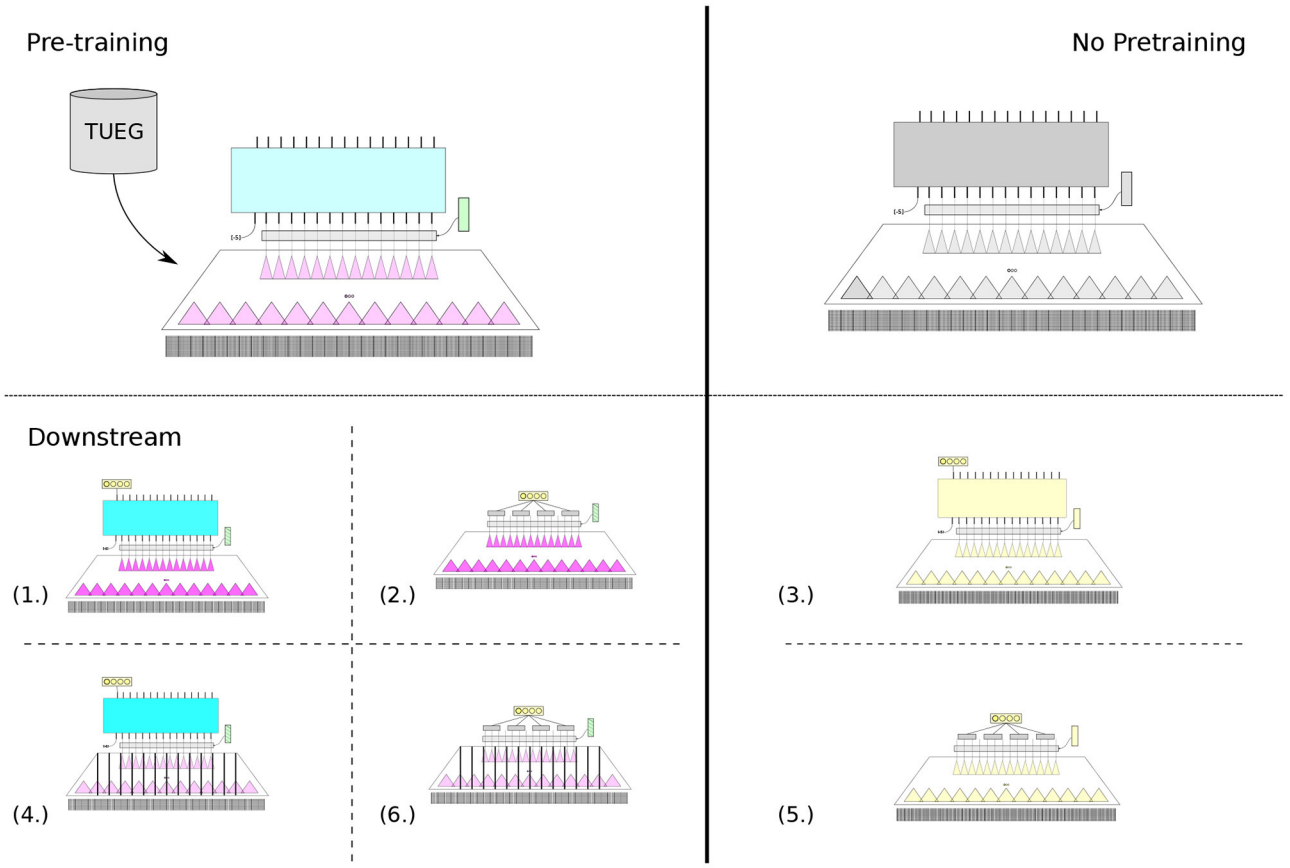

FIGURE 2 | Six different permutations of the model architecture were trained with conventional fully supervised training (in a leave-one/multi-subject-out fashion, see **Table 1**) for each downstream task. Indicated here is the portion of the overall architectures used (see **Figure 1**), and how pre-training model weights were leveraged for a four-way classification task (rectangle with four circles in it). Four tasks (left half) leveraged model weights that were first developed through pre-training. All *yellow* modules here indicate randomly initialized weights. Color that progresses in intensity (from pre-training to downstream) indicates further training, while added bars indicate weights that were kept unchanged during that training stage.

all 6 fine-tuning configurations, contiguous sections of 10% of the entire BENDR of a trial were masked with the mask token learned during pre-training (not changed after pre-training) at a probability of 0.01. In other words, this was the likelihood of a sample being the beginning of a contiguous masked section, as in pre-training. Additionally across the BENDR (throughout each vector in the sequence), a similar procedure dropped features to 0, where contiguous sections of 10% of the channels (51) were dropped with a probability of 0.005.

The P300, ERN, and SSC datasets all had imbalanced class distributions; during training, we adjusted for these imbalances by *undersampling* points uniformly of the more frequent classes with replacement so that the number of samples drawn—per epoch—of each class was equal to the number of examples of the least frequent target class. As the test conditions then were imbalanced, test performance was evaluated using metrics that accounted for this, and followed previous work (Baevski et al., 2020; Kostas and Rudzicz, 2020b). Metrics are specified by dataset in **Table 2**.

3. RESULTS

3.1. Pre-training Generalization

Figure 3 shows how accurate the transformer stage is at producing an appropriately similar BENDR. There are two key observations in this figure, the first is that there is little variability across the first four datasets, and within each of the five datasets. The latter point implies that this accuracy is not radically variable across different subjects (though, when fine-tuning for classification, this variability returns; see **Figure 5**). This could be because (a) the transformer adequately learns a general model of how BENDR sequences of novel persons and equipment progressed; (b) the BENDR themselves are invariant to different people, hardware, and tasks; (c) some combination of the last two possibilities; or (d) the problem is being solved via some non-signal characteristics. We return to this question shortly. The second observation was already alluded: the P300 dataset distinctly under-performs the other downstream datasets. However, this coincided with the shortest evaluation sequence. Looking at **Figure 4**, we see that all five datasets have consistently

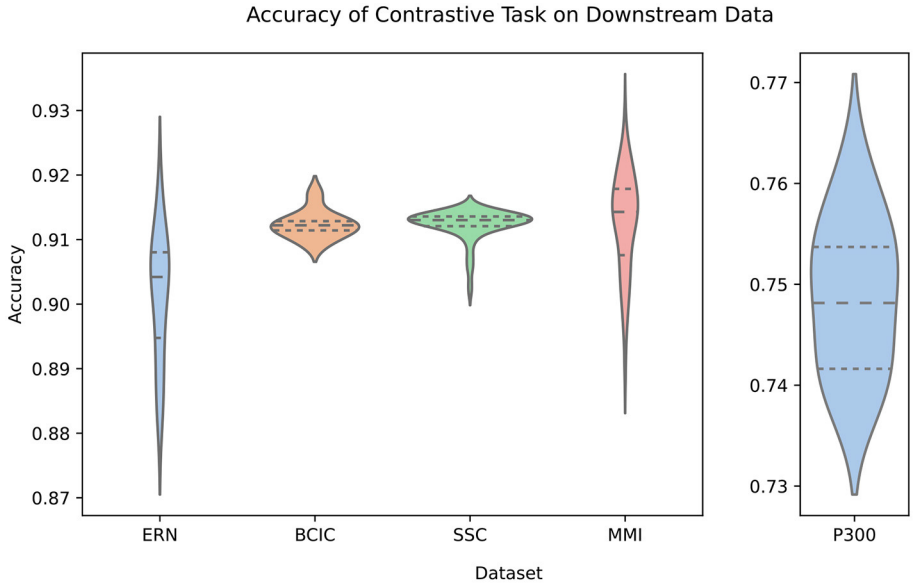

FIGURE 3 | Violin plot (inner lines for quartile divisions) of test subject-wise accuracy for each downstream dataset. Specifically, accuracy of the sequence model (transformer stage) at creating a representation that is closest to the correct representation at masked sequence positions. The P300 dataset is distinctly lower performing (note the adjusted *Y*-axis) than the remaining datasets, though this was likely due to its shorter evaluation context (see **Figure 4**). Nonetheless, there is minimal test subject-wise variation, particularly when compared to classifier performance generally.

similar performance when evaluated with 20 s of data, so the dip in P300 performance of **Figure 3** seems less remarkable. Taken together, **Figures 3**, **4** clearly indicate that a longer evaluation context makes the contrastive task easier. This suggests that the contrastive task is, in fact, solved by learning signal-relevant features, rather than some more crude solution like interpolation, or by simply creating a sequence of recognizable position representations (both of which have no reason to exhibit this dependence on sequence length). We believe that the most likely explanation for the *rise* in performance with more context is that local representations are more difficult distractors, implying that the new effective sampling rate remains too high (and there is still redundant information encoded in local BENDR). Notwithstanding, there is a strong uniformity of performance across datasets and subjects (in both **Figures 3**, **4**), meaning this scheme develops features (whether through the transformer itself, or the BENDR) that generalize to novel subjects, hardware, and tasks, though their applicability to downstream contexts remains to be seen.

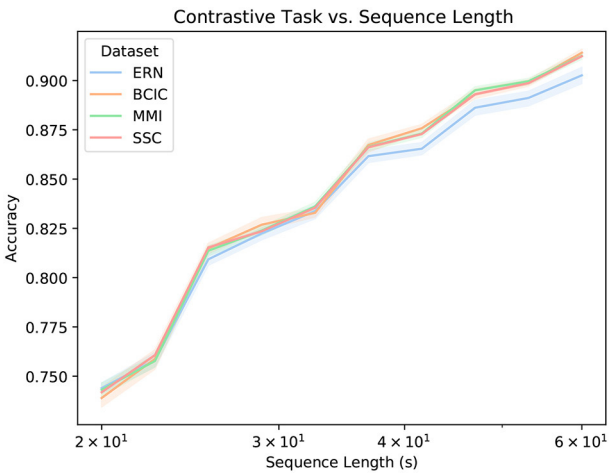

FIGURE 4 | Contrastive accuracy vs. evaluation length in seconds (*x*-axis logarithmic). Performance is distinctly similar for all datasets, rising for longer sequences. We suggest that this implies that samples that are further apart are easier to distinguish between than neighboring samples. Thus, while BErt-inspired Neural Data Representations (BENDR) encode local signal characteristics well, there is redundancy.

3.2. Downstream Fine-Tuning

Figure 5 and **Table 2** present a picture of how effectively BENDR could be adapted to specific tasks. Overall, the fine-tuned linear classification (the downstream configuration in **Figure 2.2**) that bypassed the transformer entirely after pre-training was highest performing four out of five times, although using the transformer for classification (**Figure 2.1**) performed consistently similarly (confidence intervals always overlapped), and surpassed the bypassed transformer (**Figure 2.2**) with the P300 dataset (and was highest performing for this dataset). Deploying the full network (initial stage and transformer) *without pre-training* was generally ineffective, though this was not the case with the SSC dataset, which may have been due to the larger amount of data available for fully supervised learning. In fact, for both the full and linear model architectures trained with the SSC data, fine-tuning the pre-trained model is mostly on par with the fully supervised counterpart. Considering our results with the SSC data relative to those of Banville et al. (2019) proposed

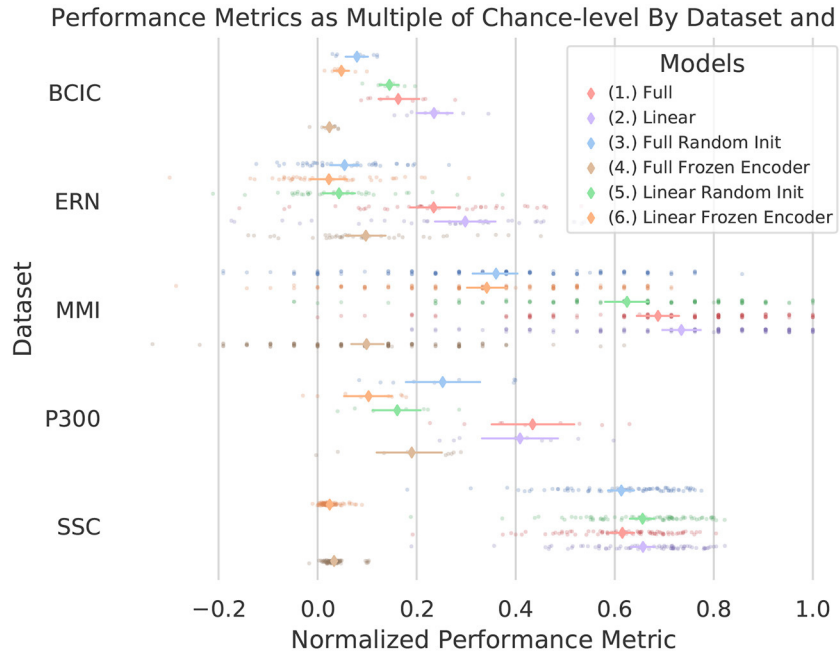

FIGURE 5 | Performance of all downstream datasets for each of the six model configurations considered. Metrics vary by dataset, see **Table 2**. Metrics were normalized to range from chance (0) to perfect (1). Individual translucent points are performances of single subjects (within each test fold), solid diamonds indicate mean performance across all subjects/folds, with surrounding bars showing 0.95 confidence intervals using $n = 1000$ bootstrap sampling. The discretized pattern of the MMI dataset is due to the limited trials *per subject*, which resulted in limited distribution of performance levels. Notably here, (1) or (2) was consistently among the best performing, yet both remained within the confidence levels of each other and aside from a few cases with the ERN dataset did not result in subject performances that were worse than chance. (3) and especially (4) and (6) often stayed marginally above chance, indicating that the pre-trained features were not sufficient without further training. The randomly initialized average-pooled BENDR with linear classifier (5) also performed well, though less consistently, suggesting pre-training was needed for *consistent* performance. Model configurations are numbered in accordance with the list presented in section 2.4.2.

contrastive learning for sleep staging (described in section 1.1), their reported results show that the fine-tuned variants of our own model (1 and 2) achieved a higher mean balanced accuracy relative to their two proposed schemes. Taken in concert with our own approach's wider applicability and more fine-grained temporal feature development, we believe this demonstrates that ours is a promising alternative. Interestingly, with and without pre-training (**Figure 2.2, 2.5**) achieved similar performance to Banville et al.'s fully supervised results (where our configurations and their architecture employ similar 1D convolution-based schemes), which is notable as with this dataset, both their "temporal-shuffling" and "relative-positioning" tasks underperformed this full supervision performance level (though we cannot speak to statistical significance of this comparison).

Our fine-tuned approaches similarly appear reasonably competitive with prior work on the MMI dataset (Dose et al., 2018; Kostas and Rudzicz, 2020b), particularly when considering that only 19 channels (rather than the full set of 64) were being used. Outside of the MMI and SSC dataset, remaining results are not competitive with more targeted solutions (Kostas and Rudzicz, 2020b). Whenever pre-training was not used, despite heavy regularization (and the very low learning rates) the randomly initialized parameters were consistently prone to overfitting, all the more so with the full model architecture. Conversely, the pre-trained networks were slow to fit to the downstream training data (under the exact same training scheme for fine-tuning). Despite that these results were not necessarily state of the art, this single pre-training scheme nonetheless shows a breadth of transferability that is apparently unique, and aside from the SSC dataset, consistently here outperforms the fully supervised counterparts.

4. DISCUSSION

We are unaware of any prior work assessing transformer-based (Vaswani et al., 2017) DNNs with EEG data (raw or otherwise). This is perhaps consistent with the ineffectiveness we observed with the *randomly initialized* full architecture (**Figure 2.3**) and could imply that effective use of this powerful emerging architecture *requires* pre-training (or at least enough data, given the better looking SSC performance). This may be due to the large number of parameters that these models require, making training difficult without sufficient hardware resources. The total number of parameters trained in configuration (1) is over one billion parameters. Future work should continue to evaluate this architecture, particularly as it appears to be more widely applicable than the NLP applications it was originally proposed for (Baevski et al., 2020; Dosovitskiy et al., 2020).

We believe that our approach can be improved through adjusting the neural network architecture and pre-training configuration such that it becomes more data-domain (EEG) appropriate. Future work will prioritize effective integration of

spatial information, likely by better isolating temporal and spatial operations. Evaluation using large downstream datasets that *also* feature many channels, such as the *Montreal Archive of Sleep Studies* (MASS)[13] will be considered. Though available for public access at the time of writing, these data were unavailable while experiments were prepared and conducted. Prior work shows that DNN approaches effective for EEG leverage spatial information (Chambon et al., 2018), and it is presently unclear to what degree this is the case with BENDR. In terms of data-appropriate temporal modeling, which we have considered with relatively more zeal in this work, recall that **Figure 4** presents the possibility that local representations may be retaining redundant information, further improvements therefore may be found in better compressing the temporal resolution of BENDR. Future work will consider larger downsampling factors in the initial stage, along with longer sequences, balancing the more difficult problem of summarizing more data (in effect, further data *compression*), with the apparent increased effectiveness of the contrastive task (as observed in **Figure 4**) on longer sequences. A small but potentially fruitful avenue for further improvement includes reconsidering the additive convolutional layer as a substitute for explicit position encodings, which are in fact more common (Vaswani et al., 2017; Devlin et al., 2019; Raffel et al., 2020). Recall that this was originally for two reasons: `wav2vec 2.0` did the same, and we felt it best to limit excessive changes to the architecture on a first iteration, and also because it seamlessly supported flexible input lengths. This latter point comes however, with a trade-off: our particular position encoder had a receptive field of 25 (stride of 1), which means a little over 9 s of input. While it seems that convolutional position encodings offer better performance (Mohamed et al., 2019), this input width exceeded the *entire* length of all but the sleep classification task (the length we chose was optimized for pre-training behavior).

After considering these possible avenues for improving BENDR, we still do not fully discount the validity of some of the transfer learning paths we appear to exclude above in our introduction. We will reconsider these paths in future work. Particularly, given the success we had in crossing boundaries of hardware in this work, and in prior work (Kostas and Rudzicz, 2020a), it may be possible to construct an *aggregate* dataset featuring a variety of EEG classification tasks toward better ImageNet-like pre-training. The construction of a more coherent label set that crosses several BCI paradigms would no doubt be a significant effort (e.g., problems may include: is a rest period before one task paradigm the same as rest before another? What about wakeful periods in sleep?). This would no doubt be imbalanced; the labels would be distributed in a long-tailed or Zipfian distribution that would likely require well thought-out adjustment (Cao et al., 2019; Tang et al., 2020).[13] Furthermore, the value of ImageNet pre-training *seems to be* localized to very early layers and the internalization of domain-relevant data statistics (Raghu et al., 2019; Neyshabur et al., 2020). Future work could look into which of these may be leveraged with a new aggregate (multiple subjects *and* tasks) pre-training, or the common subject-specific fine-tuning. This may provide insight into better weight initialization, or integration of explicit early layers similar to Raghu et al. (2019) (one could also argue that SincNet layers Ravanelli and Bengio, 2018 are some such layers that could factor here). Additionally, as temporally minded reconstruction losses continue to develop (Rivest and Kohar, 2020), reconsidering the effectiveness of signal reconstruction as a pre-training objective (and/or regularization) is warranted, whether this is within an MLM-like scheme similar to BENDR, or a seq2seq model (Graves, 2012).

[13] http://massdb.herokuapp.com/en/

5. CONCLUSION

We have proposed MLM-like training as a self-supervised pre-training step for BCI/EEG DNNs. This is in the interest of diversifying the investigations into successful transfer learning schemes for DNNs applied to BCI and EEG with possible applicability to neuroimaging more generally. While previous approaches fashioned DNN transfer learning after ImageNet pre-training, we find this approach inadequate as there is limited applicable data availability and it is questionably analogous to its forebear. While our proposed alternative might similarly suffer from this latter point to some degree (the most distinct MLM success is with discrete sequences, not continuous ones), it is more conducive to leveraging potentially immense amounts of unlabeled data, it is not limited to long-term feature developments as with previous proposals, and it seems to produce representations equally suited to different users and sessions, which is a problem previous work appears less suited to solving. In summary, we see strong paths for the effective deployment of powerful computation and massive data scales with EEG and BCI. Effective solutions in these specific applications could help drive application *and* analysis solutions in neuroimaging and perhaps physiological signal analysis generally.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

DK conceived of the presented idea, designed experiments, performed the analysis, drafted the manuscript, and designed the figures. DK developed implementation with assistance from SA-O. SA-O and FR edited manuscript. FR provided supervision throughout. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the Electronics and Telecommunications Research Institute (South Korea) [20ZS1100, Core Technology Research for Self-Improving Integrated Artificial Intelligence System] and NSERC Discovery [435874]. Rudzicz holds a CIFAR Chair in AI.

REFERENCES

Ahn, M., and Jun, S. C. (2015). Performance variation in motor imagery brain-computer interface: a brief review. *J. Neurosci. Methods* 243, 103–110. doi: 10.1016/j.jneumeth.2015.01.033

Aroca-Ouellette, S., and Rudzicz, F. (2020). "On Losses for Modern Language Models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics)*, 4970–4981. Available online at: https://www.aclweb.org/anthology/2020.emnlp-main.403

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. (2019). "A theoretical analysis of contrastive unsupervised representation learning," in *36th International Conference on Machine Learning, ICML 2019* (Long Beach, CA), Vol. 2019-June, 9904–9923.

Baevski, A., and Mohamed, A. (2020). "Effectiveness of self-supervised pre-training for ASR," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona: IEEE), 7694–7698.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. T. Lin. Available online at: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

Banville, H., Albuquerque, I., Hyvarinen, A., Moffat, G., Engemann, D.-A., and Gramfort, A. (2019). "Self-supervised representation learning from electroencephalography signals," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)* (Pittsburgh, PA: IEEE), 1–6.

Banville, H., Chehab, O., Hyvärinen, A., Engemann, D.-A., and Gramfort, A. (2020). Uncovering the structure of clinical EEG signals with self-supervised learning. *J. Neural Eng.* 18:046020. doi: 10.1088/1741-2552/abca18

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. T. Lin. Available online at: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. *Adv. Neural Inf. Proc. Syst.* 32, 1–18. Available online at: https://proceedings.neurips.cc/paper/2019/hash/621461af90cadfdaf0e8d4cc25129f91-Abstract.html

Chambon, S., Galtier, M. N., Arnal, P. J., Wainrib, G., and Gramfort, A. (2018). A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 758–769. doi: 10.1109/TNSRE.2018.2813138

Chen, K., Wang, J., Chen, L.-C., Gao, H., Xu, W., and Nevatia, R. (2016). ABC-CNN: an attention based convolutional neural network for visual question answering. *arXiv*.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. (2020). Big self-supervised models are strong semi-supervised learners. *arXiv* 1–18.

Chung, Y.-A., Tang, H., and Glass, J. (2020). "Vector-quantized autoregressive predictive coding," in *Interspeech 2020*, Vol. arXiv (Shanghai: ISCA), 3760–3764.

Cimtay, Y., and Ekmekcioglu, E. (2020). Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset eeg emotion recognition. *Sensors* 20, 1–20. doi: 10.3390/s20072034

Citi, L., Poli, R., and Cinel, C. (2010). Documenting, modelling and exploiting P300 amplitude changes due to variable target delays in Donchins speller. *J. Neural Eng.* 7:056006. doi: 10.1088/1741-2560/7/5/056006

Citi, L., Poli, R., and Cinel, C. (2014). Erp-based brain-computer interface recordings. doi: 10.13026/C2101S

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *CVPR09* (Miami Beach, FL).

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, eds J. Burstein, C. Doran, and T. Solorio (Minneapolis, MN: Association for Computational Linguistics), 4171–4186. doi: 10.18653/v1/n19-1423

Ditthapron, A., Banluesombatkul, N., Ketrat, S., Chuangsuwanich, E., and Wilaiprasitporn, T. (2019). Universal joint feature extraction for P300 EEG classification using multi-task autoencoder. *IEEE Access* 7, 68415–68428. doi: 10.1109/ACCESS.2019.2919143

Dose, H., Møller, J. S., Iversen, H. K., and Puthusserypady, S. (2018). An end-to-end deep learning approach to MI-EEG signal classification for BCIs. *Exp. Syst. Appl.* 114, 532–542. doi: 10.1016/j.eswa.2018.08.031

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv arXiv:2010.11929*.

Fahimi, F., Zhang, Z., Goh, W. B., Lee, T.-S., Ang, K. K., and Guan, C. (2019). Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI. *J. Neural Eng.* 16:026007. doi: 10.1088/1741-2552/aaf3f6

Fan, A., Grave, E., and Joulin, A. (2019). Reducing transformer depth on demand with structured dropout. *arXiv* 103, 1–15. Available online at: https://openreview.net/forum?id=SylO2yStDr

Gemein, L. A., Schirrmeister, R. T., Chrabąszcz, P., Wilson, D., Boedecker, J., Schulze-Bonhage, A., et al. (2020). Machine-learning-based diagnostics of EEG pathology. *Neuroimage* 220:17021. doi: 10.1016/j.neuroimage.2020.117021

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). PhysioBank, physioToolkit, and physioNet: components of a new research resource for complex physiologic signals. *Circulation* 101:E215–E220. doi: 10.1161/01.cir.101.23.e215

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*. Berlin; New York: Springer, c2012.

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., et al. (2020). "Bootstrap your own latent - a new approach to self-supervised learning," *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. T. Lin. Available online at: https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html

He, K., Girshick, R., and Dollar, P. (2019). "Rethinking imageNet pre-training," in *Proceedings of the IEEE International Conference on Computer Vision* (Seoul), 4917–4926.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016* (Las Vegas, NV: IEEE Computer Society), 770–778. doi: 10.1109/CVPR.2016.90

Hénaff, O. J. (2020). "Data-efficient image recognition with contrastive predictive coding," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, Vol. 119 (PMLR), 4182–4192. Available online at: http://proceedings.mlr.press/v119/henaff20a.html

Hendrycks, D., and Gimpel, K. (2016). Bridging nonlinearities and stochastic regularizers with Gaussian error linear units. *arXiv arXiv:1606.08415*.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* (Honolulu, HI: IEEE Computer Society), 2261–2269. doi: 10.1109/CVPR.2017.243

Huang, X. S., Perez, F., Ba, J., and Volkovs, M. (2020). "Improving transformer optimization through better initialization," in *Proceedings of Machine Learning and Systems 2020*, 9868–9876.

Huh, M., Agrawal, P., and Efros, A. A. (2016). What makes imageNet good for transfer learning? *CoRR* 1–10.

Jiang, D., Li, W., Zhang, R., Cao, M., Luo, N., Han, Y., et al. (2020). A further study of unsupervised pre-training for transformer based speech recognition. *arXiv arXiv:2005.09862*.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). SpanBERT: improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguist.* 8, 64–77. doi: 10.1162/tacl_a_00300

Jurcak, V., Tsuzuki, D., and Dan, I. (2007). 10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems. *Neuroimage* 34, 1600–1611. doi: 10.1016/j.neuroimage.2006.09.024

Kemp, B., Zwinderman, A., Tuk, B., Kamphuisen, H., and Oberyé, J. (2018). The sleep-edf database [expanded]. doi: 10.13026/C2X676

Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A., and Oberyé, J. J. (2000). Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Trans. Biomed. Eng.* 47, 1185–1194. doi: 10.1109/10.867928

Kingma, D. P., and Ba, J. L. (2015). "Adam: a method for stochastic optimization," *3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings* (San Diego, CA), 1–15.

Kornblith, S., Shlens, J., and Le, Q. V. (2019). "Do better imagenet models transfer better?" in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 2656–2666.

Kostas, D., Pang, E. W., and Rudzicz, F. (2019). Machine learning for MEG during speech tasks. *Sci. Rep.* 9:1609. doi: 10.1038/s41598-019-38612-9

Kostas, D., and Rudzicz, F. (2020a). Dn3: an open-source python library for large-scale raw neurophysiology data assimilation for more flexible and standardized deep learning. *bioRxiv*. doi: 10.1101/2020.12.17.423197

Kostas, D., and Rudzicz, F. (2020b). Thinker invariance: enabling deep neural networks for BCI across more people. *J. Neural Eng.* 17:56008. doi: 10.1088/1741-2552/abb7a7

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural Networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems, Vol. 1*, NIPS'12 (Lake Tahoe: Curran Associates Inc.), 1097–1105.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *J. Neural Eng.* 15:aace8c. doi: 10.1088/1741-2552/aace8c

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lin, Y.-P., and Jung, T.-P. (2017). Improving EEG-based emotion classification using conditional transfer learning. *Front. Hum. Neurosci.* 11:334. doi: 10.3389/fnhum.2017.00334

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., et al. (2018). A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update. *J. Neural Eng.* 15:031005. doi: 10.1088/1741-2552/aab2f2

Margaux, P., Emmanuel, M., Sébastien, D., Olivier, B., and Jérémie, M. (2012). Objective and subjective evaluation of online error correction during P300-Based spelling. *Adv. Hum. Comput. Interact.* 2012, 1–13. doi: 10.1155/2012/578295

Mohamed, A., Okhonko, D., and Zettlemoyer, L. (2019). Transformers with convolutional context for ASR. *arXiv*.

Mousavi, S., Afghah, F., and Acharya, U. R. (2019). SleepEEGNet: automated sleep stage scoring with sequence to sequence deep learning approach. *PLoS ONE* 14:e0216456. doi: 10.1371/journal.pone.0216456

Neyshabur, B., Sedghi, H., and Zhang, C. (2020). "What is being transferred in transfer learning?," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. T. Lin. Available online at: https://proceedings.neurips.cc/paper/2020/hash/0607f4c705595b911a4f3e7a127b44e0-Abstract.html

Ngiam, J., Peng, D., Vasudevan, V., Kornblith, S., Le, Q. V., and Pang, R. (2018). Domain adaptive transfer learning with specialist models. *arXiv*.

Obeid, I., and Picone, J. (2016). The temple university hospital EEG data corpus. *Front. Neurosci.* 10:196. doi: 10.3389/fnins.2016.00196

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140:1–140:67.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. *arXiv*.

Ravanelli, M., and Bengio, Y. (2018). Interpretable convolutional filters with sincNet. *Arxiv*.

Rivest, F., and Kohar, R. (2020). A new timing error cost function for binary time series prediction. *IEEE Trans. Neural Netw. Learn. Syst.* 31, 174–185. doi: 10.1109/TNNLS.2019.2900046

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng.* 16:051001. doi: 10.1088/1741-2552/ab260c

Sannelli, C., Vidaurre, C., Müller, K.-R., and Blankertz, B. (2019). A large scale screening study with a SMR-based BCI: categorization of BCI users and differences in their SMR activity. *PLoS ONE* 14:e0207351. doi: 10.1371/journal.pone.0207351

Schalk, G., Mcfarland, D. J., Hinterberger, T., Birbaumer, N., Wolpaw, J. R., and Technology, A. B.-C. I. B. C. I. (2004). BCI2000 : a general-purpose brain-computer interface (BCI) system. *IEEE Trans. Biomed. Eng.* 51, 1034–1043. doi: 10.1109/TBME.2004.827072

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* 38, 5391–5420. doi: 10.1002/hbm.23730

Schwemmer, M. A., Skomrock, N. D., Sederberg, P. B., Ting, J. E., Sharma, G., Bockbrader, M. A., et al. (2018). Meeting brain-computer interface user performance expectations using a deep neural network decoding framework. *Nat. Med.* 24, 1669–1676. doi: 10.1038/s41591-018-0171-y

Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proc. Natl. Acad. Sci. U.S.A.* 117, 30033–30038. doi: 10.1073/pnas.1907373117

Tang, K., Huang, J., and Zhang, H. (2020). Long-tailed classification by keeping the good and removing the bad momentum causal effect. *NeurIPS* 1–12.

Tangermann, M., Müller, K. R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., et al. (2012). Review of the BCI competition IV. *Front. Neurosci.* 6:55. doi: 10.3389/fnins.2012.00055

van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv arXiv:1807.03748*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, eds I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett (Long Beach, CA), 5998–6008. Available online at: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vidaurre, C., and Blankertz, B. (2010). Towards a cure for BCI illiteracy. *Brain Topography* 23, 194–198. doi: 10.1007/s10548-009-0121-6

Wu, Y., and He, K. (2020). Group normalization. *Int. J. Comput. Vis.* 128, 742–755. doi: 10.1007/s11263-019-01198-w

Xu, G., Shen, X., Chen, S., Zong, Y., Zhang, C., Yue, H., et al. (2019). A deep transfer convolutional neural network framework for EEG signal classification. *IEEE Access* 7, 112767–112776. doi: 10.1109/ACCESS.2019.2930958

Yosinski, J., Clune, J., Nguyen, A. M., Fuchs, T. J., and Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv arXiv:1506.06579*.

