Artificial Intelligence In Medicine 147 (2024) 102738

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

An in-depth survey on Deep Learning-based Motor Imagery Electroencephalogram (EEG) classification

Xianheng Wang [a,*], Veronica Liesaputra [a], Zhaobin Liu [b], Yi Wang [c], Zhiyi Huang [a]

[a] Department of Computer Science, University of Otago, Dunedin, New Zealand
[b] College of Information Science and Technology, Dalian Maritime University, Liaoning, PR China
[c] Laboratory for Circuit and Behavioral Physiology, RIKEN Center for Brain Science, Wako-shi, Saitama, Japan

ARTICLE INFO

Keywords:
Motor imagery electroencephalogram
classification
Deep learning
Survey

ABSTRACT

Electroencephalogram (EEG)-based Brain–Computer Interfaces (BCIs) build a communication path between human brain and external devices. Among EEG-based BCI paradigms, the most commonly used one is motor imagery (MI). As a hot research topic, MI EEG-based BCI has largely contributed to medical fields and smart home industry. However, because of the low signal-to-noise ratio (SNR) and the non-stationary characteristic of EEG data, it is difficult to correctly classify different types of MI-EEG signals. Recently, the advances in Deep Learning (DL) significantly facilitate the development of MI EEG-based BCIs. In this paper, we provide a systematic survey of DL-based MI-EEG classification methods. Specifically, we first comprehensively discuss several important aspects of DL-based MI-EEG classification, covering input formulations, network architectures, public datasets, etc. Then, we summarize problems in model performance comparison and give guidelines to future studies for fair performance comparison. Next, we fairly evaluate the representative DL-based models using source code released by the authors and meticulously analyse the evaluation results. By performing ablation study on the network architecture, we found that (1) effective feature fusion is indispensable for multi-stream CNN-based models. (2) LSTM should be combined with spatial feature extraction techniques to obtain good classification performance. (3) the use of dropout contributes little to improving the model performance, and that (4) adding fully connected layers to the models significantly increases their parameters but it might not improve their performance. Finally, we raise several open issues in MI-EEG classification and provide possible future research directions.

1. Introduction

Brain–Computer Interfaces (BCIs), as communication bridges between human brain and external devices, have been widely applied in various areas, including rehabilitation training [1,2], robotics control [3], sport training, smart live [4,5], game industries [6,7] and person identification [8]. By decoding brain signals, e.g., Electroencephalograph (EEG), BCI systems can identify the user's intentions and give the corresponding control commands to external devices. As the most common technique to record brain activity, EEG is well known for its high temporal resolution, low cost for data collection, good mobility and low sensitivity to movement [9]. Currently, many types of EEG signals have been used in BCI systems, where Motor Imagery (MI) is one of the most popular ones. Fig. 1 shows a standard MI EEG-based BCI system, which consists of five major parts: MI-EEG data acquisition, preprocessing, feature extraction, classification and application interface. In these five components, most researchers mainly focus on the feature extraction part and the classification part, which are the most challenging.

With the great success of Deep Learning (DL) achieved in Computer Vision (CV) and Natural Language Processing (NLP), more and more researchers have turned their attention to DL and developed DL-based models for MI-EEG classification. Many of them, such as [10–12], can outperform previous traditional machine learning-based methods. In this paper, we comprehensively survey DL-based methods for MI-EEG classification. Our study covers various aspects of DL-based MI-EEG methods. Specifically, we systematically categorize and summarize input formulations for the existing DL-based methods, network architectures, commonly-used regularization methods, public datasets and common metrics. We also discuss several issues in model comparison, propose guidelines for fair performance comparison and evaluate 13 typical DL-based MI-EEG decoding models using the source code released by the authors[1]. Besides, by conducting ablation studies, we

* Corresponding author.
E-mail addresses: wanhe541@student.otago.ac.nz (X. Wang), veronica.liesaputra@otago.ac.nz (V. Liesaputra), zhiyi.huang@otago.ac.nz (Z. Huang).

[1] https://github.com/Henrywang621/DL-based-MI-EEG-models

https://doi.org/10.1016/j.artmed.2023.102738
Received 17 December 2022; Received in revised form 16 October 2023; Accepted 28 November 2023
Available online 2 December 2023

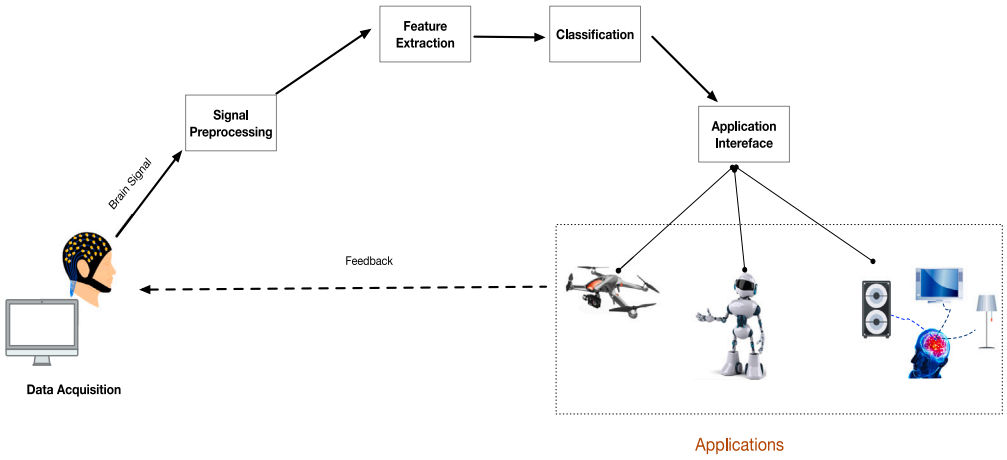

Fig. 1. A schematic diagram for the typical MI-EEG system.

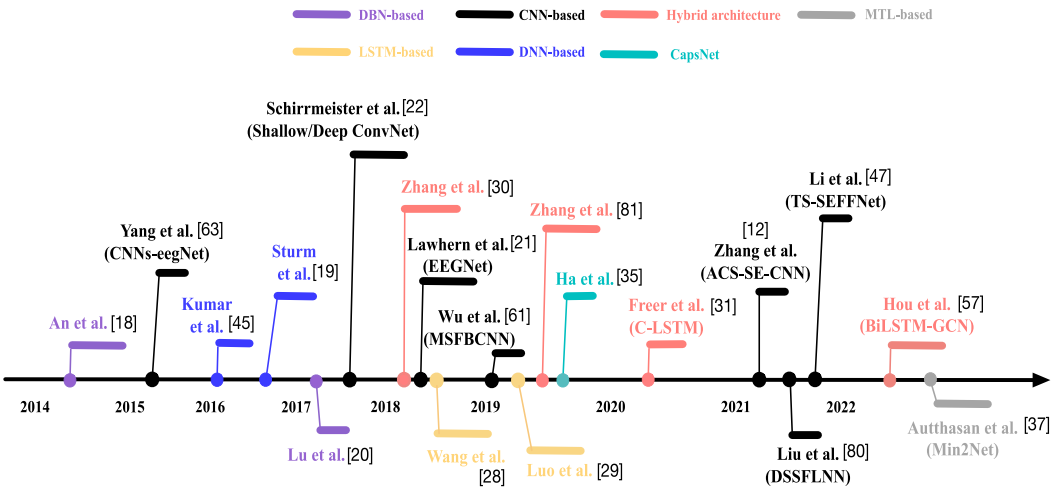

Fig. 2. A brief chronology of deep learning for MI-EEG classification.

explore the effect of some design factors on model performance and reveal important design factors that can influence classification performance. To foster future studies, we also discuss open issues in DL-based MI-EEG classification and provide potential research directions.

1.1. A brief history of deep learning-based MI-EEG classification methods

Hundreds of MI-EEG classification methods have been proposed over the past two decades. Most early works mainly utilized hand-crafted features [13] along with traditional machine learning (ML) classifiers, e.g., Support Vector Machine (SVM) [14], Naïve Bayesian classifier [15] and Linear Discriminant Analysis(LDA) [16], for MI-EEG classification. Due to the limitations brought by hand-crafted features [17], the performance of traditional algorithms is usually not satisfactory.

Benefiting from the rapid development of Deep Learning (DL) techniques, many researchers have developed more effective MI-EEG decoding methods based on DL. Compared with the traditional methods, DL-based methods can automatically extract more discriminative and relevant features from MI-EEG data with Low Signal-to-Noise-Ratios (SNR). This usually contributes to better classification results. The earlier DL-based models, such as [18–20], were generally built based on Deep Belief Networks (DBNs) or Deep Neural Networks (DNNs). The architectures of these earlier models are relatively simple and shallow, but they can often outperform most previous traditional non-deep learning methods. To further improve DL-based methods performance, many studies [21–23] emulated the computational steps of the Filter-Bank Common Spatial Pattern (FBCSP) [15] when constructing their decoding models. FBCSP is a State-Of-The-Art (SOTA) non-DL method that has won several EEG decoding competitions. Compared with FBCSP, these DL-based models can finish several computational steps, e.g., feature extraction and classification, in a unified framework.

In the same time period, some researchers developed spectrogram-based CNN models [10,24,25] for MI-EEG classification. They first utilized time–frequency approaches, i.e., Short-Time Fourier Transform (STFT), to transform raw EEG data into time–frequency representations. These time–frequency images were then inputted into a newly proposed CNN model, or a pre-trained CNN model that has achieved success in computer vision for classification (e.g., VGG16 [26]).

Some methods [27–29] adopted Long-Short Term Memory (LSTM) to build their models because of its ability to capture temporal dependencies in signals. The reported results show these LSTM-based methods can achieve better results than many previous non-DL algorithms and some earlier DL-based methods.

To further improve performance, more and more works [30–32] proposed hybrid architectures by combining different DL algorithms, where the combination of CNN and LSTM is the most common design choice. Besides, some recent studies [33] also utilized Capsule Neural Network (CapsNet) [34,35] to develop MI-EEG decoding models. The

reported results illustrate that their performance is superior than some previous CNN-based methods.

Most recently, some researchers have turned their attention to multi-task learning (MTL) and developed MTL-based MI-EEG classification models [36–38]. These works aim to utilize other related tasks, e.g., input reconstruction [37], to assist MI-EEG classification. Some of them [36,37] show superior performance over several SOTA models, e.g., EEGNet [21], Deep ConvNet [22], etc. Fig. 2 shows these representative DL-based methods proposed from 2014 to now.

1.2. Existing reviews on MI-EEG classification

There were some previous review works on MI-EEG classification. As far as we know, the earliest review can be traced back to 2013. Hwang et al. [13] reviewed various previous EEG-based BCI papers, all of which are not using DL. Another early review article [39] mainly focused on Sensorimotor Rhythm (SMR)-based BCI and its applications. Likewise, the included works in the review do not use deep learning. Several recent reviews started to pay more attention to DL-based methods due to the use of more DL in BCIs. Lotte et al. [40] reviewed EEG classification methods for BCIs proposed from 2007 to 2017, but only a few of them used DL. Two later review articles [9,41] focused on DL-based EEG classification methods. They involved several different application domains, such as BCI, epilepsy, and sleep. However, the number of works included on MI-EEG classification in these articles is very limited.

Most recently, Saegh et al. [42] discussed 40 papers related to DL-based MI-EEG classification. Although this survey covers many aspects of this field, including input formulation, the DL techniques used, common frequency ranges and so on, it has the following drawbacks. (1) It just gave high-level summary of deep learning techniques used in MI-EEG classification. It lacks an in-depth analysis of the DL techniques, such as how these DL techniques are used to construct classification models, what design factors can affect the performance of specific DL architectures, and so on. (2) The authors just cited the results reported in the original papers for comparison. Since different works often adopted different datasets for training and testing [30,43,44], directly citing and comparing the reported results from different papers could result in unfair comparison and invalid conclusions. (3) This survey did not cover many typical and latest decoding models, such as [11,12,19,45–48].

Different from the previous reviews, our work provides a systematic and in-depth review of DL-based MI-EEG classification methods, covering 67 papers related to DL for MI-EEG classification. The criteria used to select these papers are explained in Section 2.1. We reviewed many most typical and most recent DL-based MI-EEG decoding methods. We not only systematically categorize input formulation and network architecture, but also discuss typical design patterns and common input formulations for different network architectures. Moreover, we select 13 representative models for evaluation and discussion. These models cover most common network architectures. Instead of citing the reported results, we use the source code provided by the authors to test their performance in our evaluation. For fair comparison, we use the hyper parameters adopted by the authors, which give the best performance of the models. Through ablation studies on the network architecture, we explore the effect of some common design factors on model performance and obtain several important insights as follows. (1) Effective feature fusion is essential for developing accurate multi-stream CNN architectures. (2) LSTM alone could not be used for classifying MI-EEG signals; it should be combined with spatial feature extraction techniques. (3) Dropout does not have a significant effect in increasing model performance. (4) Researchers should avoid using fully connected layers in their decoding models except the output layer. These discoveries could provide insights to researchers in their design and implementation of new models. Finally, we shed light on challenges and future research directions in this field.

1.3. Our contributions

This paper has the following contributions:

- We comprehensively review DL-based MI-EEG classification models, including network architectures, systematic categorizations, summaries of input formulations, and datasets.
- We select, evaluate and discuss 13 typical DL-based MI-EEG decoding models. They cover most common network architectures. We also discuss existing problems in performance comparison and give guidelines to future studies in terms of fair performance comparison.
- We conduct ablation studies to investigate the effect of design factors on several common network architectures. According to our experimental results, we provide suggestions to researchers in their design of new models.
- We discuss several challenges and open issues in DL-based MI-EEG classification and discuss potential directions of future studies.

The rest of this paper is organized as follows. Section 2 presents the proposed taxonomies of input formulations and network architectures. It also introduces commonly-used regularization techniques, datasets and metrics in MI-EEG classification. Section 3 describes the problems that can lead to unfair performance comparison and provides guidelines for fair comparison in future. We evaluate 13 typical DL-based decoding models, covering several common architectural choices, and analyse design factors that affect model performance in Section 3. Section 4 discusses open issues and directions of future research. Finally, Section 5 concludes this paper.

2. Deep learning-based motor imagery MI-EEG classification methods

Benefiting from DL, MI-EEG classification has achieved great progress. For DL-based MI-EEG classification methods, there are several aspects that deserve our attention, including input formulation, network architectures, commonly-used public datasets and evaluation metrics. In this section, we comprehensively discuss these important aspects. Section 2.1 presents our criteria for selecting reviewed papers related to DL-based MI-EEG classification. Section 2.2 classifies and summarizes typical input formulations. In Section 2.3, we briefly discuss data normalization in DL-based MI-EEG classification. Section 2.4 classifies and summarizes typical network architectures. Section 2.5 introduces and summarizes common regularization techniques in DL-based MI-EEG classification. In Section 2.6, we review and discuss commonly-used MI-EEG datasets and popular metrics.

2.1. Selection criteria of related articles

Since a large number of DL-based MI-EEG classification methods have been proposed, it is unrealistic to review all of these works in one article. We use a systematic review and meta-analysis procedure named PRISMA [49] to choose papers. Specifically, we first input keywords (see Fig. 3) into a well-known multidisciplinary database, i.e., Web of Science.[2] Then, we manually discard some search results unrelated to MI. Specifically, some search results that are only related to other EEG-based BCI paradigms, such as event-related potential and steady-state visually evoked potential. Next, we further discard search results that are not peer reviewed, e.g. arXiv papers, or if the papers solely focus on non-classification tasks, such as feature selection.

Fig. 3 shows the diagram of our PRISMA-based article selection in detail. By adopting PRISMA, 60 related works are chosen. We have also manually included some typical and latest works [36–38,50–53] that are not chosen by our PRISMA-based article selection. Finally, we choose 67 related works for review.

[2] https://www.webofknowledge.com/.

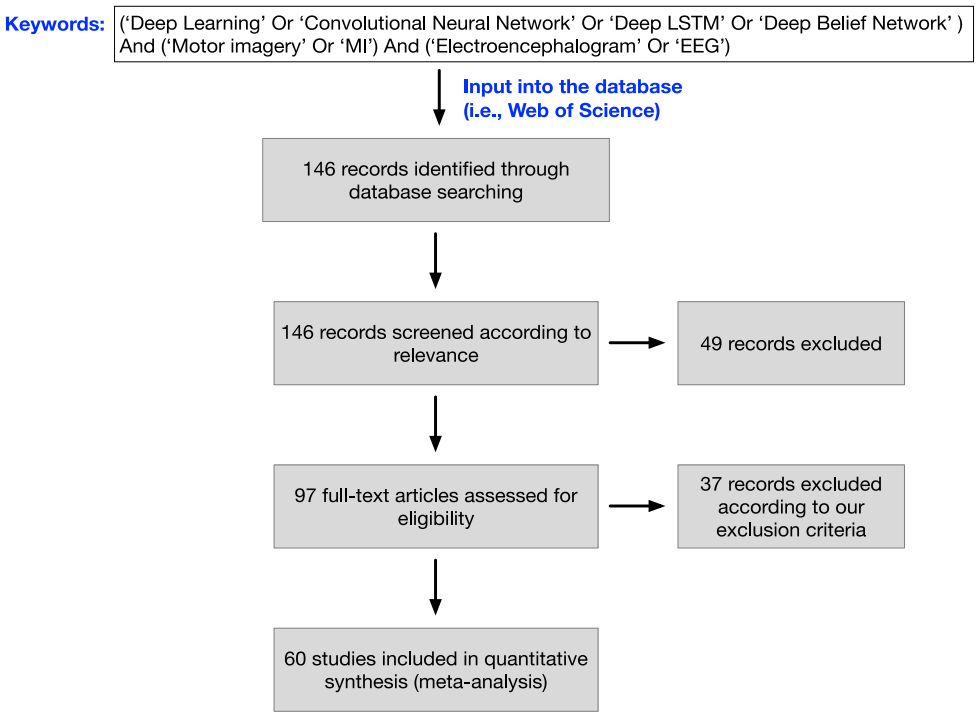

Fig. 3. The diagram of PRISMA-based article selection. This search was conducted in July, 2022.

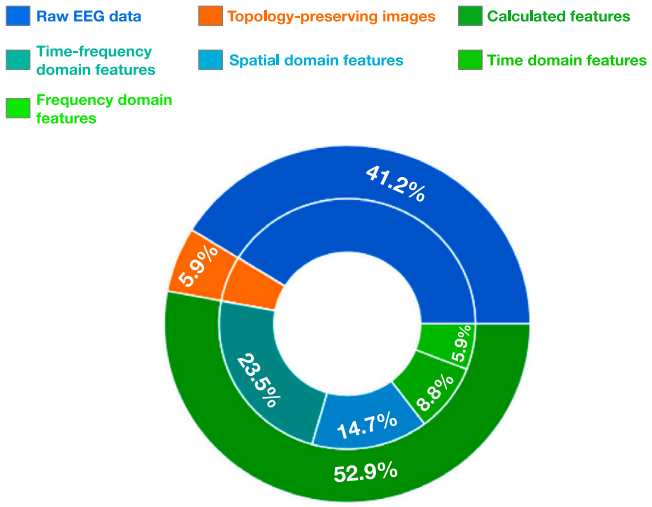

Fig. 4. The proportion of typical input formulations used in the reviewed papers. Note that 100% in the pie chart represents all the models in the reviewed papers.

Table 1
Classification of the reviewed studies according to input formulations.

Category		Publications
Raw EEG data		[17,21,22,31,47,54–58] [4,11,19,25,48,59–63] [23,36–38,50–53,64]
Topology-preserving images		[30,44,65,66]
Calculated features	(1) Time–frequency features	[10,12,24,43,67–72]
	(2) Spatial features	[20,25,73–76]
	(3) Time features	[45,46,63,77–83]
	(4) Frequency features	[27,28,84–87]
		[18,27,88,89]

2.2. Typical input formulations

Formulating or choosing the suitable input formulation is one of the important factors in designing an accurate classification model. So far, various input formulations have been developed and adopted for DL-based MI-EEG classification methods. According to the characteristics of these input formulations, they can be classified into three categories: raw EEG data (see Section 2.2.1), topology-preserving images (see Section 2.2.2) and calculated features (see Section 2.2.3). Fig. 4 shows the proportion of these three categories being used as input in the reviewed models. Table 1 summarizes representative publications for different categories of input formulations. Since input formulation is usually closely linked with network architectures, we also briefly discuss common input formulations used for different network architectures in Section 2.4.

2.2.1. Raw EEG data

Since DL techniques are good at extracting effective features from data, many works, such as [21,22,54], directly use raw EEG data as the input of their DL models. Compared with other types of input formulations, the main advantage of using raw EEG data as input is that no additional computational steps is required to process the data. The input raw data is generally in the form of 2D matrices, where the rows often represent time and the columns represent EEG channels (Refer to Fig. 5(a)). So far, a large proportion (around 41.2%) of the existing DL methods have taken raw EEG data as input, covering various types of network architectures (see Table 2), such as Convolutional Neural Network-based architecture, Long Short-Term Memory-based architecture, Hybrid Deep Network-based architecture, etc.

2.2.2. Topology-preserving images

Although raw EEG data in the form of 2D matrices has been widely used as the input of DL-based MI methods, it neglects spatial relationships between the EEG channels [30]. To address this problem, some methods [30,44,66] utilized the position information of the EEG electrodes to construct the input of their networks. For convenience, we call this type of input formulation as topology-preserving images.

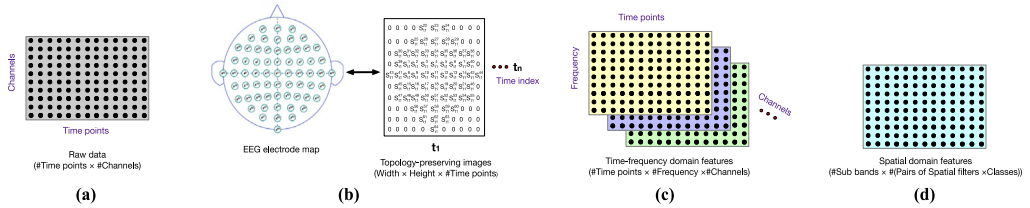

Fig. 5. Examples of several commonly-used input formulations. (a) Raw EEG data in the form of the 2D matrix. (b) Topology-preserving images. (c)–(d) examples of two types of popular calculated features.

Table 2
Classification of the reviewed papers in terms of network architecture. RD, CF and TP-Img represent raw EEG data, calculated features and topology-preserving images respectively.

Category	Subcategory	Input formulations	Publications
CNN-based	(1) Single-stream	RD or CF	[11,21,22,24,54,86,87,90] [25,63,68,69,71,75,78,79] [10,12,23,67,70,73,74,91]
	(2) Multi-stream	RD, CF or TP-Img	[17,47,58–61,64,76], [43,50,55,62,66,77,82]
LSTM-based		RD or CF	[4,25,27–29,92]
Hybrid	(1) Hybrid CNN-LSTM	RD or TP-Img	[30,31,56,81]
	(2) Parallel CNN-MLP	CF	[46,83]
	(3) Other hybrid architecture	RD or CF	[32,57]
DNN-based		RD or CF	[19,45,85]
DBN-based		RD or CF	[18,20,88,89]
MTL-based		RD	[36–38,51–53]

As shown in Fig. 5(b), topology-preserving images can be regarded as a sequence of 2D data segments, each of which is built according to the EEG electrode map. To construct topology-preserving images, zero padding is used to extend the rows with lower number of sensors in the EEG electrode map to make their length the same as the rows with the largest number of sensors. More details about how to obtain topology-preserving images could be found in [30]. Compared with the raw EEG data, topology-preserving images preserve the complete position information of the electrode distribution. However, constructing topology-preserving images is also more time-consuming, and we need to know the electrode distribution of the corresponding EEG electrode cap. For this type of input formulation, it is often used as the input of CNN-based methods and hybrid CNN-LSTM methods (see Table 2).

2.2.3. Calculated features

Apart from raw EEG data and topology-preserving images, many studies utilized the features extracted from the EEG data as the input of their models. These calculated features can be roughly categorized as: spatial domain features, time domain features, frequency domain features and time–frequency domain features. Time–frequency domain features are the most common input formulation. Wavelet Transform (WT) [67,69,70] and Short-Time Fourier Transform (STFT) [25,73] are often used to transfer EEG signals to time–frequency feature maps. Fig. 5(c) shows an example of time–frequency features. Since CNNs are good for image data, the image-like time–frequency feature maps were generally fed into CNN-based models for classification. The second most common calculated features are spatial domain features (Refer to Fig. 5(d)), which can be obtained by using Common Spatial Patterns (CSP) [45,63,77] or the variants [15,78,93] of CSP. For example, the work by [63] first extracted augmented CSP features from EEG signals, and then the extracted CSP features are fed into a 5-layer CNN model for MI-EEG classification. Besides, some DL-based methods also took time domain features (e.g., numerical measures [27,28,85]), frequency domain features (e.g., Power Spectral Density (PSD) [27,88] and Fast Fourier Transform [18,89]) as input.

2.3. Normalization

Data normalization aims to transform the values of the dataset into the same scale. Due to the high variability of EEG signals across different sessions and different subjects, data normalization is often applied to MI-EEG decoding models' input as a standard pre-processing step [27,61]. The most frequently used normalization function is Z-score normalization, which can be formulated as below.

$$Z_{score} = \frac{X - \mu}{\sigma} \tag{1}$$

where X is the input of a model. μ and σ represent the mean and the standard deviation calculated over X respectively. Apicella et al. [94] investigated and evaluated the effect of data normalization on different EEG tasks, and they concluded that data normalization can result in a significant improvement on various EEG tasks, including MI-EEG models' classification performance.

2.4. Representative network architectures

Over hundreds of Deep Learning-based models have been proposed to classify MI-EEG signals. In terms of network architecture, we categorize the existing DL-based models into six categories: CNN-based, LSTM-based, Hybrid Deep Network-based, Deep Neural Network-based, Deep Belief Network-based and MTL-based. Table 2 lists the corresponding publications of each category of network architecture and summarizes the input formulations used in different network architectures.

2.4.1. Convolutional Neural Network (CNN)-based methods

Convolutional Neural Network is one of the most popular DL algorithms. It has been broadly applied in computer vision [95], natural language processing [96] and speech recognition [97]. According to the network architecture, we further classify CNN-based decoding models into two categories, namely single-stream CNN-based networks (see Fig. 7(a)) and multi-stream CNN-based networks (refer to Fig. 7(b)).

Single-stream CNN-based network generally consists of convolution layers, pooling layers, activation functions and fully connected

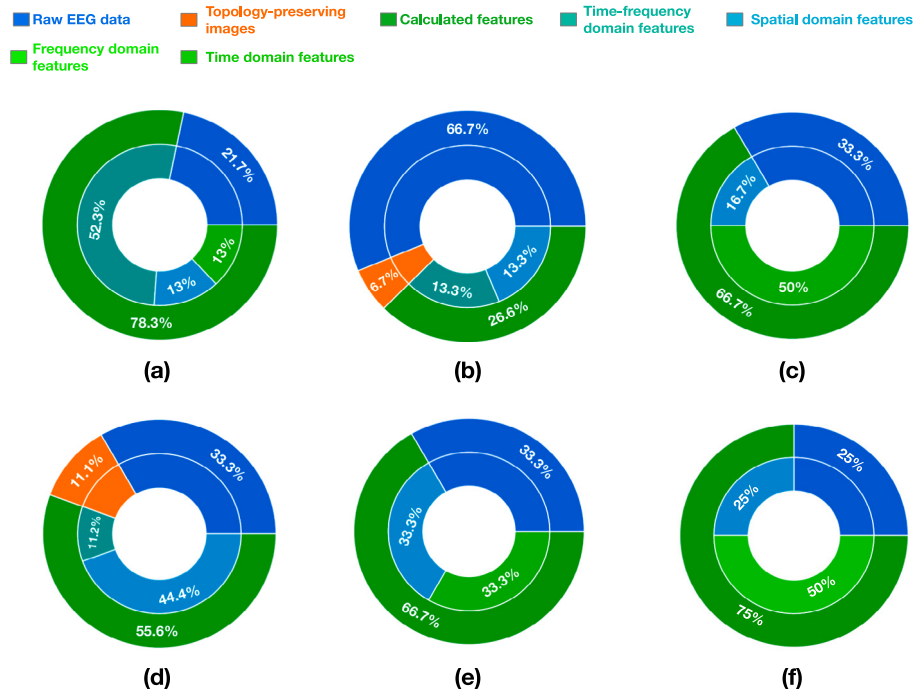

Fig. 6. The proportion of input formulations used for each type of network architectures. (a) Single-stream CNN-based network. (b) Multi-stream CNN-based network. (c) LSTM-based network. (d) Hybrid deep network-based models. (e) DNN-based network. (f) DBN-based network. Note that 100% in each pie chart represents all the reviewed models based on the corresponding network architectures. We omit the multi-task learning-based architecture here, since all reviewed studies based on this architecture use raw EEG data as input.

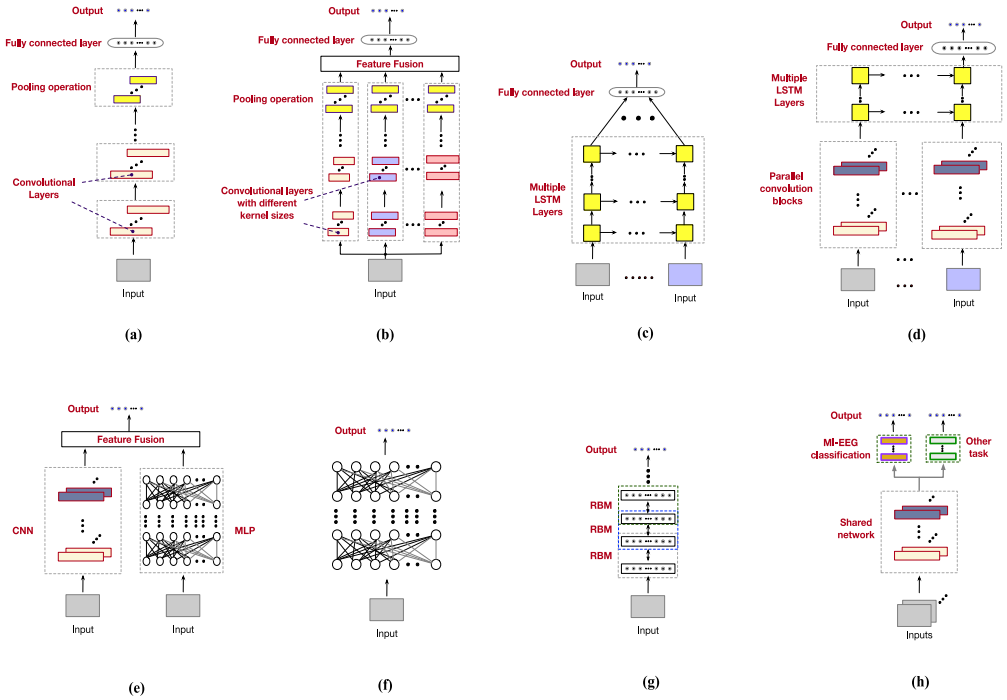

Fig. 7. Representative network architectures for MI-EEG classification. (a) Single-stream CNN-based network. (b) Multi-stream CNN-based network. (c) LSTM-based network. (d) Hybrid CNN-LSTM network. (e) Parallel CNN-MLP network. (f) DNN-based network. (g) DBN-based network. (h) MTL-based network.

(FC) layers. These components are stacked to form a single-stream network (refer to Fig. 7(a)). As shown in Table 2, the reviewed single-stream CNN-based methods used calculated features (e.g., time–frequency domain features, time domain features, etc.) or raw EEG data as input. time–frequency domain features and raw EEG data are the most popular ones (see Fig. 6(a)). The network design of a single-stream CNN model that takes image-like time–frequency features as input often draws on the experience of DL models for computer vision. For example, Xu et al. [73] first converted EEG data into time–frequency images by using their STFT-based processing approach. Then, the generated

image-like representations are fed into the proposed MI-EEG decoding framework, which is based on a well-known deep learning model (i.e., VGG-16 [26]) for image classification. When raw EEG signals or time domain features are chosen as input formulation, researchers often used computational steps of FBCSP [15], which is a SOTA traditional algorithm, as the guidelines for network design. For instance, inspired by FBCSP, Schirrmeister et al. [22] designed a single-stream CNN-based model, namely Shallow ConvNet. It first performs a temporal convolution and a spatial convolution to simulate the bandpass and CSP filters in FBCSP. Next, a series of operations (e.g., squaring nonlinear transformation, mean pooling and logarithmic function) are adopted to simulate other corresponding computational steps in FBCSP.