Zanini, P., Congedo, M., Jutten, C., Said, S., and Berthoumieu, Y. (2018). Transfer learning: a riemannian geometry framework with applications to brain-computer interfaces. *IEEE Trans. Biomed. Eng.* 65, 1107–1116. doi: 10.1109/TBME.2017.2742541

Zhang, D., Chen, K., Jian, D., and Yao, L. (2020a). Motor imagery classification via temporal attention cues of graph embedded EEG signals. *IEEE J. Biomed. Health Informat.* 24, 2570–2579. doi: 10.1109/JBHI.2020.2967128

Zhang, X., Yao, L., Wang, X., Monaghan, J. J. M., Mcalpine, D., and Zhang, Y. (2020b). A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers. *J. Neural Eng.* 18:031002. doi: 10.1088/1741-2552/abc902

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

Downstream hyperparameters

TABLE A1 | Hyperparameters that varied between datasets, and these were not changed between different model configurations (see list in section 2.4.2).

Dataset	Batch Size	Epochs	Learning Rate
MMI	4	7	1×10^{-5}
BCIC	60	15	5×10^{-5}
ERN	32	15	1×10^{-5}
P300	80	20	1×10^{-5}
SSC	64	40	5×10^{-5}

frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 08 July 2021
doi: 10.3389/fnhum.2021.655840

A Lightweight Multi-Scale Convolutional Neural Network for P300 Decoding: Analysis of Training Strategies and Uncovering of Network Decision

Davide Borra[1*†], *Silvia Fantozzi*[1,2†] *and Elisa Magosso*[1,2,3†]

[1] Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi" (DEI), University of Bologna, Cesena, Italy, [2] Interdepartmental Center for Industrial Research on Health Sciences & Technologies, University of Bologna, Bologna, Italy, [3] Alma Mater Research Institute for Human-Centered Artificial Intelligence, University of Bologna, Bologna, Italy

OPEN ACCESS

Edited by:
Minkyu Ahn,
Handong Global University,
South Korea

Reviewed by:
Heung-Il Suk,
Korea University, South Korea
Seungchan Lee,
Korea Electronics Technology Institute, South Korea

***Correspondence:**
Davide Borra
davide.borra2@unibo.it

†ORCID:
Davide Borra
orcid.org/0000-0003-3791-8555
Silvia Fantozzi
orcid.org/0000-0002-0660-7204
Elisa Magosso
orcid.org/0000-0002-4673-2974

Specialty section:
This article was submitted to Brain-Computer Interfaces, a section of the journal Frontiers in Human Neuroscience

Received: *19 January 2021*
Accepted: *17 May 2021*
Published: *08 July 2021*

Citation:
Borra D, Fantozzi S and Magosso E (2021) A Lightweight Multi-Scale Convolutional Neural Network for P300 Decoding: Analysis of Training Strategies and Uncovering of Network Decision. Front. Hum. Neurosci. 15:655840. doi: 10.3389/fnhum.2021.655840

Convolutional neural networks (CNNs), which automatically learn features from raw data to approximate functions, are being increasingly applied to the end-to-end analysis of electroencephalographic (EEG) signals, especially for decoding brain states in brain-computer interfaces (BCIs). Nevertheless, CNNs introduce a large number of trainable parameters, may require long training times, and lack in interpretability of learned features. The aim of this study is to propose a CNN design for P300 decoding with emphasis on its lightweight design while guaranteeing high performance, on the effects of different training strategies, and on the use of *post-hoc* techniques to explain network decisions. The proposed design, named MS-EEGNet, learned temporal features in two different timescales (i.e., multi-scale, MS) in an efficient and optimized (in terms of trainable parameters) way, and was validated on three P300 datasets. The CNN was trained using different strategies (within-participant and within-session, within-participant and cross-session, leave-one-subject-out, transfer learning) and was compared with several state-of-the-art (SOA) algorithms. Furthermore, variants of the baseline MS-EEGNet were analyzed to evaluate the impact of different hyper-parameters on performance. Lastly, saliency maps were used to derive representations of the relevant spatio-temporal features that drove CNN decisions. MS-EEGNet was the lightest CNN compared with the tested SOA CNNs, despite its multiple timescales, and significantly outperformed the SOA algorithms. *Post-hoc* hyper-parameter analysis confirmed the benefits of the innovative aspects of MS-EEGNet. Furthermore, MS-EEGNet did benefit from transfer learning, especially using a low number of training examples, suggesting that the proposed approach could be used in BCIs to accurately decode the P300 event while reducing calibration times. Representations derived from the saliency maps matched the P300 spatio-temporal distribution, further validating the proposed decoding approach. This study, by specifically addressing the aspects of lightweight design, transfer learning, and interpretability, can contribute to advance the development of deep learning algorithms for P300-based BCIs.

Keywords: electroencephalography, P300, convolutional neural networks, transfer learning, decision explanation, brain-computer interfaces

INTRODUCTION

The P300 response is an attention-dependent event-related potential (ERP) first reported in electroencephalographic (EEG) signals by Sutton et al. (1965). This wave is characterized by a positive deflection that peaks within the time window between 250 and 500 ms after stimulus onset, and it is mostly distributed on the scalp around the midline EEG electrodes (Fz, Cz, Pz), increasing its magnitude from the frontal to the parietal sites (Polich, 2007). The P300 can be evoked in an oddball paradigm (Farwell and Donchin, 1988), where an infrequent deviant stimulus immersed in a sequence of frequent standard stimuli is presented to the user while he/she is attending to it (e.g., by counting how many times a rare event occurs). Rare events induce the P300 response; this response can be used as a neural signal in EEG-based brain-computer interfaces (BCIs), enabling direct communication between the brain and surroundings without the involvement of peripheral nerves or muscles (Nicolas-Alonso and Gomez-Gil, 2012). One of the first P300-based BCIs was developed by Farwell and Donchin (1988) using a visual stimulation in the oddball paradigm. These systems could be especially beneficial for patients suffering from motor neuron disease (Rezeika et al., 2018) to provide alternative ways of communication. Furthermore, they may represent viable training tools for patients with attention deficits as recently reported in Amaral et al. (2018) where a P300-based BCI paradigm was tested in patients suffering from autism spectrum disorder (ASD) to improve their social attention.

Of course, a crucial aspect of a P300-based BCI is the decoding algorithm that translates brain signals into classes (e.g., P300 and non-P300 classes). Machine learning (ML) techniques have been recognized to be powerful tools in learning discriminative patterns from brain signals. In recent years, deep learning, a branch of ML originally proposed in computer vision (Guo et al., 2016; Ismail Fawaz et al., 2019), has been applied to decoding problems of physiological signals, such as electroencephalography, electromyography, electrocardiography, and electrooculography (Faust et al., 2018). At variance with more traditional ML approaches characterized by a separation between feature extraction, selection and classification stages (LeCun et al., 2015), deep learning techniques automatically learn features from raw or light pre-processed inputs to maximize between-class discriminability and finalize the decoding task in an end-to-end fashion.

Among deep learning techniques for classification, convolutional neural networks (CNNs) are widely used. These are specialized feed-forward neural networks involving the convolution operator to process data with a grid-like topology and are inspired by the hierarchical structure of the ventral stream of the visual system. Stacking neurons with a local receptive field on top of others creates receptive fields of individual neurons that increase in size in deeper layers of the CNN and increases the complexity of the features to which the neurons respond (Lindsay, 2020), realizing different levels of feature abstraction. This way, CNNs automatically learn hierarchically structured features from the input data, finalized to the classification. However, CNNs have some weaknesses: they introduce a large number of trainable parameters (consequently requiring a large number of training examples), they introduce many hyper-parameters (i.e., parameters that define the functional form of decoder), and learned features are difficult to be interpreted.

The field of EEG classification (and in particular P300 classification) has been widely exploiting the advantages of CNNs (Faust et al., 2018; Craik et al., 2019). At the same time, solutions to mitigate the weaknesses of these algorithms have been proposed within this field, as reported in the state-of-the-art (SOA) description below.

In CNN-based EEG classification, EEG signals can be arranged into a 2D representation with electrodes along a dimension and time steps along the other, and fed as input to the CNN that predicts the corresponding label. CNN designs for EEG classification include both shallow and deep neural networks, and solutions have been proposed either by performing spatial and temporal convolutions together (i.e., mixed spatio-temporal feature learning) or separately (i.e., unmixed spatio-temporal feature learning). Among the latter, several have been successfully applied to P300 classification (Cecotti and Graser, 2011; Manor and Geva, 2015; Lawhern et al., 2018; Liu et al., 2018; Shan et al., 2018; Farahat et al., 2019) and generally have been proved to outperform traditional ML approaches. Cecotti and Graser (2011) designed a CNN comprising two convolutional and two fully-connected layers to decode the P300 event. Remarkably, this was also the first attempt of CNN-based P300 decoding. Extensions of this architecture mainly focused on the increase of depth, and inclusion of batch normalization and dropout (Manor and Geva, 2015; Liu et al., 2018). Moreover, Farahat et al. (2019) proposed a dual-branched CNN (BranchedNet) that learns temporal features in two different timescales with parallel temporal convolutions, reporting an increase in performance with respect to a single-scale convolution. While these CNNs performed better than traditional ML techniques in P300 decoding, two aspects deserve attention: (i) they learn spatial features (i.e., spatial convolution, performed across electrodes) and then temporal features in the next layers (i.e., temporal convolution, performed across time samples); (ii) they do not address the challenge of reducing the number of trainable parameters. Regarding the first aspect, Shan et al. (2018) pointed out that these architectures may lose useful raw temporal information related to the P300 event since temporal features are learned from spatially filtered signals instead of from raw inputs. The authors proved that an architecture with the first layer performing a mixed spatio-temporal convolution (OCLNN) improved the decoding performance compared with the architecture proposed by Cecotti and Graser (2011) and other variants (Manor and Geva, 2015; Liu et al., 2018). Regarding the second aspect, recently, Lawhern et al. (2018) have designed a shallow CNN for EEG decoding, which is also applied to P300 detection (EEGNet). This design, besides performing temporal convolution in the first layer, uses separable and depthwise convolutions, i.e., convolutions specifically devoted to reducing the number of trainable parameters (Chollet, 2016).

Remarkably, recently, we have proposed a CNN (Borra et al., 2020a) based on the design of EEGNet that won the P300 decoding challenge issued by the International Federation of Medical and Biological Engineering (IFMBE) in 2019, where the dataset (BCIAUT-P300) was a large multi-participant and multi-session collection of data. The solution of the authors outperformed significantly a CNN derived from Manor and Geva (2015) with a spatial convolutional layer as the first layer, long short-term memories, and traditional ML approaches (Simões et al., 2020). These results further substantiate that CNNs, which include a temporal convolutional layer as the first layer, can represent advantageous solutions for P300 decoding compared with traditional approaches and other CNN designs.

Techniques have been proposed for interpreting and understanding what the CNN has learned (Montavon et al., 2018); in the field of EEG classification, they are fundamental to validate correct learning, checking that the learning system does not rely on artifactual sources but on neurophysiological features. These techniques explain the decoding decision taken by the CNN, i.e., features on which the CNN mainly relies to discriminate among classes. In this way, they represent tools to explore and analyze the underlying neurophysiology potentially characterizing new features (unknown so far) and gaining insights into neural correlates of the underlying phenomena. Montavon et al. (2018) provided a definition for *explanation of CNN decision*: "the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g., classification or regression)." Among the explanation techniques proposed in the computer vision domain (Montavon et al., 2018), saliency maps (Simonyan et al., 2013), simple representations reporting the gradient of a target class score with respect to each input pixel, have been recently transposed to P300 decoding (Farahat et al., 2019). Furthermore, other techniques were adopted to understand CNNs for P300 decoding, such as temporal and spatial kernel visualizations (Cecotti and Graser, 2011; Lawhern et al., 2018), and kernel ablation tests (Lawhern et al., 2018). In addition to these techniques, interpretable layers (where the learned features are directly interpretable without the need for *ad hoc* techniques) were recently applied to EEG decoding tasks (Zhao et al., 2019; Borra et al., 2020b,c).

Within this field of research, the aim of this study is to further contribute to the development of CNNs for EEG-based P300 decoding and to their analysis, with particular emphasis on the following aspects: keeping limited the number of trainable parameters (also referred to as model size) to realize lightweight CNNs suitable also for small datasets; assessing the effects of different learning strategies (including transfer learning) in view of the practical usage of these algorithms in BCIs; explaining the CNN decision i.e., the neurophysiological aspects that resulted in an optimal discriminability between classes. Specifically, the main contribution points are the following:

i) The realization of a CNN named MS-EEGNet combining two designs previously proposed in the literature with unique characteristics but treated separately, with the aim of jointly exploiting their respective strengths (see section MS-EEGNet). On one hand, we adopted a branched architecture in order to extract features in two different timescales, since this may improve the performance of P300 decoding (as suggested by Farahat et al., 2019). On the other hand, the branched solution would tend to increase the number of convolutional layers (since convolutions are replicated along each branch) and consequently the number of trainable parameters. Therefore, we adopted solutions to keep limited the number of trainable parameters by limiting the overall number of convolutional layers (designing a shallow network) and at the same time implementing computationally efficient convolutions, such as depthwise and separable convolutions (as adopted by Lawhern et al., 2018). The latter are characterized by a reduced number of required multiplications, hence by a lower computational cost, and by a reduced number of trainable parameters compared with conventional convolutions (as those adopted by Farahat et al., 2019). In addition, learning compressed temporal representations in MS-EEGNet helped to further reduce the overall model size. In this way, we proposed a multi-scale lightweight design. The so obtained network was then thoroughly analyzed to evaluate its performance and potentialities in view of practical applications (see points below).
ii) Analysis of the main hyper-parameters of the architecture, evaluating variant designs to investigate the role of multi-scale temporal feature learning (see section Alternative Design Choices of MS-EEGNet: Changing Hyper-parameters in the MST Block).
iii) Application of MS-EEGNet to three different datasets, to evaluate the proposed approach on variable-sized datasets and on differently elicited P300 responses, comparing the performance with other SOA algorithms, including both CNNs and a traditional ML pipeline (see sections Data and Pre-processing and State-of-the-Art Algorithms).
iv) Training of MS-EEGNet with different strategies that include transfer learning. Transfer learning is of relevance as it could provide important benefits in practical BCI applications, alleviating the need for a large training set and reducing training times when using the CNN on a new user (see section Training).
v) Application of an explanation technique based on saliency maps to derive the spatial and temporal features that drove MS-EEGNet decision (see section Explaining P300 Decision: Gradient-Based Representations).

MATERIALS AND METHODS

In this section, first, we introduce the problem of EEG decoding *via* CNNs. Then, we describe the proposed architecture in its baseline and variant versions, P300 datasets, re-implemented SOA algorithms, training strategies, and CNN explanation. Lastly, we illustrate the adopted statistical analyses.

Convolutional neural networks were developed in PyTorch (Paszke et al., 2017) and trained using a workstation equipped with an AMD Threadripper 1900X, NVIDIA TITAN V, and 32 GB of RAM. Codes of MS-EEGNet are available at https://github.com/ddavidebb/P300_decoding_MS-EEGNet.

EEG Decoding *via* CNNs

Let us consider an EEG dataset collected from many participants and recording sessions. Each single participant- and session-specific dataset is composed of many trials collected by epoching the continuous EEG recording with respect to the onset of the stimulus (e.g., standard or deviant stimulus). Thus, each trial is associated with a specific class (e.g., non-P300 or P300 class), with a total of N_c classes. Indicating with $M^{(s,r)}$ the total number of trials for the s-th subject and the r-th recording session, the corresponding dataset can be formalized as: $D^{(s,r)} = \left\{\left(X_0^{(s,r)}, y_0^{(s,r)}\right), \ldots, \left(X_i^{(s,r)}, y_i^{(s,r)}\right), \ldots, \left(X_{M^{(s,r)}-1}^{(s,r)}, y_{M^{(s,r)}-1}^{(s,r)}\right)\right\}$. $X_i^{(s,r)} \in \mathbb{R}^{C\times T}$ represents the pre-processed EEG signals of the i-th trial ($0 \leq i \leq M^{(s,r)} - 1$), with C indicating the number of electrodes and T indicating the number of time steps. $y_i^{(s,r)}$ is the label associated with $X_i^{(s,r)}$, i.e., $y_i^{(s,r)} \in L = \{l_0, \ldots, l_{N_c-1}\}$. In the particular case of P300 decoding, i.e., discrimination between standard and deviant trials, $N_c = 2$ and $L = \{l_0, l_1\} = \{\text{“}non - P300, \text{“}P300''\}$.

The objective decoding problem can be formalized as the optimization of a parametrized classifier f implemented by a CNN, $f\left(X_i^{(s,r)}; \theta\right) : \mathbb{R}^{C\times T} \rightarrow L$, with parameters θ, learning from a training set to assign the correct label to unseen EEG trials. Therefore, in the following, we refer to $X_i^{(s,r)}$ as the CNN input, represented as a 2D matrix of shape(C, T) with time steps along the width and electrodes along the height. Lastly, each dataset $D^{(s,r)}$ was divided into a training set used to optimize the parameters contained in array θ, and a test set used to evaluate the algorithm on unseen data. Furthermore, a separate validation set needs to be extracted from the training set to define a stop criterion of the optimization. As described in section Data and Pre-processing, here, we used three datasets: dataset 1 was a large public dataset where each participant performed different recording sessions, while datasets 2 and 3 were two small private datasets where each participant performed a single recording session.

The Proposed Convolutional Neural Network and Its Variants

MS-EEGNet

The proposed shallow architecture was composed of three fundamental blocks, each consisting of many layers. A schematic representation of the CNN is reported in **Figure 1**. The spatio-temporal (ST) block extracted temporal and spatial features from the input EEG signals *via* temporal and spatial convolutional layers, respectively. Downstream, the multi-scale temporal (MST) block used lightweight parallel temporal convolutions to extract temporal patterns in different scales from the feature maps provided by the previous block. Lastly, multi-scale activations were provided to the fully-connected (FC) block that finalized the decoding task using a single fully-connected layer.

In all the layers except for the last two, the output was a collection of spatio-temporal feature maps and its shape can be described by a tuple of three integers, with the first integer indicating the number of feature maps, and the second and third integers representing the number of spatial and temporal samples within each map, respectively. In the following, to describe the CNN, we will refer to the hyper-parameters of the involved layers. Each convolutional layer is characterized by the number of convolutional kernels (K), kernel size (F), stride size (S), and padding size (P). In addition, depthwise convolution introduced also a depth multiplier (D) specifying the number of kernels to learn for each input feature map. Hyper-parameters will be denoted by a superscript and a subscript. The superscript indicates the specific block to which the layer belongs using acronyms “ST,” “MST_0,” “MST_1,” and “FC,” where the index in the MST block discriminates between the two scales (in general MST_i, where $0 \leq i \leq N_b - 1$ and N_b denotes the number of parallel branches). The subscript indicates to which convolutional layer inside the block the hyper-parameters refer (convolutional layers inside each block were labeled with an increasing index, starting from 0). Lastly, pooling layers were described by pool size (F_p) and pool stride (S_p), with the corresponding superscript. Both convolutions and poolings were 2D; therefore, F, S, P, F_p, and S_p were tuples of two integers: the first referred to the spatial dimension, while the second referred to the temporal dimension. Lastly, the number of time samples changed across pooling operators and was denoted with T_p. Regarding the single fully-connected layer included in the classification block, the number of neurons was denoted with N^{FC} and represented the number of classes to decode (N_c).

MS-EEGNet was analyzed in a baseline version and in many variants by adopting a *post-hoc* hyper-parameter evaluation procedure on the main MST block hyper-parameters. The baseline version is described in the current section, where the structure and function of each block are presented, while the variants are described in section Alternative Design Choices of MS-EEGNet: Changing Hyper-parameters in the MST Block.

i. Spatio-temporal block. This was designed to learn temporal and spatial features separately. At first, a temporal convolutional layer was included, learning ${K_0}^{ST} = 8$ temporal kernels with filter size ${F_0}^{ST} = (1, 65)$, unitary stride and zero padding ${P_0}^{ST} = (0, 32)$ to preserve the number of input temporal samples. Then, the ${D_1}^{ST} = 2$ spatial filters of size $(C, 1)$ were learned for each temporal feature map in a spatial depthwise convolutional layer, with unitary stride and without zero padding (Lawhern et al., 2018; Borra et al., 2020a). Thus, a total number of ${K_1}^{ST} = {K_0}^{ST} \cdot {D_1}^{ST} = 16$ spatial filters were learned and constrained to have a norm upper bounded by $c = 1$ (kernel max-norm constraint) as in previous studies (Lawhern et al., 2018; Borra et al., 2020a; Vahid et al., 2020). The feature maps of this layer were not fully connected with the feature maps of the previous layer. This not only reduced the number of trainable parameters but also allowed more straightforward spatio-temporal feature learning. Indeed, each group of D_1 spatial filters was related to a specific temporal filter (Lawhern et al., 2018) (i.e., to specific spectral information). Furthermore, the output activations of the temporal and spatial convolutional layers were normalized *via* batch normalization (Ioffe and Szegedy, 2015). Downstream the spatial depthwise

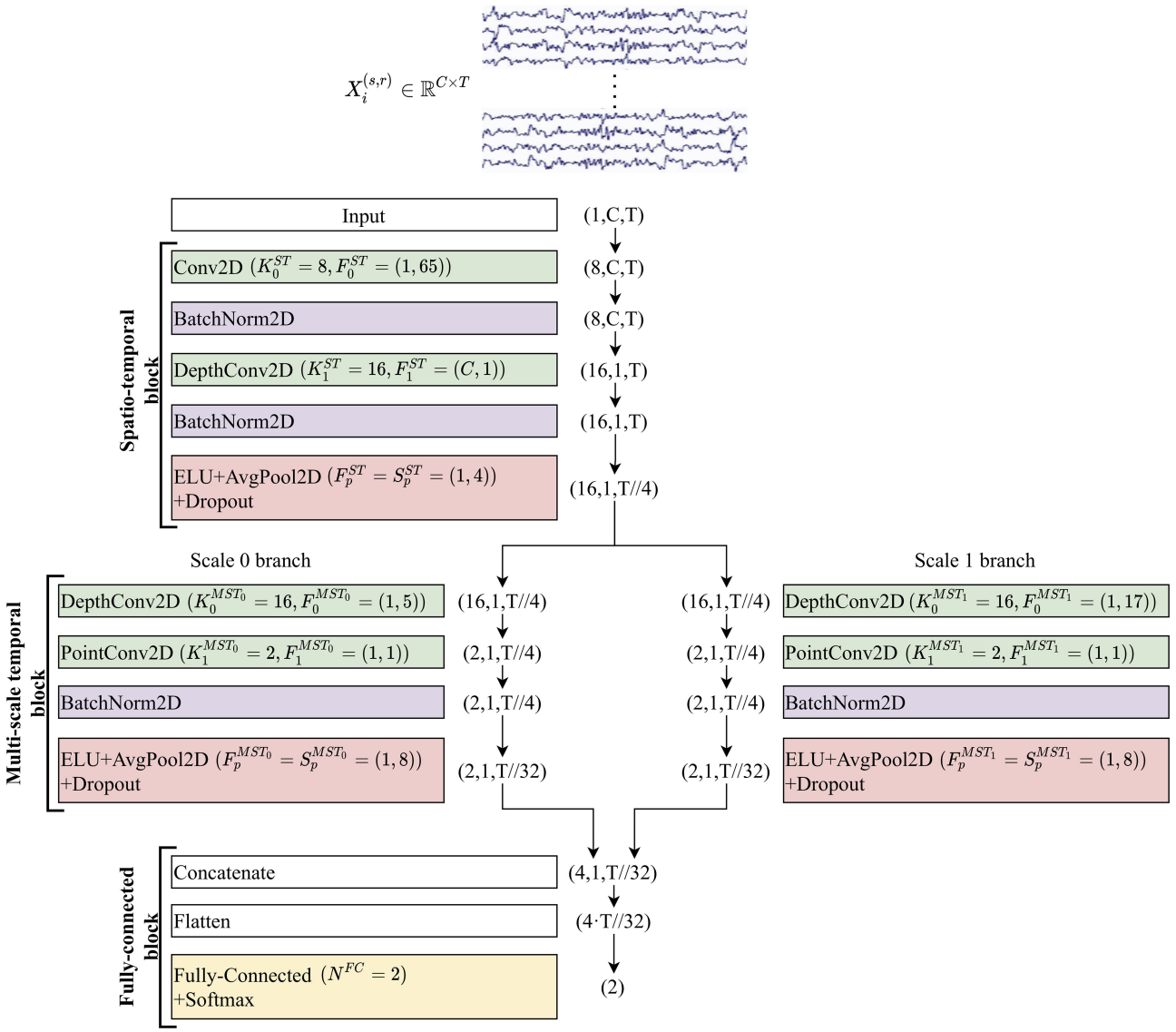

FIGURE 1 | Structure of MS-EEGNet. Layers are represented by colored rectangles, reporting the layer name and main hyper-parameters. The tuple outside each rectangle represents the output shape of each layer. For all outputs except the last two (Flatten and Fully-connected + Softmax), the tuples are composed of three numbers representing the number of feature maps (channel dimension), number of spatial samples, and number of temporal samples within each map. The input layer provides an output of shape $(1,\ C, T)$, as it just replicates the original input matrix with shape (C, T), providing a single feature map as output. The temporal dimension changed from T to $T//32$ along the entire CNN (where the symbol // indicates the floor division operator) due to average pooling operations. See sections EEG Decoding via CNNs and The Proposed Convolutional Neural Network and Its Variants for the meaning of symbols, and see **Table 1** for further details.

convolution and its associated batch normalization, the neurons were activated *via* an exponential linear unit (ELU) non-linearity (Clevert et al., 2015), i.e., $f(x) = x,\ x > 0$ and $f(x) = \alpha\left(\exp(x) - 1\right),\ x \leq 0$. We adopted this activation function, since it was proved to allow faster and more noise-robust learning than other non-linearities (Clevert et al., 2015) and to outperform other activation functions when using CNNs with EEG signals (Schirrmeister et al., 2017). The α hyper-parameter controls the saturation value for negative inputs, and $\alpha = 1$ was set here. Then, an average pooling layer was introduced to reduce the size of the activations along the temporal dimension from T to ${T_p}^{ST}$, with a pool size of ${F_p}^{ST} = (1, 4)$ and pool stride of ${S_p}^{ST} = (1, 4)$, providing activations sampled in 1/4 of the sampling frequency of the signals (32 Hz when using signals extracted from dataset 1 and approximately 31.3 Hz from datasets 2 and 3). Lastly, a dropout layer (Srivastava et al., 2014) (with a different dropout rate p depending on the training strategy adopted, see section Training) was added.

ii. Multi-scale temporal block. This block was designed to learn how to summarize along the temporal dimension the feature maps provided by the ST block. Differently from EEGNet

where features in a single timescale were learned at this stage, here the features were learned in N_b different timescales, inspired by the design of the Inception modules (Szegedy et al., 2015). In the baseline MS-EEGNet $N_b = 2$, thus, two different sets of short and large kernels were separately learned in the two parallel branches. This was accomplished *via* two parallel temporal depthwise convolutional layers with a unitary depth multiplier, i.e., ${D_0}^{MST_0} = {D_0}^{MST_1} = {D_0}^{MST} = 1$ and ${K_0}^{MST_0} = {K_0}^{MST_1} = {K_0}^{MST} = {K_1}^{ST} \cdot {D_0}^{MST}$, and with different kernel sizes in the two branches extracting a summary of roughly 150 ms $[{F_0}^{MST_0} = (1, 5)]$ and 500 ms $[{F_0}^{MST_1} = (1, 17)]$, for each input feature map. That is, each output feature map was a sort of weighted moving average of the input feature map using moving windows of two different lengths, ∼150 and 500 ms (referred to as scales). The large kernel size was chosen to match the temporal kernel size used in the single-scale branch of EEGNet (Borra et al., 2020a). The small kernel size was chosen so that the ratio between the small and large kernels was approximately the same as that in BranchedNet ($r^{MST} = \frac{1}{4}$), keeping odd kernel size (i.e., 500 ms/4 = 125 ms = four samples at 32 Hz, approximated to five samples to have an odd integer). The small and large temporal filters should be able to learn high and low-frequency patterns from the input, respectively (Supratak et al., 2017). Here, unitary stride and zero-padding of ${P_0}^{MST_0} = (0, 2)$ and ${P_0}^{MST_1} = (0, 8)$ were adopted, preserving the number of the input temporal samples. After each depthwise convolutional layer, a pointwise convolutional layer $[{F_1}^{MST_0} = {F_1}^{MST_1} = {F_1}^{MST} = (1, 1)]$ was added to learn how to optimally combine the feature maps in a specific timescale with unitary stride and without zero-padding. At variance with BranchedNet (Farahat et al., 2019) where convolutions were not designed to keep limited the number of trainable parameters, the proposed multi-scale temporal block was designed using separable convolutions (i.e. depthwise convolution followed by pointwise convolution) with the specific aim of reducing the training parameters. In this same perspective, the number of output feature maps was set as low as ${K_1}^{MST_0} = {K_1}^{MST_1} = {K_1}^{MST} = 2$ in each branch, learning a compressed representation of the input feature maps (i.e., the 16 input feature maps provided by the depthwise convolutional layer were recombined into only two different feature maps, for each branch). Then, for each branch, the output activations of the pointwise convolutional layer were normalized *via* batch normalization (Ioffe and Szegedy, 2015) and activated with an ELU non-linearity ($\alpha = 1$). Finally, an average pooling layer was introduced with a pool size of ${F_p}^{MST_0} = {F_p}^{MST_1} = {F_p}^{MST} = (1, 8)$ and pool stride of ${S_p}^{MST_0} = {S_p}^{MST_1} = {S_p}^{MST} = (1, 8)$ to reduce the temporal dimension from ${T_p}^{ST}$ to ${T_p}^{MST}$, followed by a dropout layer (Srivastava et al., 2014) (with different dropout rate p depending on the training strategy adopted, see section Training).

iii. Fully-connected block. This block was devoted to produce output probabilities from the feature maps provided by the multi-scale temporal block. The input feature maps were concatenated together along the feature map dimension and unrolled along a single dimension *via* a flatten layer. Then, this multi-scale feature vector was given as input to a fully-connected layer with $N^{FC} = N_c = 2$ neurons (associated with the P300 and non-P300 classes). These two outputs were transformed via a Softmax activation function to obtain conditional probabilities $p\left(l_k \middle| X_i^{(s)}\right), k = 0, 1$.

A more detailed description of the structural hyper-parameters and of the number of trainable parameters of the baseline version of MS-EEGNet can be found in **Table 1**. The overall number of trainable parameters (or model size) and the training time (or computational time) of the baseline MS-EEGNet are reported in **Table 2**. Note that in this table, these variables are reported also for the variant designs of MS-EEGNet (see section Alternative Design Choices of MS-EEGNet: Changing Hyper-parameters in the MST Block) and for the examined SOACNNs (see section State-of-the-Art Algorithms).

Alternative Design Choices of MS-EEGNet: Changing Hyper-Parameters in the MST Block

In addition to the baseline MS-EEGNet described previously, we evaluated other alternative designs to better investigate the behavior of the proposed MST block by modifying some hyper-parameters (HPs) one at a time. In the following, the alternative designs are described and indicated *via* the modified HP: $HP_{variant}$ vs. $HP_{baseline}$.

i. $N_b = \{1, 3\}$ vs. $N_b = 2$: use of one or three branches. In this *post-hoc* analysis, we studied whether the proposed dual-scale temporal feature learning was beneficial compared with the traditional single-scale learning ($N_b = 1$) and which scale was able to learn more relevant class-discriminative temporal features. To this aim, MS-EEGNet was modified either by removing the short scale (scale 0), leaving only the large-scale branch $[N_b = 1(large)]$ or the large scale (scale 1) leaving only the short-scale branch $[N_b = 1(short)]$. It is worth noticing that single-scale variant design $N_b = 1$ (large) did not correspond to the EEGNet adaptation used in Borra et al. (2020a), since here we adopted compressed representations in separable convolutional layers. In addition, we studied whether a third timescale ($N_b = 3$) could be useful by modifying MS-EEGNet by the inclusion of an additional timescale between the ones of the baseline version: and this variant learned summaries of about 125, 250, and 500 ms, corresponding to kernel sizes in the MST block of ${F_0}^{MST_0} = (1, 5)$, ${F_0}^{MST_1} = (1, 9)$, and ${F_0}^{MST_2} = (1, 17)$, respectively.

ii. ${F_0}^{MST_0} = (1, 9)$ vs. ${F_0}^{MST_0} = (1, 5)$: enlarging the kernel size in the short-scale branch (scale 0 in **Table 1**). This was performed to evaluate the effect of a different ratio between the short- and large-scales of the MST block compared with the one adopted in the baseline MS-EEGNet. Specifically, $r^{MST} = \frac{1}{2}$ vs. $r^{MST} = \frac{1}{4}$ leading to 500 ms/2 = 250 ms = eight samples at 32 Hz, approximated to nine samples to have odd integer.