Multi-stream CNN-based network, as shown in Fig. 7(b), commonly contains multiple feature extraction branches with different configurations (e.g., the size of convolution filters, the number of filters, etc.). For the reviewed multi-stream CNN-based methods, most of them, such as [58,60,61], selected raw EEG data as model input (see Fig. 6(b)). Considering that the optimal size of convolution filters may vary from subject to subject, these methods generally adopted several groups of 1-D convolutions with different kernel sizes to better extract temporal and spatial features from raw EEG signals [59,60]. The features extracted by multi-scale convolutions are fused and further processed to produce final predictions. For example, Jia et al. [60] presented a multi-stream CNN-based network named MMCNN, which directly takes raw EEG data as input. The proposed model consists of five network branches. These network branches adopt convolutions with different kernel sizes to extract multi-scale features. To obtain discriminative features, the authors also added a Squeeze-and-Excitation block [98] to each of network branches. The generated features from different branches are then fused to make predictions.

2.4.2. Long-Short Term Memory (LSTM)-based methods

Due to the ability to detect temporal dependencies from sequential data, some works [4,25,27–29,92] have tried to build their decoding models based on LSTM. As shown in Fig. 7(c), LSTM-based models typically include an input layer, at least one LSTM layer, a fully connected layer and an output layer. Their inputs are usually the calculated features (see Fig. 6(c)). For example, Zhang et al. [27] extracted multiple types of time and frequency domain features from raw EEG data. These extracted features are utilized to train their proposed LSTM model with three hidden-layers.

2.4.3. Hybrid deep network-based methods

To design more effective decoding models, some researchers have also tried to combine different DL algorithms. According to the different combinations, the hybrid deep network-based architectures can be further categorized into hybrid CNN-LSTM, parallel CNN-MLP and other hybrid architectures. Fig. 6(d) illustrates that the proportion of different input formulations used for the reviewed hybrid deep network-based models.

Hybrid CNN-LSTM methods, as illustrated in Fig. 7(d), is the most common hybrid architecture, generally having a CNN sub-network, followed by a LSTM sub-network. This design aims to simultaneously learn spatial and temporal features from EEG data. The input formulation of hybrid CNN-LSTM models can be raw EEG data, calculated features or images. A well-known hybrid CNN-LSTM network is Cascade Model [81], which mainly consists of a 2D-CNN, two stacked LSTM layers and fully connected layers. It takes 2D data meshes that preserve position information of EEG electrodes as input. The 2D-CNN extracts spatial features from the mesh-like representations, while the stacked LSTM layers are used to learn dependencies among time steps.

Parallel CNN-MLP methods (refer to Fig. 7(e)) fuse features learned from the CNN-based and MLP-based sub-networks for MI-EEG classification. The input to different sub-networks is generally different types of calculated features. A representative Parallel CNN-MLP-based method is the work of [46], where a CNN-based sub-network is designed for refining dynamic energy features, and a MLP-based sub-network is adopted for static energy features.

Other Hybrid Architectures. Apart from the aforementioned two types of combinations, some works [32,57] proposed decoding models based on other hybrid architectures. Due to the fact that these combinations are less common, we classify them into Other Hybrid Architectures. For example, Dai et al. [32] presented a DL-based MI-EEG framework, which combines CNN and Variational Autoencoder (VAE) [99]. Most recently, Hou et al. [57] designed a novel decoding model named attention-based BiLSTM-GCN for MI-EEG classification. This is the first model that combines Bidirectional Long Short-Term Memory (BiLSTM) [100] and Graph Convolutional Neural Network (GCN) [101].

2.4.4. Deep Neural Network (DNN)-based methods

Some early works [19,45,85] developed their models based on DNN for MI-EEG classification (see Fig. 7(f)). As shown in Fig. 6(e), raw EEG data or calculated features are the most common input formulations. For instance, CSP-DNN [45], a typical DNN-based method, contains two hidden layers. This model takes CSP features extracted from raw EEG data as input. The reported results show that it can outperform some well-known non-DL methods. Compared with CNNs, DNNs with the same number of layers usually have much more trainable parameters, which can lead to expensive computational complexity and makes the models easily suffer from overfitting. Therefore, DNN-based architecture has been rarely developed for MI-EEG classification recently.

2.4.5. Deep Belief Network (DBN)-based methods

Like DNN-based models, the existing DBN-based models are generally developed by early works, such as [18,20,88]. For the reviewed DBN-based methods, they either directly used raw EEG data [88] or calculated features, e.g. frequency domain features FFT [18,20], as input (see Fig. 6(f)). As shown in Fig. 7(g), DBN-based models are typically constructed by stacking Restrict Boltzmann Machine (RBM). The features extracted by the stacked RBMs are finally classified by a classifier. For example, FDBN [20], as one of representative DBN-based models, is a four-layer network, containing three RBMs and a softmax layer. This model adopts the extracted frequency domain features as input. The reported results in [20] showed its superior performance over several classical tradition methods.

2.4.6. Multi-task learning (MTL)-based methods

All aforementioned network architectures only learn a single task, i.e., MI-EEG classification, within their networks. Unlike these models, some recent works [37,38,51] have explored multi-task learning (MTL) in MI-EEG classification. Inspired by the human brain's learning system, these works utilize other related tasks to assist MI-EEG classification. The common tasks in the MTL-based architectures that assist MI-EEG classification mainly include input reconstruction [37,51], classifying non-target EEG datasets [38,53] and discriminating source and target domains [36,52]. By leveraging on these related tasks, MTL-based model performance on the target task of MI-EEG classification can be improved. However, the additional network branches created for these other tasks might increase the complexity of the training process [53]. As shown in Fig. 7(h), a common design pattern of MTL-based architecture involves initially employing a shared network for extracting feature representations from inputs, followed by the distinct network branches for different tasks. During the training process, multiple tasks are learned simultaneously. For instance, DMTL-BCI [51], as a typical MTL-based MI-EEG classification model, first utilizes a representation module to learn features from the original input. Then, the learned features are transmitted into the classification module for MI-EEG classification and the reconstruction module for input reconstruction. Two tasks are jointly optimized simultaneously. The reported results in the original paper show superior performance over some representative MI-EEG models, such as EEGNet [21] and shallow ConvNet [22].

Table 3
Summary of public MI-EEG datasets. Note that LH (Left hand), RH (Right Hand), RF (Right Foot), BF (Both Feet), T (Tongue), BH (Both Hands) and R (Rest) represent the imagination of different movements.

Dataset	Source	Year	#Subjects	Sampling rate	#channels	MI tasks	Citations
BCI competition II 3	[107]	2003	1	128 Hz	3	LH/RH	[24,69]
BCI competition III IVa	[108]	2006	5	1000 Hz	118	LH/RH/RF	[10,45,71,85]
BCI competition IV 2a	[109]	2008	9	250 Hz	22	LH/RH/BF/T	[11,21,22,24,31,63,86] [23,46,59–61,78,79] [17,47,55,62,64,66,81] [28,29,36,50,53,83,89] [37,38,51]
BCI competition IV 2b	[110]	2008	9	250 Hz	3	LH/RH	[11,25,46,59,60,69,70] [25,29,53,61,64,74,85]
PhysioNet	[111]	2009	109	160 Hz	64	LH/RH/BH/BF/R	[27,30,44,54,58,68,92] [4,43,57]
High-Gamma	[22]	2017	14	500 Hz	128	LH/RH/BF/R	[17,22,38,47,61]

2.5. Regularization

Regularization is a technique that aims to avoid overfitting in deep learning. Several regularization methods have been widely used in DL-based MI-EEG classification, mainly dropout, batch normalization, L1/L2 regularization, network parameter initialization, sample preprocessing, data augmentation and transfer learning.

Dropout [102] is a frequently used regularization technique, where a specific ratio of neurons are randomly discarded during the training process. Several representative MI-EEG models, such as [21,22,47], use dropout during training. In Section 3.3, we investigate the effect of using dropout on the model performance through ablation study.

Batch normalization has been verified to help the networks converge faster by normalizing the inputs of their intermediate layers [103]. It has been applied in many MI-EEG classification models [21,47,78].

L1 and L2 regularization are also used by some methods [76,80] for MI-EEG classification. By adding the regularization term, the overfitting problem can be reduced during training. The more details about this type of regularization can be found in [102].

The existing works generally do not specify how their network parameters are initialized. By investigating the models with available source code (See Table 5), we found Glorot initialization [104] is the most commonly-used way to initialize network parameter. A recent survey paper [105] systematically discusses and summarizes weight initialization strategies for deep neural networks, which might be able to help future studies choose the appropriate ways to initialize network parameter.

Appropriate sample preprocessing strategies can also help reduce overfitting during training. Normalization and data filtering are the two most common sample preprocessing strategies in MI-EEG classification. The former is briefly discussed in Section 2.3. As for the latter, it aims to eliminate noise and unnecessary information. A recent survey paper [106] well summarizes data filtering in MI-EEG classification, which could provide more insights about this topic.

Data augmentation and transfer learning are also two types of regularization techniques used in MI-EEG classification. We discuss them in detail in Sections 4.2 and 4.3 respectively.

2.6. MI-EEG datasets and evaluation metrics

2.6.1. MI-EEG datasets

Many MI-EEG datasets have been collected and used so far. Here, we only focus on the six most commonly used public MI-EEG datasets, as summarizes in Table 3.

Among these public MI-EEG datasets, BCI competition IV 2a is the most popular one, which is used in 31 reviewed studies (see Fig. 8), followed by BCI competition IV 2b (14 reviewed studies) and the Physionet dataset (11 reviewed studies). The works that used these three datasets account for the majority of the total of the reviewed studies. The most recent public dataset, High-Gamma, is used in only 5 reviewed studies. Considering its publication time, it may be used more frequently in future studies.

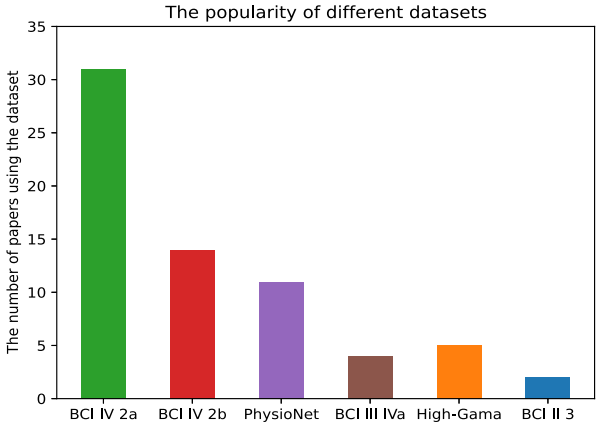

Fig. 8. The popularity of different public MI-EEG datasets used in the reviewed papers. Note that some reviewed studies, such as [22,47,61], use multiple datasets, and some other studies [12,90] only use private datasets.

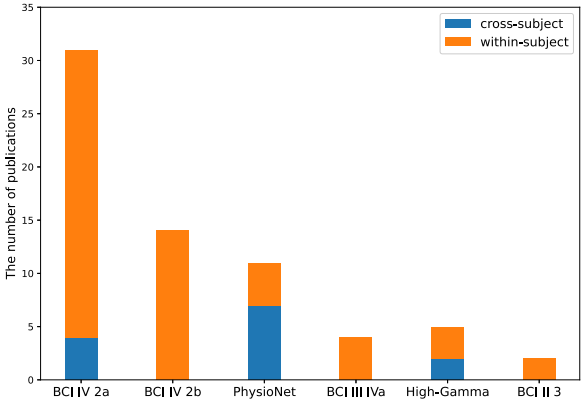

Fig. 9. Distribution of the reviewed studies that perform the within- or cross-subject classification task on six commonly-used public datasets. Note that some reviewed studies, such as [22,47,61], use multiple datasets, and some other studies [12,90] only use private datasets.

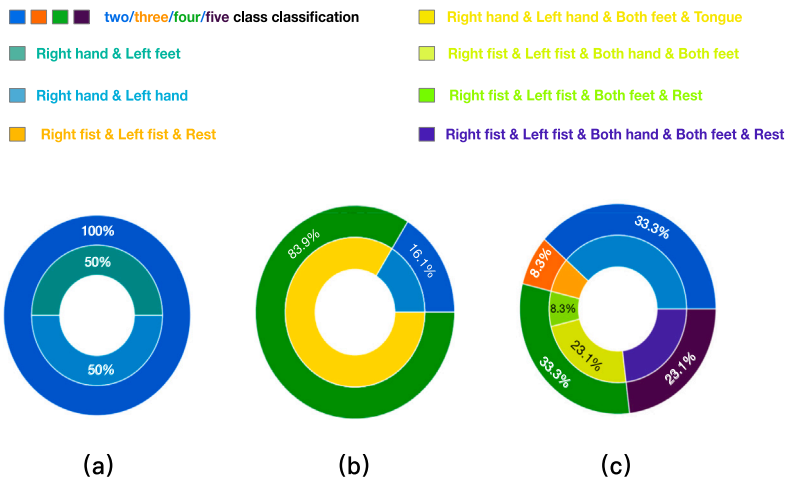

Fig. 10. The proportion of the reviewed studies that choose to perform different MI-EEG classification tasks on different public datasets. (a) BCI competition III IVa. (b) BCI competition IV 2a. (c) PhysioNet. Note that 100% in each pie chart represents all the reviewed papers using the corresponding datasets.

In terms of performance evaluation, researchers perform either within- or cross-subject classification on their selected datasets. Within-subject classification uses the data from the same subject for training and testing models, while cross-subject classification means the training data and test data are from different subjects. As shown in Fig. 9, the reviewed studies generally performed within-subject classification on some earlier datasets (BCI competition II 3, BCI competition III IVa, etc.), and cross-subject classification on newer datasets, such as the PhysioNet dataset and the High-Gamma dataset. This is because earlier datasets usually contain a small number of subjects with a large number of the collected trials for each subject. However, recent datasets usually contains more subjects enabling researchers to conduct cross-subject analysis on them.

MI-EEG datasets often contain different numbers and types of MI tasks (refer to Table 3). For BCI competition II 3 and BCI competition IV 2b, they only contain two MI tasks: imagining the movement of the left hand and imagining the movement of the right hand. Reviewed works that selected these two datasets can only perform binary classifications (left hand vs. right hand).

However, the situation is a bit more complicated for other several public MI-EEG datasets with more MI tasks. For these datasets, the existing works may perform different classification tasks on them (see Fig. 10). As shown in Fig. 10(a), all the reviewed papers that used BCI competition III IVa chose to perform binary classification. However, half of them performed right hand vs. left hand classification, while the other half did right hand vs. left feet. For BCI competition IV 2a, most of the studies (83.9%) performed four-class classification task on the dataset (refer to Fig. 10(b)), and the others chose to classify left hand vs. right hand. With regard to PhysioNet, the most common choice (33.3%) is to perform the left hand vs. right hand classification (see Fig. 10(c)). Note that Fig. 10 omits the High-Gamma dataset, since all reviewed studies using this dataset only performed four-class classification on it.

2.6.2. Metrics

By investigating the existing methods, we find that there are eight metrics for model assessment. They are Accuracy, Precision, Recall, F1-score, Area Under Curve (AUC) score, Cohen's Kappa, Specificity and Sensitivity [50,112], where Accuracy, F1-score, AUC score and Cohen's Kappa are most commonly-used ones.

The articles [112,113] provide a detailed summary of how to use these metrics to evaluate binary and multi-class classification.

3. Evaluation and analysis

In this section, we first point out several common problems in performance comparison of existing studies and provide guidelines for fair performance comparison (See Section 3.1). Then, we evaluate and analyse 13 representative DL-based MI-EEG classification models, which cover several most common network architectures in Section 3.2. Finally, we further explore the effect of some design factors on model performance by using ablation study in Section 3.3

3.1. Existing problems in performance evaluation

Performance comparison is the common way to verify the effectiveness of the proposed models. However, we find several problems that limits MI-EEG researchers in this area to directly compare the results reported in the papers due to the inherent dissimilarities of their evaluation criteria and methodologies. We discuss these common problems as follows.

Different meanings of cross-subject classification in different works. Within-subject (also called intra-subject) and cross-subject (also called inter-subject) validation are two widely-used validation schemes. Within-subject validation [47,50] uses part of a subject's data for training and the remaining data of the same subject for testing. This process is repeated for each subject in the dataset. However, we find that there are three different cross-subject validation schemes: leave-n-subjects-out, mix-up-all-subjects and random-selection. (1) *leave-n-subjects-out*. In the medicine field, such a validation scheme is also often called subject-wise data split [114]. Many existing methods, such as [21,44,54], adopt this cross-subject validation, where data from a specific number of subjects is selected as the training set, and data from the remaining subjects is used as the test set. This validation scheme is often applied in the scenario where BCI systems are used to predict unseen subjects' MI-EEG data. In fact, an ideal MI-EEG based BCI system should be able to accurately classify unseen subjects's signals [48]. (2) *mix-up-all-subjects*. In the real world, it is quite realistic to perform follow-up on the same subject after a period of time. With the mix-up-all-subjects scheme, the training data from each subject in the dataset is collated together and used for training the model. The remaining data of each subject is collated together and used to test the trained model [64,91]. By investigating the reviewed papers, we find this validation scheme is not frequently used. Compared with the leave-n-subjects-out scheme, this one is often considered as an easier classification task [91]. (3) *random-selection*. In the medicine field, it is also often called record-wise data split [114]. Specifically, some works [30,43,58] randomly select a specific ratio of data from the whole dataset for training

Table 4
Recommendations for performance comparison.

	Guidelines	Description
1	Use source code or the verified implementations of decoding models for comparison	Researchers should select the decoding models, with publicly available source code, as the baseline models. Otherwise, reimplemented version of the code should be verified by the authors of the original method.
2	Compare models under the same conditions	Researchers should make sure all the compared models are trained on the same training data and are evaluated on the same test data.
3	Avoid only evaluating the models on the private data	If the authors use their private dataset, they should also evaluate their models with public datasets.
4	Specifically give which type of cross-subject validation used.	If the authors perform the cross-subject evaluation, they should specifically illustrate the cross-subject validation scheme used.
5	Release the source code and provide detailed instructions	To facilitate future studies in our community, we advocate researchers to make their source code publicly available.

and the remaining data for testing. When this cross-subject scheme is used, there are three possible situations for the prepared training set and test set. The first one, like the leave-n-subjects-out scheme, the generated training set and test set contain data from different subjects. The second one is that both training set and test set contain part of all subjects' data, which is similar to the mix-up-all-subject scheme. The third situation is that the test set contains part of data of some subjects in the training set but also contains some subjects that are not included in the training set. If researchers do not specifically illustrate which type of cross-subject scheme they used, the inconsistency in the meaning of cross-subject classification could cause confusion in the performance evaluation. Here, we recommend the use of leave-n-subjects-out scheme because an ideal BCI system should have the ability to accurately classify MI-EEG signals of new subjects [64] and the leave-n-subjects-out cross-subject scheme can well validate the models' ability in this respect.

Different evaluation strategies used in performance comparison. Fair model comparison should ensure that all compared models perform the same classification tasks. However, there exist inconsistencies in model comparison for some studies, such as [30,43,92]. These inconsistencies could lead to invalid and unfair model comparison. For example, in these works, their proposed models and some compared models were evaluated on different classification tasks, i.e., different numbers of classes (binary vs. multi-class) and different analysis types (within-subject vs. cross-subject). Considering that the existing works adopt different evaluation strategies for model comparison, directly using and analysing the classification results reported by different works, like what was done in the most recent MI-EEG survey [42] can be misleading and it would not provide accurate and meaningful insights. For instance, the review paper [42] recommends time-series input formulation because the average classification accuracy of the reviewed models using time-series input formulation is higher than that of the reviewed works that take images or calculated features as input. Unfortunately, the authors have only used the reported results from different papers, which are often obtained by using different evaluation strategies (e.g., different ways to prepare training data and test data). Different evaluation strategies could significantly impact the model performance [31]. Thus, these results could not be directly compared or averaged—leading to incorrect conclusions.

Selected baseline models without available source code. Some DL-based MI-EEG classification models do not have their source code publicly available, and sometimes important implementation details of the models are not stated in the original papers either. To be able to do performance comparison with those models, researchers must re-implement them according to their own interpretations. In such situations, the classification performance of the re-implemented versions could significantly differ from the original ones—leading to invalid and unfair performance comparison.

Performance comparison only conducted on the private datasets. The existing DL-based studies generally used public MI-EEG datasets (refer to Section 2.6) to evaluate their methods. However, there are also some works [12,90] that only evaluated their models on private datasets. Due to the unavailability of these datasets, the quality of the private data is unknown. Thus, if the researchers only compare their proposed methods with the baseline models on their private datasets, the evaluation results could be biased. Besides, the performance of the models evaluated on these private datasets cannot be verified by other researchers, which reduces the credibility of the results.

Since the aforementioned problems could make performance comparison invalid and/or unfair, it is necessary to standardize the comparison of results in the field of DL-based MI-EEG classification. Thus, we present several guidelines (see Table 4) that other researchers can use to precisely and fairly evaluate the performance of DL-based MI-EEG decoding models.

3.2. Performance evaluation for representative DL-based MI-EEG decoding models

In our approach, we carefully select decoding models by following three criteria. The first one is that the selected models' architecture should be one of the mostly common architectures: single-stream CNN-based architecture, multi-stream CNN-based architecture, LSTM-based architecture and hybrid CNN-LSTM-based architecture. The second one is that the selected models are generally highly cited and/or published in reputable conferences or journals. Thirdly, following our recommended guidelines in Table 4, the selected models should have their source code publicly available.

According to our criteria, we select 12 representative DL-based MI-EEG decoding methods (see Table 5) from the 67 papers that we reviewed. Besides, we also acquire the source code of MBEEGNet [50] from the authors, since we need to use it to comprehensively investigate the effect of feature fusion on the performance of multi-stream CNN-based models (Refer to Section 3.3). Thus, we finally evaluate 13 models from 67 papers that we reviewed. For fair performance evaluation and comparison, all models are trained and tested on the same GPU, i.e., an Nvidia Quadro P6000. The selected benchmark datasets are the two most commonly used public MI-EEG datasets shown in Table 3, i.e. BCI Competition IV 2a and the PhysioNet; and we adopt four metrics (i.e., Accuracy, F_1-score, AUC score and Cohen's Kappa) to measure the chosen models.

3.2.1. Experiment setup

The performance evaluation setup of the selected classification models on two public datasets are illustrated in detail as follows.

Table 5
Representative MI-EEG decoding models with publicly available source code. Note that we refer to [54] as ETENet for the sake of convenience.

Model category	Year	Framework	Model	Github link
Single-stream CNN	2017	Pytorch	ShallowConvNet [22]	https://github.com/braindecode/braindecode
Single-stream CNN	2017	Pytorch	DeepConvNet [22]	https://github.com/braindecode/braindecode
Single-stream CNN	2018	Tensorflow	EEGNet-4,2 [21]	https://github.com/vlawhern/arl-eegmodels
Single-stream CNN	2018	Tensorflow	EEGNet-8,2 [21]	https://github.com/vlawhern/arl-eegmodels
Single-stream CNN	2018	Tensorflow	ETENet [54]	https://github.com/hauke-d/cnn-eeg
Single-stream CNN	2019	Tensorflow	pCNN model [25]	https://github.com/gumpy-bci/gumpy-deeplearning
Multi-stream CNN	2020	Tensorflow	EEGNet fusion [58]	https://github.com/rootskar/EEGMotorImagery
Multi-stream CNN	2021	Pytorch	TS-SEFFNet [47]	https://github.com/LianghuiGuo/TS-SEFFNet
LSTM	2019	Tensorflow	LSTM model [25]	https://github.com/gumpy-bci/gumpy-deeplearning
Hybrid CNN-LSTM	2018	Tensorflow	Cascade Model [30]	https://github.com/dalinzhang/Cascade-Parallel
Hybrid CNN-LSTM	2018	Tensorflow	Parallel model [30]	https://github.com/dalinzhang/Cascade-Parallel
Hybrid CNN-LSTM	2020	Pytorch	C-LSTM [31]	https://github.com/dfreer15/DeepEEGDataAugmentation

Table 6
Classification performance of representative deep learning-based MI-EEG models on PhysioNet and BCI competition IV 2a. Each of the metrics listed in the table is averaged over all evaluation sets. Bold values and italic values indicate the best performance and chance-level performance respectively. Note that the model performance reported in this table might differ from the reported results in other papers because we have utilized different evaluation strategies to the ones used by the model's original authors.

Model	#Parameter	PhysioNet (Two classes)				BCI IV 2a (Four classes)			
		Acc (%)	F1-score	AUC	Kappa	Acc (%)	F1-score	AUC	Kappa
Shallow ConvNet [22]	105.7K	75.33	0.746	0.838	0.506	56.29	0.559	0.792	0.417
Deep ConvNet [22]	303.9K	**77.89**	0.766	**0.876**	**0.558**	42.53	0.411	0.689	0.234
EEGNet-4,2 [21]	1.39K	77.44	0.776	0.863	0.549	56.66	0.548	0.787	0.422
EEGNet-8,2 [21]	2.47K	77.51	0.777	0.859	0.55	52.37	0.479	0.78	0.365
ETENet [54]	305.5K	76.05	0.758	0.847	0.521	38.5	0.371	0.652	0.18
pCNN [25]	170.7K	57.60	0.567	0.613	0.152	36.84	0.354	0.625	0.158
Cascade Model [30]	1.49M	58.62	0.583	0.631	0.185	34.31	0.332	0.589	0.107
Parallel Model [30]	1.47M	56.73	0.547	0.591	0.149	35.42	0.346	0.621	0.13
LSTM Model [25]	99.1K	*49.93*	*0.494*	*0.507*	*−0.003*	*26.08*	*0.246*	*0.506*	*0.001*
EEGNet fusion [58]	22.9K	76.18	**0.784**	0.849	0.523	44.31	0.381	0.759	0.258
MBEEGNet [50]	9.4K	75.99	0.755	0.853	0.52	41.07	0.337	0.76	0.21
C-LSTM [31]	66.1K	74.72	0.736	0.797	0.495	**69.1**	**0.679**	**0.889**	**0.587**
TS-SEFFNet [47]	359.8K	67.08	0.667	0.759	0.342	42.59	0.39	0.74	0.235

MI-EEG decoding models evaluated on the PhysioNet dataset. As described in Section 2.6, there are various classification tasks that we could perform on this dataset (see Figs. 9 and 10). However, the most commonly used evaluation setup is to conduct cross-subject analysis on left hand vs. right hand classification [27,44]. Thus, we use it as well in our evaluation. Furthermore, like the previous works [30,43,44], we remove the low-quality recordings, i.e., the MI-EEG data of subject #88, #89, #92, #100 and #104.

In our evaluation, we use MI-EEG data (MI movement of left hand and right hand) of 104 subjects. To conduct cross-subject analysis, 70% of the subjects are randomly chosen for training, 10% for validation and 20% for testing. To reduce the chance of selection bias, we prepare five different evaluation sets, each of which contains a training set, a validation set and a test set. During the training process, early stopping is used to monitor the validation loss, and the hyperparameter "patience" is set to 30.

MI-EEG decoding models evaluated on the BCI competition IV 2a dataset. BCI Competition IV 2a dataset contains EEG data recorded from 9 subjects. As explained in Section 2.6, existing works generally performed within-subject four-class classification on this dataset (see Figs. 9 and 10). Thus, we too evaluate the representative methods under within-subject four-class classification on this dataset. Considering the amount of EEG data for each subject, we adopt the similar validation scheme as [21], where four-fold cross-validation is used for within-subject analysis on the BCI competition IV 2a dataset. Specifically, two of the four folds are chosen for training, one fold for validation and the final fold for testing. To prevent models from overfitting, we use early stopping during training, where training will be terminated when validation loss does not decrease in 30 epochs.

3.2.2. Classification results

Table 6 shows classification performance of 13 representative MI-EEG decoding models on the two widely-used public datasets. Firstly, it can be seen that a simple LSTM-based model [25] can only achieve around chance-level accuracy on two benchmark datasets. One possible reason for such a poor classification performance is the insufficient network depth, since this model only has one LSTM layer. To investigate whether stacking more LSTM layers could improve model performance, we further evaluate the classification performance after adding more layers to this model. More details and discussion will be presented in our ablation study in Section 3.3.

Secondly, our evaluation results show that Shallow ConvNet, EEGNet-4,2 and EEGNet-8,2 achieve very competitive classification performance on both datasets. Among them, the two different configurations of EEGNet not only perform well but also are highly compact (see Table 6). This shows the potential of developing accurate yet lightweight decoding models. Although Deep ConvNet and ETENet can achieve the same level of performance as the aforementioned three single-stream CNN based models on the PhysioNet dataset, they perform significantly poorly (See Table 7) on BCI IV competition 2a, as shown in Table 6. Studying these two models carefully, we find that they are closely related to Shallow ConvNet. Deep ConvNet can be regarded as the "deeper" version of Shallow ConvNet, containing more convolutional layers. Considering the size of training data and how Shallow ConvNet could perform better than Deep ConvNet, we think the main reason for Deep ConvNet poor performance on BCI competition IV2a is due to insufficient data, which hinders this deeper CNN to reach its full potential [22]. As for ETENet, its network architecture is fundamentally Shallow ConvNet with an additional fully connected (FC) layer before the output layer and with no Dropout after the pooling layer. These differences seem to be the reason for performance difference between Shallow ConvNet and ETENet. We will further investigate the effect of these differences on performance in our ablation study in Section 3.3.

Note that all the aforementioned single-stream CNN-based models take raw EEG data as input. To investigate the performance of models

Table 7
Wilcoxon Signed-Rank test for comparing several representative single-stream CNN-based models.

	P-value (PhysioNet)	P-value (BCI IV 2a)
ETENet vs. EEGNet-4,2	0.556	0.008
ETENet vs. EEGNet-8,2	0.588	0.011
ETENet vs. Shallow ConvNet	0.225	0.008
Deep ConvNet vs. EEGNet-4,2	0.434	0.01
Deep ConvNet vs. EEGNet-8,2	0.465	0.019
Deep ConvNet vs. Shallow ConvNet	0.043	0.015

using time–frequency domain features as input, we also evaluate a representative single-stream CNN-based model (i.e., pCNN [25]) whose input is time–frequency spectrogram images. Our evaluation results (See Table 6) show pCNN performs significantly poor ($p-values < 0.05$, Wilcoxon Signed-Rank test) on PhysioNet and BCI competition IV 2a datasets, compared to other single-stream CNN models like EEGNet-4,2 and Shallow ConvNet. A possible explanation for its poor performance is that the authors of pCNN only use three EEG channels corresponding to the electrodes C3, C4 and Cz to generate spectrogram images. Thus, some useful information that exists in other channels cannot be utilized by the model to better classify different MI-EEG signals. Unfortunately, we could not investigate the performance of other models, such as [12,74,82], which also used time–frequency domain features as input because their source code is unavailable. This could be explored in future studies when more related works publicly release their code.

Apart from several single-stream CNN-based models, we also evaluate three representative multi-stream CNN-based networks, namely EEGNet fusion, MBEEGNet and TS-SEFFNet. Although these three models have more complex network architectures, they do not show better classification performance (Refer to Table 6) than some simple single-stream CNN-based models, i.e., EEGNet-4,2 and EEGNet-8,2. By studying their network architecture, we find that EEGNet fusion and MBEEGNet are closely related to EEGNet. In fact, both EEGNet fusion and MBEEGNet have three feature extraction branches, which are different configurations of EEGNets without the output layer. The features from three different feature extraction branches are simply concatenated for the final predictions. Considering the architectural relationship among EEGNet, EEGNet fusion and MBEEGNet and their classification performance, we can reasonably suspect that effective feature fusion may be a key factor to affect the performance of multi-stream CNN-based models. In our ablation study in Section 3.3, we will explore the effect of feature fusion on multi-stream CNN-based decoding models.

Finally, for the three hybrid CNN-LSTM models, C-LSTM shows superior performance ($p-values < 0.03$, Wilcoxon Signed-Rank test) over two classical hybrid CNN-LSTM models (i.e., Cascade Model and Parallel Model). Especially on BCI IV 2a, C-LSTM achieves the best classification accuracy (69.1%) among all evaluated models. In fact, this model is an expanded version of Shallow ConvNet. The authors utilized an additional LSTM layer to capture the temporal dependencies between features for making better decisions. For Cascade Model and Parallel Model, although both of them adopted seemingly more sophisticated input formulation, i.e., topology-preserving images and contain much more learnable parameters, their performance is mediocre on the two datasets.

3.2.3. Efficiency for representative models

Apart from model performance, model efficiency is also an important metric. In this part, we further evaluate and analyse the training and testing time of the representative MI-EEG models. Table 8 shows the runtime performance of 13 representative models on two public datasets. Here, three points need to be noted. (1) The work [115] demonstrates that the same deep learning model exhibits varying runtimes when implemented with different deep learning frameworks. Thus, some results with close values in Table 8 may not be directly comparable because the source code of different models could be based on different deep learning frameworks (e.g., tensorflow and pytorch). Table 5 lists the deep learning framework that each model's source code is based on. (2) A small number of epochs do not mean that the models take less training time, e.g., Cascade Model, LSTM Model, etc. (3) To better evaluate runtime performance of the representative models, we test each model five times and then calculate the mean values and the standard deviations.