TABLE 1 | Architecture details of MS-EEGNet.

Block	Layer name	Hyper-parameters	Number of trainable parameters
	Input	$K_0 = 1$	0
ST	Conv2D	$K_0{}^{ST} = 8$, $F_0{}^{ST} = (1, 65)$, $P_0{}^{ST} = (0, 32)$	$F_0{}^{ST}[0] \cdot F_0{}^{ST}[1] \cdot K_0{}^{ST} \cdot K_0$
	BatchNorm2D	$m = 0.99$	$2 \cdot K_0{}^{ST}$
	Depthwise-Conv2D	$D_1{}^{ST} = 2$, $K_1{}^{ST} = K_0{}^{ST} \cdot D_1{}^{ST}$, $F_1{}^{ST} = (C, 1)$, kernel max norm=1	$F_1{}^{ST}[0] \cdot F_1{}^{ST}[1] \cdot K_0{}^{ST} \cdot D_1{}^{ST}$
	BatchNorm2D	$m = 0.99$	$2 \cdot K_1{}^{ST}$
	ELU	$\alpha = 1$	0
	AvgPool2D	$F_p{}^{ST} = S_p{}^{ST} = (1, 4)$	0
	Dropout	$p = 0.25$ or $p = 0.5$	0
MST scale 0	Depthwise-Conv2D	$D_0{}^{MST_0} = 1$, $K_0{}^{MST_0} = K_1{}^{ST} \cdot D_0{}^{MST_0}$, $F_0{}^{MST_0} = (1, 5)$, $P_0{}^{MST_0} = (0, 2)$	$F_0{}^{MST_0}[0] \cdot F_0{}^{MST_0}[1] \cdot K_1{}^{ST} \cdot D_0{}^{MST_0}$
	Pointwise-Conv2D	$K_1{}^{MST_0} = 2$, $F_1{}^{MST_0} = (1, 1)$	$F_1{}^{MST_0}[0] \cdot F_1{}^{MST_0}[1] \cdot K_1{}^{MST_0} \cdot K_0{}^{MST_0}$
	BatchNorm2D	$m = 0.99$	$2 \cdot K_1{}^{MST_0}$
	ELU	$\alpha = 1$	0
	AvgPool2D	$F_p{}^{MST_0} = S_p{}^{MST_0} = (1, 8)$	0
	Dropout	$p = 0.25$ or $p = 0.5$	0
MST scale1	Depthwise-Conv2D	$D_0{}^{MST_1} = 1$, $K_0{}^{MST_1} = K_1{}^{ST} \cdot D_0{}^{MST_1}$, $F_0{}^{MST_1} = (1, 17)$, $P_0{}^{MST_1} = (0, 8)$	$F_0{}^{MST_1}[0] \cdot F_0{}^{MST_1}[1] \cdot K_1{}^{ST} \cdot D_0{}^{MST_1}$
	Pointwise-Conv2D	$K_1{}^{MST_1} = 2$, $F_1{}^{MST_1} = (1, 1)$	$F_1{}^{MST_1}[0] \cdot F_1{}^{MST_1}[1] \cdot K_1{}^{MST_1} \cdot K_0{}^{MST_1}$
	BatchNorm2D	$m = 0.99$	$2 \cdot K_1{}^{MST_1}$
	ELU	$\alpha = 1$	0
	AvgPool2D	$F_p{}^{MST_1} = S_p{}^{MST_1} = (1, 8)$	0
	Dropout	$p = 0.25$ or $p = 0.5$	0
FC	Concatenate Flatten		0
	Fully-Connected	$N^{FC} = 2$	$N^{FC} \cdot (T_p{}^{MST_0} \cdot K_1{}^{MST_0} + T_p{}^{MST_1} \cdot K_1{}^{MST_1} + 1)$
	Softmax		0

Each layer is provided with its name, main hyper-parameters and the number of trainable parameters. See sections EEG Decoding via CNNs and The Proposed Convolutional Neural Network and Its Variants for the meaning of symbols. The total number of trainable parameters was 1,154 when using signals from dataset 1 and 1,210 when using signals from datasets 2 and 3. In all layers, unless otherwise noted stride (S) and padding (P) were set to (1, 1) *and* (0, 0)*, respectively.*

iii. $K_1{}^{MST} = \{1, 8, 16\}$ vs. $K_1{}^{MST} = 2$: different number of feature maps in the pointwise convolutions. In particular, $K_1{}^{MST}$ was set to 1 in each branch in order to analyze whether the learning of a single recombination of the input feature maps was enough to provide an accurate decoding performance. In addition, $K_1{}^{MST}$ was set to 8 in each branch in order to analyze another compressed representation, while maintaining the total number of feature maps across the two different timescales unchanged as in the MST input (i.e., eight feature maps in each branch, resulting in 16 feature maps across the two scales, as in the input of the MST block). Lastly, $K_1{}^{MST}$ was set to 16 in each branch, corresponding to a condition where no compressed representation was learned in either branch.

iv. Deep MST vs. MST: increasing the depth of the MST block. This was performed to evaluate the effect on the performance of an increased depth in the MST block (and thus, learning more non-linear dependencies) while maintaining the same overall receptive field of the neurons in the temporal domain. In each branch, we added another depthwise convolutional layer after the first one. However, in order to maintain the same receptive field as when using a single depthwise convolutional layer in the baseline MST block, the kernel size of each depthwise convolutional layer was halved with respect to the baseline values, i.e., $F_0{}^{MST_0} = F_1{}^{MST_0} = (1, 3)$ and $F_0{}^{MST_1} = F_1{}^{MST_1} = (1, 9)$. After the second depthwise convolutional layer, the pointwise convolutional layer was added [$F_2{}^{MST_0} = F_2{}^{MST_1} = F_2{}^{MST} = (1, 1)$], and the rest of the block was maintained unchanged as in the baseline version.

Overall, eight variants were designed by changing a specific hyper-parameter value of MS-EEGNet while keeping all the other hyper-parameters as in the baseline MS-EEGNet. These alternative designs were trained with a within-participant and within-session strategy (as it is the most common strategy adopted in the literature) and compared with MS-EEGNet trained with the same strategy. Lastly, the number of trainable parameters and training time are reported in **Table 2** for each variant design.

Data and Pre-processing

Dataset 1

The first dataset is BCIAUT-P300, a public benchmark dataset released for the IFMBE 2019 scientific challenge (available at https://www.kaggle.com/disbeat/bciaut-p300) (Simões et al.,

TABLE 2 | Number of trainable parameters, also denoted as model size in the text, and training time (referred to the WS strategy), also denoted as computational time, of the baseline MS-EEGNet, MS-EEGNet variants, and SOA CNNs when using signals from dataset 1 and datasets 2–3.

Algorithm	Trainable parameters (dataset 1/datasets 2–3)		Training time (dataset 1/datasets 2–3)	
	Value	Δ (%)	Value (ms/epoch)	Δ (%)
Baseline MS-EEGNet	1,154/1,210	–	220/45.5	–
MS-EEGNet variants				
$N_b = 1$(*large*)	1,022/1,082	−11.4/−10.6	195/38.1	−11.4/−16.3
$N_b = 1$(*short*)	830/890	−28.1/−26.5	172/38.4	−21.8/−15.6
$N_b = 3$	1,350/1,402	17.0/15.9	282/50.8	28.2/11.6
${F_0}^{MST_0} = (1, 9)$	1,218/1,274	5.5/5.3	221/46.3	0.5/1.8
${K_1}^{MST} = 1$	1,102/1,162	−4.5/−4.0	224/45.0	1.8/−1.1
${K_1}^{MST} = 8$	1,466/1,498	27.0/23.8	287/46.1	30.5/1.3
${K_1}^{MST} = 16$	1,882/1,882	63.1/55.5	240/45.0	9.0/−1.1
deepMST	1,202/1,258	4.2/4.0	295/47.0	34.1/3.3
SOA CNNs				
EEGNet	1,386/1,418	20.1/17.2	186/40.5	−15.5/−11.0
BranchedNet	5,418/7,954	369/557	250/50.3	13.6/10.5
OCLNN	1,650/1,874	43.0/54.9	96.2/22.9	−56.3/−49.7

These values were reported for deep learning-based decoders to provide a more complete comparison between the proposed CNN and SOA CNNs. For each CNN, between dataset 1 and datasets 2–3, the different number of parameters resulted from the different number of EEG channels ($C = 8$ for dataset 1 and $C = 12$ for datasets 2–3, see section Data and Pre-processing) and time samples considered ($T = 140$ for dataset 1 and $T = 113$ for datasets 2–3, see section Data and Pre-processing), while the different training time resulted from the different number of training examples (1,280 trials and 240 trials for each participant and each session, respectively, for dataset 1 and datasets 2–3, see section Data and Pre-processing). In addition, the percentage difference (Δ) of trainable parameters and training time between SOA CNNs or MS-EEGNet variants ("other" condition) and the baseline MS-EEGNet ("baseline" condition) is reported, i.e., $100 \cdot (value_{other} - value_{baseline})/value_{baseline}$.

2020) consisting of a larger number of examples than other public benchmarks (Blankertz et al., 2004, 2006) or private (Lawhern et al., 2018; Farahat et al., 2019; Solon et al., 2019) datasets. Signals were recorded from 15 participants (all males, age of 22 ± 5 years, mean ± standard deviation) with ASD during seven recording sessions (for a total of 4 months) while testing a P300-based BCI (Amaral et al., 2018). The paradigm consisted of the participants paying attention to one of eight objects randomly flashing in a virtual scene, with P300 stimuli corresponding to the flashing of the attended object (this was repeated several times for each different attended object). For each participant and recording session, 1,600 trials were recorded during the calibration stage (training set), and 2,838 trials were recorded during the online stage (test set), on average.

Signals were recorded at 250 Hz from eight electrodes: C3, Cz, C4, CPz, P3, Pz, P4, and POz. The reference was placed at the right ear and the ground at AFz. These signals were acquired notch filtered at 50 Hz and then pass-band filtered between 2 and 30 Hz (Simões et al., 2020). EEG signals were pre-processed as in previous studies (Amaral et al., 2017; Borra et al., 2020a). In particular, epochs were selected from −100 to 1,000 ms relative to the event stimulus, and signals were downsampled to 128 Hz to reduce the number of time steps to be processed in the CNN. Architectures were trained as described in section Training using the training set of the competition for each session, while the test set was used to test the algorithms. From each participant- and session-specific training set, a validation set of 20% of the total training set was extracted (corresponding to 320 trials) to perform early stopping, while the remaining percentage of the total training set (corresponding to 1,280 trials) was used to optimize the architectures.

Datasets 2 and 3

The second dataset was collected from seven participants (all males, age 25 ± 8 years) recorded in an auditory oddball study during a single recording session, and the third dataset was collected from seven participants (5 males, age 22 ± 0.4 years, different from dataset 2 participants) recorded in a visual oddball study during a single recording session. All the participants were healthy volunteers not reporting psychological or hearing disorders. Both experiments were approved by the Bioethics Committee of the University of Bologna (file number 29146, year 2019) and were conducted in a controlled laboratory environment.

The auditory oddball paradigm consisted of 400 tones presented to the participants through a speaker, with the standard and deviant stimuli differing by the frequency of tones (500 and 1,000 Hz, respectively). The visual oddball paradigm consisted of 400 stimuli presented to the participants through a bicolor LED with the standard and deviant stimuli differing by the LED color (blue and red, respectively). In both paradigms, each stimulus was reproduced for 56 ms followed by a pause of 944 ms (inter-stimuli interval); thus, each trial lasted 1 s. This paradigm was similar to the one adopted by Justen and Herbert (2018). Furthermore, in each paradigm, a total number of 325 standard and 75 deviant stimuli were presented to participants in a randomized order. Thus, for each participant, a total number of 400 trials were available, with a class imbalance ratio of 75:325 for the P300 and non-P300 classes. While listening to the tones or while looking at the LED, the participants were seated in a comfortable chair in front of a button with their eyes opened, and they were instructed to respond to the deviant stimuli by pressing a button with their right index finger as quickly as possible, minimizing other movements.

Signals of both datasets 2 and 3 were recorded at 125 Hz using a portable EEG recording system (OpenBCI system, using Cyton and Daisy Biosensing boards) from 12 electrodes: C3, Cz, C4, CP5, CP1, CP2, CP6, P3, Pz, P4, PO3, and PO4. The reference was placed at the right earlobe and the ground at the left earlobe. The same pre-processing was adopted for datasets 2 and 3. In particular, signals were band-pass filtered between 2 and 30 Hz with a zero-phase second-order filter, and epochs were extracted from −100 to 800 ms relative to the stimulus onset. For datasets 2 and 3, the architectures were trained as described in section Training using a 4-fold cross-validation scheme. Therefore, in each fold, each participant-specific dataset was divided into a training (75%) and a test (25%) set, corresponding to 300 and 100 trials, respectively. Lastly, a validation set of 20% of the training set (corresponding to

60 trials) was extracted to perform early stopping, while the remaining percentage (corresponding to 240 trials) was used to optimize the architectures.

As described in section EEG Decoding via CNNs, $X_i^{(s,r)} \in \mathbb{R}^{C\times T}$ represented the CNN input. From the previous dataset descriptions, $C = 8$ for dataset 1 and $C = 12$ for datasets 2 and 3, while $T = 140$ for dataset 1 and $T = 113$ for datasets 2 and 3.

State-of-the-Art Algorithms

The proposed baseline architecture was compared with other SOA algorithms, such as the winning algorithm of the IFMBE 2019 challenge based on EEGNet (Borra et al., 2020a), BranchedNet (Farahat et al., 2019), and OCLNN (Shan et al., 2018). The first was a single-branched CNN performing the temporal convolution in the first layer. The second one, was a dual-branched CNN exploiting parallel temporal convolutions but at variance with the architecture proposed here, performed spatial convolution in the first layer and did not use optimized convolutions aimed to keep limited the number of trainable parameters, resulting in a less parsimonious multi-scale CNN. OCLNN was a CNN performing a mixed spatio-temporal convolution in the first layer without using optimized convolutions. To allow for a more complete comparison between MS-EEGNet and other deep learning-based decoders, the number of trainable parameters and training time of SOA CNNs are summarized in **Table 2**.

In addition to these SOA CNNs, we re-implemented xDAWN+RG, an ML pipeline for P300 decoding. In particular, this solution included a combination of xDAWN spatial filtering (Rivet et al., 2009; Barachant and Congedo, 2014), Riemannian Geometry (Barachant et al., 2012), L_1 feature regularization, and classification based on an Elastic Net regression.

Details about SOA CNNs and xDAWN+RG can be found in sections 1 and 2 in **Supplementary Materials**.

Training

MS-EEGNet was trained with different training strategies.

i. Within-participant and within-session training (WS). For each participant and session, EEG signals (see section Data and Pre-processing) were used to train, validate, and test a participant-specific and session-specific CNN. In addition, we also trained CNNs using only a fraction of the participant- and session-specific training set, simulating practical cases of reduced numbers of available calibration trials, and investigated how the performance changed; this is an important issue from the perspective of limiting the calibration time in practical applications. Reduced training sets were defined by extracting 15, 30, 45, and 60% of the total training set in the corresponding session (corresponding to 192, 384, 576, and 768 training trials for dataset 1, and 48, 96, 144, and 192 trials for datasets 2 and 3, maintaining the class imbalance characterizing each dataset. For each architecture, 105 (15 participants * 7 sessions per participant) CNNs were trained for dataset 1, while seven (7 participants * 1 session per participant) CNNs were trained for datasets 2 and 3. The WS strategy (with 100% of training trials) was adopted also with SOA algorithms to perform *post-hoc* hyper-parameter evaluation.

ii. Within-participant and cross-session training (CS). This training strategy was adopted only for the dataset 1 because of its multi-session dimension and used in the winning solution of the authors in the IFMBE 2019 challenge (Borra et al., 2020a) using the same dataset. For each participant, an overall training set and an overall validation set were obtained by considering all the session-specific training and validation sets belonging to that particular participant. Then, these overall sets were used to train and validate a participant-specific CNN incorporating inter-session variability. It is worth noticing that this participant-specific CNN was then tested separately over each session-specific test set (relative to that participant) for consistency with the test procedure adopted in i). For each architecture, 15 CNNs were trained for dataset 1. This strategy was adopted also with SOA algorithms.

iii. Leave-one-subject-out training (LOSO). The EEG signals of one participant (i-th participant) were held back, and the training and validation sets were obtained by collecting EEG signals from all the session-specific training and validation sets of the remaining participants (j-th participants $\forall j,\ j \neq i$). Thus, for each held back participant ($\forall i$) an architecture was trained and validated with signals extracted from 14 participants for dataset 1 and from six participants for datasets 2 and 3. The so obtained network was then tested separately over each session-specific test set of the held back participant, consistently with the testing procedure in (i) and (ii). The residual signals of the held back participant not used in the testing procedure remained unused (i.e., 0% of the dataset of the held back participant was used to train and validate the model); that is, LOSO models did not learn from the examples of the held back participant. This training strategy led to a CNN incorporating inter-participant and (in case of dataset 1) inter-session variabilities. For each architecture, 15 CNNs were trained for dataset 1, while seven CNNs were trained for datasets 2 and 3. This strategy was adopted also with SOA algorithms. Lastly, to design LOSO models incorporating the knowledge from a variable number of participants, we additionally performed trainings extracting signals from a random subset of participants, i.e., using 10, six, and two participants for dataset 1, and using four and two participants for datasets 2 and 3. Thus, the performed LOSO strategy was named "LOSO-M," where M is the number of participants used ($M = \{14, 10, 6, 2\}$ when using signals from dataset 1, while $M = \{6, 4, 2\}$ for datasets 2 and 3). It is worth noticing that the LOSO-14 strategy for dataset 1 and LOSO-6 strategy for datasets 2 and 3 corresponded to the conventional LOSO strategies for these datasets.

iv. Transfer learning (TL) on single sessions (WS). As in the WS strategy (point i), for each participant and session, EEG signals (see section Data and Pre-processing) were used to train, validate, and test a participant- and session-specific CNN. Differently from the WS strategy where the trainable parameters were initialized randomly, in the TL-WS strategy,

the parameters were initialized from the ones obtained with LOSO trainings when the specific participant of interest was held back. Therefore, the knowledge learned in the LOSO strategy (using training examples sampled from many participants except the held back participant) was transferred to the held back participant. Then, a fraction of the session-specific training set of the held back participant was used as training set, using the same percentages as in the WS strategy (point i). In this way, we compared the performance of the WS and TL-WS strategies to investigate if and to what extent the TL-WS strategy outperformed the WS strategy with a reduced number of calibration trials. For each architecture, 105 CNNs were trained for dataset 1, while seven CNNs were trained for datasets 2 and 3.

The transfer learning strategy reflects a practical situation in which a new user approaches the BCI system in a new session, and a calibration phase, as short as possible, is needed to obtain an accurate participant-specific decoder. Therefore, a pre-trained model that incorporates both inter-participant and inter-session variabilities as obtained with the LOSO strategy could be a better initialization point with respect to the random one (as used in the WS training strategy), leading to performance improvement especially when using only a small number of training examples of a new user in a new recording session.

The adopted training strategies had a different definition of the training set. However, in all cases, CNNs were tested on the same participant-specific and session-specific test sets, allowing a fair comparison across different training strategies. In this study, the adopted metric to quantify the performance for the P300 decoding task at the trial level was the area under the ROC curve (AUC), as done previously (Lawhern et al., 2018), and was computed on each participant- and session-specific test set.

EEG signals of the training, validation, and test sets were standardized by computing the mean and variance on the training set. Regarding the TL-WS strategy, the first and second moments were computed on the training set used to train the pre-trained models. Except for the TL-WS strategy in which the trainable parameters were initialized from the pre-trained models, in the other training approaches, the weights were randomly initialized by adopting a Xavier uniform initialization scheme (Glorot and Bengio, 2010), and biases were initialized to zero.

The optimization was performed by minimizing the negative log likelihood or, equivalently, the cross-entropy between the empirical probability distribution defined by the training labels and the probability distribution defined by the model. Adaptive moment estimation (Adam) (Kingma and Ba, 2014) was used as an optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ for computing the running averages of the gradient and its square, and $\varepsilon = 10^{-8}$ to improve numerical stability. The learning rate was set to $lr = 10^{-3}$for the WS, CS, and LOSO strategies, while for the TL-WS strategy the optimizer state was the same as the one of the pre-trained models. To address class imbalance, a single mini-batch of data was composed by a proportion of 50–50% of the two classes, randomly selecting the trials within the dataset as done in Borra et al. (2020a). The mini-batch size and the maximum number of epochs were set to 64 and 500, respectively, and early stopping was performed by interrupting the optimization when validation loss did not decrease for 50 consecutive epochs.

In addition to early stopping, which acts as a regularizer, other regularizer mechanisms were integrated into MS-EEGNet as mentioned in section MS-EEGNet, comprising batch normalization (Ioffe and Szegedy, 2015) with a momentum term of $m = 0.99$ and $\varepsilon = 1e-3$ for numerical stability, dropout (Srivastava et al., 2014) with a dropout probability of 0.5 for WS and TL-WS trainings and 0.25 for CS and LOSO trainings, and kernel max-norm constraint.

Explaining P300 Decision: Gradient-Based Representations

The MS-EEGNet decision was explained using the saliency maps and *post-hoc* (i.e., obtained once the CNN training has ended) gradient-based representations proposed by Simonyan et al. (2013) to quantify the importance of neurons belonging to a target layer of interest (commonly the input layer) for a specific class. These representations are commonly used to explain CNN decisions when decoding EEG (Farahat et al., 2019; Borra et al., 2020c; Vahid et al., 2020) and offer the advantage of requiring the sole computation of backpropagation. Of course, other more advanced techniques, such as layer-wise relevance propagation (LRP), can represent a valid alternative but they introduce many factors that affect representations, such as the propagation rule (e.g., αβ rule) and propagation parameters (e.g., α and β) (Montavon et al., 2018), whose setting would require preliminary deep investigations. Hence, we preferred to adopt the saliency maps. Here, these were computed by backpropagating the gradient of the P300 class score (i.e., the output related to the P300 neuron, immediately before Softmax activation) back to the input layer (i.e., the neurons corresponding to the input spatio-temporal samples), when P300 trials belonging to the test set were fed as input to the CNN. Thus, each resulting saliency map was a spatio-temporal representation associated with a test trial, quantifying how much each spatio-temporal input sample affects the P300 class score, i.e., how much the P300 class score changes with respect to a small change in the input EEG signals. For each dataset, these representations were computed using MS-EEGNet trained with the LOSO strategy, as this strategy was more likely to enhance input samples relevant to the decoding task compared with WS/CS trainings (Farahat et al., 2019). Indeed, during LOSO trainings, the models were fed with signals recorded from multiple participants and multiple recording sessions. Therefore, the neural networks were more prone to learn optimal inter-participant and inter-session features to generalize properly. Conversely, during WS/CS trainings, the neural networks were more prone to learn optimal session-specific/participant-specific features. Thus, representations associated with the LOSO models were more likely to visualize general task-relevant spatio-temporal features, while those related to the WS/CS models were more likely to include also session-specific/participant-specific and task-irrelevant features.

The saliency maps were computed for each deviant trial (containing the P300 response) belonging to each participant-

TABLE 3 | AUC (% mean ± SEM) obtained with MS-EEGNet and the re-implemented SOA algorithms adopting the WS, CS, and LOSO strategies.

Algorithm	Dataset 1			Dataset 2		Dataset 3	
	WS	CS	LOSO	WS	LOSO	WS	LOSO
MS-EEGNet	**83.52 ± 1.67**	**86.38 ± 1.60**	75.40 ± 1.81	**89.60 ± 1.73**	74.82 ± 3.04	**92.63 ± 1.77**	**86.09 ± 1.88**
EEGNet	82.53 ± 1.83 **	85.88 ± 1.63 **	75.76 ± 1.71	87.98 ± 2.65	75.15 ± 3.01	91.22 ± 1.92 *	83.30 ± 2.53
BranchedNet	77.43 ± 1.65 ***	84.20 ± 1.82 ***	**76.03 ± 1.86**	83.34 ± 2.12 ***	72.39 ± 2.89	91.60 ± 1.53	84.84 ± 1.46
OCLNN	75.95 ± 1.64 ***	81.28 ± 1.65 ***	71.40 ± 1.42 **	79.92 ± 2.78 ***	**75.21 ± 3.14**	89.01 ± 2.03 ***	83.73 ± 1.59
xDAWN+RG	79.17 ± 1.43 ***	80.89 ± 1.32 ***	67.05 ± 1.71 **	82.63 ± 2.07 ***	73.83 ± 2.71	90.03 ± 1.87 *	82.40 ± 2.77

*The results of the performed Wilcoxon signed-rank tests (see section Statistics-i) are also reported (*p < 0.05, **p < 0.01, ***p < 0.001, corrected for multiple tests). Within each column, the bold characters are used to denote the best performance among the tested algorithms.*

and session-specific test set. Then, these maps were averaged across trials and folds (only for datasets 2 and 3), obtaining an average participant-specific and session-specific representation, named *spatio-temporal representation*. Then, by averaging spatio-temporal representations across sessions (seven sessions for dataset 1 and a session for datasets 2 and 3), a participant-specific representation was computed normalized between [−1, 1], and finally averaged across the participants, resulting in a *grand average (GA) spatio-temporal representation*. This representation could be useful to study similarities between the temporal course of gradients related to more relevant electrodes and the grand average ERPs of those specific electrodes. Additionally, the absolute value of each saliency map was also computed, and the absolute saliency maps were then averaged across trials, folds (only for datasets 2 and 3), and either the spatial or the temporal dimension to obtain an *absolute temporal or spatial representation*, respectively, for each participant and session. Then, by averaging the absolute temporal/spatial representation across sessions, a participant-specific representation was computed, normalized between [0, 1], and finally averaged across the participants, resulting in a *GA absolute temporal/spatial representation*. These absolute representations allowed the evaluation of more class-discriminative time samples and electrodes for the P300 class.

Statistics

Before performing the statistical analyses, AUCs were computed for each participant- and session-specific test set and then averaged across sessions (seven sessions for dataset 1 and 1 session for datasets 2 and 3), in order to compare the performance metric at the level of participant. The following statistical comparisons were performed on the performance metric.

i. Pairwise comparisons between MS-EEGNet and the SOA algorithms (EEGNet, BranchedNet, OCLNN, xDAWN+RG) trained with the WS, CS, and LOSO strategies. AUCs were compared between the contrasted conditions separately for each dataset.
ii. Pairwise comparisons between the baseline MS-EEGNet and each of its variants, trained with the WS strategy. The AUCs were merged together across different datasets and compared between the contrasted conditions using CNNs trained with the WS strategy; a similar procedure was adopted in Schirrmeister et al. (2017) and Borra et al. (2020c) in order to evaluate the overall effect of the hyper-parameters of interest with the *post-hoc* evaluation.
iii. Pairwise comparisons between MS-EEGNet trained with the WS and TL-WS strategies, for each percentage of training examples of the new user and for each number of participants (M) from whom the knowledge was transferred to the new user (see section Training-iv). This test was performed in order to evaluate the effect of the TL-WS strategy on the performance as a function of the percentage of training examples and M. In these pairwise comparisons, the AUCs were compared between the contrasted conditions separately for each dataset.

The statistical analysis performed was the same as that used in Schirrmeister et al. (2017) and Borra et al. (2020c). In particular, Wilcoxon signed-rank tests were used to check for statistically significant differences between the contrasted conditions. To correct for multiple tests, a false discovery rate correction at $\alpha = 0.05$ using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) was applied.

RESULTS

Performance

MS-EEGNet and State-of-the-Art Algorithms

Table 3 reports the AUCs at the participant level (mean ± standard error of the mean, SEM) obtained with MS-EEGNet and with SOA algorithms using WS, CS, and LOSO strategies, together with the results of the performed statistical tests.

MS-EEGNet scored an AUC of 83.52 ± 1.67%, 89.6 ± 1.73%, and 92.63 ± 1.77% when using signals from datasets 1–3 adopting the WS strategy. The proposed architecture significantly outperformed all the tested SOA algorithms when using dataset 1, and significantly outperformed BranchedNet, OCLNN, and

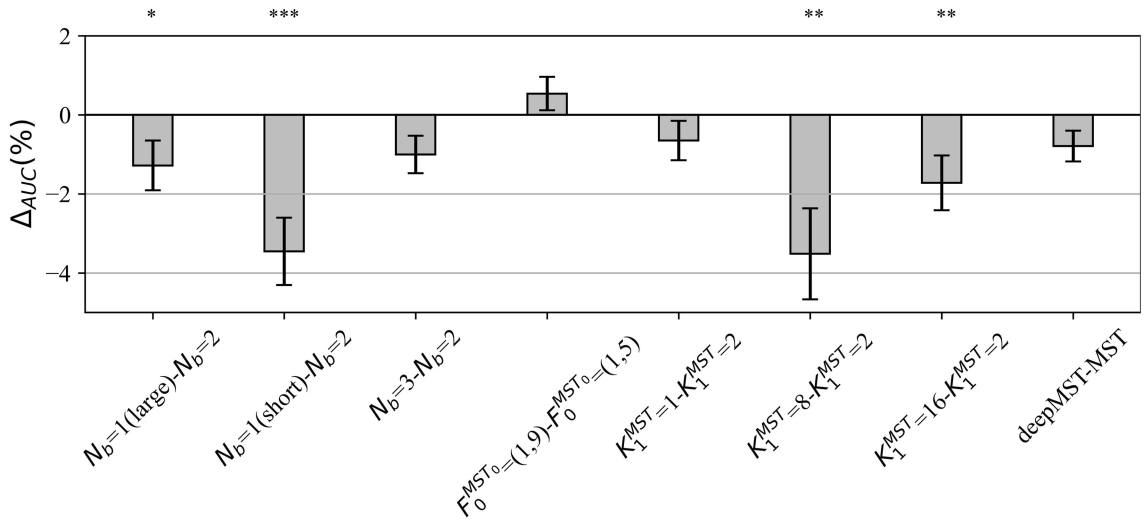

FIGURE 2 | Impact of alternative design choices of MS-EEGNet on the performance metric. The figure reports the difference between the AUC scored with the variant and the baseline design (i.e., $\Delta_{AUC} = AUC_{variant} - AUC_{baseline}$) for each condition of the hyper-parameter (*HP*) tested, reported on the x-axis as "$HP_{variant} - HP_{baseline}$." The height of each gray bar represents the mean value across the participants of Δ_{AUC}, while the error bar (black lines) represents the standard error of the mean. The results of Wilcoxon signed-rank tests (see section Statistics-ii) are also reported (*$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, corrected for multiple tests) on top of the figure.

xDAWN+RG with dataset 2, and EEGNet, OCLNN, and xDAWN+RG with dataset 3. In addition, adopting the CS strategy, MS-EEGNet confirmed its decoding improvement with respect to the SOA, scoring an AUC of 86.38 ± 1.6%, outperforming significantly all the SOA algorithms. Lastly, adopting the LOSO strategy, MS-EEGNet scored an AUC of 75.4 ± 1.81%, 74.82 ± 3.04%, and 86.09 ± 1.88% when using signals from datasets 1–3. In this strategy, the proposed solution did not perform significantly better than the other SOA solutions (see section Performance of MS-EEGNet and Comparison With State-of-the-Art Algorithms) except for dataset 1 where MS-EEGNet outperformed OCLNN and xDAWN+RG.

Design Choices of MS-EEGNet

In the *post-hoc* hyper-parameter evaluation, we investigated the effect of particular design aspects of MS-EEGNet on the decoding performance, by statistically evaluating the difference in the AUCs between each variant MS-EEGNet and baseline MS-EEGNet ($\Delta_{AUC} = AUC_{variant} - AUC_{baseline}$). The results are reported in **Figure 2**. In particular, the adoption of $N_b = 1$(large), $N_b = 1$(short), $K_1{}^{MST} = 8$, $K_1{}^{MST} = 16$ significantly worsened the performance, with an average drop in performance of 1.28, 3.46, 3.51, and 1.72%.

Variable Number of Training Examples: Within-Session and Transfer Learning Strategies

The performance obtained by MS-EEGNet in the WS strategy as a function of the percentage of training examples (reported on the x-axis) is reported in **Figures 3A–C** (white bars) for datasets 1–3. In all the datasets, a percentage of training trials of 30–45% was sufficient to obtain performance only a few points below that obtained with the entire training set, and in particular close or above 80%.