According to Table 8, we have several observations. Firstly, apart from LSTM Model, all compared models that take raw EEG data as input are trained faster than the models, e.g., Cascade Model, Parallel Model and pCNN, whose inputs are topology-preserving images or time–frequency spectrogram images. This might be attributed to two reasons: (1) Models that utilize topology-preserving or spectrogram images as input often employ more complex network architectures, such as hybrid CNN-LSTM networks. However, the models that use raw EEG data as input (e.g., EEGNet and Shallow ConvNet) are often based on CNN architectures, which are usually simple and contain less trainable parameters. (2) Some of these models that take the calculated features as input, e.g., pCNN, incorporate an additional preprocessing step to convert raw EEG data into images. However, the models that use raw EEG data as input do not have such an additional step. Secondly, CNN-based models typically require less training time compared to both LSTM-based models, where the LSTM unit performs intricate computations at each time point, and hybrid CNN-LSTM models, such as Parallel Model, which often feature complex network architectures. For example, a representative CNN-based model, i.e., ETENet, only spends around 131.6 s and 15.7 s finishing the training process on PhysioNet and BCI IV 2a respectively, while LSTM model needs around 15,614 s and 2416.6 s to finish the training process on the two datasets respectively. However, this does not mean non-CNN-based models will always be slower to train than the CNN-based models. C-LSTM is a typical example. This model only spends about 219.65 s and 38 s finishing the training process on two datasets respectively, which is faster than many CNN-based models, such as pCNN, EEGNet fusion, TS-SEFFNet, and so on. As illustrated in Section 3.2.2, C-LSTM simply modifies a compact model, i.e., Shallow ConvNet. This can explain why it only needs relatively less time for training. Thirdly, all evaluated models only require a small amount of time to make a prediction on the test dataset (we called this testing time). Specifically, most of models can complete the prediction of 115 EEG trials within one second, except several models with complex architectures and/or a large number of parameters, e.g., EEGNet fusion, Cascade Model, etc. LSTM model is the slowest at making a prediction, but its testing time is still acceptable (about 11 s).

3.3. Ablation study for representative models

In this section, we explore the effect of some design factors on the performance of typical architectural designs using ablation study. Before we discuss and analyse the results of our ablation study, we first introduce the three types of ablation studies that are commonly conducted in DL-based MI-EEG classification: network architecture ablation, feature ablation and channel ablation. (1) The network architecture ablation [30,116] analyzes the effect of specific parts of the DL model on the model performance by removing the corresponding

Table 8
Runtime performance of representative DL-based MI-EEG models. Bold values indicate the minimum time cost or minimum number of training epochs, while italic values represent the maximum time cost or maximum number of training epochs. All models are evaluated on the same machine with a GPU, i.e., Nvidia Quadro P6000. Testing time represents the total time the models take to predict 115 EEG trials. s means seconds.

Model	PhysioNet		BCI IV 2a		Testing time (s)
	#Epoch	Training time (s)	#Epoch	Training time (s)	
Shallow ConvNet	67 ± 10	194.8 ± 27.5	92 ± 55	48.8 ± 29.8	**0.14 ± 0.01**
Deep ConvNet	55 ± 12	167.6 ± 26.9	69 ± 4	38.0 ± 21.6	0.152 ± 0.01
EEGNet-4,2	*153 ± 46*	460.8 ± 148.1	*229 ± 50*	252.6 ± 54.6	0.199 ± 0.12
EEGNet-8,2	122 ± 29	367.8 ± 87.9	212 ± 44	39.3 ± 8.0	0.217 ± 0.07
ETENet	43 ± 9	**131.6 ± 30.0**	37 ± 7	**15.7 ± 1.0**	0.156 ± 0.02
pCNN	60 ± 5	1523.6 ± 127.1	87 ± 18	615.6 ± 123.4	0.794 ± 0.02
Cascade Model	**32 ± 1**	2710 ± 132.8	36 ± 4	352.7 ± 37.8	4.0 ± 0.07
Parallel Model	35 ± 3	3261.5 ± 266.9	41 ± 7	389.9 ± 53.2	3.55 ± 0.04
LSTM Model	38 ± 6	*15,614.8 ± 2054.6*	62 ± 31	*2416.6 ± 1223.4*	*11.07 ± 0.31*
EEGNet fusion	55 ± 6	605.6 ± 67.1	73 ± 21	183.7 ± 52.7	1.35 ± 0.08
MBEEGNet	51 ± 5	307 ± 28.6	75 ± 36	26.75 ± 12.7	0.347 ± 0.03
C-LSTM	63 ± 3	219.65 ± 9.6	78 ± 20	38.0 ± 2.0	0.24 ± 0.05
TS-SEFFNet	41 ± 3	580.3 ± 46.7	**32 ± 2**	23.4 ± 1.6	0.45 ± 0.04

Table 9
Wilcoxon Signed-Rank test for comparing EEGNet fusion with its EEGNet branches. Baseline is EEGNet fusion. The structure of EEGNet fusion-B1, -B2 and -B3 are shown in Fig. 11(a).

	P-value (PhysioNet)	P-value (BCI IV 2a)
Baseline vs. EEGNet fusion-B1	0.685	0.26
Baseline vs. EEGNet fusion-B2	0.345	0.313
Baseline vs. EEGNet fusion-B3	0.786	0.173

Table 10
Wilcoxon Signed-Rank test for comparing MBEEGNet with its EEGNet branches. Baseline is MBEEGNet. The structure of MBEEGNet-B1, -B2 and -B3 are shown in Fig. 11(b).

	P-value (PhysioNet)	P-value (BCI IV 2a)
Baseline vs. MBEEGNet-B1	0.054	0.066
Baseline vs. MBEEGNet-B2	0.068	0.314
Baseline vs. MBEEGNet-B3	0.343	0.859

parts from the model. (2) Feature ablation [117] is a procedure where each input feature is replaced by a given reference, and its performance difference to the original performance is calculated and examined. (3) Channel ablation (CA) [24] investigates the effect of the specific EEG channels on model performance by removing some EEG channels or only using some specific channels.

Here, we mainly focus on network architecture ablation to shed light on the future design of DL models.

- **Effective feature fusion could be indispensable for designing accurate multi-stream CNN-based models.** Two of the multi-stream CNN-based models we evaluate, EEGNet fusion and MBEEGNet, consist of an input layer, three different branches, a feature concatenation layer and an output layer. Each of their branches has the same architecture as EEGNet without the output layer (See Fig. 11). As mentioned in Section 3.2, these two models (i.e., EEGNet fusion and MBEEGNet) do not show superior performance over two configurations of EEGNet on the two datasets.
  This leads to a question: *Are the two aforementioned multi-stream CNN-based models (with multiple branches) more accurate than three different configurations of EEGNet corresponding to each of their branches respectively?* If not, it means that simply concatenating features extracted by different branches is not a good design choice.
  To answer the question, we evaluate the classification performance of each branch of EEGNet fusion and MBEEGNet respectively on the two public datasets. Specifically, we use the corresponding source code of each branch of the two models and add an output layer to each of them. By doing so, we obtain three EEGNet models with different configurations based on EEGNet fusion and MBEEGNet respectively (See Fig. 11). Then, we compare the classification performance of EEGNet fusion and MBEEGNet with the obtained corresponding EEGNet models respectively.
  Tables 11 and 12 show the results of the ablation study. It can be seen that some branches of EEGNet fusion outperform the complete model in some cases (Refer to Table 11). Similarly, Table 12 shows some branches of MBEEGNet perform better than the complete model in some cases. To evaluate whether the differences of average accuracy shown in Tables 11 and 12 are random or not, we further conduct Wilcoxon Signed-Rank test shown in Tables 9 and 10.
  From Table 9, we see there is no statistically significant difference between EEGNet fusion and its EEGNet branches ($p-values > 0.15$) on the two datasets. Likewise, we do not see a significantly performance difference between MBEEGNet and its EEGNet branches ($p-values > 0.05$) in Table 10. This illustrates that simply fusing features from different feature extraction branches cannot bring improvement on classification performance. The main reason might be that fusing different features without feature selection often leads to accumulation of irrelevant information [118], which can negatively affect model performance. Thus, feature fusion layers should be carefully designed for multi-stream CNN-based models.
- **LSTM should be combined with techniques that can extract spatial features when using raw EEG data as input.** As shown in Section 3.2, LSTM with one hidden layer could only achieve chance-level accuracy on the two public datasets. To investigate if this poor performance is caused by insufficient number of LSTM layers, we test the classification performance of LSTM with more hidden layers. The performance results are shown in Table 13. According to the Wilcoxon Signed-Rank test shown in Table 14, there is no significant increase of performance after stacking more LSTM layers (up to 5 hidden layers). This means that simply adding more LSTM layers to a LSTM model cannot bring significant improvement on classification performance. The main reason could be that pure LSTM models generally lack the ability to capture spatial dependencies between EEG channels, although LSTM layers are capable of learning temporal features. A feasible strategy to improve model performance is to combine LSTM with other techniques that can learn spatial dependencies between EEG channels, such as CNN [30,81] or CSP [15].
  To find out whether adding additional spatial feature extraction can improve model performance, we conduct an ablation study on two typical hybrid CNN-LSTM models: Cascade model and C-LSTM. For Cascade Model, we remove the spatial feature extraction part, a 2D-CNN and a fully connected layer, from it. The

Table 11
Classification results of EEGNet fusion and the EEGNet branches of EEGNet fusion. EEGNet fusion-B1, -B2 and -B3 represent respectively the first, second and third branch of EEGNet fusion with an output layer.

Model	PhysioNet				BCI IV 2a			
	Acc (%)	F1-score	AUC	Kappa	Acc (%)	F1-score	AUC	Kappa
EEGNet fusion-B1	76.01	0.753	0.849	0.52	**52.58**	**0.514**	**0.785**	**0.368**
EEGNet fusion-B2	**76.55**	0.76	**0.851**	**0.531**	48.92	0.454	0.766	0.318
EEGNet fusion-B3	76.33	0.758	0.848	0.527	41.4	0.359	0.72	0.219
EEGNet fusion	76.18	**0.784**	0.849	0.523	44.31	0.381	0.759	0.258

Table 12
Classification results of MBEEGNet and the EEGNet branches of MBEEGNet. MBEEGNet-B1, -B2 and -B3 represent respectively the first, second and third branch of MBEEGNet with an output layer.

Model	PhysioNet				BCI IV 2a			
	Acc (%)	F1-score	AUC	Kappa	Acc (%)	F1-score	AUC	Kappa
MBEEGNet-B1	73.83	0.737	0.812	0.467	**52.47**	**0.489**	**0.795**	**0.37**
MBEEGNet-B2	74.79	0.751	0.832	0.496	44.70	0.385	0.759	0.26
MBEEGNet-B3	**76.51**	**0.757**	0.846	**0.53**	38.73	0.305	0.743	0.183
MBEEGNet	75.99	0.755	**0.853**	0.52	41.07	0.337	0.76	0.21

Table 13
Classification performance of LSTM models with different number of hidden layers.

Models	PhysioNet				BCI IV 2a			
	Acc (%)	F1-score	AUC	Kappa	Acc (%)	F1-score	AUC	Kappa
1-layer LSTM	49.93	0.494	0.507	−0.003	26.08	0.246	0.506	0.001
2-layer LSTM	52.74	0.492	0.542	0.055	26.18	**0.248**	**0.513**	0.015
3-layer LSTM	53.47	**0.575**	0.562	0.07	25.56	0.24	0.507	0.007
4-layer LSTM	**54.58**	0.526	**0.569**	**0.09**	**26.49**	0.245	0.512	**0.02**
5-layer LSTM	52.24	0.527	0.545	0.045	26.01	0.24	0.51	0.013

Table 14
Wilcoxon Signed-Rank test for comparing the 1-layer LSTM model with LSTM models with more layers.

	P-value (PhysioNet)	P-value (BCI IV 2a)
1-layer LSTM vs. 2-layer LSTM	0.685	0.953
1-layer LSTM vs. 3-layer LSTM	0.345	0.401
1-layer LSTM vs. 4-layer LSTM	0.786	0.441
1-layer LSTM vs. 5-layer LSTM	0.806	0.859

remaining part is a two-hidden-layer LSTM model. Similarly, with regard to C-LSTM, we mainly keep its LSTM part and evaluate the performance of the trimmed C-LSTM. For these two trimmed models, we take the data from the same segments used in the original ones as input. Tables 15 and 16 show the results of the ablation study. It can be seen that the removal of the CNN spatial feature extraction from the two hybrid CNN-LSTM models leads to a significant performance degradation ($p - values < 0.05$) on two benchmark datasets. This verifies that LSTM should be enhanced with additional feature extraction techniques to better decode MI-EEG signals.

- **The use of Dropout may make a limited contribution to the improvement of model performance.** As discussed in Section 3.2, among the single-stream CNN-based models, ETENet is the only model that does not adopt dropout [119] in the training process. In fact, due to the effectiveness of dropout for reducing overfitting, it has been widely used in DL-based MI-EEG decoding models, such as [21,22,47]. Therefore, it would be interesting to explore the impact of dropout on model performance.
  To this end, we carry out the following experiments. (1) We add a dropout layer after the pooling layer of ETENet, like many typical single-stream CNN-based models [21,22,31]. Then, we compare its classification performance with the original model. The dropout rate is set to 0.5, which is the same as many existing methods [21,22,31]. (2) We remove dropout layers from several typical models (i.e., Shallow ConvNet, Deep ConvNet, EEGNet-8,2 and TS-SEFFNet) and compare them with their corresponding original model. Table 17 shows the performance results.
  Although applying dropout in the training process can improve model performance in most cases, there is not a significant improvement on classification performance for all five typical decoding models according to Tables 17 and 18. This verifies that the contribution made by dropout in model training is limited.
- **Adding fully connected layers to the model may not be a good design choice.** As discussed in Section 3.2, the main difference between the network architecture of ETENet and Shallow ConvNet is that ETENet has an additional Fully Connected (FC) layer before the output layer. Through further investigation, we find that this additional FC layer accounts for 66% (around 201.6k parameters) of the total number (around 305.5k) of parameters of ETENet. Since ETENet also does not show superior performance over Shallow ConvNet on the two public datasets (see Table 6), it draws our interest to explore the effect of the FC layer on model performance.
  To this end, in our ablation study, we evaluate the performance of two representative models (i.e., ETENet and Cascade Model) without the FC layer. The results are shown in Table 19. From the table we see that there is no significant performance difference ($p - values > 0.4$) between these two models and them without FC. This shows that the FC layer could be redundant for these two models. Considering that FC layers usually account for a large proportion of the total number of trainable parameters, we suggest the future studies should add FC layers with care. At least, we should verify that the added FC layers have positive effect on model performance, especially when the training datasets are not large enough.

Note that network architectures mentioned in this paper are used in various other EEG classification tasks (e.g., EEG emotion recognition [120], EEG depression diagnosis [121], etc.), thus our aforementioned conclusions could inform the network design of those EEG classification tasks as well. Future studies need to further verify this.

Table 15
Ablation studies for Cascade Model on two public MI-EEG datasets.

Models	PhysioNet				BCI IV 2a			
	Acc (%)	F1-score	AUC	Kappa	Acc (%)	F1-score	AUC	Kappa
Cascade Model w/o CNN	54.19	0.55	0.564	0.084	27.96	0.258	0.541	0.049
Cascade Model	**58.62**	**0.583**	**0.631**	**0.185**	**34.31**	**0.332**	**0.589**	**0.107**

Table 16
Ablation studies for C-LSTM on two public MI-EEG datasets.

Models	PhysioNet				BCI IV 2a			
	Acc (%)	F1-score	AUC	Kappa	Acc (%)	F1-score	AUC	Kappa
C-LSTM w/o CNN	51.11	0.511	0.521	0.022	25.19	0.241	0.502	0.003
C-LSTM	**74.72**	**0.736**	**0.797**	**0.495**	**69.1**	**0.679**	**0.889**	**0.587**

Table 17
The classification results of several typical models with and without dropout.