In addition, the performance obtained by MS-EEGNet in the TL-WS strategy is also reported as a function of: (i) the number of participants (M) adopted to design the LOSO-M model (gray and hatched bars); and (ii) the percentage of training examples. Lastly, the AUC difference between the TL-WS strategy and the WS strategy using the same percentage of training examples is shown in the lower panels of **Figures 3A–C** ($\Delta_{AUC} = AUC_{TL-WS} - AUC_{WS}$).

In the case of dataset 1, the TL-WS strategy provided higher performance compared with the WS strategy (see the distributions of Δ_{AUC} reported in **Figure 3**) for each percentage of training examples $\forall M$. This occurred also in the case of dataset 3 except for a couple of conditions ($M = 2$ using 30 and 60% of the training examples of the held back participant). Using dataset 2, TL was found beneficial only with the lowest number of training examples (i.e., 15%) $\forall M$ and using 60% of training examples with $M = 4$.

Explaining P300 Decision: Gradient-Based Representations

In this section, we analyze the features of the input variables that most strongly supported the P300 classification decision in MS-EEGNet.

Spatio-Temporal Representations

Figures 4A–C display the grand average spatio-temporal representation of MS-EEGNet trained with the LOSO strategy using signals from datasets 1–3. From these figures, the more class-discriminative electrodes can be identified, i.e., P4, Pz, and CP1 for datasets 1–3, respectively. The grand average ERPs for the standard and deviant stimuli of these representative electrodes are displayed in **Figures 4D–F**.

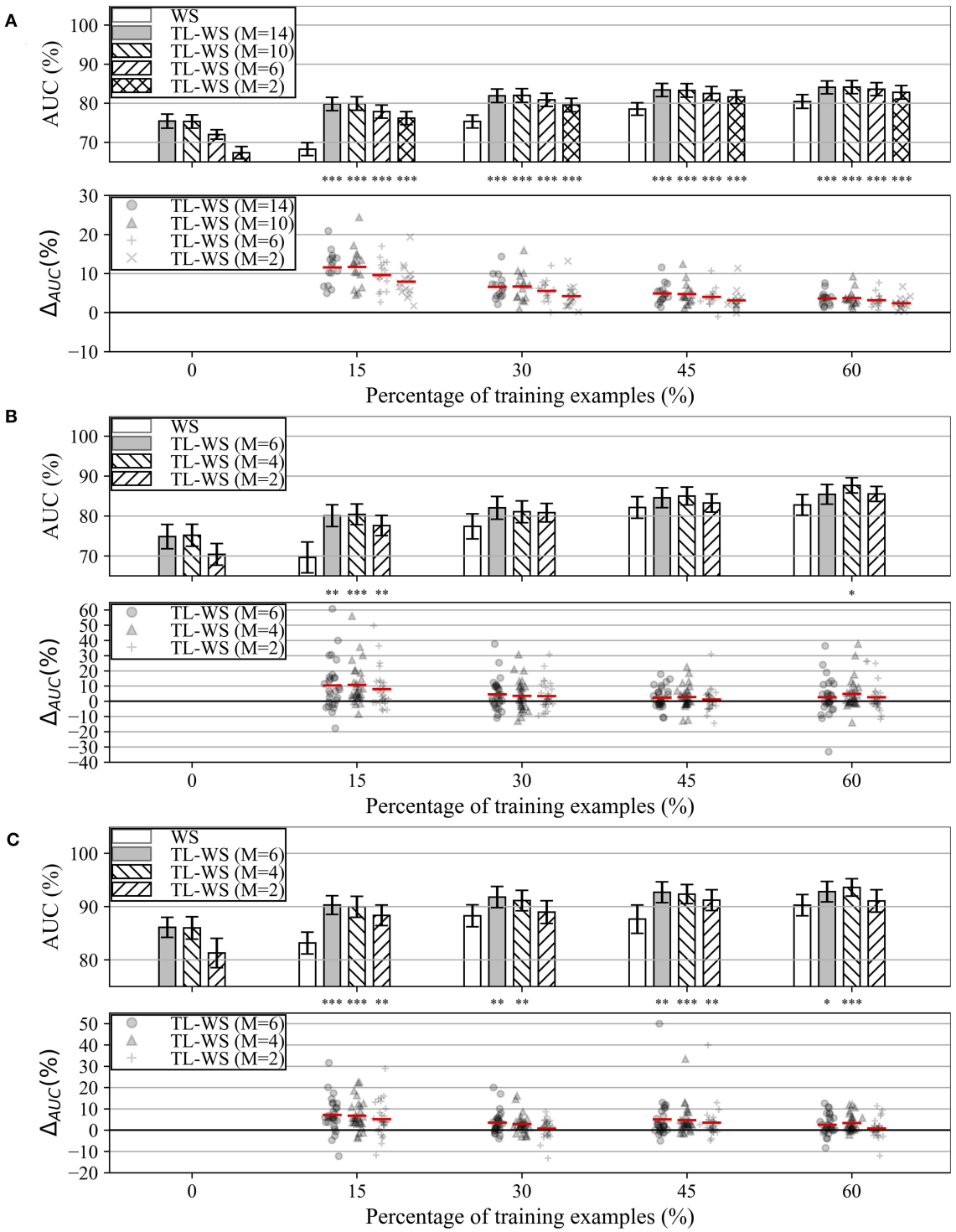

FIGURE 3 | AUC obtained with MS-EEGNet trained with the WS and TL-WS strategies for datasets 1–3 (panels **A–C**, respectively). Top plot in each panel: The AUC obtained in WS (white bars) is reported as a function of the percentage of training examples (reported on the x-axis), while the AUC obtained in TL-WS is reported also as a function of the number of participants (*M*) used to optimize the LOSO-M models (gray and hatched bars). The height of each bar represents the mean value of the performance metric across the participants, while the error bar (black lines) represents the standard error of the mean. Bottom plot in each panel: The AUC difference between the TL-WS and WS strategies (i.e., $\Delta_{AUC} = AUC_{TL-WS} - AUC_{WS}$) using the same percentage of training examples is reported using markers, and a red line denotes the mean value. For each percentage, a Wilcoxon signed-rank test was performed (see section Statistics-iii) to compare TL-WS vs. WS strategy, and the statistical significance is reported (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, corrected for multiple tests) on top of each plot.

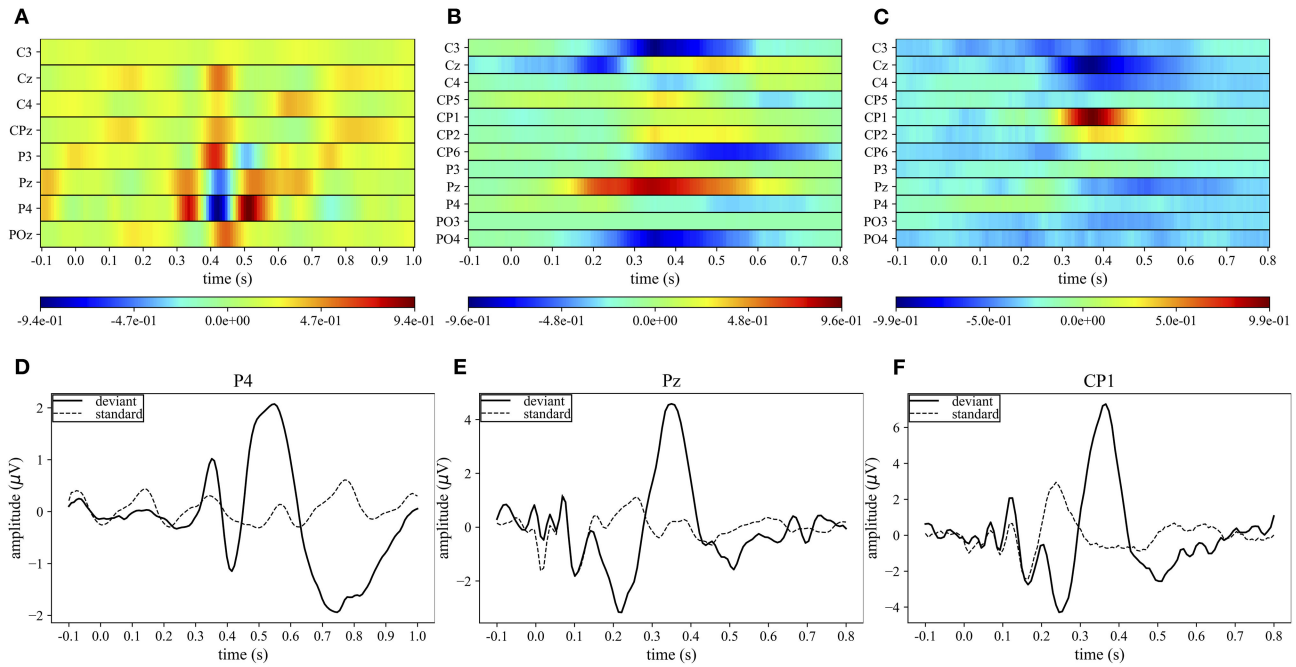

FIGURE 4 | Grand average spatio-temporal representations. The top panels **(A–C)** show the grand average spatio-temporal representation of MS-EEGNet trained with the LOSO strategy using signals from datasets 1–3. Positive gradients are shown in red, while negative gradients are shown in blue. The bottom panels **(D–F)** show the grand average ERP for the deviant (black lines) and standard (dashed black lines) stimuli associated with the most relevant electrode (the one with the largest gradient values) for datasets 1–3.

In the case of dataset 1, P4 appeared as the most important electrode, in particular from 300 to 550 ms. Three main peaks can be identified: two positives at 350 and 510 ms, and a negative at ∼410 ms (**Figure 4A**). These peaks correspond to the peaks in the grand average ERP of the deviant stimulus at approximately the same times (**Figure 4D**). In the cases of datasets 2 and 3, the most important sites were Pz from 300 to 400 ms and CP1 from 350 to 400 ms, respectively. In these cases, a single positive peak occurred in the spatio-temporal maps at about 350 and 390 ms, respectively (**Figures 4B,C**) and was associated with the peak in the grand average ERP of the deviant stimuli at approximately the same time (**Figures 4E,F**).

In the following sections, the interpretation of the relevant input features driving the MS-EEGNet P300 decision is analyzed separately in the temporal and spatial domains.

Absolute Temporal Representations

Figure 5 displays the grand average absolute temporal representations of MS-EEGNet trained with the LOSO strategy using signals from datasets 1–3 (**Figures 5A–C**). These patterns highlight, by means of local and global peaks, the more class-discriminative time samples for the P300 class across all spatial sites. These waveforms confirm the highest importance of time samples approximately between 300 and 550 ms in all the cases, with the peak at about 410, 350, and 390 ms for datasets 1–3, in agreement with the results shown in **Figure 4**. Interestingly, these waveforms synthetically highlight how the network learns different temporal profiles of sample relevance depending on the dataset, e.g., more regular waveforms in the cases of datasets 2 and 3 (but more spiking in the case of dataset 3) and more irregular waveforms in the case of dataset 1 (with several local maxima, two in particular just next to the global one, i.e., at 350 and 510 ms). These differences may be linked to the different sensory modalities involved (visual vs. auditory), different participants (healthy vs. pathological), or different paradigms used to elicit P300 (oddball paradigm vs. flashing the object under fixation).

Absolute Spatial Representations

Besides the investigation of the more P300-discriminative temporal features, it is also interesting to evidence the more P300-discriminative spatial features. To this aim, **Figure 6** shows the grand average absolute spatial representations of MS-EEGNet trained with the LOSO strategy using signals from datasets 1–3 (**Figures 6A–C**), emphasizing the different spatial profiles of sample relevance.

The three more class-discriminative electrode sites across all the time samples were (in increasing order of relevance) Pz, P3, and P4 when the CNN was trained on dataset 1; C3, Cz, and Pz when the CNN was trained on dataset 2; and Cz, CP2, and CP1 when the CNN was trained on dataset 3. Again, these differences can be associated with differences in sensory modality, participants, and paradigms adopted across the three datasets.

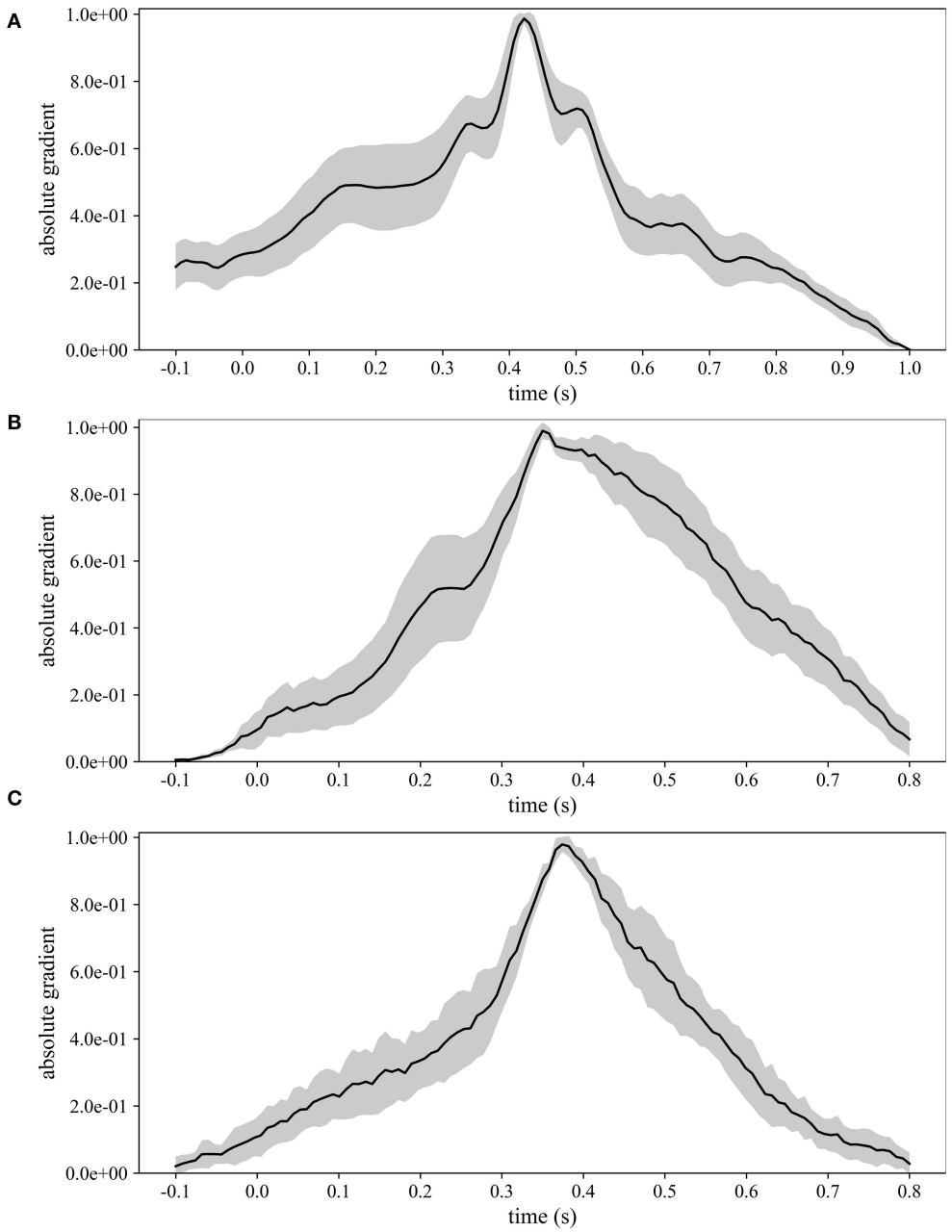

FIGURE 5 | Grand average absolute temporal representations of MS-EEGNet trained with the LOSO strategy using signals from datasets 1–3 **(A–C)**; the mean value (black line) ± standard deviation (gray shaded areas) across participants are represented.

Progressive Changes in Spatio-Temporal Sample Relevance While Increasing Training Examples

Lastly, the absolute temporal and spatial representations were also used to analyze the progressive change in the importance of the spatio-temporal samples while increasing the percentage of training examples included when training MS-EEGNet with the TL-WS and WS strategies. For the TL-WS condition, only CNNs initialized from LOSO models with the largest number of participants were considered. The absolute temporal and spatial representations are reported in **Figure 7**, in case of a representative participant and session belonging to dataset 1.

In particular, **Figures 7A,B** report the absolute temporal and spatial representations as obtained in the LOSO strategy. **Figures 7C–G** show the effects of the TL-WS strategy, as the percentage of training examples from the held back participant increased. While transferring the knowledge from the other participants and sessions, the CNN inherited the importance

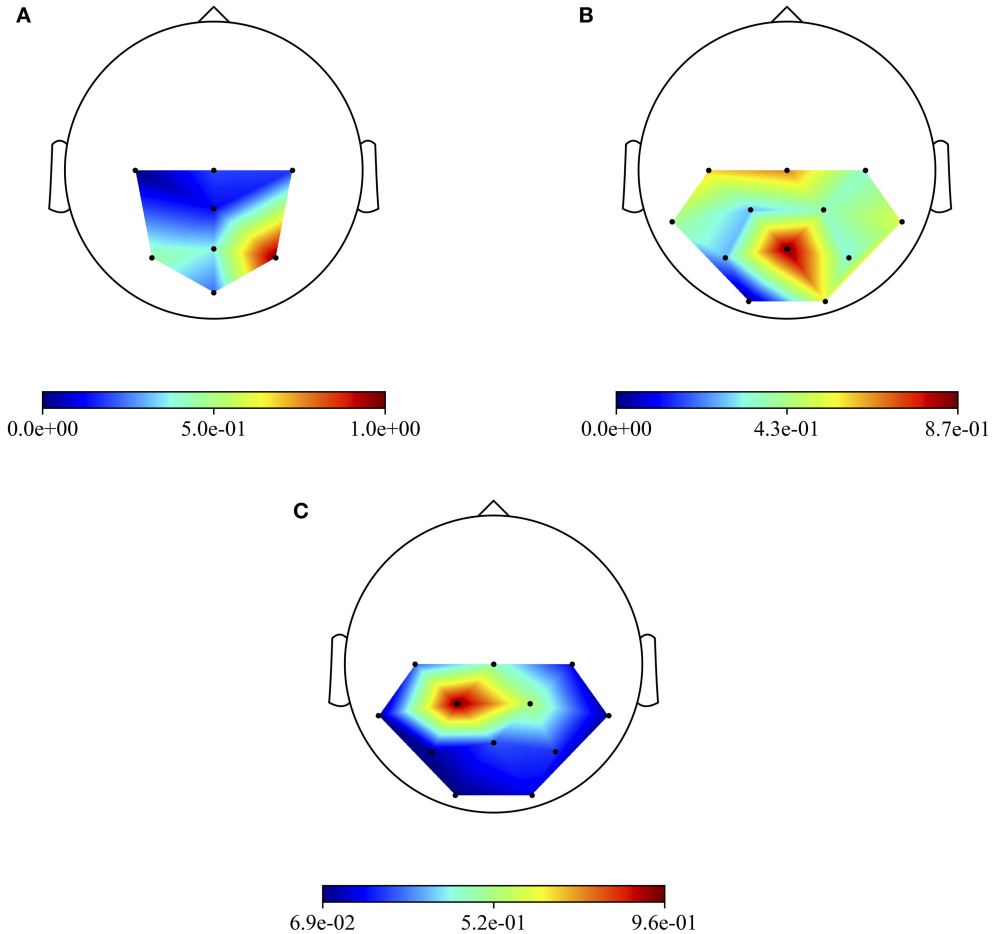

FIGURE 6 | Grand average absolute spatial representations of MS-EEGNet trained with the LOSO strategy using signals from datasets 1–3 **(A–C)**.

profile from the pre-trained condition. Thus, for each percentage of training examples (**Figures 7C–G**), the temporal and spatial profiles did not change substantially their shape from the LOSO condition, since the importance in the temporal and spatial domains was already learned in the LOSO training. Nevertheless, the amplitude increased both in the temporal and spatial domains while increasing the percentage of training examples, indicating progressive accumulation of the importance. Conversely, adopting the WS strategy (**Figures 7H–L**), the CNN was randomly initialized and, therefore, had to learn from scratch the more class-discriminative spatio-temporal samples. Thus, the temporal and spatial profiles changed more with respect to TL-WS as the percentage of training examples increased. In particular, temporal profiles changed from a nearly flat profile (e.g., 15% in **Figure 7H**) to profiles more focused on time samples in the range of 300–550 ms (e.g., 45, 60% in **Figure 7H**) peaking at approximately 410 ms. Furthermore, the spatial profiles changed from a diffused distribution (**Figure 7I**) to distributions more focused on parietal electrodes (in particular P3, Pz, and P4 in **Figures 7J–L**). However, the absolute gradients resulted lower than in the TL-WS condition, in particular in correspondence of the more class-discriminative temporal (i.e., 350, 410, 510 ms) and spatial (P3, Pz, and P4) samples.

DISCUSSION

In this study, a lightweight multi-scale CNN design for EEG decoding named MS-EEGNet was proposed and applied to decode the P300 event from three different datasets. This CNN merges the multi-scale temporal learning proposed by Farahat et al. (2019) with lightweight characteristics originally proposed in EEGNet (Lawhern et al., 2018), operating even a further decrease in the number of trainable parameters while learning multi-scale features. MS-EEGNet was compared with many SOA algorithms, such as CNNs (EEGNet, BranchedNet, OCLNN) and a traditional ML pipeline (xDAWN+RG). To better analyze the multi-scale feature learning as operated by MS-EEGNet, we performed a *post-hoc* analysis on the hyper-parameters. In addition, MS-EEGNet was extensively evaluated under four training conditions, each one reflecting

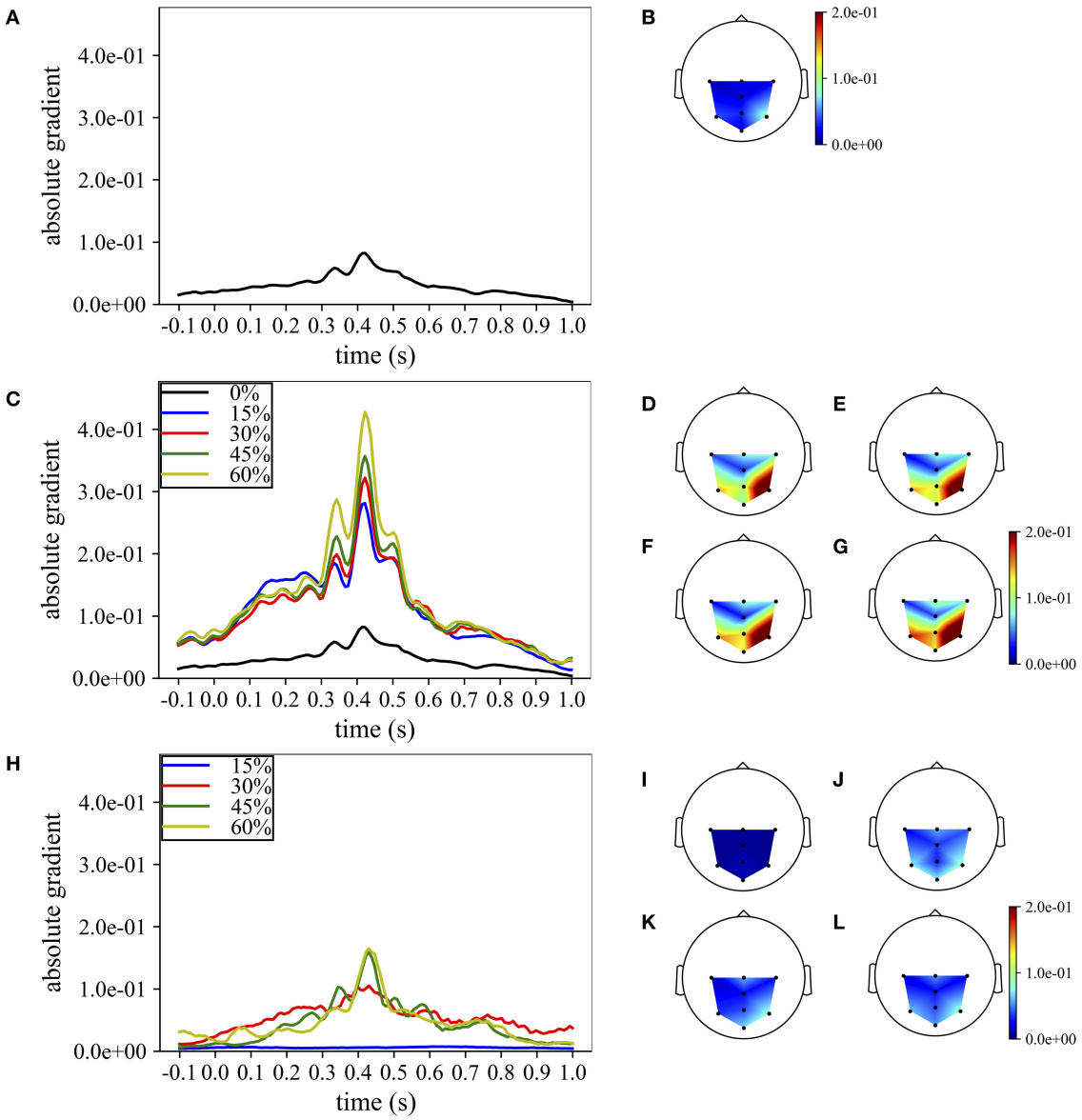

FIGURE 7 | Grand average temporal and spatial absolute representations of MS-EEGNet trained on dataset 1 for a representative participant and session, adopting the LOSO, TL-WS, and WS strategies. In particular, the representations obtained using the LOSO strategy in the temporal and spatial domains are reported in **(A,B)**, respectively. The representations obtained using the TL-WS strategy in the temporal and spatial domains are reported in **(C)** (colored lines) and **(D–G)**, as the percentage of training examples of the new participant increased (15, 30, 45, 60%, from **D–G**). The representations obtained using the WS strategy in the temporal and spatial domains are reported in **(H)** (colored lines) and **(I–L)**, as the percentage of training examples of the participant increased (15, 30, 45, 60%, from **I–L**). Note that in order to maintain the same scale across the strategies in the spatial absolute representations, in **(D–G)**, the maximum gradient value represented (2.0e−1) was below the real maximum gradient value (3.3e−1), saturating the value in particular around P4.

a different practical scenario: (i) using participant-specific signals of single recording sessions (WS); (ii) using participant-specific signals of multiple recording sessions (CS); (iii) using signals from the other participants (LOSO); and (iv) using a fraction of participant-specific signals from a pre-trained cross-participant CNN (TL-WS). Lastly, we exploited the saliency maps to obtain representations aimed to explain the MS-EEGNet decision by visualizing the relevant samples in the input domain. Both the proposed architecture and the performed analyses represent significant expansion compared with the previous study (Borra et al., 2020a), limited to the application of a design based on EEGNet to solve the P300 task proposed by the IFMBE 2019 scientific challenge (corresponding to dataset 1 here). In the following, the performance of MS-EEGNet and the results of the performed analyses are critically discussed.

Performance of MS-EEGNet and Comparison With State-of-the-Art Algorithms

The performance of MS-EEGNet using the WS strategy was above 80% for all the datasets, reaching higher values for datasets 2 and 3 compared with dataset 1 (**Table 3**). This difference could depend on several factors, such as different paradigms, stimuli, and populations (ASD vs. healthy), possibly leading to different P300 responses, e.g., with lower or higher amplitude. Regarding this, **Figures 4D–F** show that the P300 response to the deviant stimulus in dataset 1 was indeed characterized by a lower amplitude, perhaps increasing the difficulty in discriminating between standard/deviant stimuli. Other contributing factors could be the lower proportion between training and test examples, and the lower number of electrodes in dataset 1 vs. datasets 2 and 3. It is worth noticing that this same difference in the WS performance across the datasets was notable in the other algorithms, too. Using the CS strategy, the performance improved compared with the WS strategy for all the algorithms, and this result is in line with Simões et al. (2020). When comparing MS-EEGNet to the other algorithms, the design exhibited the highest performance on each dataset, adopting the WS and CS strategies. Interestingly, among the tested CNNs, OCLNN (which uses a mixed spatio-temporal convolution) and BranchedNet (which performs a spatial convolution first) performed generally lower than MS-EEGNet and EEGNet (which perform temporal convolution first). This is in line with Simões et al. (2020), where the previous design adapted from EEGNet outperformed significantly a CNN design inspired by Manor and Geva (2015) that used a first spatial convolutional layer. Therefore, these results suggest that a CNN design trained on participant-specific signals and based on a first temporal filtering of EEG signals leads to higher P300 decoding performance than other solutions that use first mixed spatio-temporal or first spatial filtering of the input signals. Hence, higher performance could be achieved learning temporal features directly from raw EEG signals (exploiting useful raw temporal information related to the P300 event) instead from signals with a higher level of abstraction. Overall, among the tested SOA CNNs, EEGNet is the one exhibiting the closest performance to MS-EEGNet; and this can be explained by the derivation of MS-EEGNet from EEGNet with the addition of multi-scale temporal feature learning and compressed representation learning. However, the results denote that the changes included in MS-EEGNet can significantly improve the high performance already achieved by EEGNet, especially using session-specific (WS) and participant-specific (CS) input distributions, see datasets 1 ($p = 2e{-}3$ and $p = 3e{-}3$ with the WS and CS strategies, respectively) and 3 ($p = 4e{-}2$) in **Table 3**, using a lower number of trainable parameters.

As expected, adopting the LOSO strategy caused an overall drop of the performance metric across all the tested approaches, with respect to the WS and CS strategies; and the different approaches generally provided similar performance (MS-EEGNet only performed significantly better than xDAWN+RG and OCLNN in dataset 1).

Hence, overall, MS-EEGNet performed better than the other SOA algorithms in the WS and CS strategies and behaved similarly with the other SOA algorithms in the LOSO strategy. This becomes more relevant considering that MS-EEGNet is the lightest CNN among the tested ones, as EEGNet, BranchedNet, and OCLNN introduced more trainable parameters (see **Table 2**). Indeed, this is particularly important, as in practice it is common to deal with small EEG datasets. Thus, keeping limited the number of trainable parameters is crucial when designing CNNs for EEG decoding in order to avoid overfitting. Likely, the lightweight design of MS-EEGNet may explain the absence of higher performance in the LOSO strategy due to the peculiarities of the LOSO training. In this case, class-discriminative features are learned from input distributions with very large variability, involving different participants and possibly different sessions (e.g., with dataset 1). Thus, the CNN, besides needing more training examples, may need more capacity (i.e., more layers/more parameters) to solve the task with higher performance. Considering that the CNN is the lightest among the tested ones (see **Table 2**), obtaining performance similar with that of the other CNNs should not be surprising (and rather can be still considered a satisfactory result). In the LOSO strategy, MS-EEGNet significantly outperformed the traditional ML approach only for dataset 1. This may indicate that in the LOSO strategy MS-EEGNet can learn more relevant cross-participant features, leading to significant higher performance, than an ML pipeline when a larger dataset is used, as in the case of dataset 1. Lastly, besides performance and parameters to fit, considerations about the training time are relevant for practical usage. The multi-scale SOA CNN (BranchedNet) was slower to train with respect to MS-EEGNet, while single-scale SOA CNNs (EEGNet and OCLNN) were faster to train. Overall, compared with SOA CNNs, MS-EEGNet represented a good compromise between performance, model size and computational time.

Performance of MS-EEGNet: *post-hoc* Hyper-Parameter Evaluation

We performed a *post-hoc* hyper-parameter evaluation of eight variant design choices of MS-EEGNet by varying four different hyper-parameters of the multi-scale temporal block (**Figure 2**). Using a single-scale variant ($N_b = 1$) that includes only the large or the short scale, a reduction in trainable parameters and in training time was observed with respect to the baseline MS-EEGNet (see **Table 2**). At the same time, the performance significantly worsened in both cases, indicating the benefit of the multi-scale temporal feature learning with respect to single-scale feature learning for P300 decoding, at the expense of an increased number of trainable parameters and computational time. In addition, the different impact on the performance observed in the design $N_b = 1$ (large) and $N_b = 1$ (short) suggests that the temporal features learned in the large-scale branch were more class-discriminative. Interestingly, using an additional intermediate timescale (three-branched variant $N_b = 3$), a non-significant difference in performance was observed compared with the baseline MS-EEGNet, while more parameters and training time were required (see **Table 2**). These results

about the number of branches of MS-EEGNet suggest that the dual-branched design represented good compromise between performance, model size, and training time.

Furthermore, alternative ratio $r^{MST} = \frac{1}{2}$ between the two timescales obtained with ${F_0}^{MST_0} = (1, 9)$ (corresponding to learning summaries of about 500 and 250 ms), resulted in a small, not significant ($p = 0.06$) increase in performance with respect to the baseline MS-EEGNet ($r^{MST} = \frac{1}{4}$), requiring few more parameters and training time. In addition, variants learning more feature maps (${K_1}^{MST} = 8$ and ${K_1}^{MST} = 16$), with respect to the compressed representation exploited in the baseline MS-EEGNet (${K_1}^{MST} = 2$) not only required more parameters to fit and were slower (see **Table 2**) but worsened the performance significantly. This suggests that learning compressed representations could be beneficial in terms of performance, model size, and training time for P300 decoding. Remarkably, the variant architecture including the most extreme compressed representation (${K_1}^{MST} = 1$), i.e., learning only a feature map for each timescale, scored similar performance as the baseline MS-EEGNet while lightly reducing the model size and requiring the same training time (see **Table 2**), suggesting that future architectures could also exploit this design to further reduce the model size without hampering the performance. Lastly, increasing the depth of the MST block did not provide any significant improvement in performance, introduced more parameters to fit, and required more training time (see **Table 2**). Thus, these last results suggest that a shallower and lightweight MST design, as provided in the baseline MS-EEGNet, is preferable for P300 decoding.

Performance of MS-EEGNet: Transfer Learning Strategy and Variable Number of Training Trials

MS-EEGNet was capable to deal with a reduced number of training trials when trained from scratch (WS), although not at the smallest percentage of training trials (**Figure 3**). The performance increased in TL-WS. Indeed, transferring knowledge using the smallest percentage of training examples of the held back participant (i.e., 15%) resulted in a beneficial effect, compared with WS across all the datasets and regardless of the number of participants from whom the knowledge was transferred (**Figure 3**). This beneficial effect of the TL-WS strategy was also found when using more training examples (30, 45, and 60%) of the held back participant on datasets 1 and 3. As expected, the worst performance was obtained when transferring knowledge from the LOSO models trained on the smallest subset of participants ($M = 2$) for all datasets and percentages. However, this condition produced a significant increase in performance compared with randomly initialized models especially when using a small number of signals belonging to the new user (i.e., 15%). Therefore, pre-trained models do not necessarily need to be optimized on a large set of participants in order to significantly outperform randomly initialized models, especially when using a small amount of data during transfer learning (see also section 3 in **Supplementary Materials** for comparison between TL-WS and WS with 100% of training trials).

Overall, these results suggest that the proposed approach could be used to accurately decode the P300 event even with a reduced number of standard/deviant stimuli presented to the user during the calibration stage.

Explaining P300 Decision

The proposed approach achieved high performance, outperforming the SOA algorithms. As stated by Montavon et al. (2018), in practice it is also crucial to verify that the decoding performance results from a proper problem representation and not from the exploitation of artifacts in the input data. Therefore, in this study, we explained the MS-EEGNet decision for P300 decoding *via* the saliency maps, providing GA spatio-temporal, GA absolute temporal, and GA absolute spatial representations of the relevance of the input samples.

The GA spatio-temporal representations of MS-EEGNet (**Figures 4A–C**) evidenced higher values (both positive and negative) of the gradients, corresponding to more class-discriminative input samples, within time intervals (roughly between 300 and 550 ms) matching the P300 temporal occurrence for all the datasets. The positive/negative peaks in these gradient patterns corresponded to peaks in the GA ERPs of the deviant stimulus (**Figures 4D–F**). Indicating with i and j are the row and column indices, respectively; and the positive and negative gradients in the (i, j) location shown in **Figures 4A–C** represent the direction in which change in the (i,j) input feature increased the P300 class score and, consequently, the CNN decision toward the P300 class. Thus, for example, analyzing the gradients related to P4 obtained from dataset 1 (**Figure 4A**), two positive peaks and a negative peak were found. As the P4 input signal of a deviant trial increased its value at the two positive peaks (at about 350 and 510 ms), the deviant condition differed more than the standard condition, resulting in the deviant class being easier to distinguish and providing a higher score to it. Therefore, these peaks in the deviant GA ERP were associated with positive gradient peaks. Conversely, as the P4 input signal of a deviant trial reduced its value at the local minimum (at ∼410 ms), the negative peak resulted more distant from the standard condition, leading to a higher score for the deviant class (negative gradient peak). This consideration can be extended to datasets 2 and 3, by analyzing the Pz and CP1 electrodes. Therefore, as already obtained in Farahat et al. (2019), higher differences in the ERP between deviant and standard stimuli are reflected onto the saliency maps by means of positive and negative gradients.

When computing the absolute value of the saliency maps, the absolute gradient at the spatio-temporal sample (i,j) reflects how much a change in this sample affects the P300 class score. We analyzed the absolute saliency maps separately in the time and spatial domains (**Figures 5**, **6**) in order to evidence the more discriminative temporal samples and electrodes independently on the direction (positive or negative) they contributed to the decoding result. The GA temporal absolute profile for each dataset peaked approximately in correspondence with the peak of the P300 response. Interestingly, the absolute

temporal representations exhibit different patterns for the three datasets, evidencing that they are able to detect differences embedded in the P300 response across the three datasets. Lastly, the GA absolute spatial distributions represented in a topological map allowed a direct analysis of the more P300 discriminative electrodes of MS-EEGNet. These were mainly distributed in the parietal and centro-parietal areas. This may provide practical hints to reduce the number of electrodes in the design of P300-BCIs. Overall, the various gradient-based representations (**Figures 4–6**) matched the P300 spatio-temporal distribution, confirming that MS-EEGNet was able to capture meaningful task-related features, without exploiting artifactual/noisy input sources.

Interestingly, using a representative example, we show that while transferring knowledge the importance of temporal and spatial samples gradually increased from the LOSO condition (**Figures 7A,B**) as the percentage of training examples increased. In particular, it appears that the more task-relevant temporal and spatial samples were already learned in the LOSO strategy. However, during transfer learning (**Figures 7C–G**), the LOSO temporal and spatial profiles (template profiles) were modeled on the new participant- and session-specific training distribution, giving progressively more importance to particular temporal intervals/electrode sites starting from the template profiles. The availability of these template profiles allowed rapid learning of the relevant participant-specific and session-specific input samples (i.e., needing a low number of training examples of the new participant). Conversely, when training CNNs from scratch with the WS strategy, the profile distribution rapidly changed its shape both in the temporal (**Figure 7H**) and spatial (**Figures 7I–L**) domains but reached lower importance values compared with the TL-WS strategy. When transferring knowledge, the profile was more focused on interval 300–550 ms with three distinct main peaks and on sites P4 > P3 > Pz already at the lowest percentage (15%, **Figures 7C,D**); while at the same percentage, the WS strategy was characterized by more flat and homogeneous distributions (**Figures 7H,I**). These considerations could explain the performance improvement obtained in the TL-WS strategy (**Figure 3**): the parameters learned using the LOSO strategy overall represented a better initialization point in the parameter space compared with a random one.

CONCLUDING REMARKS

In conclusion, we wish to stress that this study aims to contribute to uncovering the enormous potentialities of deep learning via CNNs for EEG decoding and to their exploitation in practice adopting different training strategies, reflecting different scenarios. The multi-scale design was the most lightweight and at the same time outperformed many SOA algorithms when using three different P300 datasets, indicating that care has to be taken to design CNNs for EEG decoding, keeping limited the parameters to fit, especially when handling small datasets (not as large as the ones adopted in the computer vision field, e.g., > 100 K of examples). In addition, the hyper-parameter *post-hoc* analysis confirmed that the innovative aspects of the architecture, i.e., the design of a lightweight multi-scale temporal block implemented *via* separable convolutions and the use of compressed representation learning were beneficial. Crucially, the capability of MS-EEGNet to transfer knowledge with high performance even with a small number of training examples could be highly useful in practice to reduce the calibration time of P300-based BCIs on a new user.

Saliency maps confirmed their utility to explain the neural network decision in P300 decoding tasks; the derived spatial and temporal representations resulted to match the P300 spatio-temporal distribution. However, the utility of these representations is not limited to provide an additional validation of the algorithm. Indeed, the CNN ability to learn automatically the most meaningful features to perform classification gives the possibility to use these algorithms as data-driven EEG analysis tools. Then, the use of the saliency maps (or similar representations) allows the interpretation of the CNN decision, and it is possible to take advantage of these interpretations for increasing the comprehension of brain dynamics underlying decoded events (e.g., P300 response). For example, representations derived from saliency maps (in the time and/or spatial domain) could be used to study the variability between participants (i.e., which features of the input samples are more/less consistent across participants) and within-participant (i.e., by comparing representations associated with early and late trials, e.g., to investigate the effects of training or treatment). Furthermore, the analysis of between-participants and within-participant variabilities could be useful, in perspective, to develop biomarkers to diagnose and monitor neurological or psychiatric disorders (Farahat et al., 2019), e.g., P300 amplitude, latency, and topographical alterations in mild cognitive impairment (Medvidovic et al., 2013), dementia (Vecchio and Määttä, 2011), and schizophrenia (Jeon and Polich, 2003). In addition, identifying the more class-discriminative temporal and spatial input features can also have a relevant practical impact on the design of BCIs. For example, the identification of a small subset of more relevant electrodes (as we found here) may drive the definition of BCI systems with a very small electrode montage, increasing the comfort of a participant and reducing preparation time. It is worth noticing that by performing this analysis on within-participant CNNs, the optimal electrode montage could also be identified on an individual basis.

Overall, this study, by specifically addressing the aspects of lightweight design, transfer learning, and interpretability of the proposed CNN, can contribute to advance the development of deep learning-based decoders for P300-BCIs. Future developments include the application of the proposed architecture to other ERP decoding tasks, and the adoption of interpretable and more lightweight layers, such as the sinc-convolutional layer, to perform band-pass filtering (Ravanelli and Bengio, 2018; Borra et al., 2020b,c). In addition, automatic hyper-parameter search (Snoek et al., 2012) will be exploited to further improve the MS-EEGNet design and other explanation techniques, such as layer-wise relevance propagation, will be investigated, carefully analyzing the effect of different propagation rules and parameters for EEG decoding.

DATA AVAILABILITY STATEMENT

The dataset 1 used in this study can be found online at https://www.kaggle.com/disbeat/bciaut-p300. The datasets 2 and 3 used in this study are available under request to the corresponding author DB. Codes are available at https://github.com/ddavidebb/P300_decoding_MS-EEGNet.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Bioethics Committee, University of Bologna (datasets 2 and 3). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DB and EM conceived and designed the methodology and wrote the original draft. EM designed the recording protocol for datasets 2 and 3 and contributed to acquiring signals. DB processed data. DB, SF, and EM critically analyzed the data and reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was part of the Department of Excellence (L. 232 of 01/12/2016) Project of the Department of Electrical, Electronic and Information Engineering, University of Bologna, funded by the Italian Ministry of Education, Universities, and Research (MIUR).

ACKNOWLEDGMENTS

The authors would like to thank Lorenzo Giunchi and Federico Babini who contributed to recording datasets 2 and 3 used in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2021.655840/full#supplementary-material

REFERENCES

Amaral, C., Mouga, S., Simões, M., Pereira, H. C., Bernardino, I., Quental, H., et al. (2018). A feasibility clinical trial to improve social attention in Autistic Spectrum Disorder (ASD) using a brain computer interface. *Front. Neurosci.* 12:477. doi: 10.3389/fnins.2018.00477

Amaral, C. P., Simões, M. A., Mouga, S., Andrade, J., and Castelo-Branco, M. (2017). A novel Brain Computer Interface for classification of social joint attention in autism and comparison of 3 experimental setups: a feasibility study. *J. Neurosci. Methods* 290, 105–115. doi: 10.1016/j.jneumeth.2017.07.029

Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. (2012). Multiclass brain–computer interface classification by riemannian geometry. *IEEE Trans. Biomed. Eng.* 59, 920–928. doi: 10.1109/TBME.2011.2172210

Barachant, A., and Congedo, M. (2014). A plug&play P300 BCI using information geometry. *arXiv [Preprint]*. arXiv:1409.0107.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Blankertz, B., Muller, K.-R., Curio, G., Vaughan, T. M., Schalk, G., Wolpaw, J. R., et al. (2004). The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials. *IEEE Trans. Biomed. Eng.* 51, 1044–1051. doi: 10.1109/TBME.2004.826692

Blankertz, B., Muller, K. R., Krusienski, D. J., Schalk, G., Wolpaw, J. R., Schlogl, A., et al. (2006). The BCI competition III: validating alternative approaches to actual BCI problems. *IEEE Trans. Neural Syst. Rehabil. Eng.* 14, 153–159. doi: 10.1109/TNSRE.2006.875642

Borra, D., Fantozzi, S., and Magosso, E. (2020a). "Convolutional neural network for a P300 brain-computer interface to improve social attention in autistic spectrum disorder," in *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019*, eds J. Henriques, N. Neves, and P. de Carvalho (Cham: Springer International Publishing), 1837–1843.

Borra, D., Fantozzi, S., and Magosso, E. (2020b). "EEG motor execution decoding via interpretable sinc-convolutional neural networks," in *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019*, eds J. Henriques, N. Neves, and P. de Carvalho (Cham: Springer International Publishing), 1113–1122.

Borra, D., Fantozzi, S., and Magosso, E. (2020c). Interpretable and lightweight convolutional neural network for EEG decoding: application to movement execution and imagination. *Neural Netw. Soc.* 129, 55–74. doi: 10.1016/j.neunet.2020.05.032

Cecotti, H., and Graser, A. (2011). Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 433–445. doi: 10.1109/TPAMI.2010.125

Chollet, F. (2016). "Xception: deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI), 1800–1807.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv [Preprint]*. arXiv:1511.07289.

Craik, A., He, Y., and Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. *J. Neural Eng.* 16:031001. doi: 10.1088/1741-2552/ab0ab5

Farahat, A., Reichert, C., Sweeney-Reed, C., and Hinrichs, H. (2019). Convolutional neural networks for decoding of covert attention focus and saliency maps for EEG feature visualization. *J. Neural Eng.* 16:066010 doi: 10.1088/1741-2552/ab3bb4

Farwell, L. A., and Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalogr. Clin. Neurophysiol.* 70, 510–523. doi: 10.1016/0013-4694(88)90149-6

Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., and Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: a review. *Comput. Methods Prog. Biomed.* 161, 1–13. doi: 10.1016/j.cmpb.2018.04.005

Glorot, X., and Bengio, Y. (2010). "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Sardinia), 249–256.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S. (2016). Deep learning for visual understanding: a review. *Neurocomputing* 187, 27–48. doi: 10.1016/j.neucom.2015.09.116

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning Proceedings of Machine Learning Research*, eds F. Bach and D. Blei (Lille: PMLR), 448–456.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2019). Deep learning for time series classification: a review.

Data Min. Knowl. Disc. 33, 917–963. doi: 10.1007/s10618-019-00619-1

Jeon, Y.-W., and Polich, J. (2003). Meta-analysis of P300 and schizophrenia: patients, paradigms, and practical implications. *Psychophysiology* 40, 684–701. doi: 10.1111/1469-8986.00070

Justen, C., and Herbert, C. (2018). The spatio-temporal dynamics of deviance and target detection in the passive and active auditory oddball paradigm: a sLORETA study. *BMC Neurosci.* 19:25. doi: 10.1186/s12868-018-0422-3

Kingma, D. P., and Ba, J. (2014). Adam: a *method for stochastic* optimization. *arXiv [Preprint]*. arXiv:1412.6980.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural Eng.* 15:056013. doi: 10.1088/1741-2552/aace8c

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521:436. doi: 10.1038/nature14539

Lindsay, G. (2020). Convolutional neural networks as a model of the visual system: past, present, and future. *J. Cogn. Neurosci.* 1–15. doi: 10.1162/jocn_a_01544. [Epub ahead of print].

Liu, M., Wu, W., Gu, Z., Yu, Z., Qi, F., and Li, Y. (2018). Deep learning based on Batch Normalization for P300 signal detection. *Neurocomputing* 275, 288–297. doi: 10.1016/j.neucom.2017.08.039

Manor, R., and Geva, A. B. (2015). Convolutional neural network for multi-category rapid serial visual presentation BCI. *Front. Comput. Neurosci.* 9:146. doi: 10.3389/fncom.2015.00146

Medvidovic, S., Titlic, M., and MarasSimunic, M. (2013). P300 evoked potential in patients with mild cognitive impairment. *Acta Inform. Med.* 21:89. doi: 10.5455/aim.2013.21.89-92

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* 73, 1–15. doi: 10.1016/j.dsp.2017.10.011

Nicolas-Alonso, L. F., and Gomez-Gil, J. (2012). Brain computer interfaces, a review. *Sensors* 12, 1211–1279. doi: 10.3390/s120201211

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). "Automatic differentiation in PyTorch," in *NIPS-W* (Long Beach, CA).

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Ravanelli, M., and Bengio, Y. (2018). "Speaker recognition from raw waveform with SincNet," in *2018 IEEE Spoken Language Technology Workshop (SLT)* (Athens), 1021–1028.

Rezeika, A., Benda, M., Stawicki, P., Gembler, F., Saboor, A., and Volosyak, I. (2018). Brain–computer interface spellers: a review. *Brain Sci.* 8:57. doi: 10.3390/brainsci8040057

Rivet, B., Souloumiac, A., Attina, V., and Gibert, G. (2009). xDAWN algorithm to enhance evoked potentials: application to brain–computer interface. *IEEE Trans. Biomed. Eng.* 56, 2035–2043. doi: 10.1109/TBME.2009.2012869

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* 38, 5391–5420. doi: 10.1002/hbm.23730

Shan, H., Liu, Y., and Stefanov, T. (2018). "A simple convolutional neural network for accurate P300 detection and character spelling in brain computer interface," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence* IJCAI'18 (Stockholm: AAAI Press), 1604–1610.

Simões, M., Borra, D., Santamaría-Vázquez, E., GBT-UPM; Bittencourt-Villalpando, M., Krzemiński, D., et al. (2020). BCIAUT-P300: a multi-session and multi-subject benchmark dataset on autism for P300-based brain-computer-interfaces. *Front. Neurosci.* 14:568104. doi: 10.3389/fnins.2020.568104

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv [Preprint]*. arXiv:1312.6034.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *arXiv [Preprint]*. arXiv:1206.2944.

Solon, A. J., Lawhern, V. J., Touryan, J., McDaniel, J. R., Ries, A. J., and Gordon, S. M. (2019). Decoding P300 variability using convolutional neural networks. *Front. Hum. Neurosci.* 13:201. doi: 10.3389/fnhum.2019.00201

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Supratak, A., Dong, H., Wu, C., and Guo, Y. (2017). DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* 25, 1998–2008. doi: 10.1109/TNSRE.2017.2721116

Sutton, S., Braren, M., Zubin, J., and John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science* 150, 1187–1188. doi: 10.1126/science.150.3700.1187

Szegedy, C., Wei, L., Yangqing, J., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA: IEEE), 1–9.

Vahid, A., Mückschel, M., Stober, S., Stock, A.-K., and Beste, C. (2020). Applying deep learning to single-trial EEG data provides evidence for complementary theories on action control. *Commun. Biol.* 3:112. doi: 10.1038/s42003-020-0846-z

Vecchio, F., and Määttä, S. (2011). The use of auditory event-related potentials in Alzheimer's disease diagnosis. *Int. J. Alzheimer's Dis.* 2011, 1–7. doi: 10.4061/2011/653173

Zhao, D., Tang, F., Si, B., and Feng, X. (2019). Learning joint space–time–frequency features for EEG decoding on small labeled data. *Neural Netw.* 114, 67–77. doi: 10.1016/j.neunet.2019.02.009

Conflict of Interest: The authors declare that this study received materials from NVIDIA Corporation with the donation of the TITAN V used for this research. The provider was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

frontiers in Human Neuroscience

SYSTEMATIC REVIEW
published: 25 November 2021
doi: 10.3389/fnhum.2021.772837

Artificial Intelligence Algorithms in Visual Evoked Potential-Based Brain-Computer Interfaces for Motor Rehabilitation Applications: Systematic Review and Future Directions

Josefina Gutierrez-Martinez[1], Jorge A. Mercado-Gutierrez[1], Blanca E. Carvajal-Gámez[2], Jorge L. Rosas-Trigueros[2] and Adrian E. Contreras-Martinez[2]*

[1] División de Investigación en Ingeniería Médica, Instituto Nacional de Rehabilitación Luis Guillermo Ibarra Ibarra, Mexico City, Mexico, [2] Escuela Superior de Cómputo, Instituto Politécnico Nacional, Mexico City, Mexico

OPEN ACCESS

Edited by:
Hohyun Cho, Washington University School of Medicine in St. Louis, United States

Reviewed by:
Caterina Cinel, University of Essex, United Kingdom
Andrej Savic, University of Belgrade, Serbia

***Correspondence:**
Josefina Gutierrez-Martinez
josefina_gutierrez@hotmail.com

Specialty section:
This article was submitted to Brain-Computer Interfaces, a section of the journal Frontiers in Human Neuroscience

Received: *08 September 2021*
Accepted: *04 November 2021*
Published: *25 November 2021*

Citation:
Gutierrez-Martinez J, Mercado-Gutierrez JA, Carvajal-Gámez BE, Rosas-Trigueros JL and Contreras-Martinez AE (2021) Artificial Intelligence Algorithms in Visual Evoked Potential-Based Brain-Computer Interfaces for Motor Rehabilitation Applications: Systematic Review and Future Directions. Front. Hum. Neurosci. 15:772837. doi: 10.3389/fnhum.2021.772837

Brain-Computer Interface (BCI) is a technology that uses electroencephalographic (EEG) signals to control external devices, such as Functional Electrical Stimulation (FES). Visual BCI paradigms based on P300 and Steady State Visually Evoked potentials (SSVEP) have shown high potential for clinical purposes. Numerous studies have been published on P300- and SSVEP-based non-invasive BCIs, but many of them present two shortcomings: (1) they are not aimed for motor rehabilitation applications, and (2) they do not report in detail the artificial intelligence (AI) methods used for classification, or their performance metrics. To address this gap, in this paper the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology was applied to prepare a systematic literature review (SLR). Papers older than 10 years, repeated or not related to a motor rehabilitation application, were excluded. Of all the studies, 51.02% referred to theoretical analysis of classification algorithms. Of the remaining, 28.48% were for spelling, 12.73% for diverse applications (control of wheelchair or home appliances), and only 7.77% were focused on motor rehabilitation. After the inclusion and exclusion criteria were applied and quality screening was performed, 34 articles were selected. Of them, 26.47% used the P300 and 55.8% the SSVEP signal. Five applications categories were established: Rehabilitation Systems (17.64%), Virtual Reality environments (23.52%), FES (17.64%), Orthosis (29.41%), and Prosthesis (11.76%). Of all the works, only four performed tests with patients. The most reported machine learning (ML) algorithms used for classification were linear discriminant analysis (LDA) (48.64%) and support vector machine (16.21%), while only one study used a deep learning algorithm: a Convolutional Neural Network (CNN). The reported accuracy ranged from 38.02 to 100%, and the Information Transfer Rate from 1.55 to 49.25 bits per minute. While LDA is still the most used AI algorithm, CNN has shown promising results, but due to their high technical implementation requirements, many researchers

do not justify its implementation as worthwile. To achieve quick and accurate online BCIs for motor rehabilitation applications, future works on SSVEP-, P300-based and hybrid BCIs should focus on optimizing the visual stimulation module and the training stage of ML and DL algorithms.

Keywords: BCI, visual stimulation, classification, performance metrics, steady state visually evoked potentials, P300, functional electrical stimulation, virtual reality

INTRODUCTION

One of the most traditional neurorehabilitation strategies aimed at restoring motor functions lost due to various lesions of the nervous system [stroke, spinal cord injury (SCI), cerebral palsy, among others] is based on the neurofacilitation approach for proprioceptive stimulation and guidance of brain plasticity processes (Carr and Shephered, 2006; Hindle et al., 2012). These techniques involve passive stretching, contraction and relaxation of specific muscles groups in order to improve their flexibility and to stimulate the sensory function, muscle tone and recovery of movement patterns. Some key elements for motor and sensory functional recovery (Jang, 2013) are repetition of movement patterns (Zbogar et al., 2017), somatosensory stimulation (Hara, 2008) and the application of stimuli outside the motor and sensory pathways (visual, auditory, or proprioceptive) (Bach-y-Rita and Kercel, 2003; Bento et al., 2012; Takeuchi and Izumi, 2012; Galińska, 2015). These neurorehabilitation strategies make possible to re-educate neural tissue that is not completely damaged or to reactivate other areas to form new synaptic connections (Gordon, 2005).

To this end, various technologies (devices and strategies) have been developed to offer therapies that help patients to recover impaired motor functions. Brain-Computer Interface (BCI), Functional Electrical Stimulation (FES), and Neuroprostheses are devices proposed to improve motor and neurological functions (Iosa et al., 2012). The theoretical argument is that therapeutic interventions based on these neurorehabilitation technologies take advantage of the preserved neuro-muscular structures and functions, and that they can help to compensate or re-learn the functions previously performed by the damaged areas, thus improving the sensory-motor function (Iosa et al., 2012; Altaf, 2019).

Principles of Brain-Computer Interfaces

The main objective of BCIs is to decipher the user's intentions, registered from electrical, magnetic, thermal or chemical signals generated by the brain, and translate them into orders that are interpreted and translated by a computer into commands, in order to establish direct communication between the brain and external devices. These systems allow the user to interact with their environment, without using the peripheral nervous system or the muscular system, and when used in combination with proper motor or sensory stimuli and functional tasks, they can be used to assist, increase or help repair cognitive or sensory-motor functions. BCIs can be classified as invasive and non-invasive, according to the sensors that they use to collect brain signals, and as endogenous and exogenous, depending on if their experimental strategy requires external stimuli or not. Each type of BCI has advantages and disadvantages regarding its temporal and spatial resolution, computational cost, training requirements, and clinical application (Wolpaw et al., 2000; Birbaumer and Cohen, 2007).

Invasive BCIs have a high signal-to-noise ratio (SNR) that allows accurate pattern recognition or continuous decoding of kinematic parameters. However, this BCI approach face the risk of surgical complications and infections, short-term and long-term signal instabilities that degrade neural decoding of intent (Perge et al., 2013), and the challenge of maintaining stable chronic recordings (Meng et al., 2016). Due to their ease, non-invasive nature, high temporal resolution, portability and low cost, most BCIs use the surface electroencephalography (EEG) as the preferred method to obtain BCI control signals (Radaman and Vasilakos, 2017). To implement EEG-based BCI systems several protocols and paradigms (e.g., imagery or visual tasks) have been used to modulate the subject's brain electrical activity (Abiri et al., 2019; Bonci et al., 2021).

Currently, several research centers are focused on studying the advantages of endogenous EEG based-BCIs to decode movement intention. To this end they use paradigms such as motor imagery to modulate sensorimotor rhythms of the EEG, which are recorded in the scalp over the sensorimotor brain area (Ramos et al., 2013; Thomas et al., 2013; Müller-Putz, 2018; Aggarwal and Chugh, 2019; Baniqued et al., 2021). Despite the advantages of endogenous BCIs based on motor related tasks (Aggarwal and Chugh, 2019), they generally need of a long training period to achieve voluntary control of the sensorimotor brain signals. Moreover, they present moderate performance for multiclass decoding (Boernama et al., 2021) and limited information transfer rate (ITR) (Choi et al., 2020). These shortcomings, combined with a relatively high inter-individual variability can limit the use of those systems outside of a controlled laboratory environment. Unlike endogenous BCIs, exogenous BCIs operate with brain signals known as event related potentials (ERPs) or steady state evoked potentials, which can be spawned by auditory, visual o somatosensorial stimuli (Wang et al., 2008). In the category of exogenous BCI paradigms the most widely used are those based on visually evoked potentials (VEPs)O VEPs are generated in response to visual stimuli, such as flashing lights presented to the subject quickly and repeatedly. These potentials can be controlled and characterized with relative ease, and their properties depend closely on the type and features of the visual stimulus (Kubler et al., 2001).

Brain-Computer Interfaces Based on Visual Paradigms

If a visual stimulus is presented repeatedly at a fixed frequency in the 1–100 Hz range, a very stable response over time (in amplitude and phase) is elicited in the occipital area (Müller-Putz et al., 2005; Won et al., 2015). Those responses are called steady state visually evoked potentials (SSVEP) (Vialatte et al., 2010; Norcia et al., 2015). Recently, SSVEP-based BCIs have received increased attention because they can provide relatively high bit rates of up to 325 bits/min, while requiring little training (Vialatte et al., 2010; Gao et al., 2014; Nakanishi et al., 2018). In addition, SSVEPs are highly robust to artifacts produced by blinks and eye movements (Perlstein et al., 2003) and to electromyographic noise contamination.

On the other hand, exogenous ERPs can also be elicited when infrequent visual stimuli are interspersed with other more frequent or routine stimuli. In this case a positive peak called P300, is evoked at about 300 ms after the stimulus (Blankertz et al., 2011; Yeom et al., 2014), which can be recorded mainly at parietal and occipital zones over the scalp. P300 ERPs are typically elicited during an oddball target detection task, when a target or relevant stimulus is presented infrequently in a background of frequent standard stimuli. Its latency reflects processing speed or efficiency during stimulus evaluation, independent of the motor preparation time (Kutas et al., 1977). Many BCI applications based on the P300 ERP use graphical interfaces operating under the row/column paradigm, that evoke the P300 potential when the elements attended by the user are visually intensified (the target stimuli) (Philip and George, 2020). This paradigm requires the subject to focus his/her attention only in the target stimulus and not in any other stimuli (Polich, 2007; Guo et al., 2019; Riggins and Scott, 2019), which implies the ability to inhibiting attention drifts to irrelevant stimuli.

P300-based and SSVEP-based BCIs have been widely studied since they are considered robust systems with high ITR (Cheng et al., 2002; Rupp, 2014; Naeem et al., 2020) and good accuracy. In both cases the selected parameters of the stimulation pattern led to a trade-off between ITR and accuracy (Cecotti, 2011). Moreover, both BCI approaches have a high potential for clinical use, since they require few subject's EEG data for training classification models. This makes them feasible for practical applications with short-term training (Polich, 2007; Yao et al., 2012), few recording channels and therefore lower computational cost than other BCI modalities (Müller-Putz et al., 2005; Kluge and Hartmann, 2007; McCane et al., 2015; Kundu and Ari, 2017; Nagel et al., 2017; Han et al., 2018). In this regard, it has been shown previously that technologies based on these two BCI modalities, can be transferred to be used not only in the clinical environment, but even at the patient's home (Sellers et al., 2010).

Artificial Intelligence Algorithms in Brain-Computer Interfaces

Traditional machine learning (ML) methods have been widely used in BCI applications, such as Artificial Neural Networks, Support Vector Machine (SVM) or Linear Discriminant Analysis (LDA). This classic ML approach require the use of namely manually designed techniques for EEG feature extraction (e.g., temporal, spectral and time-frequency methods, to name a few). The feature extraction plus ML technique approach presents the following problems: (1) it can only learn the features that researchers focus on, but ignores other potentially informative ones (Lecun et al., 2015); (2) methods performing well on certain subjects (with similar age or occupation) may not give a satisfactory performance on others, yielding a high subject-to-subject variability in EEG signals. For these reasons, different deep neural networks (DNN) have been proposed to overcome the challenges of ML techniques in BCI, allowing automatic feature extraction and classification, while achieving competitive performance on the target tasks. Hence, DNN have become an useful method to improve classification performance of BCI systems using EEG signals (Craik and Contreras, 2019) and evoked potentials (Kwak et al., 2017), with reduced computational cost and improved usability.

Visual Brain-Computer Interface for Motor Related Applications

Currently, there is a growing interest in the application of VEP- and VERP-based BCI systems for people with disabilities. Systematic reviews have shown the potential of VEP-BCIs for motor rehabilitation purposes (Kaufmann et al., 2013; Lazarou et al., 2018). These systems allow the control of orthoses, prostheses, or FES devices to assist disabled patients during therapy (Stan et al., 2015; Zhao et al., 2016). The most common application of these BCI systems is for spellers (at least 30% of papers), but for the device control there are wheelchairs (Zhang et al., 2014, 2016; Turnip et al., 2015; Lopes et al., 2016; Waytowich and Krusienski, 2017; Yu et al., 2017; Chen et al., 2020), robots (Zhao et al., 2015; Çiğ et al., 2017; Venuto et al., 2017; Erkan and Akbaba, 2018; Yuan et al., 2018; Khadijah et al., 2019; Wang et al., 2020), and domotics tools (Venuto and Mezzina, 2018; Hossain et al., 2020; Lee T. et al., 2020).

Although several papers have been published on BCI applications based on visual paradigms, many of them do not report the performance of the Artificial Intelligence (AI) algorithms used for detection and classification of evoked potentials (P300 or SSVEP). Likewise, although numerous BCI papers are focused on studying and analyzing the performance of the classification algorithms, most of them do not report online tests with a specific application, either for communication, or for the control of motor assistive or rehabilitation technologies.