Model		PhysioNet				BCI IV 2a			
		Acc (%)	F1-score	AUC	Kappa	Acc (%)	F1-score	AUC	Kappa
ETENet	w/o Dropout	76.23	0.762	0.85	0.525	**38.5**	**0.371**	**0.652**	**0.18**
	+ Dropout	**78.53**	**0.787**	**0.868**	**0.571**	35.38	0.336	0.627	0.138
EEGNet	w/o Dropout	77.14	0.761	0.853	0.541	55.79	**0.552**	**0.793**	0.411
	+ Dropout	**77.51**	**0.777**	**0.859**	**0.55**	**56.66**	0.548	0.787	**0.422**
Shallow ConvNet	w/o Dropout	74.67	0.741	0.828	0.494	55.79	0.552	**0.793**	0.411
	+ Dropout	**75.33**	**0.746**	**0.838**	**0.506**	**56.29**	**0.559**	0.792	**0.417**
Deep ConvNet	w/o Dropout	74.81	0.748	0.838	0.496	38.2	0.375	0.654	0.176
	+ Dropout	**77.89**	**0.766**	**0.876**	**0.558**	**42.53**	**0.411**	**0.689**	**0.234**
TS-SEFFNet	w/o Dropout	66.96	0.661	0.729	0.339	40.16	**0.393**	0.663	0.202
	+ Dropout	**67.08**	**0.667**	**0.759**	**0.342**	**42.59**	0.39	**0.74**	**0.235**

Table 18
Wilcoxon Signed-Rank test for evaluating whether the difference of classification performance between MI-EEG decoding models is significant. "w/o" means without.

	P-value (PhysioNet)	P-value (BCI IV 2a)
ETENet vs. ETENet with Dropout	0.223	0.011
EEGNet vs. EEGNet w/o Dropout	0.893	0.314
Shallow ConvNet vs. Shallow ConvNet w/o Dropout	0.313	0.678
Deep ConvNet vs. Deep ConvNet w/o Dropout	0.345	0.213
TS-SEFFNet vs. TS-SEFFNet w/o Dropout	0.374	0.893

Table 19
Ablation studies for ETENet and Cascade Model on fully connected layers.

Models	PhysioNet				BCI IV 2a			
	Acc (%)	F1-score	AUC	Kappa	Acc (%)	F1-score	AUC	Kappa
ETENet w/o FC	**77.03**	**0.763**	**0.858**	**0.541**	37.05	0.312	**0.668**	0.162
ETENet	76.23	0.762	0.85	0.525	**38.5**	**0.371**	0.652	**0.18**
Cascade Model w/o FC	**60.16**	**0.596**	**0.634**	**0.188**	**37.05**	0.312	**0.668**	0.162
Cascade Model	58.62	0.583	0.631	0.185	34.31	0.332	0.589	0.107

4. Open issues and future research direction

DL-based models have achieved significant improvement for MI-EEG classification, but there are still some limitations that slowed the development of this field. In this section, we discuss some open issues and potential research directions.

4.1. Network architecture design

Although various network architectures have been developed so far (see Section 2.4), how to design more effective decoding networks to further improve classification performance is still a challenging issue.

Based on the evaluation of 13 typical models in Section 3.2, we can see that the FBCSP-like single-stream CNN-based architecture shows great potential, and some researchers (e.g., [21]) have developed accurate yet extremely compact model based on this design pattern.

Apart from the FBCSP-like single-stream CNN-based architecture, using some newly emerged network architectures, e.g., CapsuleNet [34], could be a promising research direction. However, further study is needed as there is limited literature for this direction. Also there is no source code available for the new architecture at present.

More importantly, existing DL-based MI-EEG classification models are generally designed empirically. This limits the discovery of better network architectures. A promising solution is to utilize neural architecture search (NAS) to automatically discover optimal network architectures, as some studies [122,123] have shown that the networks found by NAS perform better than the handcrafted ones in fields such as image classification.

4.2. Imbalanced MI-EEG data

Due to some factors, e.g., data missing, subjects' withdrawal from the experiment, etc, it is quite common there is data imbalance in the number of MI trials for each subject and/or each class in MI-EEG datasets, which has a negative effect on the performance of DL-based models [124].

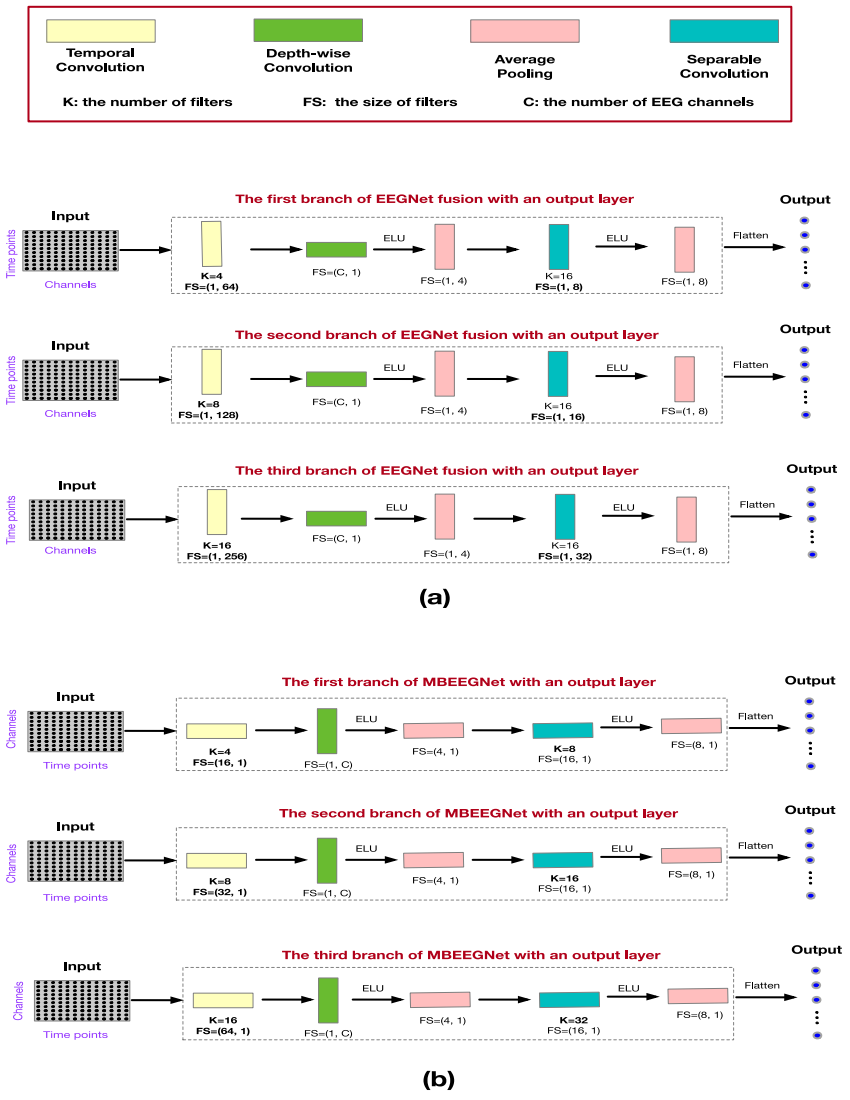

Fig. 11. The details of different configurations of EGGNet corresponding to all branches of two multi-stream CNN-based models. (a) EEGNet fusion and (b) MBEEGNet.

A common solution to the data imbalance problem is to use data augmentation which can keep the number of trials of different classes and subjects the same. Many data augmentation techniques, such as [11,31,59], have been proposed and used in MI-EEG classification. Some of them are inspired by image classification tasks, such as adding noise [31], flipping the data [31], and so on. In fact, these techniques (i.e., adding noise and flipping the data) are also often used in other EEG classification tasks [125,126]. Besides, a recent study [127] also tries to augment data by using generative adversarial network (GAN), which is a potential research direction for solving data imbalance in MI-EEG classification.

4.3. Data distribution discrepancy

Existing DL-based methods can generally achieve good classification performance when the training set and test set have the same data distribution [21]. However, it is often difficult to meet this ideal condition in the real world. For example, the distributions of the collected MI-EEG data vary significantly across subjects [64] because of different EEG patterns between subjects, different placements of EEG caps on subjects' scalps, etc. When the data collected from different subjects is spilt into the training set and the test set, there is a data distribution discrepancy between the two sets, which could result in negative transfer [64,128]. Negative transfer means the use of source domain data for training weakens the classification performance of models in target domain data for testing, mainly due to the dissimilar data distributions between the source domain and the target domain.

A possible solution is to adopt effective transfer learning. Until now, some recent DL-based works, such as [36,52,73], have applied transfer learning in EEG classification. Transfer learning in these works can be classified as fine-tuning adaptation [73,129] and domain adversarial adaptation [36,52,53].

Fine-tuning adaptation used in MI-EEG classification usually consists of three steps. The first step is to select an effective pre-trained model (e.g., VGG16 [130], ResNet50 [131], Deep ConvNet [22], etc.). The second one is to replace the output layer of the pre-trained model with a new output layer, where the number of neurons is equal to the number of classes of the target task. The third step is to fine-tune the revised pre-trained model using the target EEG data or the spectrogram images based on the target EEG data, which enables the model to be applied in the target EEG classification task.

Domain adversarial adaptation has attracted much attention recently. Some studies [36,52,53] have developed domain adversarial neural networks (DANN) to tackle the distribution discrepancy between the source domain and the target domain in MI-EEG classification. These DANN-based methods usually contain a feature extractor, a classifier and a domain discriminator. The feature extractor is used to obtain deep representations from source domain data and target

domain data. The job of the classifier is to decode the obtained feature representations. As for the domain discriminator, it is designed to predict the domain labels. During the training period, the discriminator tries to predict the origin of the feature representations extracted by the feature extractor, while the feature extractor strives to fool the discriminator, making the discriminator fail to predict the domain labels of the feature representations. This forces the feature extractor to extract more general features, which can alleviate the distribution discrepancy between the source and target domains.

All these transfer learning-based methods report their superior performance over some traditional methods (e.g., DNN, standard CNN, etc.) in their original papers. Especially for the methods based on domain adversarial adaptation, the reported results [52,53] show that they can outperform some SOTA MI-EEG classification models (e.g., EEGNet). This shows domain adversarial adaptation-based models have a promising future. However, there are still several open issues for domain adversarial adaptation-based methods, such as the loss of the domain-specific features caused by the shared feature extractor [36,52] and the complex training process due to the additional discriminator [36,52] or classifier [53]. These aforementioned issues need to be further studied.

A most recent study [132] has comprehensively surveyed negative transfer and presented a reliable transfer learning scheme to alleviate negative transfer. The authors give some solutions to avoid negative transfer according to the results of domain similarity estimation, which might be helpful for future studies, especially for cross-subject training.

4.4. Real-world application of EEG-based MI-BCI system

Recently researchers [57,60,78] have a tendency to design more and more complex network architectures. Although these recent models show superior classification performance over many previous methods, their high model complexity can influence the inference speed, which is crucial for real-world application [133]. In fact, most of existing works only focus on classification performance and ignore other factors, such as inference speed and model size, that are indispensable for practical application.

Future studies should consider the model complexity, as the BCI systems need to be deployed in mobile devices in real-world scenarios and need to produce real-time predictions. A possible solution is to use network pruning techniques [134], which can remove superfluous parameters from a trained model with minimal loss in classification performance. This type of techniques might be able to help some high-performing but complex models be adopted in real-world applications. In fields such as image segmentation and object detection, studies [135,136] have shown the effectiveness of network pruning methods for developing lightweight models with high classification accuracy.

5. Conclusions

Rapid advances in deep learning have largely facilitated the development of MI-EEG classification. Currently, using deep learning to develop MI-EEG decoding models becomes dominant because of the ability to automatically execute feature engineering. Although various deep learning techniques have been applied in MI-EEG classification, the majority of the existing studies are generally based on several network architectures, e.g., CNN, LSTM, hybrid deep network, etc.

Performance comparison is commonly used to verify the effectiveness of models. However, we find that there are several problems in the performance comparison of many existing works. (1) Different meanings for cross-subject classification. (2) Different evaluation strategies (e.g., performing different classification tasks to evaluate the baseline and proposed models) used in performance comparison. (3) Different interpretation of baseline models when there is no publicly available source code. (4) Performance comparison only conducted on the private datasets. These problems could cause invalid and/or unfair comparison of results. In this paper, we present several guidelines that future studies should use to overcome the aforementioned problems.

Among the deep learning techniques used, CNN is the most popular design choice. The models based on CNN often use raw EEG data or extracted time–frequency domain features as input. When the input formulation is the former, the computational steps of a SOTA traditional method, i.e., FBSCP, are often adopted as the guidelines for model design. When the latter is chosen as input, the network design often draws on the experience of models for computer vision. By evaluating and comparing typical decoding models on the two benchmark datasets, we are surprised to find that an extremely simple FBCSP-like single-stream CNN-based decoding model (i.e., EEGNet.) can achieve better performance than many recent models with more complex network architecture. This reveals that FBCSP-like CNN-based architecture is a promising design choice. Moreover, through ablation studies, we find and verify that effective feature fusion plays an indispensable role for developing accurate multi-stream CNN-based models. LSTM is another widely used deep learning technique for MI-EEG classification. However, our experiment results illustrate the models that simply stack LSTM layers cannot well classify raw EEG signals. Although a combination of CNN and LSTM model (i.e., C-LSTM) can achieve very competitive performance, the other two representative hybrid CNN-LSTM models we evaluate only achieve mediocre performance, which may be because of their complex network architecture and the insufficient data. Recently, some researchers also explore other possible combinations of deep learning techniques, such as the combination of GCN and BiLSTM. Due to the limited literature, further study on these hybrid architectures is needed.

DL-based methods have made some progress in MI-EEG classification. However, there remain some challenges such as data distribution discrepancy and lack of attention to the importance of real-world applications. By utilizing our performance comparison guidelines and our findings of influential design factors on the performance of typical MI-EEG classification architectural designs, we hope that future studies could solve these issues fairly and expeditiously.

Declaration of competing interest

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Papanastasiou G, Drigas A, Skianis C, Lytras M. Brain computer interface based applications for training and rehabilitation of students with neurodevelopmental disorders. A literature review. Heliyon 2020;6(9):e04250. http://dx.doi.org/10.1016/j.heliyon.2020.e04250, URL https://www.sciencedirect.com/science/article/pii/S240584402031094X.

[2] Rohm M, Schneiders M, Müller C, Kreilinger A, Kaiser V, Müller-Putz GR, Rupp R. Hybrid brain–computer interfaces and hybrid neuroprostheses for restoration of upper limb functions in individuals with high-level spinal cord injury. Artif Intell Med 2013;59(2):133–42. http://dx.doi.org/10.1016/j.artmed.2013.07.004, URL https://www.sciencedirect.com/science/article/pii/S0933365713001176. Special Issue: Brain-computer interfacing.

[3] Irimia D, Ortner R, Krausz G, Guger C, Poboroniuc M. BCI application in robotics control. IFAC Proc Vol 2012;45(6):1869–74. http://dx.doi.org/10.3182/20120523-3-RO-2023.00432, URL https://www.sciencedirect.com/science/article/pii/S1474667016334231. 14th IFAC Symposium on Information Control Problems in Manufacturing.

[4] Zhang X, Yao L, Huang C, Sheng QZ, Wang X. Intent recognition in smart living through deep recurrent neural networks. In: Neural information processing. Cham: Springer International Publishing; 2017, p. 748–58.

[5] Leeb R, Perdikis S, Tonin L, Biasiucci A, Tavella M, Creatura M, Molina A, Al-Khodairy A, Carlson T, d.R. Millán J. Transferring brain–computer interfaces beyond the laboratory: Successful application control for motor-disabled users. Artif Intell Med 2013;59(2):121–32. http://dx.doi.org/10.1016/j.artmed.2013.08.004, URL https://www.sciencedirect.com/science/article/pii/S0933365713001218. Special Issue: Brain-computer interfacing.

[6] Ahn M, Lee M, Choi J, Jun SC. A review of brain-computer interface games and an opinion survey from researchers, developers and users. Sensors 2014;14(8):14601–33. http://dx.doi.org/10.3390/s140814601, URL https://www.mdpi.com/1424-8220/14/8/14601.

[7] Holz EM, Höhne J, Staiger-Sälzer P, Tangermann M, Kübler A. Brain–computer interface controlled gaming: Evaluation of usability by severely motor restricted end-users. Artif Intell Med 2013;59(2):111–20. http://dx.doi.org/10.1016/j.artmed.2013.08.001, URL https://www.sciencedirect.com/science/article/pii/S0933365713001140. Special Issue: Brain-computer interfacing.

[8] Zhang X, Yao L, Kanhere SS, Liu Y, Gu T, Chen K. MindID: Person identification from brain waves through attention-based recurrent neural network. Proc ACM Interact Mob Wearable Ubiq Technol 2018;2(3). http://dx.doi.org/10.1145/3264959.