Traditionally, manually designed feature extraction techniques and machine learning algorithms have been used to detect and classify P300 and SSVEP signals within BCI systems (Bashashati et al., 2007; Lin et al., 2007). Common examples of feature extraction algorithms are spectral parameters, time-frequency representations, parametric models, cross-correlation and canonical correlation analysis (CCA), and matched filtering. Regarding ML classifiers used to detect EEG states or activity in BCI systems, examples are support vector machine (SVM), Linear Discriminant Analysis (LDA), fuzzy logic algorithms, and artificial neural networks, Unfortunately, these classification techniques can only learn from the features the designer focuses

on, missing out on others that might be useful to improve their performance. Therefore, in recent years, deep learning techniques such as convolutional neural networks (CNN), recurrent neural networks (RNN), or deep belief networks (DBN) have been used in BCIs to overcome the aforementioned shortcomings of traditional ML methods (Cecotti, 2011; Cecotti and Graser, 2011; Manor and Geva, 2015; Liu et al., 2018; Shan et al., 2018).

The performance of the AI algorithms used in BCI-based spelling applications (Huang and Huang, 2017) has been evaluated through metrics such as accuracy, precision and ITR. On the one hand, BCI spellers based on SSVEP signals have reported ITR values as high as 4.5 bpm (91.04%) ITR (accuracy) (Chen et al., 2015), 325 bpm (89.83%) (Nakanishi et al., 2018) or 701 bpm (74.9%) (Nagel and Spüler, 2019). On the other hand, BCI spellers based on P300 signals have reported ITR values of 20.259 bpm (79%) (Lin et al., 2018). For hybrid spelling systems that integrate P300 and SSVEP, authors have reported an online classification accuracy of up to 93.85%, with ITR of 56.44 bpm (Yin et al., 2013). Despite the extensive number of published studies on P300-based and SSVEP-based BCI systems, only a few are focused on the rehabilitation or assistance of movements. Moreover, they generally do not report the same performance metrics used in spelling systems. Such is the case of Kaplan et al. (2016), who developed a P300-based BCI system to control phantom fingers using visual stimuli placed over them, as an "ideomotor training simulator." On the other hand, Giménez et al. (2011) presented the electronic design of a functional electrical stimulation (FES) system and its interface with a BCI based on P300. However, these works focus on the integration of the BCI commands with the actuator, but there is a lack of information about the feature extraction methods, the AI-based classifiers, and the performance metrics they used.

Objectives and Structure of the Paper

To address this gap, in this paper we applied the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology for a systematic literature review (SLR). The main aim of this review is to gather all relevant published works that cover the current state-of-the-art in P300 and SSVEP-based BCI systems, with an emphasis on those used for motor rehabilitation applications and the AI algorithms used for detection and classification by analyzing a large number of recent publications. It provides a general overview of the topic of interest, from traditional ML techniques to cutting-edge DL trends and underlines future challenges in the field.

The review is organized as follows: Section "Introduction" introduces key concepts and critical issues in SSVEP-based and P300-based BCI systems, and details the objectives of the review; section "Materials and Methods" describes how the systematic review was conducted, and how the studies were selected, assessed and analyzed; section "Results" focuses on presenting the papers that reported the most important performance and efficiency (accuracy and ITR) metrics of the selected studies, and describes current trends and promising approaches in this type of BCI systems. Finally, section "Discussion" discusses challenges in VEP-based BCI systems for motor rehabilitation and provides recommendations for future research.

MATERIALS AND METHODS

The SLR is based on the PRISMA methodology. To ensure data quality, we searched in the scientific databases PubMed/MEDLINE, IEEE Xplore, ScienceDirect, Scopus, Embase, and Google Scholar. The search was performed in article titles, abstracts, and keywords of works published in English language. There was no lower limit for the publication date, but the databases were searched up to June 2021. Additional records were identified through other literature sources and patent search engines like Google Patents, WIPO, and SIGA.

Search Strategy and Selection Criteria

This SLR covers the current state-of-the-art in BCI systems based on P300 or SSVEP signals, and hybrid modalities, used in motor rehabilitation applications. In particular, the SLR is focused on the AI algorithms used for classification and the reported performance metrics in the context of the BCI applications. Three reviewers from our team carried out the search of papers to reduce the risk of selection errors and selection bias.

The three steps involved in the manual literature search process are summarized in the PRISMA flow diagram (Page et al., 2021) in **Figure 1**. In the first step (Step 1- Identification) the title of articles reporting AI algorithms for SSVEP-, P300-based BCIs, as well as hybrid SSVEP/P300 BCI systems, were identified from electronic databases. Then, data extraction from abstracts and keywords was performed, and duplicate records, unrelated studies and articles published before 2011 were removed. The second step was a more detailed review of the full text articles (according to the inclusion and exclusion criteria), to assess the eligibility of the selected papers (Step 2-Screening). If the abstract did not indicate clearly whether the inclusion and exclusion criteria were met, the full text paper was also read. Papers not involving a motor rehabilitation application were removed. In the last step (Step 3-Included), the studies considered relevant and of recent advances were selected for further analysis in this SLR. The last filtering was applied to papers after reading the full text, taking into consideration whether they did not report any performance metric or did not involve a P300- or SSVEP-based BCI strategy.

Research Questions

The goals of the SLR were translated into a set of research questions (RQ), to better explain and summarize the evidence about the AI algorithms used in P300- and SSVEP-based BCIs. In this context, the following research questions (RQs) were proposed.

RQ1: What type of evoked potential (P300 or SSVEP) is involved in the BCI's visual paradigm?

RQ2: Is the purpose of the BCI system aimed to some motor rehabilitation application, including orthosis, prosthesis, virtual reality (VR) or FES?

RQ3: Is the classification algorithm based on AI methods?

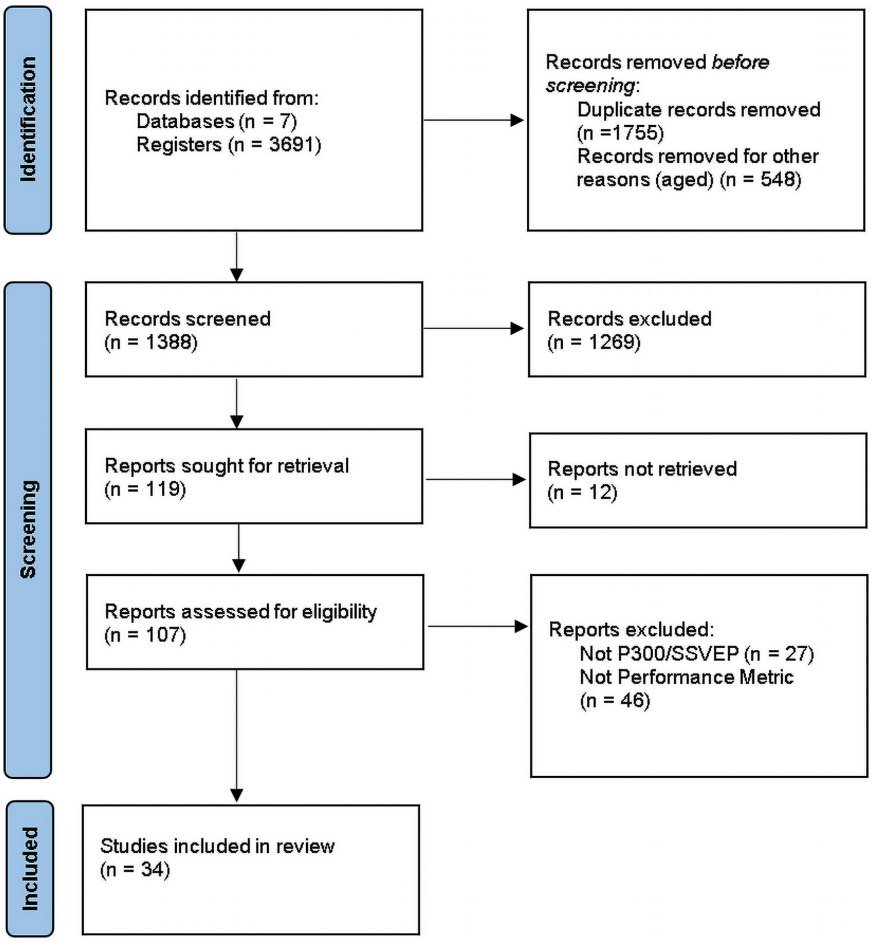

FIGURE 1 | PRISMA flow diagram for the Systematic Literature Review.

RQ4: Are the validation methods mentioned?

RQ5: Does the paper report the performance metrics values (accuracy, ITR, etc.) of the algorithms?

RQ6: Are patients or healthy subjects involved in the study?

RQ7: What are the future challenges foreseen by the authors?

Inclusion and Exclusion Criteria

The following medical and technical search terms were used to query the databases: "BCI," "P300," "SSVEP," "brain computer interface," "FES," "evoked potential visual," "neurorehabilitation," "functional electrical stimulation." These search terms were further combined with "artificial intelligence," "machine learning," "deep learning," and "artificial neural network," among others. Articles were also explored based on performance-related terms such as accuracy and ITR. Articles were discarded if they were not thematically relevant to the scope of this paper or they did not include tests with patients or healthy subjects. In addition to the structured literature search, a manual search of works cited in the articles included in the SLR was also conducted. Thus, some articles not identified by the original search were included in this review, if all other requirements were met. The level of evidence was not graded due to the exploratory nature of many of the studies.

Data Extraction and Analysis

According to the proposed taxonomy, described in **Figure 2**, only two types of articles were considered: originals and reviews. The selected articles were divided into three major categories, the first one being the AI methods cluster, which provides a general overview of the used AI algorithms. The second category is a four-tiered research cluster, related to BCIs involving motor rehabilitation applications. Tier 1 contains articles involving FES systems, tier 2 provides articles related to prostheses, tier 3 considers orthoses application and tier 4 included studies aimed to the use of VR. The third category is the performance measurement cluster, which comprises the metrics employed for performance assessment of the classification algorithms.

RESULTS

Three thousand six hundred and ninety one studies were retrieved from the electronic databases (Step 1-Identification), as

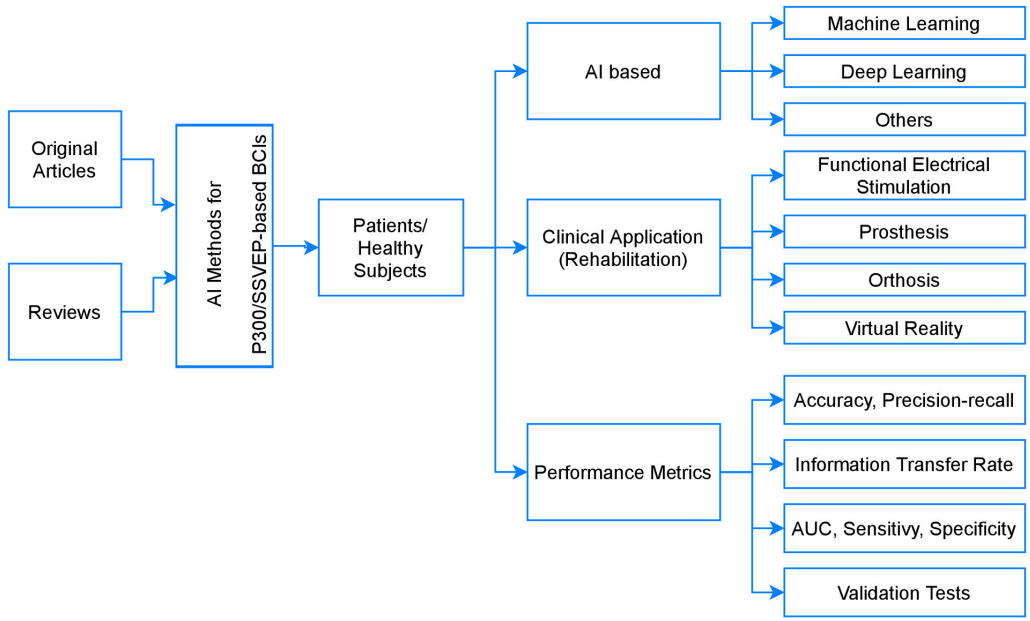

FIGURE 2 | Taxonomy of the SLR: AI methods used in BCI-based P300/SSVEP systems for motor rehabilitation applications.

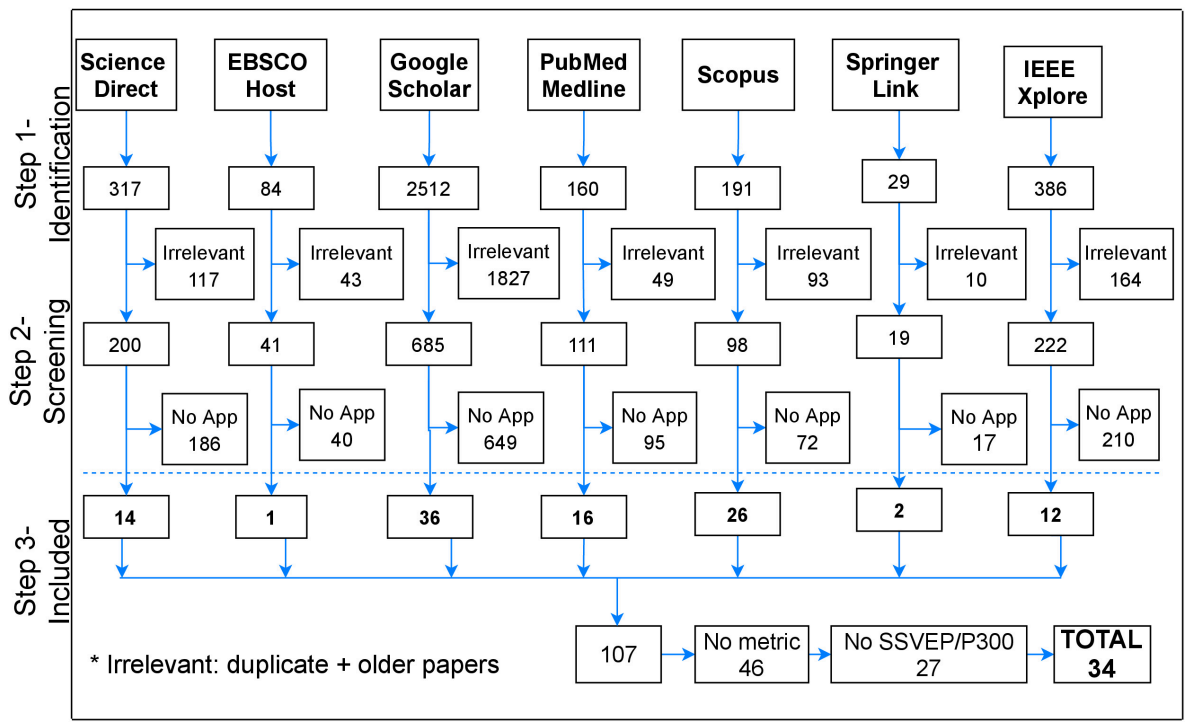

FIGURE 3 | Number of records identified from each database for the Systematic Literature Review.

shown in **Figure 3**; the first filtering step was based on the title, abstract, and keywords of the articles. After the exclusion criteria were applied, 2303 articles were discarded due to duplication or publication date prior to 2011. Of the total articles published after 2011 (1388), 1269 were excluded during full text review (Step 2-Screening) because 51.02% (702) refer to implementation and offline analysis of diverse classification algorithmic strategies, without using them in an actual application. In contrast, 28.48% (392) deal with BCI (P300- or SSVEP-based) used as speller, and 12.73% (175) for diverse applications to control wheelchairs,

home appliances, robots or video games; only the remaining 7.77% (107) are focused on applying (P300- or SSVEP-based) BCIs for motor rehabilitation purposes.

The remaining 107 articles underwent a quality screening where 27 studies were eliminated, because they did not refer to either P300 or SSVEP BCIs; also 46 studies were eliminated for not specifying the performance metrics of AI algorithms. Finally, the remaining 34 papers were included as relevant to this SLR and then selected for data extraction and further analysis (Step 3-Included).

Categorization of the Results

Table 1 shows the 34 papers considered as relevant for this SLR, of which 26.47% (9) refer to P300, 55.8% (19) to SSVEP strategy and 17.64% (6) to the hybrid BCI modality. Of the six hybrid BCIs articles, three combined P300 and SSVEP signals, and the other three combined SSVEP (2) or P300 (1) with the motor imagery paradigm. The papers were divided in five major categories, corresponding to the actuator device controlled by the P300- or SSVEP-based BCI system: FES (17.64%, $n = 6$), VR (23.52%, $n = 8$), Orthosis (29.41%, $n = 10$), Prosthesis/Exoskeleton (11.76%, $n = 4$), and RRS (Robotic Rehabilitation System) (17.64%, $n = 6$). The main application of the selected works is rehabilitation of the hand (52.94%, $n = 18$) and the lower limb (26.47%), in the latter case by means of exoskeletons and rehabilitation systems. In the VR category, objects and proprioceptive stimulation (Tidoni et al., 2017) are controlled in a virtual smart home environment (Edlinger et al., 2011).

All VEP-based BCI systems were tested on healthy subjects, and only 4 (11.76%) of them included both abled-bodied participants and patients, mainly with SCI and amyotrophic lateral sclerosis (ALS). The remaining (30) works tested their systems exclusively with healthy subjects. Nine of the identified studies tested the BCI system in more than ten able-bodied subjects (Brunner et al., 2011; Horki et al., 2011; Sakurada et al., 2013; Kwak et al., 2015; Chen et al., 2018; Delijorge et al., 2020; Son et al., 2020; Zhu et al., 2020). Of the four studies that recruited both healthy subjects and patients (Sakurada et al., 2013; Tidoni et al., 2017; Okahara et al., 2018; Delijorge et al., 2020), only one (Sakurada et al., 2013) reported the classification accuracy for both patients (88.46%) and healthy subjects (81.1%). Moreover, all of them used a different number of EEG electrodes (3–8), BCI paradigms (P300, SSVEP and hybrid), and visual stimulation patterns. Also, the four studies were focused on upper limb, but they used different actuators: neuroprosthesis, orthosis, VR and rehabilitation system.

Prior to classification, some feature selection algorithm is commonly applied to (i) reduce redundancy, (ii) choose the features more related to the target mental states in the BCI, (iii) reduce the number of parameters to be optimized by the classifier, or (iv) produce faster predictions for new data. Power Spectral Density (PSD), Short-time Fourier Transform (STFT), Common Spatial Patterns (CSP), and Independent Component Analysis (ICA) are commonly used algorithms for feature extraction, but amplitude/spectral power (37.83%) and CCA (10.81%) were the most reported methods in this SLR.

Regarding the use of AI methods for classification, the most reported ML algorithms were LDA (48.64%) and SVM (16.21%), with reported accuracy range from 38.02 to 100% and ITR from 1.55 to 49.25 bpm. The best ITR (49.25 bpm) was for the SSVEP paradigm using an ensemble classifier (Chen et al., 2018). Only one study used a DL algorithm: CNN, with excellent classification accuracy (99.28 and 94.93% in static and dynamic conditions) but unspecified ITR (Kwak et al., 2017). On the other hand, only five papers reported other performance metrics besides classification accuracy: true positive rate, positive predictive value, false positive rate, Area under the ROC Curve (AUC), sensitivity and specificity. Finally, less than one out of three of the selected papers reported the validation method they used: k-fold cross-validation (29.41%, $n = 10$) and leave one-out cross validation (2.94%, $n = 1$).

Other Results

As mentioned, hybrid VEP-based BCI systems were also found, which use two BCI control signals, each one for a specific task. For example, the hybrid SSVEP/MI system reported by Savić et al. (2012) is used to active a FES system, where the SSVEP signal is used for target selection and the MI strategy for activation of the FES-assisted reach-to-grasp of a certain object. Other hybrid BCI systems using P300 and SSVEP signals have been reported, one for controlling a smart home environment, where a SSVEP-based toggle switch was implemented to activate and deactivate the P300 BCI (Edlinger et al., 2011). Another hybrid BCI allows subjects to simultaneously imagine themselves moving both hands or both feet, while fixing the sight on one of two oscillating visual stimuli to activate an SSVEP BCI system (Brunner et al., 2011).

Regarding EEG electrodes, SSVEP and P300 BCI systems used a minimum of two recording channels for SSVEP (Li et al., 2018) and 1 for P300 (Bhattacharyya et al., 2014), and it goes up to a maximum of 19 for SSVEP (Son et al., 2020) and 32 for P300 BCIs (Duvinage et al., 2012; Huang et al., 2019). They are placed predominantly over the parietal and occipital (visual cortex) regions, in the positions P3, Pz, P4, PO3, PO4, T5, T6, O1, Oz, and O2 of the 10–20 International system for EEG electrode placement.

A key component in P300-/SSVEP-based BCI systems is the visual stimulation module. Although this element is not considered in detail in this paper, it is worth mentioning that there is a great variety of visual stimulation patterns (Amaral et al., 2017; Choi et al., 2019), ranging from flashes with variable duration (tens or hundreds of ms), with matrices of different types (LEDs, characters, or icons) to evoke P300 signals, and a range of frequencies (from 5 to 25 Hz) to produce SSVEP signals. For P300 BCIs, two strategies were used to improve the performance, 3D virtual visual stimuli (Huang et al., 2019), and overlay of smiley faces over targets (Delijorge et al., 2020).

However, if a low visual stimulation frequency is used by the visual stimulation module, the system's ITR may be limited. To overcome this limitation, diverse stimuli colors and flickering frequencies have been proposed for hybrid BCI's. With these variations of the visual stimulation paradigm, a good trade-off is achieved between accuracy (92.30%) and ITR (82.38 bpm),

TABLE 1 | Artificial Intelligence Algorithms applied for detection and classification of P300 or SSVEP signals in BCI Applications for motor rehabilitation.

First author, year	BCI signal	Application/actuator	Subjects		# Electrodes	Visual stimulation pattern	Feature extraction method	Classifier	Performance		Validation Method
			Impaired	Healthy					Accuracy (%)	ITR (bpm)	
Stan et al., 2015	P300	Hand orthosis	None	9	8	Flashes: 75 ms Flash-time: 100 ms	NS	LDA	100	NS	NS
Kwak et al., 2017	SSVEP	Lower limb exoskeleton	None	7	8	5 LEDs flashing at 9, 11, 13, 15, 17 Hz with 50% DC	NR	CNN	Static: 99.28, Ambulatory: 94.93	NS	10-fold CV
Kwak et al., 2015	SSVEP	Lower limb exoskeleton	None	11	8	5 LEDs: 9, 11, 13, 15, 17 Hz with 50% DC	CCA	k-nearest neighbors	91.3	32.9	5-fold CV
Delijorge et al., 2020	P300	Robotic hand orthosis	8 ALS	18	8	2–30 random flashes	CCA	RLDA	Offline: 78.7 (target), 85.7 (non-target). Online: 89.83	18.13	5-fold CV
Zhao et al., 2016	SSVEP	FES, upper limb rehabilitation	None	5	14	Squares flashing at 12, 15, 20 Hz	Power spectrum	LDA	Offline: 79.37–85.13 Online: 54.32–87.5	Offline: 27.54	10-fold CV
Tidoni et al., 2017	SSVEP	VR, Propioceptive Stimulation	3 SCI	18	8	3 × 3 grid. flash-time: 133.33 ms dark-time: 83.34 ms	NS	LDA	83.33	1.55	NS
Yao et al., 2011	SSVEP	FES, upper limb rehabilitation	None	4	8	White blocks of lights flickering at 6.82, 7.5, 8.33, 9.37, and 12.5 Hz	5 flickering frequencies and their harmonic components	LDA	Online: 82.22	Ns	Ns
Brunner et al., 2011	Hybrid: SSVEP + P300	Moving both hand or both feet	None	12	SSVEP: 2. MI: 3.	LEDs flickering at 8 Hz (top) and LED at 13 Hz (bottom)	logarithmic band power: SSVEP and ERD	LDA	ERD: 79.9 SSVEP: 98.1 Hybrid: 96.5	ERD: 3.2. SSVEP (6.1) hybrid (6.3)	CV
Edlinger et al., 2011	Hybrid: SSVEP + P300	VR, control of virtual smart home environment	None	3	SSVEP: 8. parietal/ occipital. P300: 8 frontal, central occipital, parietal	P300: rectangular matrix with characters or icons, flashed in a random order SSVEP: flickering lights (LEDs) or flickering symbols (5 -25 Hz)	SSVEP: minimum energy (ME) algorithm, P300: NA	P300: LDA, SSVEP: LDA	P300: 100	NS	NS
Su et al., 2011	Hybrid: P300 + MI	VR	None	4	P300: 14. MI: 22.	NS	P300: piecewise cubic spline interpolation+ Butterworth filter + average. MI: multiple band-pass filters	P300: SVM, MI: FLDA	Offline (MI): 92.5–100	NS	NS

(Continued)

TABLE 1 | (Continued)

First author, year	BCI signal	Application/actuator	Subjects		# Electrodes	Visual stimulation pattern	Feature extraction method	Classifier	Performance		Validation Method
			Impaired	Healthy					Accuracy (%)	ITR (bpm)	
Sakurada et al., 2013	SSVEP + P300	Upper limb rehabilitation,. Occupational therapy	3 (upper cervical SCI)	12	SSVEP: 3	SSVEP: 3 LEDs flickering at 8 Hz (green and blue). P300: Flash matrix	power spectrum (FFT) + CCA	SVM	Healthy: 88.46. Patients: 81.19	NS	NS
Choi et al., 2016	SSVEP + MI	FES, hand-wrist rehabilitation. SSVEP to stop FES	None	4	MI: 3 central. SSVEP: 2 occipital.	SSVEP: LED flickering at 9 Hz	MI: ERD/ERS, SSVEP: averaged Pearson's correlation (*r*-value)	MI: FLDA SSVEP: CCA	MI: 90.485	NS	10-fold CV
Yao et al., 2012	SSVEP	FES, knee rehabilitation (movement training system)	None	2	8	a red horizontal bar, flickering light at 6.82, 8.33 and 12.5 Hz	Power spectrum	LDA	Online: 80.36–96.4	NS	10-fold CV
Duvinage et al., 2012	P300	Lower limb rehabilitation. Foot lifting orthosis	None	5	32	NS	xDAWN + two epochs average	LDA	94.30	NS	NS
Ortner et al., 2011	SSVEP	Hand Orthosis	None	7	1: O1	2 LEDS, flickering at 8 and 13 Hz	PSD	HSD	78	NS	NS
Rohani et al., 2014	P300	VR	None	5	4	NS	NS	SVM	NS	NS	NS
Son et al., 2020	SSVEP	FES, upper limb rehabilitation	None	11	19	flickering action video at 15 Hz	STFT, Power average	CSP (discriminating 2 class)	93.51	NS	10-fold CV
Chen et al., 2019	High-frequency SSVEP	Robotic arm	None	10	9: parietal or occipital	Flicker: 30, 31, 32, and 33 Hz	Spectral amplitude	FBCCA	Online: 97.75	Online: 17	NS
Li et al., 2018	SSVEP	Hand prosthesis	None	6	2: occipital	Scene graph paradigm -drinking & eating-, (8, 9.24, 10.9, and 12 Hz)	Time-frequency spectra, STFT	CCA	94.58	19.55	NS
Horki et al., 2011	SSVEP + MI	Prosthesis: artificial upper limb, elbow control	None	12	26: occipital and central	2 bars of red LEDs, flickering at 8 and 13 Hz	Sequential floating forward selection	CCA	Offline: 91	NS	10-fold CV
Koo et al., 2015	SSVEP	VR	None	3	8: central, parietal and occipital	Flickering lights at 5.5, 6.7, 7.5, and 8.6 Hz	NS	CCA for SSVEP detection	100	24.58	NS
Chu et al., 2018	SSVEP	Robotic rehabilitation system	None	6	14: frontal, parietal, occipital	Three squares flashing at 12, 15, 20 Hz	Power spectrum	LDA (voting)	82.30	27.40	NS

(Continued)

TABLE 1 | (Continued)

First author, year	BCI signal	Application/actuator	Subjects		# Electrodes	Visual stimulation pattern	Feature extraction method	Classifier	Performance		Validation Method
			Impaired	Healthy					Accuracy (%)	ITR (bpm)	
Gui et al., 2015	SSVEP	Lower limb rehabilitation system (hip and knee)	None	6	4: occipital and parietal	Flickering at 6.82, 7.5, 8.33, and 12.5 Hz	Spectral amplitude	LDA	92.40	NS	NS
Huang et al., 2019	P300	VR	None	6	32	3D stereo visual stimuli	NS	BLDA	96	42.51	10-fold CV
Yao et al., 2019	SSVEP	VR	None	10	9: parietal and occipital	2 stimulus presentation methods. 3D stimulus at 9, 10, 11, 12, 45 Hz	NS	FBCCA	Static mode: 92	Static mode: 22.49	Leave one-out CV
Touyama and Sakuda, 2017	Collaborative SSVEP	VR	None	8	2: parieto-occipital	two virtual cubes flickering at 6 and 8 Hz	Spectral amplitude	FLDA	95.2	NS	NS
Bhattacharyya et al., 2014	P300	Robot arm control for prosthetics application	None	5	1: Pz	Oddball-like paradigm	(Temporal) Average of 4 epochs	SVM (linear kernel)	Offline: 95.2. Online: 81.5	Online: 23.83	NS
Chen et al., 2018	SSVEP	Robotic arm control	None	12	10: P3, Pz, P4, PO3, PO4, T5, T6, O1, Oz, O2	15 targets (8–15 Hz in 0.5 Hz steps)	FBCCA for EEG decomposition	Ensemble Classifier	Robotic movement task: 92.78	49.25	NS
Casey et al., 2019	P300	Robotic arm control	None	4	6: Pz, P3, P4, PO3, PO4, and Oz	P300 speller programmed to control a robotic arm	Minimum and maximum amplitudes in the frequency domain (6 features per electrode)	2 classifiers: SVM (RBF kernel), and Random Forest	38.023	NS	NS
Achanccaray et al., 2019	P300	Robotic arm Control	None	8	16	Two images flashing randomly: a wheelchair and a robotic arm	CSP	BLDA	Training: 91.6. Test: 82.6.	NS	NS
Ding-Guo and Ying, 2012	SSVEP	FES, lower limb	None	6	NS	NS	Frequency-domain	LDA	85	NS	NS
Huang et al., 2013	P300	Elbow rehabilitation robot	None	NS	NS	Panel with 25 commands	NS	SVM	Online: 90.82	NS	NS
Okahara et al., 2018	SSVEP	Neuro-prosthesis	3-ALS	NS	1: Oz	4 × 4 LED flicker at 32–54 Hz	PSD	Classification Threshold	Online 83.3	NS	NS
Xu et al., 2021	SSVEP	Upper Limb Exoskeleton	None	5	6: O1, O2, Oz, P3, Pz, P4	4 Flickering squares at 8.57, 10, 12, 15 Hz	Frequency domain	CCA	Offline: 86.1	NS	NS

BLDA, Bayesian linear discriminant analysis; CCA, canonical correlation analysis; CSP, common spatial patterns; DC, duty cycle; FLDA, Fisher's Linear discriminant analysis; LDA, linear discriminant analysis; NS, non-specified; SVM, support vector machine; VR, virtual reality; CV, cross validation; FES, Functional Electrical Stimulation; MI, motor imagery; SSVEP, steady state visually evoked potentials; SCI, spinal cord injury; HSD, harmonic sum decision; STFT, short-time Fourier Transform.

enhancing the potential to develop P300/SSVEP-based BCIs for the control of rehabilitation devices (Katyal and Singla, 2020).

DISCUSSION

The results of the SLR are discussed according to the Research Questions stated in section "Research Questions."

RQ1: What Type of Evoked Potential (P300 or Steady State Visually Evoked Potentials) Is Involved in the Brain-Computer Interface Visual Paradigm?

As shown in this SLR, despite the large number of articles related to BCI systems based on VEPs, most of them report the implementation and analysis of diverse algorithmic strategies to train and test their classification performance, without any actual application, such as motor rehabilitation. We found that using either P300 or SSVEP signals, it is possible to operate a BCI system by performing visual attention tasks. EEG signal features in those systems are extracted in the time or frequency domain, without compromising greatly the system's accuracy and requiring little or no training.

The SSVEP signal has some advantages over the P300: (1) no mental task is required to induce the intended potential, (2) enables subjects to use the paradigm without requiring great mental load, and (3) it achieves higher ITR. However, the number of command choices in an SSVEP paradigm is generally represented by frequencies within the band of 5–20 Hz (Katyal and Singla, 2020).