[9] Craik A, He Y, Contreras-Vidal JL. Deep learning for electroencephalogram (EEG) classification tasks: a review. J Neural Eng 2019;16(3):031001. http://dx.doi.org/10.1088/1741-2552/ab0ab5.

[10] Chaudhary S, Taran S, Bajaj V, Sengur A. Convolutional neural network based approach towards motor imagery tasks EEG signals classification. IEEE Sens J 2019;19(12):4494–500. http://dx.doi.org/10.1109/JSEN.2019.2899645.

[11] Yang L, Song Y, Ma K, Xie L. Motor imagery EEG decoding method based on a discriminative feature learning strategy. IEEE Trans Neural Syst Rehabil Eng 2021;29:368–79. http://dx.doi.org/10.1109/TNSRE.2021.305195.

[12] Zhang H, Zhao X, Wu Z, Sun B, Li T. Motor imagery recognition with automatic EEG channel selection and deep learning. J Neural Eng 2021;18(1):016004. http://dx.doi.org/10.1088/1741-2552/abca16.

[13] Hwang H-J, Kim S, Choi S, Im C-H. EEG-based brain-computer interfaces: A thorough literature survey. Int J Hum–Comput Interact 2013;29(12):814–26. http://dx.doi.org/10.1080/10447318.2013.780869, arXiv:https://doi.org/10.1080/10447318.2013.780869.

[14] Kousarrizi MRN, Ghanbari AA, Teshnehlab M, Shorehdeli MA, Gharaviri A. Feature extraction and classification of EEG signals using wavelet transform, SVM and artificial neural networks for brain computer interfaces. In: 2009 international joint conference on bioinformatics, systems biology and intelligent computing. Shanghai: IEEE; 2009, p. 352–5. http://dx.doi.org/10.1109/IJCBS.2009.100.

[15] Ang KK, Chin ZY, Zhang H, Guan C. 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). Hong Kong: IEEE; 2008, p. 2390–7. http://dx.doi.org/10.1109/IJCNN.2008.4634130.

[16] Venkatachalam K, Devipriya A, Maniraj J, Sivaram M, Ambikapathy A, Amiri Iraj S. A novel method of motor imagery classification using EEG signal. Artif Intell Med 2020;103:101787. http://dx.doi.org/10.1016/j.artmed.2019.101787, URL https://www.sciencedirect.com/science/article/pii/S0933365719304816.

[17] Li Y, Zhang X-R, Zhang B, Lei M-Y, Cui W-G, Guo Y-Z. A channel-projection mixed-scale convolutional neural network for motor imagery EEG decoding. IEEE Trans Neural Syst Rehabil Eng 2019;27(6):1170–80. http://dx.doi.org/10.1109/TNSRE.2019.2915621.

[18] An X, Kuang D, Guo X, Zhao Y, He L. A deep learning method for classification of EEG data based on motor imagery. In: Huang D-S, Han K, Gromiha M, editors. Intelligent computing in bioinformatics. Cham: Springer International Publishing; 2014, p. 203–10.

[19] Sturm I, Lapuschkin S, Samek W, Müller K-R. Interpretable deep neural networks for single-trial EEG classification. J Neurosci Methods 2016;274:141–5. http://dx.doi.org/10.1016/j.jneumeth.2016.10.008, URL https://www.sciencedirect.com/science/article/pii/S0165027016302333.

[20] Lu N, Li T, Ren X, Miao H. A deep learning scheme for motor imagery classification based on restricted Boltzmann machines. IEEE Trans Neural Syst Rehabil Eng 2017;25(6):566–76. http://dx.doi.org/10.1109/TNSRE.2016.2601240.

[21] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. J Neural Eng 2018;15(5):056013.

[22] Tibor SR, Tobias SJ, Josef FLD, Martin G, Katharina E, Michael T, Frank H, Wolfram B, Tonio B. Deep learning with convolutional neural networks for EEG decoding and visualization. Hum Brain Map 2017;38(11):5391–420. http://dx.doi.org/10.1002/hbm.23730.

[23] Deng X, Zhang B, Yu N, Liu K, Sun K. Advanced TSGL-EEGNet for motor imagery EEG-based brain-computer interfaces. IEEE Access 2021;9:25118–30.

[24] Xu B, Zhang L, Song A, Wu C, Li W, Zhang D, Xu G, Li H, Zeng H. Wavelet transform time-frequency image and convolutional network-based motor imagery EEG classification. IEEE Access 2019;7:6084–93. http://dx.doi.org/10.1109/ACCESS.2018.2889093.

[25] Tayeb Z, Fedjaev J, Ghaboosi N, Richter C, Everding L, Qu X, Wu Y, Cheng G, Conradt J. Validating deep neural networks for online decoding of motor imagery movements from EEG signals. Sensors 2019;19(1):210. http://dx.doi.org/10.3390/s19010210.

[26] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, http://dx.doi.org/10.48550/ARXIV.1409.1556, URL https://arxiv.org/abs/1409.1556.

[27] Zhang G, Davoodnia V, Sepas-Moghaddam A, Zhang Y, Etemad A. Classification of hand movements from EEG using a deep attention-based LSTM network. IEEE Sens J 2020;20(6):3113–22. http://dx.doi.org/10.1109/JSEN.2019.2956998.

[28] Wang P, Jiang A, Liu X, Shang J, Zhang L. LSTM-based EEG classification in motor imagery tasks. IEEE Trans Neural Syst Rehabil Eng 2018;26(11):2086–95.

[29] Luo T-J, Zhou C, Chao F. Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network. BMC Bioinformatics 2018;19. http://dx.doi.org/10.1186/s12859-018-2365-1.

[30] Zhang D, Yao L, Zhang X, Wang S, Chen W, Boots R, Benatallah B. Cascade and parallel convolutional recurrent neural networks on EEG-based intention recognition for brain computer interface. Proc AAAI Conf Artif Intell 2018;32(1):1703–10.

[31] Freer D, Yang G-Z. Data augmentation for self-paced motor imagery classification with C-LSTM. J Neural Eng 2020;17(1):016041. http://dx.doi.org/10.1088/1741-2552/ab57c0.

[32] Dai M, Zheng D, Na R, Wang S, Zhang S. EEG classification of motor imagery using a novel deep learning framework. Sensors 2019;19(3):19. http://dx.doi.org/10.3390/s19030551, URL https://www.mdpi.com/1424-8220/19/3/551.

[33] Ha K-W, Jeong J-W. Decoding two-class motor imagery EEG with capsule networks. In: 2019 IEEE international conference on big data and smart computing (BigComp). Kyoto: IEEE; 2019, p. 1–4. http://dx.doi.org/10.1109/BIGCOMP.2019.8678917.

[34] Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. 2017, http://dx.doi.org/10.48550/ARXIV.1710.09829, URL https://arxiv.org/abs/1710.09829.

[35] Ha K-W, Jeong J-W. Motor imagery EEG classification using capsule networks. Sensors 2019;19(13):2854. http://dx.doi.org/10.3390/s19132854, URL https://www.mdpi.com/1424-8220/19/13/2854.

[36] Zhao H, Zheng Q, Ma K, Li H, Zheng Y. Deep representation-based domain adaptation for nonstationary EEG classification. IEEE Trans Neural Netw Learn Syst 2021;32(2):535–45. http://dx.doi.org/10.1109/TNNLS.2020.3010780.

[37] Autthasan P, Chaisaen R, Sudhawiyangkul T, Rangpong P, Kiatthaveephong S, Dilokthanakul N, Bhakdisongkhram G, Phan H, Guan C, Wilaiprasitporn T. MIN2net: End-to-end multi-task learning for subject-independent motor imagery EEG classification. IEEE Trans Biomed Eng 2022;69(6):2105–18. http://dx.doi.org/10.1109/TBME.2021.3137184.

[38] Xie Y, Wang K, Meng J, Yue J, Meng L, Yi W, Jung T-P, Xu M, Ming D. Cross-dataset transfer learning for motor imagery signal classification via multi-task learning and pre-training. J Neural Eng 2023. URL http://iopscience.iop.org/article/10.1088/1741-2552/acfe9c.

[39] Yuan H, He B. Brain–computer interfaces using sensorimotor rhythms: Current state and future perspectives. IEEE Trans Biomed Eng 2014;61(5):1425–35. http://dx.doi.org/10.1109/TBME.2014.2312397.

[40] Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A, Yger F. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. J Neural Eng 2018;15(3):031005. http://dx.doi.org/10.1088/1741-2552/aab2f2.

[41] Roy Y, Banville H, Albuquerque I, Gramfort A, Falk TH, Faubert J. Deep learning-based electroencephalography analysis: a systematic review. J Neural Eng 2019;16(5):051001. http://dx.doi.org/10.1088/1741-2552/ab260c.

[42] Al-Saegh A, Dawwd SA, Abdul-Jabbar JM. Deep learning for motor imagery EEG-based classification: A review. Biomed Signal Process Control 2021;63:102172. http://dx.doi.org/10.1016/j.bspc.2020.102172, URL https://www.sciencedirect.com/science/article/pii/S1746809420303116.

[43] Fang Z, Wang W, Ren S, Wang J, Shi W, Liang X, Fan C-C, Hou Z-G. Learning regional attention convolutional neural network for motion intention recognition based on EEG data. In: Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20. Yokohama: International Joint Conferences on Artificial Intelligence Organization; 2020, p. 1570–6. http://dx.doi.org/10.24963/ijcai.2020/218.

[44] Zhang D, Chen K, Jian D, Yao L. Motor imagery classification via temporal attention cues of graph embedded EEG signals. IEEE J Biomed Health Inf 2020;24(9):2570–9. http://dx.doi.org/10.1109/JBHI.2020.2967128.

[45] Kumar S, Sharma A, Mamun K, Tsunoda T. A deep learning approach for motor imagery EEG signal classification. In: 2016 3rd Asia-Pacific world congress on computer science and engineering (APWC on CSE). Nadi, Fiji: IEEE; 2016, p. 34–9. http://dx.doi.org/10.1109/APWC-on-CSE.2016.017.

[46] Sakhavi S, Guan C, Yan S. Parallel convolutional-linear neural network for motor imagery classification. In: 2015 23rd European signal processing conference (EUSIPCO). Nice, France: IEEE; 2015, p. 2736–40. http://dx.doi.org/10.1109/EUSIPCO.2015.7362882.

[47] Li Y, Guo L, Liu Y, Liu J, Meng F. A temporal-spectral-based squeeze-and-excitation feature fusion network for motor imagery EEG decoding. IEEE Trans Neural Syst Rehabil Eng 2021;29:1534–45. http://dx.doi.org/10.1109/TNSRE.2021.309990.

[48] Wei X, Ortega P, Faisal AA. Inter-subject deep transfer learning for motor imagery EEG decoding. In: 2021 10th international IEEE/EMBS conference on neural engineering (NER). 2021, p. 21–4.

[49] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009;339. http://dx.doi.org/10.1136/bmj.b2535, arXiv:https://www.bmj.com/content/339/bmj.b2535.full.pdf. URL https://www.bmj.com/content/339/bmj.b2535.

[50] Altuwaijri GA, Muhammad G. A multibranch of convolutional neural network models for electroencephalogram-based motor imagery classification. Biosensors 2022;12(1). http://dx.doi.org/10.3390/bios12010022, URL https://www.mdpi.com/2079-6374/12/1/22.

[51] Song Y, Wang D, Yue K, Zheng N, Shen Z-JM. EEG-based motor imagery classification with deep multi-task learning. In: 2019 international joint conference on neural networks (IJCNN). 2019, p. 1–8. http://dx.doi.org/10.1109/IJCNN.2019.8852362.

[52] Liu D, Zhang J, Wu H, Liu S, Long J. Multi-source transfer learning for EEG classification based on domain adversarial neural network. IEEE Trans Neural Syst Rehabil Eng 2023;31:218–28. http://dx.doi.org/10.1109/TNSRE.2022.3219418.

[53] Li H, Zhang D, Xie J. MI-DABAN: A dual-attention-based adversarial network for motor imagery classification. Comput Biol Med 2023;152:106420. http://dx.doi.org/10.1016/j.compbiomed.2022.106420.

[54] Dose H, Møller JS, Iversen HK, Puthusserypady S. An end-to-end deep learning approach to MI-EEG signal classification for BCIs. Expert Syst Appl 2018;114:532–42. http://dx.doi.org/10.1016/j.eswa.2018.08.031.

[55] Amin SU, Alsulaiman M, Muhammad G, Mekhtiche MA, Shamim Hossain M. Deep learning for EEG motor imagery classification based on multi-layer CNNs feature fusion. Future Gener Comput Syst 2019;101:542–54. http://dx.doi.org/10.1016/j.future.2019.06.027.

[56] Yang J, Yao S, Wang J. Deep fusion feature learning network for MI-EEG classification. IEEE Access 2018;6:79050–9. http://dx.doi.org/10.1109/ACCESS.2018.2877452.

[57] Hou Y, Jia S, Lun X, Zhang S, Chen T, Wang F, Lv J. Deep feature mining via the attention-based bidirectional long short term memory graph convolutional neural network for human motor imagery recognition. Front Bioeng Biotechnol 2022;9. http://dx.doi.org/10.3389/fbioe.2021.706229, URL https://www.frontiersin.org/article/10.3389/fbioe.2021.706229.

[58] Roots K, Muhammad Y, Muhammad N. Fusion convolutional neural network for cross-subject EEG motor imagery classification. Computers 2020;9(3):72. http://dx.doi.org/10.3390/computers9030072.

[59] Dai G, Zhou J, Huang J, Wang N. HS-CNN: a CNN with hybrid convolution scale for EEG motor imagery classification. J Neural Eng 2020;17(1):016025. http://dx.doi.org/10.1088/1741-2552/ab405f.

[60] Jia Z, Lin Y, Wang J, Yang K, Liu T, Zhang X. MMCNN: A multi-branch multi-scale convolutional neural network for motor imagery classification. In: Machine learning and knowledge discovery in databases, Vol. 12459. Cham: Springer International Publishing; 2021, p. 736–51.

[61] Wu H, Niu Y, Li F, Li Y, Fu B, Shi G, Dong M. A parallel multiscale filter bank convolutional neural networks for motor imagery EEG classification. Front Neurosci 2019;13:1275. http://dx.doi.org/10.3389/fnins.2019.01275.

[62] Amin SU, Alsulaiman M, Muhammad G, Bencherif MA, Hossain MS. Multilevel weighted feature fusion using convolutional neural networks for EEG motor imagery classification. IEEE Access 2019;7:18940–50. http://dx.doi.org/10.1109/ACCESS.2019.2895688.

[63] Yang H, Sakhavi S, Ang KK, Guan C. On the use of convolutional neural networks and augmented CSP features for multi-class motor imagery of EEG signals classification. In: 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC). Milan: IEEE; 2015, p. 2620–3. http://dx.doi.org/10.1109/EMBC.2015.7318929.

[64] Wei X, Ortega P, Faisal AA. Inter-subject deep transfer learning for motor imagery EEG decoding. 2021, arXiv:2103.05351.

[65] Liao JJ, Luo JJ, Yang T, So RQY, Chua MCH. Effects of local and global spatial patterns in EEG motor-imagery classification using convolutional neural network. Brain-Comput Interfaces 2020;7(3–4):47–56. http://dx.doi.org/10.1080/2326263X.2020.1801112, arXiv:https://doi.org/10.1080/2326263X.2020.1801112.

[66] Zhao X, Zhang H, Zhu G, You F, Kuang S, Sun L. A multi-branch 3D convolutional neural network for EEG-based motor imagery classification. IEEE Trans Neural Syst Rehabil Eng 2019;27(10):2164–77. http://dx.doi.org/10.1109/TNSRE.2019.2938295.

[67] Mammone N, Ieracitano C, Morabito FC. A deep CNN approach to decode motor preparation of upper limbs from time–frequency maps of EEG signals at source level. Neural Netw 2020;124:357–72.

[68] Hou Y, Zhou L, Jia S, Lun X. A novel approach of decoding EEG four-class motor imagery tasks via scout ESI and CNN. J Neural Eng 2020;17(1):016048. http://dx.doi.org/10.1088/1741-2552/ab4af6.

[69] Lee HK, Choi Y-S. Application of continuous wavelet transform and convolutional neural network in decoding motor imagery brain-computer interface. Entropy 2019;21(12):1199. http://dx.doi.org/10.3390/e21121199.

[70] Li F, He F, Wang F, Zhang D, Xia Y, Li X. A novel simplified convolutional neural network classification algorithm of motor imagery EEG signals based on deep learning. Appl Sci 2020;10(5):1605.