SSVEP-based BCIs can encode multiple commands without any extensive user training and show potential for high-speed communication. For example, Chen et al. (2015) reported an ITR of 267 bpm in a 45-target system (Chen et al., 2015) and in Nakanishi et al. (2018) was reported an ITR of 325.33 bpm in a 40-target system. Although the efficiency and performance of different algorithms for detecting the P300 and the SSVEP in BCI applications have already been evaluated in a variety of laboratory demonstrations (Kluge and Hartmann, 2007; Kundu and Ari, 2017), many difficulties are still faced to implement this type of BCI systems for the control of devices with clinical purposes. One of these problems is the limitation in the number of available stimulation frequencies (Müller-Putz et al., 2005). One limitation of those papers is that not all of them report a full set of technical descriptions, such as the signal processing techniques for feature extraction and performance metrics of the classification algorithms, in most cases they only report classification accuracy. However, from the reported online performance of SSVEP-based BCIs (**Table 1**), it is clear they provide effective communication speed with good average accuracy after a very short training period (Guger et al., 2012). However, flickering lights could be disturbing for some people. In the other hand, P300-based BCIs are less accurate than SSVEP-based BCIs but are more suitable for people suffering epilepsy or people having difficulties with accurate control of the eye muscles (Allison et al., 2010).

RQ2: Is the Purpose of the BCI System Aimed to Some Motor Rehabilitation Application, Including Orthosis, Prosthesis, Virtual Reality or Functional Electrical Stimulation?

As shown in **Table 1** and **Figure 2**, SSVEP- and P300-based BCIs have been used in motor rehabilitation applications to drive primarily four types of actuators and then facilitate brain plasticity in patients with limb motor dysfunction. They are (1) Orthosis (Ortner et al., 2011; Duvinage et al., 2012; Stan et al., 2015; Delijorge et al., 2020) and exoskeleton (Gui et al., 2015; Kwak et al., 2015; Bhagat et al., 2016), used to perform sequences of movements to activate the hand, wrist, arm, leg or foot. (2) FES, which has been reported to be of help to regain coordination and improve performance in functional tasks (Do et al., 2011; Ding-Guo and Ying, 2012; Yao et al., 2012; McCabe et al., 2015; van Dokkum et al., 2015; Choi et al., 2016; Osuagwu et al., 2016; Zhao et al., 2016; Son et al., 2020). (3) Prosthesis (Li et al., 2018), and (4) VR (VEP-based BCI systems immersed in virtual environment) (Su et al., 2011; Koo et al., 2015; Tidoni et al., 2017; Touyama and Sakuda, 2017; Choi et al., 2019; Huang et al., 2019; Yao et al., 2019).

RQ3: Is the Classification Algorithm Based on Artificial Intelligence Methods?

Most algorithms for classification of VEP-based BCI signals are based on AI methods. The advantages and disadvantages of each of them depend on the signal and the application. A simple and efficient ML algorithm, LDA, was among the best methods in terms of classification accuracy and ITR used in P300-based (ACC = 100% orthosis) (Stan et al., 2015) (ACC = 94.3%) (Duvinage et al., 2012), and SSVEP-based BCI systems selected in the SLR (ACC = 79%, ITR = 27.54 bpm-FES) (Zhao et al., 2016), (ACC = 83.33%, ITR = 1.55 bpm -VR) (Tidoni et al., 2017) (ACC = 82.22% -FES) (Yao et al., 2011), (ACC = 80-96% -FES) (Yao et al., 2012), (ACC = 82.30%, ITR = 27.4 bpm) (Chu et al., 2018) (ACC = 92.4%) (Gui et al., 2015), (ACC = 85% -FES) (Ding-Guo and Ying, 2012). Moreover, classification Accuracy obtained with LDA in P300-based BCI is slightly higher than with SSVEP-based BCI. Hence, LDA can be considered a first-choice ML classification algorithm for BCIs based on visual paradigms for rehabilitation applications.

Some ML classifiers such as FBCCA, FLDA, and BLDA have been proposed to improve the trade-off between accuracy and ITR of VEP-based BCI systems. They presented accuracies over 90% for both modalities (P300 and SSVEP) (Touyama and Sakuda, 2017; Achanccaray et al., 2019; Chen et al., 2019; Yao et al., 2019). The FBCCA and BLDA algorithms were superior to LDA in terms of ITR; for example, using a FBCCA (Chen et al., 2018) achieved an ACC = 92.78% with a high ITR (49.25 bpm), when an SSVEP signal was

used to control a robotic arm. In the other hand, a BLDA-based classification algorithm was applied in a P300-based BCI coupled to the VR environment; in this case ACC = 96% and ITR = 42.51 bpm were achieved (Huang et al., 2019). The filter bank CCA (FBCCA) method has been extensively studied by Chen et al. (2018). This method incorporates the fundamental and harmonic frequency components to improve the detection of SSVEPs and has demonstrated its superiority over the standard CCA method (Chen et al., 2015, 2019).

The only work in the SLR that used a DL algorithm (CNN) for signal classification was (Kwak et al., 2017). In that study, the authors reported a BCI system for control of a lower limb exoskeleton via a visual stimulus generator that produced five different frequencies for SSVEP signals. They used CCA, Multivariate Synchronization Index (MSI) and CCA with k-Nearest Neighbors (CCA-KNN) to compare the classification result with three different classification methods. Using CNN-1 (three-layer network), they achieved an accuracy of up to 91.3% and an ITR of 32.9 bpm.

Beyond the works included in this SLR, DL methods have some advantages for classification of SSVEP and P300 BCI signals in comparison with the traditional ML algorithms, including:

(1) Higher Classification Accuracy (Thomas et al., 2017).
(2) DL methods reduce the dependence on manually designed feature extraction.
(3) As the size of the dataset increases, DL techniques tend to perform better than traditional classifiers (Kwak et al., 2017; Lee J. et al., 2020).
(4) The development of new powerful GPUs (graphics processing units) and cloud-based AI services have improved the cost-effectiveness of DL systems.

Despite those advantages, DL techniques have some disadvantages compared to ML algorithms:

(1) They are complex, computationally expensive, and require a large amount of data to be trained.
(2) Configuration of the different parameters of DL systems is still a major challenge.
(3) DL methods have not yet shown convincing improvements over state-of-the-art ML classification algorithms for BCI (Lotte et al., 2018).

RQ4: Are the Validation Methods Mentioned?

Regarding validation methods, about one third of the studies reported the type of cross validation they used (1: leave one-out, 2: 5-fold, and 7: 10-fold). This data is relevant as an indicator of the robustness and confidence on the reported performance (accuracy) of the of the AI-based classification algorithms, and of their generalization ability. When the validation methods are not explicitly reported, the certainty about the results may be questionable (Abdulaal et al., 2018).

RQ5: Does the Paper Report the Obtained Values of the Performance Metrics Values (Accuracy, Information Transfer Rate, etc.) of the Algorithms?

Two important performance criteria for classification algorithms in BCI systems are accuracy and ITR. According to BCI literature (Hwang et al., 2013), an accuracy greater that 70% must be achieved by any subject to be able to use a BCI system effectively for the control of external devices. The average classification accuracy of SSVEP-based BCI systems was 90.3% (n = 20), while for P300 was 85.9% (n = 9), and 93.41% (n = 6) for hybrid systems. In contrast, few works report the ITR, with a mean value of 20.88 bpm for SSVEP (n = 10), 28.15 bpm for P300 (n = 3), and 6.3 bpm for the only hybrid BCI that reported it (Brunner et al., 2011). It is worth mentioning that the average accuracy of P300 systems was lower than for SSVEP due to a single paper (Casey et al., 2019) that reported 38% classification accuracy. Without taking into consideration that article (n = 8) the average accuracy of P300 would be very similar to SSVEP (91.88%).

However, the above comparisons must be taken with reserve, since the number of works reporting the metrics varies a lot across modalities. Moreover, there is a high heterogeneity in different aspects of their experimental paradigms, visual stimulation features (frequencies, colors, signs, figures), subjects (healthy or patients), rehabilitation application (FES, prosthesis, orthosis, VR, etc.), length of data analysis windows, signal acquisition hardware, type (passive, active) and number of electrodes, etc. Each of those aspects affect different parts of the system that influence performance metrics, such as the complexity and execution time of the signal processing and classification algorithms.

Despite all the differences across the articles in technical and human aspects that can affect performance metrics, it is noticeable the high similarity in the average accuracy for the three BCI types considered. Regarding hybrid BCIs, they did not significantly increase the classification accuracy in comparison with single modality BCIs, as was the case for Brunner et al. (2011), with 96.5% for SSVEP and 98.1% for the MI/SSVEP hybrid modality. Moreover, most hybrid BCIs did not report the ITR value. A possible reason for this is, that in comparison with single modality VEP-BCIs, hybrid BCIs have relatively low ITRs due to more complex setups, involving one operation stage for each BCI signal, each one with a signal processing block, plus the necessary pauses between operation stages. For these and other reasons, when the users present motor imagery BCI illiteracy, single modality VEP-based BCI systems could be a better option that hybrid ones (SSVEP + MI), as suggested by Brunner et al. (2011) for SSVEP.

RQ6: Are Patients or Healthy Subjects Involved in the Study?

All VEP-based BCI systems included abled-bodied and only a handful of them included both healthy subjects and patients with SCI or ALS disease. Several human factors directly related with the experimental setup of the BCIs, such as reaction times,

mental load and fatigue, and user engagement and motivation, could have impacted the performance metrics results. Those factors become especially relevant in users with severe motor impairments. Regarding P300-based BCIs, it has been reported that the P300's latency is higher for disabled subjects (around 500 ms) when compared to able-bodied ones (around 300 ms), and that the amplitude at the P300 peak is smaller for disabled (around 1.5 μV) than for the able-bodied subjects (around 2 μV) (Hoffmann et al., 2008). As an example, Sakurada et al. (2013) presented a hybrid (SSVEP + P300) BCI system, that compared the classification accuracy of healthy subjects (88.46%, $n = 12$) and SCI patients (81.1%, $n = 3$). These differences can be explained, at least in part, by the difficulty of patients to control eye gaze, and head or trunk posture during the BCI sessions, which could have in turn exacerbated physical and mental fatigue.

Beyond Research Questions

Other topics of interest were identified during the development of the SLR, that fall outside the scope of the above Research Questions. These topics are discussed in the following subsections.

Visual Stimulation Patterns

Some studies have suggested that different visual stimuli patterns produce variations in the VEP signals, and thus have an impact on the BCI performance (Speier et al., 2017; Li et al., 2020). Mainly, low- (up to 10 Hz) and medium-frequency (13–25 Hz) stimuli have been adopted in SSVEP (Kuś et al., 2013). Although stimulation in these frequency ranges evoke SSVEPs with a large amplitude, it can be annoying or tiring for some users. A possible solution to this problem is to use high-frequency stimulation. High-frequency stimuli can decrease visual fatigue caused by flickering, thus making the SSVEP-based BCI a more comfortable system (Wang et al., 2005; Diez et al., 2011; Volosyak et al., 2011). Other visual stimulation techniques have been proposed to enhance SSVEP BCIs performance, like amplitude modulation (Chang et al., 2014), variation of the duty cycle (Shyu et al., 2013) or interpolation techniques (Andersen and Müller, 2015).

For P300-based BCIs, variations in color and arrangement of the visual stimuli (Guo et al., 2019) and overlay of targets with pictures of faces of famous people (Kaufmann et al., 2011), have shown to increase the classification performance for spelling applications. Flashing elements can change the color from blue to green at the time of intensification, (Takano et al., 2009), or 3D virtual visual stimuli can also be presented to the subject (Huang et al., 2019). However, if a low visual stimulation frequency (interstimulus interval) is used by the visual stimulation module, the system's ITR may be limited (Mainsah et al., 2015). To overcome this limitation, diverse stimuli colors and flickering frequencies have been proposed for hybrid BCI's achieving a good trade-off between accuracy (92.30%) and ITR (82.38 bpm) (Katyal and Singla, 2020). These approaches have the potential to enhance the development and performance of P300/SSVEP-based BCIs for the control of rehabilitation devices.

Electrode Setup

The configuration of electrodes (number and placement) determines the suitability of the system for daily use. In the SLR systems with 4–32 electrodes were found, predominantly located over the parietal-occipital area for SSVEP and widespread from frontal to occipital areas for P300. VEP-based BCI systems using fewer electrodes require also shorter donning times and are more user friendly than systems with many electrodes. However, if too few electrodes are used, there is a risk of not capturing all necessary features for accurate classification. This has been shown previously for both P300 (McCann et al., 2015) and SSVEP (Carvalho et al., 2015; Ravi et al., 2019) BCI systems, in studies that find optimal subsets of channels, that enhance classification accuracy. Although small subsets of electrodes (even with one or two) are selected as optimal for some users and feature extraction algorithms (McCann et al., 2015), in most cases a third or more of all available electrodes are selected through channel selection algorithms (Carvalho et al., 2015) to work properly. Interestingly, in users with low SSVEP responses (BCI illiteracy) the electrode subsets chosen through channel selection algorithms may include preferentially those located in regions (central and frontal) not typical (occipital and parietal) for this BCI modality (Carvalho et al., 2015). Likewise, it has been proposed in (visually evoked) P300-based BCIs (McCann et al., 2015) the search of non-standard sets of electrodes, to optimize the performance in individuals with motor impairments, who have little or no control of eye movements.

Steady State Visually Evoked Potentials and P300 Brain-Computer Interfaces for Motor Rehabilitation

Although both SSVEP and P300 BCI systems based on visual stimuli were found in this SR, there are fundamental and technical aspects of each one that can influence their suitability to be incorporated in rehabilitation applications, to name: the experimental paradigm, the degree of cognitive and sensory requirements, covert and over attention, and synchronous/asynchronous operation. These aspects are further discussed below.

First, their experimental paradigms and neurophysiological basis are essentially distinct. On the one hand, SSVEP signals directly reflects the (fixed) frequency of presentation of visual stimuli in EEG oscillations. These signals are recorded typically in occipital electrodes over the visual cortex area (Müller-Putz et al., 2005), and they reflect the sensory processing of visual stimuli. On the other hand, P300-based BCIs based on visual stimuli are designed around the oddball paradigm, in which a series of stimuli (one relevant, or target, and other irrelevant, and ignored) are presented repeatedly in random order. In this case, the key variables are the probability of occurrence of the target stimulus and the inter-stimuli interval, which can be varied randomly.

A main difference in the experimental paradigm of SSVEP and P300 BCIs is the task required for the subject while looking at the target. For the SSVEP, the only requirement is to maintain the gaze fixed on the target visual stimulus. Generally, time windows of 1–3 s are enough to identify when the subject is visually attending the target (Liu et al., 2020). In the P300 case, the user is asked to perform some mental activity for each flashing

of the visual target that he or she acknowledges consciously, while ignoring the non-target stimuli. Generally, this mental task involves counting mentally the number of times that the target symbol or picture is intensified (visual stimuli) (Arvaneh et al., 2019). This is performed to engage continuously the working memory, thus involving a definite cognitive activity besides the visual attention task. Thus, cognitive (N200, P300) and visual (P100, N100) potentials are often found on EEG signals from P300 BCIs (Aloise et al., 2012). In contrast, sinusoidal-like SSVEP signals directly reflect the frequency (and harmonics) and phase of the attended stimuli (Sozer, 2018), without the need of any cognitive or behavioral task. Therefore, while P300-BCIs can be more cognitive demanding, SSVEP BCIs tend to induce more visual fatigue, especially when multiple targets are presented simultaneously (Dreyer et al., 2017). The cognitive demand of P300 BCIs may explain in part the lower average accuracy of papers included in the SLR, and why more (twice) papers used SSVEP instead of P300 signals. Moreover, of the four articles in the SLR involving patients, three were based on SSVEP and only one in P300, with relatively good levels of classification accuracy (80–90%). Therefore, differences in cognitive and visual fatigue can be also a key factor when choosing a BCI approach for patients with cognitive and motor impairment, like stroke or SCI.

One shared experimental requirement of SSVEP and P300 BCIs is that, to evoke the expected EEG activity, user attention must be focused on the current visual target for some time. For both paradigms the BCI system performs better when the sight is centered on the visual target (foveal vision) (Walter et al., 2012; Ron-Angevin et al., 2019). This is known as overt attention and is one of the key differences of SSVEP-based with P300-based BCIs, the latter having proved to work well also when visual stimuli are attended covertly, through the peripheral vision (Aloise et al., 2012). Although promising efforts have also been made to develop SSVEP BCIs based on covert attention (Zhang et al., 2010; Reichert et al., 2020), their performance still is lower than with overt attention. This aspect of visual BCIs has implication for the development of applications. In the case of motor rehabilitation of users with restrained control of gaze and neck movement (such as those with ALS or high cervical SCI), the possibility of attending stimuli covertly, and still obtaining informative EEG signals, would improve its clinical feasibility.

P300-based BCIs seem to have some advantages over SSVEP ones, since multi-target systems are feasible even using covert attention (Aloise et al., 2012), while SSVEP BCIs using this approach have been limited to a couple of targets (Zhang et al., 2010). Hence, a P300-based BCI system designed for covert attention, would allow the subject to attend visual stimuli (for selection of multiple actions or commands) while performing functional motor tasks, aided by some of the actuators mentioned in the SLR (FES, orthosis, robot, etc.). In the other hand, an SSVEP BCI system, based on overt attention, would be better suited for VR-based rehabilitation applications, with the user's visual attention centered (overtly) in the visual target, since all stimuli and interactions are designed to be performed through the virtual environment. The papers analyzed in this SLR did not consider explicitly covert attention in their design, which remains an approach to be explored for visual BCI-based motor rehabilitation.

Another relevant aspect of visual BCI paradigms regarding their feasibility for motor rehabilitation is their type of operation: asynchronous or synchronous. In other words, if the system allows the user to convey commands at any moment (asynchronous) or only at times established by the system (synchronous) (Nooh et al., 2011). Clearly, this can be a key factor in the design of motor rehabilitation systems and interventions based on visual BCIs. For motor and neurologic rehabilitation systems and interventions, a key factor is the user's active engagement and participation, while performing some functional tasks by their own voluntary effort or with the help of assistive technologies. This approach to rehabilitation is known as *activity-based* (Backus, 2008), and to develop systems compatible with this approach, continuous and reliable interaction between the user and the technology is highly desirable. However, these requirements are not easy to fulfill when using BCIs for the control of rehabilitation applications. Motor related BCI paradigms, such as motor imagery and motor intention, have been used extensively for BCI-controlled rehabilitation technologies (Khan et al., 2020). However, they're limited by the number of possible commands (Lotte et al., 2010) and BCI illiteracy (Lee et al., 2019), particularly for patients with severe disability (Rupp, 2014).

SSVEP and P300-based systems have proved to obtain higher classification performance and ITR than Motor-related BCI paradigms (Rupp, 2014). Hence, the importance of developing and studying visual BCI systems for these applications or combine them with motor paradigms, like the ones found on these SLR (Horki et al., 2011; Choi et al., 2016). For P300 BCIs, multiple repetitions (5 or more) of the whole stimuli sequence are typically needed to predict accurately the user's choice (Bianchi et al., 2021). Depending on the number of possible targets and interstimulus interval, the selection time for a single command can be relatively slow (tens of s) (Mainsah et al., 2015). Therefore, P300-based BCIs are not optimal for continuous control of actuators (Prosthesis, orthosis, FES, etc.) in the context of motor rehabilitation applications. Moreover, by its own nature, P300 BCIs operate in a synchronous way, a feature that restricts the operation of the system to certain times and cues indicated by the system. Thus, P300-based systems are often used to select and convey discrete and preprogrammed commands to the actuator, as those found in this SLR to control orthoses (Stan et al., 2015), VR systems (Rohani et al., 2014), or rehabilitation robots (Achanccaray et al., 2019). Interestingly, none of the analyzed papers combined a P300 BCI with an FES system, being an interesting possibility for future developments.

Regarding SSVEP BCIs, involving steady state signals they are suitable to implement asynchronous systems by continuously presenting the visual stimuli. In such case, the user could choose to perform a target selection task at any moment, and the system would be able to recognize it. In contrast to P300 BCIs, SSVEP BCIs have generally fewer possible targets, which correspond to the number of discernible frequencies, phases, and other features of the visual stimuli (and the evoked EEG signals). However, stimuli in SSVEP BCIs must be carefully designed since the

system must be capable to identify a zero-class (non-control) besides the classes associated to the actual commands. When this is not considered, false positives are very likely to occur, like Ortner et al. who reported an SSVEP-based BCI for the control of a hand orthosis (Ortner et al., 2011). Therefore, the orthosis often opened or closed when the user did not want to convey any control signal, since the flickering lights were still within their visual field. In contrast, this would not be an issue with a P300-based BCI, that requires cognitive engagement of the subject in the task, as discussed earlier.

Challenges and Future Directions

In this SRL, a large heterogeneity was identified in the reported BCI signals (P300, SSVEP or hybrid), applications (orthosis, prosthesis, FES, VR) and feature extraction methods, while the reported performance metrics were predominantly accuracy and ITR. Regarding classification methods, classical supervised ML algorithms (LDA and SVM) and some variations prevail, letting open the opportunity for the development of DL-based classification algorithms for visual BCI-based motor rehabilitation applications. The results of this work suggest the need to develop standard protocols for assessment of classification performance, when using VEP-based BCI systems for motor rehabilitation and assistive applications.

There are few reports of prototypes in pre-clinical stages of development with online tests. Therefore, there is a great opportunity to develop VEP-based BCI systems for motor rehabilitation. In this context, classification accuracy is a key metric to improve the BCI-user interaction and facilitate their adoption in clinical settings. Hence, strategies to improve the system's performance for users with low accuracy must be implemented, and the visual interfaces must be closely adapted to the user needs. Special attention should be paid to the visual stimulation module since stimulus patterns have a direct impact on the performance of P300 or SSVEP-based BCIs.

Also, it is important to investigate further the application of VR combined with BCI systems where patients can be stimulated simultaneously through multiple sensory modalities: visual, auditory, and somatosensory. That way, patients can have a richer experience while playing an active role in effective rehabilitation interventions, that could potentially help to improve and accelerate the motor recovery processes. Furthermore, it is essential to carry out pre-clinical studies and controlled interventions that include patients with different conditions such as stroke, ALS or SCI. Once those studies are performed and clinical scales are evaluated, it will be possible to validate the use of these systems in the clinic.

Finally, future works should focus on optimizing the implementation and training of artificial intelligence algorithms (especially DL-based methods) to enhance classification performance and achieve faster and more efficient online P300-based and SSVEP-based BCI systems. Only then, these systems could enhance their potential for the development of rehabilitation interventions aimed to help in the recovery of lost motor functions.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JG-M conceived and planned the SLR methodology, performed the data collection and filtering, analyzed the manuscript, and wrote the manuscript with input from the other authors. JM-G contributed to the conception and planning of the work, performed the analysis of the literature, contributed to the manuscript writing, discussed the results, and commented on the manuscript. BC-G helped in the conception of the work and to the manuscript writing, discussed the results, and commented on the manuscript. JR-T participated in the manuscript writing, discussed the results, and commented on the manuscript. AC-M aided in the analysis and filtering of the literature and commented on the results and the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Secretaría de Educación, Ciencia, Tecnología e Innovación de la Ciudad de México (SECTEI), through grant SECTEI/183/2019.

ACKNOWLEDGMENTS

To S. Omar Reyes-Acevedo, biomedical engineering student at Universidad La Salle, for his collaboration in the Identification Phase (1) of the SRL. And to Cinthya L. Toledo-Peral for her assistance with language revision and proofreading.

REFERENCES

Abdulaal, M. J., Casson, A. J., and Gaydecki, P. (2018). "Performance of nested vs. Non-nested SVM cross-validation methods in visual BCI: validation study," in *Proceeding of the European Signal Processing Conference. 2018 Nov 29;2018-September*, (IEEE), 1680–1684.

Abiri, R., Borhani, S., Sellers, E., Jiang, Y., and Zhao, X. (2019). A comprehensive review of EEG-based brain–computer interface paradigms. *J. Neural. Eng.* 16:011001. doi: 10.1088/1741-2552/aaf12e

Achanccaray, D., Chau, J. M., Pirca, J., Sepulveda, F., and Hayashibe, M. (2019). "Assistive robot arm controlled by a P300-based brain machine interface for daily activities," in *Proceeding of the 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, (IEEE), 1171–1174.

Aggarwal, S., and Chugh, N. (2019). Signal processing techniques for motor imagery brain computer interface: a review. *Array* 1-2, 100003. doi: 10.1016/j.array.2019.100003

Allison, B., Valbuena, D., Lueth, T., Teymourian, A., Volosyak, I., Graser, A., et al. (2010). BCI demographics: (How many (and what kinds of) people can use an

SSVEP BCI? *IEEE Trans. Neural. Syst. Rehabil. Eng.* 18:107. doi: 10.1109/tnsre.2009.2039495

Aloise, F., Aricò, P., Schettini, F., Riccio, A., Salinari, S., Mattia, D., et al. (2012). A covert attention P300-based brain-computer interface: geospell. *Ergonomics* 55, 538–551. doi: 10.1080/00140139.2012.661084

Altaf, S. (2019). Technological advancements in neuro-rehabilitation. *Rehabil. J.* 3, 105–106.

Amaral, C., Simões, M. A., Mouga, S., Andrade, J., and Castelo, M. (2017). A novel Brain Computer Interface for classification of social joint attention in Autism and comparison of 3 experimental setups: a feasibility study. *J. Neurosci. Methods* 290, 105–115. doi: 10.1016/j.jneumeth.2017.07.029

Andersen, S. K., and Müller, M. M. (2015). Driving steady-state visual evoked potentials at arbitrary frequencies using temporal interpolation of stimulus presentation. *BMC Neurosci.* 16:95. doi: 10.1186/s12868-015-0234-7

Arvaneh, M., Robertson, I. H., and Ward, T. E. A. (2019). P300-based brain-computer interface for improving attention. *Front. Hum. Neurosci.* 12:524. doi: 10.3389/fnhum.2018.00524

Bach-y-Rita, P., and Kercel, S. (2003). Sensory substitution and the human-machine interface. *Trends Cogn. Sci.* 7, 541–546. doi: 10.1016/j.tics.2003.10.013

Backus, D. (2008). Activity-based interventions for the upper extremity in spinal cord injury. *Top Spinal Cord Inj. Rehabil.* 13, 1–9. doi: 10.1310/sci1304-1

Baniqued, P. D. E., Stanyer, E. C., Awais, M., Alazmani, A., Jackson, A. E., Mon-Williams, M. A., et al. (2021). Brain–computer interface robotics for hand rehabilitation after stroke: a systematic review. *J. Neuro Eng. Rehabil.* 18:15. doi: 10.1186/s12984-021-00820-8

Bashashati, A., Fatourechi, M., Ward, R., Gary, E., and Birch, G. (2007). A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. *J. Neural. Eng.* 4, R32–R57.

Bento, V., Cruz, V., Ribeiro, D., and Cunha, J. (2012). The vibratory stimulus as a neurorehabilitation tool for stroke patients: proof of concept and tolerability test. *Neurorehabilitation* 30, 287–293. doi: 10.3233/NRE-2012-0757

Bhagat, A., Venkatakrishnan, A., Abibullaev, B., Artz, J., Yozbatiran, N., Blank, A. A., et al. (2016). Design and optimization of an EEG-Based Brain Machine Interface (BMI) to an Upper-limb exoskeleton for stroke survivors. *Front. Neurosci.* 10:122. doi: 10.3389/fnins.2016.00122

Bhattacharyya, S., Konar, A., and Tibarewala, D. N. (2014). Motor imagery, P300 and error-related EEG-based robot arm movement control for rehabilitation purpose. *Med. Biol. Eng. Comput.* 52, 1007–1017. doi: 10.1007/s11517-014-1204-4

Bianchi, L., Liti, C., Liuzzi, G., Piccialli, V., and Salvatore, C. (2021). Improving P300 speller performance by means of optimization and machine learning. *Ann. Oper. Res.* 2021, 1–39.

Birbaumer, N., and Cohen, L. (2007). Brain–computer interfaces: communication and restoration of movement in paralysis. *J. Physiol.* 579, 621–636. doi: 10.1113/jphysiol.2006.125633

Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Müller, K.-R. (2011). Single-trial analysis and classification of ERP components — A tutorial. *NeuroImage* 56, 814–825. doi: 10.1016/j.neuroimage.2010.06.048

Boernama, A. W. D., Setiawan, N. A., and Wahyunggoro, O. (2021). "Multiclass classification of brain-computer interface motor imagery system: a systematic literature review," in *AIMS 2021 - International Conference on Artificial Intelligence and Mechatronics Systems. 2021 Apr 28*, (IEEE). doi: 10.1007/s11517-021-02449-0

Bonci, A., Fiori, S., Higashi, H., Tanaka, T., and Verdini, F. (2021). An introductory tutorial on brain–computer interfaces and their applications. *Electronics* 10, 1–42. doi: 10.1002/9781119332428.ch1

Brunner, C., Allison, B. Z., Altstätter, C., and Neuper, C. (2011). A comparison of three brain-computer interfaces based on event-related desynchronization, steady state visual evoked potentials, or a hybrid approach using both signals. *J. Neural. Eng.* 8:025010. doi: 10.1088/1741-2560/8/2/025010

Carr, J. H., and Shephered, R. B. (2006). The changing face of neurological rehabilitation. *Rev. Bras Fisioter* 10, 147–156.

Carvalho, S. N., Costa, T. B. S., Uribe, L. F. S., Soriano, D. C., Yared, G. F. G., Coradine, L. C., et al. (2015). Comparative analysis of strategies for feature extraction and classification in SSVEP BCIs. *Biomed. Signal Process Control.* 21, 34–42.

Casey, A., Azhar, H., Grzes, M., and Sakel, M. (2019). BCI controlled robotic arm as assistance to the rehabilitation of neurologically disabled patients. *Disabil. Rehabil.: Assist. Technol.* 16, 525–537. doi: 10.1080/17483107.2019.1683239

Cecotti, H. (2011). A time–frequency convolutional neural network for the offline classification of steady-state visual evoked potential responses. *Pattern Recogn. Lett.* 32, 1145–1153. doi: 10.1016/j.patrec.2011.02.022

Cecotti, H., and Graser, A. (2011). Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Trans. Pattern Anal. Machine Intelligence* 33, 433–445. doi: 10.1109/tpami.2010.125

Chang, M. H., Baek, H. J., Lee, S. M., and Park, K. S. (2014). An amplitude-modulated visual stimulation for reducing eye fatigue in SSVEP-based brain–computer interfaces. *Clin. Neurophysiol.* 125, 1380–1391. doi: 10.1016/j.clinph.2013.11.016

Chen, J., Wu, C. H., Lin, Y., Kuo, Y., and Kuo, C. H. (2020). "Mechatronic implementation and trajectory tracking validation of a BCI-based human-wheelchair interface," in *Proceeding of the 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, (IEEE).

Chen, X., Wang, Y., Nakanishi, M., Gao, X., Jung, T. P., and Gao, S. (2015). High-speed spelling with a noninvasive brain–computer interface. *Proc. Natl Acad. Sci. U.S.A.* 112, E6058–E6067. doi: 10.1073/pnas.1508080112

Chen, X., Zhao, B., Wang, Y., and Gao, X. (2019). Combination of high-frequency SSVEP-based BCI and computer vision for controlling a robotic arm. *J. Neural. Eng.* 16:026012. doi: 10.1088/1741-2552/aaf594

Chen, X., Zhao, B., Wang, Y., Xu, S., and Gao, X. (2018). Control of a 7-DOF robotic arm system with an SSVEP-based BCI. *Int. J. Neural. Syst.* 28:1850018. doi: 10.1142/S0129065718500181

Cheng, M., Gao, X. R., Gao, S. G., and Xu, D. F. (2002). Design and implementation of a brain-computer interface with high transfer rates. *IEEE Trans. Biomed. Eng.* 49, 1181–1186. doi: 10.1109/tbme.2002.803536

Choi, I., Bond, K., and Nam, C. (2016). "A hybrid BCI-controlled FES system for hand-wrist function," in *Proceeding of the IEEE International conference on Systems, Man and Cybernetics*, (IEEE).

Choi, J., Kim, K. T., Jeong, J. H., Kim, L., Lee, S. J., and Kim, H. (2020). Developing a motor imagery-based real-time asynchronous hybrid BCI controller for a lower-limb exoskeleton. *Sensors (Basel)* 20, 1–15. doi: 10.3390/s20247309

Choi, K. M., Park, S., and Im, C. H. (2019). Comparison of visual stimuli for steady-state visual evoked potential-based brain-computer interfaces in virtual reality environment in terms of classification accuracy and visual comfort. *Comput. Intell. Neurosci.* 2019:9680697. doi: 10.1155/2019/9680697

Chu, Y., Zhao, X., Zou, Y., Xu, W., and Zhao, Y. (2018). "Robot-assisted rehabilitation system based on SSVEP brain-computer interface for upper extremity," in *Proceeding of the IEEE International Conference on Robotics and Biomimetics (ROBIO)*, (IEEE), 1098–1103.