[71] Ortiz-Echeverri CJ, Salazar-Colores S, Rodríguez-Reséndiz J, Gómez-Loenzo RA. A new approach for motor imagery classification based on sorted blind source separation, continuous wavelet transform, and convolutional neural network. Sensors 2019;19(20):4541. http://dx.doi.org/10.3390/s19204541.

[72] Tabar YR, Halici U. A novel deep learning approach for classification of EEG motor imagery signals. J Neural Eng 2016;14(1):016003. http://dx.doi.org/10.1088/1741-2560/14/1/016003.

[73] Xu G, Shen X, Chen S, Zong Y, Zhang C, Yue H, Liu M, Chen F, Che W. A deep transfer convolutional neural network framework for EEG signal classification. IEEE Access 2019;7:112767–76. http://dx.doi.org/10.1109/ACCESS.2019.2930958.

[74] Roy S, McCreadie K, Prasad G. Can a single model deep learning approach enhance classification accuracy of an EEG-based brain-computer interface? In: 2019 IEEE international conference on systems, man and cybernetics (SMC). IEEE; 2019, p. 1317–21. http://dx.doi.org/10.1109/SMC.2019.8914623.

[75] Alazrai R, Abuhijleh M, Alwanni H, Daoud MI. A deep learning framework for decoding motor imagery tasks of the same hand using EEG signals. IEEE Access 2019;7:109612–27. http://dx.doi.org/10.1109/ACCESS.2019.2934018.

[76] Ma X, Qiu S, Wei W, Wang S, He H. Deep channel-correlation network for motor imagery decoding from the same limb. IEEE Trans Neural Syst Rehabil Eng 2020;28(1):297–306. http://dx.doi.org/10.1109/TNSRE.2019.2953121.

[77] Li D, Xu J, Wang J, Fang X, Ji Y. A multi-scale fusion convolutional neural network based on attention mechanism for the visualization analysis of EEG signals decoding. IEEE Trans Neural Syst Rehabil Eng 2020;28(12):2615–26. http://dx.doi.org/10.1109/TNSRE.2020.3037326.

[78] Chen J, Yu Z, Gu Z, Li Y. Deep temporal-spatial feature learning for motor imagery-based brain–computer interfaces. IEEE Trans Neural Syst Rehabil Eng 2020;28(11):2356–66. http://dx.doi.org/10.1109/TNSRE.2020.3023417.

[79] Li D, Wang J, Xu J, Fang X. Densely feature fusion based on convolutional neural networks for motor imagery EEG classification. IEEE Access 2019;7:132720–30. http://dx.doi.org/10.1109/ACCESS.2019.2941867.

[80] Liu C, Jin J, Xu R, Li S, Zuo C, Sun H, Wang X, Cichocki A. Distinguishable spatial-spectral feature learning neural network framework for motor imagery-based brain–computer interface. J Neural Eng 2021;18(4):0460e4. http://dx.doi.org/10.1088/1741-2552/ac1d36.

[81] Zhang R, Zong Q, Dou L, Zhao X. A novel hybrid deep learning scheme for four-class motor imagery classification. J Neural Eng 2019;16(6):066004. http://dx.doi.org/10.1088/1741-2552/ab3471.

[82] Kwon O-Y, Lee M-H, Guan C, Lee S-W. Subject-independent brain–computer interfaces based on deep convolutional neural networks. IEEE Trans Neural Netw Learn Syst 2020;31(10):3839–52. http://dx.doi.org/10.1109/TNNLS.2019.2946869.

[83] Sakhavi S, Guan C. Convolutional neural network-based transfer learning and knowledge distillation using multi-subject data in motor imagery BCI. In: 2017 8th international IEEE/EMBS conference on neural engineering (NER). Shanghai, China: IEEE; 2017, p. 588–91. http://dx.doi.org/10.1109/NER.2017.8008420.

[84] Milanés Hermosilla D, Trujillo Codorniú R, López Baracaldo R, Sagaró Zamora R, Delisle Rodriguez D, Llosas Albuerne Y, Álvarez JRN. Shallow convolutional network excel for classifying motor imagery EEG in BCI applications. IEEE Access 2021;9:98275–86. http://dx.doi.org/10.1109/ACCESS.2021.3091399.

[85] Huong NTM, Linh HQ, Khai LQ. Classification of left/right hand movement EEG signals using event related potentials and advanced features. In: Vo Van T, Nguyen Le TA, Nguyen Duc T, editors. 6th international conference on the development of biomedical engineering in Vietnam (BME6). Singapore: Springer Singapore; 2018, p. 209–15.

[86] Sakhavi S, Guan C, Yan S. Learning temporal information for brain-computer interface using convolutional neural networks. IEEE Trans Neural Netw Learn Syst 2018;29(11):5619–29. http://dx.doi.org/10.1109/TNNLS.2018.2789927.

[87] Robinson N, Lee S-W, Guan C. EEG representation in deep convolutional neural networks for classification of motor imagery. In: 2019 IEEE international conference on systems, man and cybernetics (SMC). Bari: IEEE; 2019, p. 1322–6. http://dx.doi.org/10.1109/SMC.2019.8914184.

[88] Chu Y, Zhao X, Zou Y, Xu W, Han J, Zhao Y. A decoding scheme for incomplete motor imagery EEG with deep belief network. Front Neurosci 2018;12. http://dx.doi.org/10.3389/fnins.2018.00680, URL https://www.frontiersin.org/articles/10.3389/fnins.2018.00680.

[89] Hassanpour A, Moradikia M, Adeli H, Khayami SR, Shamsinejadbabaki P. A novel end-to-end deep learning scheme for classifying multi-class motor imagery electroencephalography signals. Expert Syst 2019;36(6).

[90] Tang Z, Li C, Sun S. Single-trial EEG classification of motor imagery using deep convolutional neural networks. Optik 2017;130:11–8.

[91] Zancanaro A, Cisotto G, Paulo JR, Pires G, Nunes UJ. CNN-based approaches for cross-subject classification in motor imagery: From the state-of-the-art to DynamicNet. In: 2021 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB). 2021, p. 1–7. http://dx.doi.org/10.1109/CIBCB49929.2021.9562821.

[92] Chen W, Wang S, Zhang X, Yao L, Yue L, Qian B, Li X. EEG-based motion intention recognition via multi-task RNNs. In: Proceedings of the 2018 SIAM international conference on data mining (SDM). San Diego: SIAM; 2018, p. 279–87. http://dx.doi.org/10.1137/1.9781611975321.32.

[93] Ramoser H, Muller-Gerking J, Pfurtscheller G. Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Trans Rehabil Eng 2000;8(4):441–6. http://dx.doi.org/10.1109/86.895946.

[94] Apicella A, Isgrò F, Pollastro A, Prevete R. On the effects of data normalization for domain adaptation on EEG data. Eng Appl Artif Intell 2023;123:106205. http://dx.doi.org/10.1016/j.engappai.2023.106205, URL https://www.sciencedirect.com/science/article/pii/S0952197623003895.

[95] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE; 2016, p. 770–8.

[96] Otter DW, Medina JR, Kalita JK. A survey of the usages of deep learning for natural language processing. IEEE Trans Neural Netw Learn Syst 2021;32(2):604–24. http://dx.doi.org/10.1109/TNNLS.2020.2979670.

[97] Deng L, Hinton G, Kingsbury B. New types of deep neural network learning for speech recognition and related applications: an overview. In: 2013 IEEE international conference on acoustics, speech and signal processing. Vancouver: IEEE; 2013, p. 8599–603. http://dx.doi.org/10.1109/ICASSP.2013.6639344.

[98] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: IEEE conference on computer vision and pattern recognition (CVPR). Salt lake city: IEEE; 2018, p. 7132–41.

[99] Kingma DP, Welling M. Auto-encoding variational Bayes. 2013, http://dx.doi.org/10.48550/ARXIV.1312.6114, URL https://arxiv.org/abs/1312.6114.

[100] Thireou T, Reczko M. Bidirectional long short-term memory networks for predicting the subcellular localization of eukaryotic proteins. IEEE/ACM Trans Comput Biol Bioinform 2007;4(3):441–6. http://dx.doi.org/10.1109/tcbb.2007.1015.

[101] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. 2016, http://dx.doi.org/10.48550/ARXIV.1609.02907, URL https://arxiv.org/abs/1609.02907.

[102] Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw 2015;61:85–117. http://dx.doi.org/10.1016/j.neunet.2014.09.003, URL https://www.sciencedirect.com/science/article/pii/S0893608014002135.

[103] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Bach F, Blei D, editors. Proceedings of the 32nd international conference on machine learning. Proceedings of machine learning research, vol. 37, Lille, France: PMLR; 2015, p. 448–56, URL https://proceedings.mlr.press/v37/ioffe15.html.

[104] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Teh YW, Titterington M, editors. Proceedings of the thirteenth international conference on artificial intelligence and statistics. Proceedings of machine learning research, vol. 9, Chia Laguna Resort, Sardinia, Italy: PMLR; 2010, p. 249–56, URL https://proceedings.mlr.press/v9/glorot10a.html.

[105] Narkhede MV, Bartakke PP, Sutaone MS. A review on weight initialization strategies for neural networks. Artif Intell Rev 2021;55:291–322, URL https://api.semanticscholar.org/CorpusID:237793845.

[106] Baig MZ, Aslam N, Shum HPH. Filtering techniques for channel selection in motor imagery EEG applications: A survey. Artif Intell Rev 2020;53(2):1207–32. http://dx.doi.org/10.1007/s10462-019-09694-8.

[107] Blankertz B, Muller K-R, Curio G, Vaughan T, Schalk G, Wolpaw J, Schlogl A, Neuper C, Pfurtscheller G, Hinterberger T, Schroder M, Birbaumer N. The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials. IEEE Trans Biomed Eng 2004;51(6):1044–51. http://dx.doi.org/10.1109/TBME.2004.826692.

[108] Blankertz B, Muller K-R, Krusienski D, Schalk G, Wolpaw J, Schlogl A, Pfurtscheller G, Millan J, Schroder M, Birbaumer N. The BCI competition III: validating alternative approaches to actual BCI problems. IEEE Trans Neural Syst Rehabil Eng 2006;14(2):153–9. http://dx.doi.org/10.1109/TNSRE.2006.875642.

[109] Leeb R, Brunner C, Müller-Putz G, Schlögl A, Pfurtscheller G. BCI competition 2008-graz data set A and B. 2008, URL http://www.bbci.de/competition/iv/desc_2a.pdf.

[110] Brunner C, Leeb R, Müller-Putz G, Schlögl A, Pfurtscheller G. BCI competition 2008-graz data set A and B. 2008, URL http://www.bbci.de/competition/iv/desc_2b.pdf.

[111] Schalk G, McFarland D, Hinterberger T, Birbaumer N, Wolpaw J. BCI2000: a general-purpose brain-computer interface (BCI) system. IEEE Trans Biomed Eng 2004;51(6):1034–43. http://dx.doi.org/10.1109/TBME.2004.827072.

[112] Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. 2020, arXiv:2008.05756.

[113] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manage 2009;45(4):427–37. http://dx.doi.org/10.1016/j.ipm.2009.03.002.

[114] Chaibub Neto E, Pratap A, Perumal TM, Tummalacherla M, Snyder P, Bot B, Trister A, Friend S, Mangravite L, Omberg L. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. NPJ Digit Med 2019;2. http://dx.doi.org/10.1038/s41746-019-0178-x.

[115] Comparing the runtime of a PyTorch model vs a TensorFlow model. 2023, https://saturncloud.io/blog/comparing-the-runtime-of-a-pytorch-model-vs-a-tensorflow-model/. Accessed: 2023-10-12.

[116] Ivaturi P, Gadaleta M, Pandey AC, Pazzani M, Steinhubl SR, Quer G. A comprehensive explanation framework for biomedical time series classification. IEEE J Biomed Health Inf 2021;25(7):2398–408. http://dx.doi.org/10.1109/JBHI.2021.3060997.

[117] Manjunatha H, Esfahani ET. Extracting interpretable EEG features from a deep learning model to assess the quality of human-robot co-manipulation. In: 2021 10th international IEEE/EMBS conference on neural engineering (NER). 2021, p. 339–42. http://dx.doi.org/10.1109/NER49283.2021.9441134.

[118] Zhao J, Zhao Y, Li J, Chen X. Is depth really necessary for salient object detection? In: Proceedings of the 28th ACM international conference on multimedia. New York, NY, USA: Association for Computing Machinery; 2020, p. 1745–54.

[119] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(56):1929–58, URL http://jmlr.org/papers/v15/srivastava14a.html.

[120] Zhang Y, Chen J, Tan JH, Chen Y, Chen Y, Li D, Yang L, Su J, Huang X, Che W. An investigation of deep learning models for EEG-based emotion recognition. Front Neurosci 2020;14.

[121] Yasin S, Hussain SA, Aslan S, Raza I, Muzammel M, Othmani A. EEG based major depressive disorder and bipolar disorder detection using neural networks:A review. Comput Methods Programs Biomed 2021;202:106007. http://dx.doi.org/10.1016/j.cmpb.2021.106007, URL https://www.sciencedirect.com/science/article/pii/S0169260721000821.

[122] Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. In: 2018 IEEE/CVF conference on computer vision and pattern recognition. 2018, p. 8697–710. http://dx.doi.org/10.1109/CVPR.2018.00907.

[123] Li G, Qian G, Delgadillo IC, Müller M, Thabet A, Ghanem B. SGAS: Sequential greedy architecture search. 2019, http://dx.doi.org/10.48550/ARXIV.1912.00195, URL https://arxiv.org/abs/1912.00195.

[124] Patel V, Tailor J, Ganatra A. Handling class imbalance in electroencephalography data using synthetic minority oversampling technique. In: Singh M, Tyagi V, Gupta PK, Flusser J, Ören T, Sonawane VR, editors. Advances in computing and data sciences. Cham: Springer International Publishing; 2021, p. 12–21.

[125] Wang F, Zhong S-h, Peng J, Jiang J, Liu Y. Data augmentation for EEG-based emotion recognition with deep convolutional neural networks. In: Schoeffmann K, Chalidabhongse TH, Ngo CW, Aramvith S, O'Connor NE, Ho Y-S, Gabbouj M, Elgammal A, editors. MultiMedia modeling. Cham: Springer International Publishing; 2018, p. 82–93.

[126] Li Y, Huang J, Zhou H, Zhong N. Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks. Appl Sci 2017;7(10). http://dx.doi.org/10.3390/app7101060, URL https://www.mdpi.com/2076-3417/7/10/1060.

[127] Fan J, Sun C, Chen C, Jiang X, Liu X, Zhao X, Meng L, Dai C, Chen W. EEG data augmentation: towards class imbalance problem in sleep staging tasks. J Neural Eng 2020;17(5):056017. http://dx.doi.org/10.1088/1741-2552/abb5be.

[128] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22(10):1345–59. http://dx.doi.org/10.1109/TKDE.2009.191.

[129] Zhang K, Robinson N, Lee S-W, Guan C. Adaptive transfer learning for EEG motor imagery classification with deep convolutional neural network. Neural Netw 2021;136:1–10. http://dx.doi.org/10.1016/j.neunet.2020.12.013, URL https://www.sciencedirect.com/science/article/pii/S0893608020304305.

[130] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations (ICLR). 2015, p. 1–14.

[131] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). 2016, p. 770–8. http://dx.doi.org/10.1109/CVPR.2016.90.

[132] Zhang W, Deng L, Zhang L, Wu D. A survey on negative transfer. 2020, http://dx.doi.org/10.48550/ARXIV.2009.00909, URL https://arxiv.org/abs/2009.00909.

[133] Chamola V, Vineet A, Nayyar A, Hossain E. Brain-computer interface-based humanoid control: A review. Sensors 2020;20(13). http://dx.doi.org/10.3390/s20133620, URL https://www.mdpi.com/1424-8220/20/13/3620.

[134] Liang T, Glossner J, Wang L, Shi S, Zhang X. Pruning and quantization for deep neural network acceleration: A survey. Neurocomputing 2021;461(C):370–403. http://dx.doi.org/10.1016/j.neucom.2021.07.045.

[135] Cheng M-M, Gao S, Borji A, Tan Y-Q, Lin Z, Wang M. A highly efficient model to study the semantics of salient object detection. IEEE Trans Pattern Anal Mach Intell 2021;1.

[136] Li H, Kadav A, Durdanovic I, Samet H, Graf HP. Pruning filters for efficient ConvNets. 2016, http://dx.doi.org/10.48550/ARXIV.1608.08710, URL https://arxiv.org/abs/1608.08710.