Çiğ, H., Hanbay, D., and Tüysüz, F. (2017). "Robot arm control with for SSVEP-based brain signals in brain computer interface," in *Proceeding of the International Artificial Intelligence and Data Processing Symposium (IDAP)*, (IEEE).

Craik, A., and Contreras, J. L. (2019). Deep learning for Electroencephalogram (EEG) classification tasks: a review. *J. Neural. Eng.* 16:031001. doi: 10.1088/1741-2552/ab0ab5

Delijorge, J., Mendoza, O., Gordillo, J., and Antelis, J. (2020). Evaluation of a P300-Based brain-machine interface for a robotic hand-orthosis control. *Front. Neurosci.* 14:589659. doi: 10.3389/fnins.2020.589659

Diez, P. F., Mut, V. A., Avila, E. M., and Laciar, E. (2011). Asynchronous BCI control using high-frequency SSVEP. *J. Neuroeng. Rehabil.* 8:39. doi: 10.1186/1743-0003-8-39

Ding-Guo, Y., and Ying, W. (2012). Study on brain-computer interface controlled functional electrical stimulation system for lower limbs. *Chinese J. Biomed. Eng.* 5:008.

Do, A. H., Wang, P. T., King, C. E., Abiri, A., and Nenadic, Z. (2011). Brain-computer interface controlled functional electrical stimulation system for ankle movement. *J. Neuro Eng. Rehabil.* 8:49. doi: 10.1186/1743-0003-8-49

Dreyer, A. M., Herrmann, C. S., and Rieger, J. W. (2017). Tradeoff between User Experience and BCI classification accuracy with frequency modulated steady-state visual evoked potentials. *Front. Hum. Neurosci.* 11:391. doi: 10.3389/fnhum.2017.00391

Duvinage, M., Castermans, T., Jiménez, R., and Hoellinger, T. (2012). "A five-state P300-based foot lifter orthosis: proof of concept," in *Proceeding of the 2012 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living*, (IEEE), 1–6.

Edlinger, G., Holzner, C., and Guger, C. (2011). "A hybrid brain-computer interface for smart home control," in *Proceeding of the International Conference on Human-Computer Interaction. Human-Computer Interaction. Interact. Techniques and Environments*, (Berlin: Springer), 417–426.

Erkan, E., and Akbaba, M. (2018). A study on performance increasing in SSVEP based BCI application. *Eng. Sci. Technol. Int. J.* 21, 421–427.

Galińska, E. (2015). Music therapy in neurological rehabilitation settings. *Psychiatr. Pol.* 49, 835–846.

Gao, S., Wang, Y., Gao, X., and Hong, B. (2014). Visual and auditory brain–computer interfaces. *IEEE Trans. Biomed. Eng.* 61, 1436–1447.

Giménez, D. A., Arguissain, F. G., and Tabernig, C. B. (2011). "Interfaz BCI-FES para rehabilitación neurológica: resultados preliminares." in *Proceedings of the XVIII Congreso Argentino de Bioingeniería SABI 2011 – VII Jornadas de Ingeniería Clínica* (Mar del Plata: IEEE).

Gordon, J. (2005). "A top-down model for neurologic rehabilitation," in *Proceeding of the Linking Movement Science and Intervention. Proceedings III Step Conference*, (American Physical Therapy Association), 30–33.

Guger, C., Allison, B., Grobwindhager, B., Prückl, R., Hintermüller, C., Kapeller, C., et al. (2012). How many people could use an SSVEP BCI? *Front. Neurosci.* 6:169. doi: 10.3389/fnins.2012.00169

Gui, K., Ren, Y., and Zhang, D. (2015). "Online brain-computer interface controlling robotic exoskeleton for gait rehabilitation," in *Proceeding of the IEEE International Conference on Rehabilitation Robotics*, (IEEE), 931–936.

Guo, M., Jin, J., Jiao, Y., Wang, X., and Cichockia, A. (2019). Investigation of visual stimulus with various colors and the layout for the oddball paradigm in evoked related potential-based brain–computer interface. *Front. Comput. Neurosci.* 13:24. doi: 10.3389/fncom.2019.00024

Han, C., Xu, G., Xie, J., Chen, C., and Zhang, S. (2018). Highly interactive brain-computer interface based on flicker-free steady-state motion visual evoked potential. *Sci. Rep.* 8:5835.

Hara, Y. (2008). Neurorehabilitation with new functional electrical stimulation for hemiparetic upper extremity in stroke patients. *J. Nippon Med. Sch.* 75, 4–14. doi: 10.1272/jnms.75.4

Hindle, K., Whitcomb, T., Briggs, W., and Hong, J. (2012). Proprioceptive Neuromuscular Facilitation (PNF): its mechanisms and effects on range of motion and muscular function. *J. Hum. Kinetics* 31, 105–113. doi: 10.2478/v10078-012-0011-y

Hoffmann, U., Vesin, J. M., Ebrahimi, T., and Diserens, K. (2008). An efficient P300-based brain-computer interface for disabled subjects. *J. Neurosci. Methods* 167, 115–125. doi: 10.1016/j.jneumeth.2007.03.005

Horki, P., Solis-Escalante, T., and Neuper, C. (2011). Combined motor imagery and SSVEP based BCI control of a 2 DoF artificial upper limb. *Med. Biol. Eng. Comput.* 49, 567–577. doi: 10.1007/s11517-011-0750-2

Hossain, T., Rakshit, A., and Konar, A. (2020). "Brain-computer interface based user authentication system for personal device security (Domotic assistence)," in *Proceeding of the International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, (IEEE).

Huang, T. H., Huang, H. P., Liu, Y. H., Kang, Z. H., and Kuan, J. Y. (2013). Development of a Brain-Controlled Rehabilitation System (BCRS). *J. Neurosci. Neuroeng.* 2, 79–89. doi: 10.1166/jnsne.2013.1042

Huang, W., and Huang, Z. (2017). "A real-time distributed computing mechanism for P300 speller BCI," in *Proceeding of the 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, (IEEE).

Huang, Y., Qu, J., Xiao, J., and Li, K. (2019). "A novel P300-based interactive method for virtual reality system," in *Proceeding of the WRC Symposium on Advanced Robotics and Automation*, (IEEE), 309–314.

Hwang, H. J., Kim, S., Choi, S., and Im, C. H. (2013). EEG-based brain-computer interfaces: a thorough literature survey. *Int. J. Hum.-Comput. Interact.* 29, 814–826. doi: 10.1080/10447318.2013.780869

Iosa, M., Morone, G., Fusco, A., Bragoni, M., Coiro, P., Multari, M., et al. (2012). Seven capital devices for the future of stroke rehabilitation. *Stroke Res. Treat.* 2012:187965. doi: 10.1155/2012/187965

Jang, S. (2013). Motor function-related maladaptive plasticity in stroke: a review. *NeuroRehabilitation* 32, 311–316. doi: 10.3233/NRE-130849

Kaplan, A. Y., Zhigulskaya, D. D., and Kiriyanov, D. A. (2016). Studying the ability to control human phantom fingers in P300 brain-computer interface. *Bull. Russian State Med. Univ.* 2, 24–28. doi: 10.24075/brsmu.2016-02-0

Katyal, A., and Singla, R. (2020). A novel hybrid paradigm based on steady state visually evoked potential & P300 to enhance information transfer rate. *Biomed. Signal Process. Control* 59:101884. doi: 10.1016/j.bspc.2020.101884

Kaufmann, T., Schulz, S. M., Grünzinger, C., and Kübler, A. (2011). Flashing characters with famous faces improves ERP-based brain-computer interface performance. *J. Neural Eng.* 8:056016. doi: 10.1088/1741-2560/8/5/056016

Kaufmann, T., Holz, E., and Kübler, A. (2013). Comparison of tactile, auditory, and visual modality for brain-computer interface use: a case study with a patient in the locked-in state. *Front. Neurosci.* 7:129. doi: 10.3389/fnins.2013.00129

Khadijah, N., Aznan, N., Connolly, J., Moubayed, N., and Breckon, T. (2019). "Using variable natural environment brain-computer interface stimuli for real-time humanoid robot navigation," in *Proceeding of the International Conference on Robotics and Automation (ICRA)*, (IEEE). doi: 10.3724/sp.j.1218.2011.00129

Khan, M. A., Das, R., Iversen, H. K., and Puthusserypady, S. (2020). Review on motor imagery based BCI systems for upper limb post-stroke neurorehabilitation: from designing to application. *Comput. Biol. Med.* 23:103843. doi: 10.1016/j.compbiomed.2020.103843

Kluge, T., and Hartmann, M. (2007). "Phase coherent detection of steady-state evoked potentials: experimental results and application to brain-computer interfaces," in *Proceeding of the 3rd International IEEE/EMBS Conference on Neural Engineering*, (IEEE), 425–429.

Koo, B., Lee, H., Nam, Y., and Choi, S. (2015). "Immersive BCI with SSVEP in VR head-mounted display," in *Proceeding of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (IEEE), 1103–1106. doi: 10.1109/EMBC.2015.7318558

Kubler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J. R., and Birbaumer, N. (2001). Brain–computer communication: unlocking the locked. *Psychol. Bull.* 127, 358–375. doi: 10.1037/0033-2909.127.3.358

Kundu, S., and Ari, S. (2017). P300 Detection with brain–computer interface application using PCA and ensemble of weighted SVMs. *IETE J. Res.* 64, 406–414. doi: 10.1080/03772063.2017.1355271

Kuś, R., Duszyk, A., Milanowski, P., Łabęcki, M., Bierzyńska, M., Radzikowska, Z., et al. (2013). On the quantification of SSVEP frequency responses in human EEG in realistic BCI conditions. *PLoS One* 8:e77536. doi: 10.1371/journal.pone.0077536

Kutas, M., McCarthy, G., and Donchin, E. (1977). Augmenting mental chronometry: the P300 as a measure of stimulus evaluation time. *Science* 197, 792–795. doi: 10.1126/science.887923

Kwak, N., Müller, K., and Lee, S. (2017). A convolutional neural network for steady state visual evoked potential classification under ambulatory environment. *PLoS One* 12:e0172578. doi: 10.1371/journal.pone.0172578

Kwak, N.-S., Müller, K.-R., and Lee, S.-W. (2015). A lower limb exoskeleton control system based on steady state visual evoked potentials. *J. Neural. Eng.* 12:056009. doi: 10.1088/1741-2560/12/5/056009

Lazarou, I., Nikolopoulos, S., Petrantonakis, P., Kompatsiaris, I., and Tsolaki, M. (2018). EEG-Based brain–computer interfaces for communication and rehabilitation of people with motor impairment: a novel approach of the 21st century. *Front. Hum. Neurosci.* 12:14. doi: 10.3389/fnhum.2018.00014

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444.

Lee, M.-H., Kwon, O.-Y., Kim, Y.-J., Kim, H.-K., Lee, Y.-E., Williamson, J., et al. (2019). EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy. *Gigascience* 8, 1–16. doi: 10.1093/gigascience/giz002

Lee, T., Kim, M., and Kim, S. (2020). "Data augmentation effects using borderline-SMOTE on classification of a P300-based BCI (Control home appliances)," in

Proceeding of the 2020 8th International Winter Conference on Brain-Computer Interface (BCI), (IEEE).

Lee, J., Won, K., Kwon, M., Jun, S. C., and Ahn, M. (2020). CNN with large data achieves true zero-training in online P300 brain-computer interface. *IEEE Access* 8, 74385–74400. doi: 10.1109/access.2020.2988057

Li, R., Zhang, X., Li, H., Zhang, L., Lu, Z., and Chen, J. (2018). An approach for brain-controlled prostheses based on Scene Graph Steady-State Visual Evoked Potentials. *Brain Res.* 1692, 142–153. doi: 10.1016/j.brainres.2018.05.018

Li, S., Jin, J., Daly, I., Zuo, C., Wang, X., and Cichocki, A. (2020). Comparison of the ERP-based BCI performance among chromatic (RGB) semitransparent face patterns. *Front. Neurosci.* 14:54. doi: 10.3389/fnins.2020.00054

Lin, Z., Zhang, C., Wu, W., and Gao, X. (2007). Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs. *IEEE Trans. Biomed. Eng.* 54, 1172–1176.

Lin, Z., Zhang, C., Zeng, Y., Tong, L., and Yan, B. (2018). A novel P300 BCI speller based on the Triple RSVP paradigm. *Sci. Rep.* 8:3350.

Liu, B., Huang, X., Wang, Y., Chen, X., and Gao, X. (2020). BETA: a large benchmark database toward SSVEP-BCI application. *Front. Neurosci.* 14:627. doi: 10.3389/fnins.2020.00627

Liu, M., Wu, W., Gu, Z., Yu, Z., Qi, F., and Li, Y. (2018). Deep learning based on batch normalization for P300 signal detection. *Neurocomputing* 275, 288–297. doi: 10.1109/TBME.2018.2875024

Lopes, A. C., Rodrigues, J., Perdigão, J., Pires, G., and Nunes, U. J. (2016). A new hybrid motion planner. *IEEE Robot. Autom. Mag.* 23, 82–93.

Lotte, F., Bougrain, L., Cichocki, A., and Clerc, M. (2018). A review of classification algorithms for EEG-based brain computer interfaces: a 10 year update. *J. Neural. Eng.* 15:1005. doi: 10.1088/1741-2552/aab2f2

Lotte, F., van Langhenhove, A., Lamarche, F., Ernest, T., Renard, Y., Arnaldi, B., et al. (2010). Exploring large virtual environments by thoughts using a brain–computer interface based on motor imagery and high-level commands. *Presence Teleoperators Virtual Environ.* 19, 54–70. doi: 10.1162/pres.19.1.54

Mainsah, B. O., Collins, L. M., Colwell, K. A., Sellers, E. W., Ryan, D. B., Caves, K., et al. (2015). Increasing BCI communication rates with dynamic stopping towards more practical use: an ALS study. *J. Neural. Eng.* 12:016013. doi: 10.1088/1741-2560/12/1/016013

Manor, R., and Geva, A. (2015). Convolutional neural network for multi-category rapid serial visual presentation BCI. *Front. Computat. Neurosci.* 9:146. doi: 10.3389/fncom.2015.00146

McCabe, J., Monkiewicz, M., Holcomb, J., Pundik, J., and Daly, J. J. (2015). Comparison of robotics, functional electrical stimulation, and motor learning methods for treatment of persistent upper extremity dysfunction after stroke: a randomized controlled trial. *Phys. Med. Rehabil.* 96, 981–990. doi: 10.1016/j.apmr.2014.10.022

McCane, L. M., Heckman, S. M., McFarland, D. J., Townsend, G., Mak, J. N., Sellers, E. W., et al. (2015). P300-based brain-computer interface (BCI) event-related potentials (ERPs): people with amyotrophic lateral sclerosis (ALS) vs. age-matched controls. *Clin. Neurophysiol.* 126, 2124–2131. doi: 10.1016/j.clinph.2015.01.013

McCann, M. T., Thompson, D. E., Syed, Z. H., and Huggins, J. E. (2015). Electrode subset selection methods for an EEG-based P300 brain-computer interface. *Disabil. Rehabil. Assist. Technol.* 10:216.

Meng, J., Zhang, S., Bekyo, A., Olsoe, J., Baxter, B., and He, B. (2016). Noninvasive electroencephalogram based control of a robotic arm for reach and grasp tasks. *Sci. Rep.* 6:3856518.

Müller-Putz, G. (2018). *The MoreGrasp Project*. Graz: University of Technology. Institute of Neural Engineering. Laboratory of Brain-Computer Interfaces.

Müller-Putz, G. R., Scherer, R., Brauneis, C., and Pfurtscheller, C. (2005). Steady-state visual evoked potential (SSVEP)-based communication: impact of harmonic frequency components. *J. Neural. Eng.* 2, 123–130. doi: 10.1088/1741-2560/2/4/008

Naeem, M., Kamran, M., Kang, S., Choi, H., and Yung, M. (2020). A hybrid speller design using eye tracking and SSVEP brain–computer interface. *Sensors* 20:891. doi: 10.3390/s20030891

Nagel, S., Rosenstiel, W., and Spüler, M. (2017). "Random Visual Evoked Potentials for brain-computer interface control," in *Proceeding of the 7th Graz Brain-Computer Interface Conference*, (Graz).

Nagel, S., and Spüler, M. (2019). World's fastest brain-computer interface: combining EEG2Code with deep learning. *PLoS One* 14:e0221909. doi: 10.1371/journal.pone.0221909

Nakanishi, M., Wang, Y., Chen, X., Wang, Y. T., Gao, X., and Jung, T. P. (2018). Enhancing detection of SSVEPs for a high-speed brain speller using task-related component analysis. *IEEE Trans. Biomed. Eng.* 65, 104–112. doi: 10.1109/tbme.2017.2694818

Nooh, A. A., Yunus, J., and Daud, S. M. (2011). "A review of asynchronous electroencephalogram-based brain computer interface systems," in *Proceeding of the International Conference on Biomedical Engineering and Technology (ICBET 2011)*, (Singapore).

Norcia, A. M., Appelbaum, L. G., Ales, J. M., Cottereau, B. R., and Rossion, B. (2015). The steady-state visual evoked potential in vision research: a review. *J. Vis.* 15:4.

Okahara, Y., Takano, K., Nagao, M., Kondo, K., Iwadate, Y., Birbaumer, N., et al. (2018). Long-term use of a neural prosthesis in progressive paralysis. *Sci. Rep.* 8:16787.

Ortner, R., Allison, B. Z., Korisek, G., Gaggl, H., and Pfurtscheller, G. (2011). An SSVEP BCI to control a hand orthosis for persons with tetraplegia. *IEEE Trans. Neural. Syst. Rehabil. Eng.* 19, 1–5. doi: 10.1109/TNSRE.2010.2076364

Osuagwu, B. C., Wallace, L., Fraser, M., and Vuckovic, A. (2016). Rehabilitation of hand in subacute tetraplegic patients based on brain computer interface and functional electrical stimulation: a randomised pilot study. *J. Neural. Eng.* 13:065002. doi: 10.1088/1741-2560/13/6/065002

Page, M. J., McKenezie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372:n71. doi: 10.1136/bmj.n71

Perge, J. A., Homer, M. L., Malik, W. Q., Cash, S., Eskandar, E., Friehs, G., et al. (2013). Intraday signal instabilities affect decoding performance in an intracortical neural interface system. *J. Neural. Eng.* 10:36004. doi: 10.1088/1741-2560/10/3/036004

Perlstein, W., Cole, M., Larson, M., Kelly, K., Seignourel, P., and Keil, A. (2003). Steady-state visual evoked potentials reveal frontally-mediated working memory activity in humans. *Neurosci. Lett.* 342, 191–195. doi: 10.1016/s0304-3940(03)00226-x

Philip, J. T., and George, S. T. (2020). Visual P300 mind-speller brain-computer interfaces: a walk through the recent developments with special focus on classification algorithms. *Clin. EEG Neurosci.* 51, 19–33. doi: 10.1177/1550059419842753

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Radaman, R. A., and Vasilakos, A. V. (2017). Brain computer interface: control signals review. *Neurocomputing* 223, 26–44. doi: 10.1016/j.neucom.2016.10.024

Ramos, A., Broetz, D., Rea, M., and Läer, L. (2013). Brain–machine interface in chronic stroke rehabilitation: a controlled study. *Ann. Neurol.* 74, 100–108. doi: 10.1002/ana.23879

Ravi, A., Pearce, S., Zhang, X., and Jiang, N. (2019). "User-specific channel selection method to improve SSVEP BCI decoding robustness against variable inter-stimulus distance," in *Proceeding of the International IEEE/EMBS Conference on Neural Engineering, NER. IEEE Computer Society*, (IEEE), 283–286.

Reichert, C., Tellez Ceja, I. F., Sweeney-Reed, C. M., Heinze, H. J., Hinrichs, H., and Dürschmid, S. (2020). Impact of stimulus features on the performance of a gaze-independent brain-computer interface based on covert spatial attention shifts. *Front. Neurosci.* 14:591777. doi: 10.3389/fnins.2020.591777

Riggins, T., and Scott, L. S. (2019). P300 development from infancy to adolescence. *Psychophysiology* 57:e13346. doi: 10.1111/psyp.13346

Rohani, D. A., Sorensen, H., and Puthusserypady, S. (2014). Brain-computer interface using P300 and virtual reality: a gaming approach for treating ADHD. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2014:3606. doi: 10.1109/EMBC.2014.6944403

Ron-Angevin, R., Garcia, L., Fernández-Rodríguez, A., Saracco, J., André, J. M., Lespinet-Najib, V., et al. (2019). Impact of speller size on a visual P300 brain-computer interface (BCI) system under two conditions of constraint for eye movement. *Comput. Intell. Neurosci.* 2019:7876248. doi: 10.1155/2019/7876248

Rupp, R. (2014). Challenges in clinical applications of brain computer interfaces in individuals with spinal cord injury. *Front. Neuroeng.* 7:38. doi: 10.3389/fneng.2014.00038

Sakurada, T., Kawase, T., Takano, K., Komatsu, T., and Kansaku, K. (2013). A BMI-based occupational therapy assist suit: asynchronous control by SSVEP. *Front. Neurosci.* 7:172. doi: 10.3389/fnins.2013.00172

Savić, A., Kisic, U., and Popović, M. (2012). "Toward a Hybrid BCI for Grasp Rehabilitation," in *Proceeding of the 5th European Conference of the International Federation for Medical and Biological Engineering. IFMBE Proceedings*, Vol. 37, (Springer), 806–809. doi: 10.1088/1741-2552/aac1a1

Sellers, E. W., Vaughan, T. M., and Wolpaw, J. R. (2010). A brain-computer interface for long-term independent home use. *Amyotroph Lateral Scler* 11, 449–455. doi: 10.3109/17482961003777470

Shan, H., Liu, Y., and Stefanov, T. (2018). "A simple convolutional neural network for accurate P300 detection and character spelling in brain computer interface," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, (IEEE), 1604–1610.

Shyu, K. K., Chiu, Y. J., Lee, P. L., Liang, J. M., and Peng, S. H. (2013). Adaptive SSVEP-based BCI system with frequency and pulse duty-cycle stimuli tuning design. *IEEE Trans. Neural. Syst. Rehabil. Eng.* 21, 697–703. doi: 10.1109/TNSRE.2013.2265308

Son, J. E., Choi, H., Lim, H., and Ku, J. (2020). Development of a flickering action video based steady state visual evoked potential triggered brain computer interface-functional electrical stimulation for a rehabilitative action observation game. *Technol. Health Care* 28, 509–519. doi: 10.3233/THC-209051

Sozer, A. T. (2018). "Enhanced single channel SSVEP detection method on benchmark dataset," in *15th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE). 2018 Nov 13*, (IEEE).

Speier, W., Deshpande, A., Cui, L., Chandravadia, N., Roberts, D., and Pouratian, N. (2017). A comparison of stimulus types in online classification of the P300 speller using language models. *PLoS One* 12:e175382. doi: 10.1371/journal.pone.0175382

Stan, A., Irimia, D., Botezatu, N., and Lupu, R. (2015). "Controlling a hand orthosis by means of P300-based brain computer interface," in *Proceeding of the Conference E-Health and Bioengineering Conference*, (IEEE).

Su, Y., Qi, Y., Luo, J., Wu, B., Yang, F., and Li, Y. (2011). A hybrid brain-computer interface control strategy in a virtual environment. *J. Zhejiang Univ. Sci.* 12, 351–361. doi: 10.1631/jzus.c1000208

Takano, K., Komatsu, T., Hata, N., Nakajima, Y., and Kansaku, K. (2009). Visual stimuli for the P300 brain-computer interface: a comparison of white/gray and green/blue flicker matrices. *Clin. Neurophysiol.* 120, 1562–1566. doi: 10.1016/j.clinph.2009.06.002

Takeuchi, N., and Izumi, S. (2012). Maladaptive plasticity for motor recovery after stroke: mechanisms and approaches. *Neural. Plasticity* 2012:359728.

Thomas, E., Dyson, M., and Clerc, M. (2013). An analysis of performance evaluation for motor-imagery based BCI. *J. Neural. Eng.* 10:031001. doi: 10.1088/1741-2560/10/3/031001

Thomas, J., Maszczyk, T., Sinha, N., Kluge, T., and Dauwels, J. (2017). "Deep learning-based classification for brain-computer interfaces," in *Proceeding of the IEEE International Conference on Systems, Man, and Cybernetics*, (IEEE), 234–239.

Tidoni, E., Abu-Alqumsan, M., Leonardis, D., Kapeller, C., and Fusco, G. (2017). Local and remote cooperation with virtual and robotic agents: a P300 BCI study in healthy and people living with spinal cord injury. *IEEE Trans. Neural Syst. Rehabil. Eng.* 25, 1622–1632. doi: 10.1109/TNSRE.2016.2626391

Touyama, H., and Sakuda, M. (2017). Online control of a virtual object with collaborative SSVEP. *J. Adv. Comput. Intell. Intell. Inform.* 21, 1291–1297. doi: 10.20965/jaciii.2017.p1291

Turnip, A., Simbolon, A., Amri, F., and Agung, M. (2015). "Utilization of EEG-SSVEP method and ANFIS classifier for controlling electronic wheelchair," in *Proceeding of the International Conference on Technology, Informatics, Management, Engineering & Environment (TIME-E)*, (IEEE).

van Dokkum, L. E., Ward, T., and Laffont, I. (2015). Brain computer interfaces for neurorehabilitation – its current status as a rehabilitation strategy post-stroke. *Ann. Phys. Rehabil. Med.* 58, 3–8. doi: 10.1016/j.rehab.2014.09.016

Venuto, D., Annese, V., and Mezzina, G. (2017). "An embedded system remotely driving mechanical devices by P300 brain activity (control drive car)," in *Proceeding of the Design, Automation & Test in Europe Conference & Exhibition (DATE)*, (IEEE).

Venuto, D., and Mezzina, G. (2018). "User-centered ambient assisted living: brain environment interface," in *Proceeding of the 7th Mediterranean Conference on Embedded Computing (MECO)*, (IEEE).

Vialatte, F., Maurice, M., Dauwels, J., and Cichocki, A. (2010). Steady-state visually evoked potentials: focus on essential paradigms and future perspectives. *Progr. Neurobiol.* 90, 418–438. doi: 10.1016/j.pneurobio.2009.11.005

Volosyak, I., Valbuena, D., Lüth, T., Malechka, T., and Gräser, A. (2011). BCI demographics II: how many (and what kinds of) people can use a high-frequency SSVEP BCI? *IEEE Trans. Neural. Syst. Rehabil. Eng.* 19, 232–239. doi: 10.1109/TNSRE.2011.2121919

Walter, A., Quigley, C., Andersen, S. K., and Mueller, M. M. (2012). Effects of overt and covert attention on the steady-state visual evoked potential. *Neurosci. Lett.* 519, 37–41. doi: 10.1016/j.neulet.2012.05.011

Wang, Q., Lu, G., Pei, Z., Tang, C., Xu, L., Wang, Z., et al. (2020). "P300 recognition based on ensemble of SVMs: - BCI controlled robot contest of 2019 world robot conference," in *Proceeding of the 39th Chinese Control Conference (CCC)*, (IEEE).

Wang, Y., Gao, X., Hong, B., Jia, C., and Gao, S. (2008). *Brain-Computer Interfaces*. IEEE Engineering Medicine Biology Magazine. (IEEE), 64–71.

Wang, Y., Wang, R., Gao, X., and Gao, S. (2005). "Brain–computer interface based on the high-frequency steady-state visual evoked potential," in *Proceeding of the 1st International Conference on Neural Interface and Control*, (IEEE), 37–39.

Waytowich, N., and Krusienski, D. (2017). "Development of an extensible SSVEP-BCI software platform and application to wheelchair control," in *Proceeding of the 8th International IEEE/EMBS Conference on Neural Engineering (NER)*, (IEEE).

Wolpaw, J., Birbaumer, N., Heetderks, W., McFarland, D. J., Peckham, P. H., Schalk, G., et al. (2000). Brain–computer interface technology: a review of the first international meeting. *IEEE Trans. Rehabil. Eng.* 8, 164–173. doi: 10.1109/tre.2000.847807

Won, D.-O., Hwang, H.-J., Dähne, S., Müller, K.-R., and Lee, S.-W. (2015). Effect of higher frequency on the classification of steady-state visual evoked potentials. *J. Neural. Eng.* 13:016014. doi: 10.1088/1741-2560/13/1/016014

Xu, Y., Wu, Q., Chen, B., and Chen, X. (2021). SSVEP-based active control of an upper limb exoskeleton using a low-cost brain–computer interface. *Ind. Rob.* doi: 10.1108/IR-03-2021-0062 [Online ahead of print].

Yao, L., Zhang, D., Huang, G., and Zhu, X. (2011). "Using SSVEP based brain-computer interface to control functional electrical stimulation training system," in *Proceeding of the IEEE 5th International Conference on Cybernetics and Intelligent Systems (CIS)*, (IEEE).

Yao, L., Zhang, D., and Zhu, X. (2012). "SSVEP based brain–computer interface controlled functional electrical stimulation system for knee joint movement," in *Intelligent Robotics and Applications. Lecture Notes in Computer Science*, eds C. Y. Su, S. Rakheja, and H. Liu (Berlin: Springer), 526–535.

Yao, Z., Wang, Y., Yang, C., Pei, W., Gao, X., and Chen, H. (2019). An online brain-computer interface in mobile virtual reality environments. *Integr. Comput. Aided Eng.* 26, 345–360. doi: 10.3233/ICA-180586

Yeom, S.-K., Fazli, S., Müller, K.-R., and Lee, S.-W. (2014). An efficient ERP-based brain-computer interface using random set presentation and face familiarity. *PLoS One.* 9:111157. doi: 10.1371/journal.pone.0111157

Yin, E., Zhou, Z., and Jian, Z. (2013). A novel hybrid BCI speller based on the incorporation of SSVEP into the P300 paradigm. *J. Neural. Eng.* 10:026012. doi: 10.1088/1741-2560/10/2/026012

Yu, Y., Zhou, Z., Liu, Y., Jiang, J., Yin, E., Zhang, N., et al. (2017). "Self-Paced operation of a wheelchair based on a hybrid brain-computer interface combining motor imagery and P300 potential," in *Proceeding of the IEEE Transactions on Neural Systems and Rehabilitation Engineering*, (IEEE), 25. doi: 10.1109/TNSRE.2017.2766365

Yuan, Y., Li, Z., and Liu, Y. (2018). "Brain teleoperation of a mobile robot using deep learning technique," in *Proceeding of the 3rd International Conference on Advanced Robotics and Mechatronics (ICARM)*, (IEEE).

Zbogar, D., Eng, J., Miller, W., Krassioukov, A. V., and Verrier, M. C. (2017). Movement repetitions in physical and occupational therapy during spinal

cord injury rehabilitation. *Spinal. Cord.* 55, 172–179. doi: 10.1038/sc.2016.129

Zhang, D., Maye, A., Gao, X., Hong, B., Engel, A. K., and Gao, S. (2010). An independent brain-computer interface using covert non-spatial visual selective attention. *J. Neural. Eng.* 7:16010.

Zhang, R., Li, Y., Yan, Y., Zhang, H., and Wu, S. (2014). "An intelligent wheelchair based on automated navigation and BCI techniques," in *Proceeding of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (IEEE).

Zhang, R., Li, Y., Yan, Y., Zhang, H., Wu, S., Yu, T., et al. (2016). "Control of a wheelchair in an indoor environment based on a brain–computer interface and automated navigation," in *Proceeding of the IEEE Transactions on Neural Systems and Rehabilitation Engineering*, (IEEE), 24. doi: 10.1109/TNSRE.2015.2439298

Zhao, S., Xu, P., Li, Z., and Su, C. H. (2015). "Brain-actuated teleoperation control of a mobile robot," in *Proceeding of the IEEE International Conference on Robotics and Biomimetics (ROBIO)*, (IEEE).

Zhao, X., Chu, Y., Han, J., and Zhang, Z. (2016). SSVEP-based brain-computer interface controlled functional electrical stimulation system for upper extremity rehabilitation. *IEEE Trans. Syst. Man Cybernetics: Syst.* 46, 947–956.

Zhu, Y., Li, Y., Lu, J., and Li, P. (2020). A hybrid BCI based on SSVEP and EOG for robotic arm control. *Front. Neurorobot.* 14:583641. doi: 10.3389/fnbot.2020.583641

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership