

Toward Open-World Electroencephalogram Decoding Via Deep Learning: A Comprehensive Survey

Xun Chen, Chang Li, Aiping Liu, Martin J. McKeown, Ruobing Qian, Z. Jane Wang

Electroencephalogram (EEG) decoding aims to identify the perceptual, semantic, and cognitive content of neural processing based on non-invasively measured brain activity. Traditional EEG decoding methods have achieved moderate success when applied to data acquired in static, well-controlled lab environments. However, an open-world environment is a more realistic setting, where situations affecting EEG recordings can emerge unexpectedly, significantly weakening the robustness of existing methods. In recent years, deep learning (DL) has emerged as a potential solution for such problems due to its superior capacity in feature extraction. It overcomes the limitations of defining ‘handcrafted’ features or features extracted using shallow architectures, but typically requires large amounts of costly, expertly-labelled data – something not always obtainable. Combining DL with domain-specific knowledge may allow for development of robust approaches to decode brain activity even with small-sample data. Although various DL methods have been proposed to tackle some of the challenges in EEG decoding, a systematic tutorial overview, particularly for open-world applications, is currently lacking. This article therefore provides a comprehensive survey of DL methods for open-world EEG decoding, and identifies promising research directions to inspire future studies for EEG decoding in real-world applications.

I. INTRODUCTION

Identifying and predicting mental processes from observed patterns of neural activities have long been explored in cognitive neuroscience and brain computer interfaces [1]. Non-invasive techniques, such as electroencephalography (EEG), magnetoencephalography (MEG), near infrared spectroscopy (NIRS), and functional magnetic resonance imaging (fMRI) provide accessible ways to broadly examine brain activity without surgical intervention. In particular, the EEG is popular for brain decoding in practical applications, as it is relatively inexpensive and has advantages of safety, high temporal resolution, wide accessibility, and potential portability [2], [3]. EEG signals can be obtained by placing electrodes on the surface of the scalp, providing measurements of post-synaptic potentials – an indirect measure of neuronal activity [4]. This allows for, after digitization and suitable analyses, decoding of brain states and communication between our brain and the outside world [5].

EEG decoding methods have made great progress in recent decades. However, most existing EEG decoding methods were

designed using data collected in static or well-controlled lab environments that utilized rigorous experimental protocols and strict laboratory conditions, which are unrealistic in an open-world environment. Open-world EEG decoding refers to identifying the perceptual, semantic, and cognitive content of measured brain activity in an open-world environment [6]. With the emergence of new open-world applications in various fields, such as in entertainment, industry, and medicine, there is an urgent need to develop efficient EEG decoding methods for real-world scenarios. Additionally, such open-world applications typically adopt few-electrode portable, wearable, and wireless systems to take advantage of technological advances in hardware [7]. These recordings from complex environments tend to be heavily contaminated with artifact, making brain decoding even more challenging.

The steps involved in EEG decoding typically include preprocessing, feature extraction, and classification, and successful open-world EEG decoding requires specific considerations at each step. Even under the most stringent recording conditions, EEG signals are easily corrupted by various artifacts (*e.g.*, eye blinks, muscle artifacts, cardiac interference, and electromagnetic interference) [7]. EEGs in open-world environments are also contaminated by outdoor open-world artifacts caused by extensive movement (such as muscle and mechanical artifacts) and electromagnetic factors [7]. Historically EEG features have been extracted from time-domain (such as mean, variance, and kurtosis), frequency-domain (such as power spectral density and fast Fourier transform), and time-frequency domains (such as discrete wavelet transform). In addition, traditionally-defined features strongly depend on human expertise in a specific domain, and manual feature extraction is time-consuming. Classification of the extracted features include such techniques such as decision tree (DT), support vector machine (SVM), and linear discriminant analysis (LDA). EEG signals are temporally non-stationary, with their statistics varying over time which can make generalization of a classifier challenging based on limited amounts of data.

During feature extraction, most EEG decoding methods assume that the training and test data are identically distributed. However, high inter-subject variability, electrode shifts, and physiological state changes will inevitably lead to mismatch between training and test distributions in an open-world situation [8]. Even a small mismatch in distributions can cause significant performance degradation.

To overcome the above challenges, deep learning (DL)

TABLE I: Challenges and solutions for three typical steps of open-world EEG decoding via DL.

	Challenges	Solutions
Preprocessing	How to remove various EEG artifacts automatically with good generalization ability	CNN
		Autoencoder
		LSTM
		GAN
Feature extraction	How to solve the distribution mismatch between the source and the target EEG data	Transfer learning
	How to exploit complementary information of EEG from multiple tasks and modalities	Multi-task learning
		Multi-modal learning
	How to make the EEG models robust for adversarial attacks	Adversarial training
Classification	How to design EEG models with desirable performance under small sample size	Few-shot learning
		Semi-supervised learning
	How to recognize both the known and the unknown categories of EEG	Zero-shot learning
	How to exploit the structure of unlabeled EEG data to provide supervision	Self-supervised learning
	How to train robust EEG models in the presence of noisy label	Noisy label classification

methods, which are an automatic end-to-end learning framework, consisting of preprocessing, feature extraction, and classification, have achieved state-of-the-art performance in the field of EEG decoding [9], with better generalization abilities and more flexible applicability. DL approaches avoid time-consuming preprocessing and feature extraction by working on raw EEG signals directly to learn useful information, which can capture both discriminative high-level features and underlying dependencies.

Despite their successes, DL approaches have their own challenges. Supervised DL implicitly assumes that there exists a large number of labelled EEG training samples for DL to achieve good generalization performance [9]. However, EEG classification performance deteriorates as the number of available training samples diminishes, emphasizing the need for robust DL models under more typical small sample size scenarios [10]. Transfer learning [8], multi-task learning [11] and multi-modal learning [12] have been recently proposed to address small-sample data concerns. DL models have usually assumed that the categories of the test EEG signals have been seen during training. However, there may be test EEG signals that do not belong to any category of the training set. Zero-shot learning [13] may be a potential solution to recognize both known and unknown categories of EEG signals. Another challenge with DL approaches is that the labels are likely not perfectly assigned in the training set [14], resulting in overfitting of the incorrect labels and reducing ultimate classification accuracy when applied to unseen data. Unlabeled EEG data can be analyzed via self-supervised [15] and semi-supervised learning methods [16] to obtain important latent information, which we will expand on later. Finally, DL models can be easily fooled with adversarial examples, which are modified normal examples with small deliberate perturbations [17], so care must be given to prevent this possibility.

Although there are several review articles on EEG decoding, to the best of our knowledge, most of them focus on EEG decoding in static or well-controlled lab environments. In this article, we review open-world EEG decoding via DL, which has the capacity to develop robust EEG decoding for real

open-world applications. Our objectives are as follows: (1) present the problems and challenges faced by open-world EEG decoding, (2) provide a taxonomy of DL solutions for open-world EEG decoding, and (3) discuss potential ways to obtain more robust DL models for open-world EEG decoding.

II. OPEN-WORLD EEG DECODING VIA DL

In this section, we introduce new problems and methods to deal with open-world EEG decoding via DL on each step of EEG decoding, *i.e.*, preprocessing, feature extraction, and classification. The structure of open-world EEG decoding via DL is schematically illustrated in Table I.

A. Preprocessing

The performance of EEG decoding depends heavily on the quality of the EEG signals. Unfortunately, the recorded signals are usually contaminated by various artifacts, which can be magnified in open-world applications within complex environments [7]. Therefore, it is of both theoretical and practical significance to remove complex artifacts from contaminated EEG signals in the open-world.

Filtering and regression are traditional artifact removal methods. Filtering methods assume that artifacts and EEG signals reside in distinct frequency bands [18]. However, artifacts and EEG signals usually overlap in the frequency domain, and there is a risk that portions of the EEG signals may be eliminated during this artifact removal process. Regression methods usually assume that each EEG channel can be modeled as a linear or nonlinear superposition of clean brain activity and artifact signals that can be obtained from reference channels or artifact templates. However, regression methods only work when suitable reference channels that are available (*e.g.*, channels for measuring eye movement).

Another popular approach for EEG denoising during the preprocessing stage is blind source separation (BSS) [19], which assumes that the clean EEG and artifacts are statistically independent in the time domain, so they will be isolated into different components. This allows for removal of artifact-related components during the reconstruction process. BSS

methods typically require human intervention to identify the artifact-related components, which is subjective and time-consuming. These methods generally require that the number of channels must be larger than or equal to the number of underlying sources [18], so they are less attractive when only a few channels are available, as is typical in mobile scenarios.

Empirical mode decomposition (EMD) [20] and the wavelet transform (WT) [21] are two representative methods for denoising when only a limited number of EEG channels are available. EMD decomposes an input signal into multiple empirical modes according to the intrinsic mode function (IMF). IMFs are a set of the band-limited functions that satisfy two basic conditions: (1) the number of extreme points and the number of zero crossings must either equal or differ at most by one, and (2) at every point, the mean value of the envelopes defined by local maxima and local minima should be zero. EMD's data-driven approach capable of dealing with non-stationary stochastic processes makes it suitable for removing artifacts from contaminated EEG signals. However, EMD is time-consuming and is not suitable for online applications in the open-world. Similar to EMD, the WT first decomposes the contaminated EEG signal into different sub-bands. Then, a threshold function is used to update the coefficients related to the sub-bands that are assumed to be artifact-related. Finally, the EEG signals are reconstructed using the updated coefficients. However, selection of an incorrect threshold setting could lead to the degradation of the reconstructed EEG signals [7]. A traditional neural network with shallow layers can be used to replace the threshold function in the wavelet analysis, which has the advantage of approximating smooth nonlinear functions. Nevertheless, the approximation ability of a traditional, shallow layer neural network tends to be inferior to that of DL.

DL-based methods can be used to automatically filter out artifacts from contaminated EEG signals. One typical method is to learn a mapping between noisy EEG signals and their cleaned versions [22]. The performance of DL-based artifact removal methods relies fundamentally on the size of the training datasets. For example, Zhang *et al.* established a benchmark EEG dataset for the training and testing of DL-based artifact removal methods [22]. The EEG epochs were acquired from a motor imaginary EEG dataset, with a band-pass filter between 1 and 80 Hz applied, followed by re-sampling to 256 Hz. Then, the ICLabel method [23] was used to attenuate the artifacts. Finally, the EEG signals were segmented into epochs of 2-s. The ocular artifact epochs were acquired from open-access EEG data, band-pass filtered between 0.3 and 10 Hz, followed by re-sampling to 256 Hz, then segmented into 2-s epochs. The myogenic artifact epochs were acquired from a facial surface electromyography (EMG) dataset, band-pass filtered between 1 and 120 Hz, re-sampling to 512 Hz, then segmented into 2-s epochs. For all the categories, the epochs were standardized by subtracting their mean and dividing by their standard deviation, and then were visually checked by an expert. Finally, 4514 clean EEG epochs, 3400 ocular artifact epochs, and 5598 muscular artifact epochs were acquired. Simulated noisy signals can be generated by linearly mixing the clean EEG epochs with EOG

or EMG epochs, the SNRs for EEG epochs contaminated by ocular artifacts range from -7dB to 2dB, and the SNRs for those contaminated by myogenic artifacts range from -7dB to 4dB. In this way, the clean EEG epochs can be considered as ground truth, and the mixed epochs as contaminated EEG. This allowed the adoption of a large number of noisy EEG epochs with ground truth (clean EEG epochs) for model training and testing. Mathematically, simulated noisy EEG signals can be formulated as

$$\mathbf{Y} = \mathbf{X} + \lambda \cdot \mathbf{N}, \quad (1)$$

where \mathbf{Y} denotes the contaminated EEG signal with artifacts, \mathbf{X} denotes the clean EEG signal, \mathbf{N} denotes the artifacts, and λ denotes the relative contribution of the artifacts.

The goal of DL-based artifact removal is to learn an end-to-end nonlinear function f to map a noisy EEG signal to approximate a clean EEG signal as follows

$$\hat{\mathbf{X}} = f(\mathbf{Y}, \theta), \quad (2)$$

where $\hat{\mathbf{X}}$ denotes the approximated clean EEG signal, and θ denotes the parameters to be learned. The learning process can be realized by minimizing the objective function as follows

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{N} \sum_{i=1}^N \|\mathbf{X}_i - f(\mathbf{Y}_i, \theta)\|_F^2, \quad (3)$$

where N represents the number of training samples, and \mathbf{Y}_i and \mathbf{X}_i represent the i^{th} contaminated EEG signal and clean EEG signal, respectively. After obtaining the optimized trained parameters $\hat{\theta}$, denoised EEG signals can be obtained for the EEG test dataset.

Autoencoder, one of the major branches of DL, has also been used in artifact removal [24]. Autoencoder consists of three layers, *i.e.*, input, hidden and output layers. An autoencoder maps an input x to an output z . It takes an input x and maps it to a hidden representation y through a mapping (encoder) as follows

$$y = s(W^T x + b), \quad (4)$$

where s denotes the non-linear function such as sigmoid, W and b denote the weight and bias vector from input to hidden layer, respectively. The hidden representation is then mapped back (decoder) again to the output z of the same size as input x as follows

$$z = s(W'^T y + b'), \quad (5)$$

where W' and b' denote the weight and bias vectors from hidden layer to the output layer, respectively. The training of the network can be accomplished by measuring the reconstruction error, which can be measured with the traditional mean squared error (MSE). For example, Ghosh *et al.* proposed an automated eye blink artefact removal from EEG using SVM and autoencoder [24]. A sliding window of 0.45-s is applied to the EEG data, and each window is processed as follows: (1) Identification of artifacts: The signal within the window is input to the SVM classifier, SVM classifies whether the signal is an artifact or clean EEG. If it classifies the signal as an artifact, it is input to the autoencoder for correction. On the other hand, if the signal is classified as non-artifact,

then the window is slid forward. (2) Cleaning the artifacts: Signal window marked as artifact is then input to the pre-trained autoencoder. The autoencoder removes the artifact of contaminated EEG window and outputs the clean EEG signal. However, the encoder and decoder in [24] are the simplest case of the autoencoder, which use only one linear layer and nonlinear layer to represent both the encoder and decoder. The encoder and decoder for EEG preprocessing can have multiple layers. For example, a reversible GAN was used for the removal of BCG artifacts [25]. An autoencoder is used as a subnetwork in the single shot reversible GAN. The encoder contains multiple convolutional layers, and the decoder contains multiple transposed convolutional layers (deconvolutional layers).

Long short-term memory (LSTM) models, a popular variant of recurrent neural network (RNN), which solve the vanishing and exploding gradient problem of RNN by adding extra parameters to the RNN model [26], can also be used for EEG artifact reduction. An LSTM is composed of four different gates, which checks the input of the cell state and determines the influence of the cell state on the output. The four gates are termed as input gate, forget gate, output gate and block input gate. The input gate and the block input gate control the new information flow to the memory cell. Sigmoid and Tanh are used as activation functions for the input gate and block input gate, respectively. The forget gate controls which previous information should be retained into memory cell. The output gate determines what is to send as output from LSTM unit. A sigmoid activation function is used for both the forget gate and the output gate. Manjunath *et al.* proposed a low complexity LSTM for detecting various artifacts in multi-channel brain EEG signals [27]. It consists of one average pooling layer of size 64, one LSTM layer having 12 units, one dense layer having 50 neurons and one output layer for binary classification. The experimental results indicate that the LSTM based FPGA hardware outperforms the CNN based FPGA hardware by $1.88\times$ in terms of dynamic power consumption per classification.

Although the deep network-based methods described above have achieved desirable performance, an end-to-end network typically uses the MSE between the network output and ground truth as the loss function, leading to over-smoothing and loss of detail [28]. To overcome these problems, a generative adversarial network (GAN) can be used to remove artifacts, where a generative network is trained to map a noisy EEG signal to a clean EEG signal, and a discriminator network is trained to discriminate between real and generated EEG signals. An example where this is useful is in simultaneous EEG and functional magnetic resonance imaging (fMRI) recordings which can measure brain activity with both high temporal and spatial resolution. A ballistocardiogram (BCG) artifact exists due to cardiac activity and blood flow inside the static magnetic field of the MRI scanner [25]. However, BCG artifact removal remains challenging using DL, since it is difficult to obtain clean EEG signals and BCG-contaminated EEG signals at the same time. A paired signal-to-signal problem refers to the mapping between input data and output data using paired training data. However, it is hard to

obtain BCG-contaminated EEG signals and clean EEG signals in the same state simultaneously, so BCG artifact removal can be considered an ‘unpaired signal-to-signal problem’. A cycle-consistent generative adversarial network (CycleGAN) [29] is a technique to solve the unpaired image-to-image translation, and it can be widely used in many applications, such as unpaired image denoising, unpaired image super-resolution, and unpaired image dehazing. Since the SNR of the EEG signal is low, the direct use of CycleGAN, which performs well on the unpaired problem, still cannot easily remove the BCG artifacts in simultaneous EEG-fMRI. Lin *et al.* proposed a novel single-shot reversible GAN for the removal of BCG artifacts [25]. Being capable of bidirectional input and output, the forward model can map contaminated EEG signals to clean EEG signals, and the reverse model can achieve data conversion from clean EEG signals to contaminated EEG signals.

B. Feature extraction

1) *Transfer learning for addressing the distribution mismatch issues:* Recently, EEG recognition methods have proven successful in many applications, particularly when the training and test EEG data are drawn from the same distribution. However, this is not the case in many real open-world problems, where there is notable high inter-subject variability, electrode shift, and physiological state changes, all of which affect the generalization ability of models [8]. The performance of a classifier trained in the source domain will almost certainly drop when tested on the target domain due to the distribution mismatch between the source and target domains, limiting its practical use.

Transfer learning aims to improve the performance of learners in the target domain by fusing knowledge from one or more related, but differently-distributed source domains. For example, it may be desired to augment small-sample EEG from one institution with large data sets collected from other institutions. Domain adaptation is a special case of transfer learning that uses labeled data in one or more source domains to improve the learning performance in a target domain. Traditional domain adaptation methods in EEG signal analysis include distribution and subspace adaptation [8]. Distribution adaptation can be generally classified into marginal and conditional distribution adaptation. The objective of marginal distribution adaptation is to transfer knowledge when the marginal distributions of the source domain (X_S) and target domain (X_T) are different, *i.e.*, $P(X_S) \neq P(X_T)$. The objective of conditional distribution adaptation is to transfer knowledge when the conditional distributions of the source and target domains are different, *i.e.*, $P(Y_S|X_S) \neq P(Y_T|X_T)$. Measures of marginal distribution difference and conditional distribution difference include maximum mean discrepancy [30], Kullback–Leibler divergence [31], and Jensen–Shannon divergence [32]. Subspace adaptation transforms data in both the source and target domains into a common latent subspace in which their distributions are similar [33]. Linear subspace adaptation usually utilizes linear subspace learning algorithms for domain adaptation in EEG signal analysis, such as principal component analysis and linear discriminant analysis.

Manifold learning methods map the original high-dimensional EEG data in the source and target domains into a common low-dimensional manifold structure [8]. Statistical feature alignment aims to map the EEG data into a subspace to align the statistical features in the source and target domains [34], such as variance and median absolute deviation. However, traditional domain adaptation methods align the distribution in the source and target domains or learn shared common subspaces with shallow representations.

Recently, several studies have demonstrated that deep networks can learn more transferable representations [35]. The deep features eventually transition from general to specific, and the transferability sharply decreases in higher layers that are near to the output [36]. The most commonly-used method in EEG signal analysis is to fine-tune a pre-trained DL model [37], *i.e.*, train a base network and then copy its first n layers to those of a target network. A pre-trained model can leverage the knowledge gained from a large dataset to solve a different but similar task with a small dataset more effectively. The remaining layers of the target network are randomly initialized and trained towards the target task. One can either fine-tune the entire deep network, or freeze the first n layers. This depends on the size of the target dataset and the number of parameters in the first n layers. If the size of the target dataset is large or the number of parameters is small, it can be fine-tuned to the target domain to improve performance. If the size of the target dataset is small and the number of parameters is large, overfitting will likely occur, and hence the first n layers are often frozen. For example, Raghu *et al.* extracted features using pretrained network and used SVM for classifying the seizure type [37], and it outperformed conventional feature and clustering based approaches.

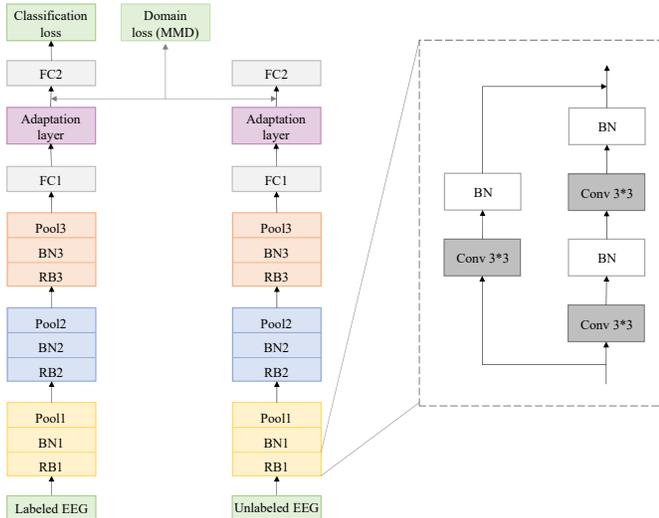


Fig. 1: Cross-subject recognition method based on CNN and DDC (Figure adapted from [38]).

Although fine-tuning is easy to implement and understand, it is less effective when there is substantial mismatch between the distributions of source and target domains. To solve this problem, Zhang *et al.* proposed a cross-subject recognition method based on a convolutional neural network (CNN) and

deep domain confusion (DDC) [38], as shown in Fig. 1. They trained the CNN using the source and target EEG data jointly to minimize the loss of classification accuracy. The DDC method can narrow the difference in feature distribution between the source and target domains by minimizing the distance between the two domains (maximizing the domain confusion). Hence, a classifier trained in the source domain can be applied to the target domain to reduce the loss of classification accuracy. The maximum mean discrepancy (MMD) for maximizing domain confusion is defined as follows

$$\text{MMD}(X_S, X_T) = \left\| \frac{1}{|X_S|} \sum_{x_s \in X_S} \Phi(x_s) - \frac{1}{|X_T|} \sum_{x_t \in X_T} \Phi(x_t) \right\|_{\mathcal{H}}^2, \quad (6)$$

where x_s and x_t denote the deep features of the source and target domains in the adaptation layer, respectively. $|X_S|$ and $|X_T|$ denote the numbers of samples in the source and target domains, respectively. Φ denotes the kernel function that maps the deep feature to a reproducing kernel Hilbert space (RKHS). The total loss is defined as follows

$$L = L_C(X_L, Y_L) + \alpha \text{MMD}(X_S, X_T), \quad (7)$$

where $L_C(X_L, Y_L)$ denotes the classification loss in the tagged source domain EEG data X_L and the corresponding ground truth labels Y_L , and α denotes the regularization parameter.

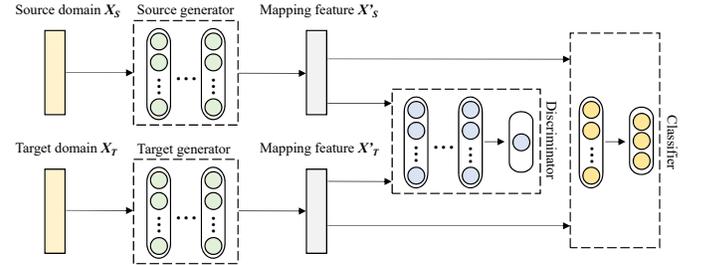


Fig. 2: Generative adversarial network domain adaptation framework (Figure adapted from [39]).

An alternative approach is to adopt generative adversarial domain adaptation [39], which is closely related to GANs. GANs have the advantage of generative ability and can be formulated as a minimax problem. The distribution of generated data is approximate to that of real data when the two-player game achieves equilibrium. Thus, as shown in Fig. 2, the generative adversarial domain adaptation [39] framework has been widely used to solve the distribution mismatch problem in computer vision as well as in EEG decoding. The source and target generators aim to map the source and target domains to a common feature space, respectively. The discriminator aims to distinguish the source and target distributions in the common feature space, and the classifier is used to recognize the EEG state.

Deep domain adaptation methods improve the model performance in the target domain by eliminating the domain shift between the source and target domains. However, most domain adaptation methods require the source domain to have the same feature space and label space as the target domain, which may not always be the case in open-world EEG-based applications.

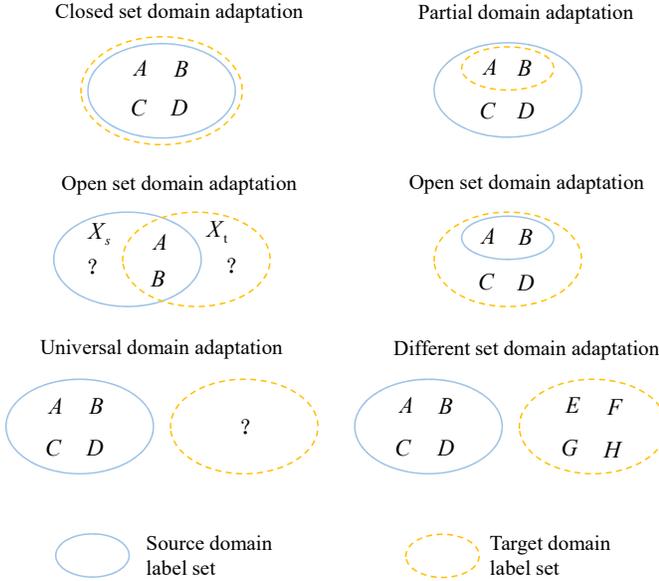


Fig. 3: Different domain adaptation scenarios [40].

In fact, the source and target domains may have different label spaces. Figure 3 shows different domain adaptation scenarios. In closed set domain adaptation, the source and target domains are assumed to have the same label space. In partial domain adaptation [41], the classes of the target domain contain only a subset of the source domain, and in open set domain adaptation, it is assumed that the source and target domains have several common classes but also several classes that are different and unknown. In open set domain adaptation considered on the left [42] of the second row in Fig. 3, the source and target domains contain some common classes, but each also contains an “unknown” class. In open set domain adaptation considered on the right [43] of the second row in Fig. 3, the source domain only contains a subset of the target domain classes. In universal domain adaptation [44], the target domain may contain several common classes with the source domain. However, it may also contain several unknown classes. In different set domain adaptation, the target domain contains partially or completely different classes from the source domain.

Domain adaptation methods require acquaintance with the target-domain data to measure the discrepancy between the source and target domains in the training stage. However, they require data collection and model training for each target domain (subject) that are high in cost and low in efficiency [45]. Domain generalization aims to learn a model using data from a single or multiply-related but different source domains so that the model can generalize well to any target domain [45]. For open-world EEG-based applications, domain generalization can extract domain-invariant features by exploiting domain differences among source subjects without access to the target subjects. Thus, domain generalization can be more robust in open-world applications when applied to unseen domains. For example, Ma *et al.* proposed a domain generalization method by applying deep adversarial networks to reduce the influence of subject variability without requiring any information from

unseen subjects, their method could generalize well to multiple test subjects compared with existing domain adaptation methods [45].

2) *Multi-task and multi-modal learning for exploring the joint and complementary information:* Some recent techniques are based on the observation that humans can learn multiple tasks simultaneously. They can use the knowledge learned in one task to help with learning of another, related task. The large numbers of annotated EEG samples typically required by DL methods for adequate recognition performance are almost impossible to obtain due to the high cost of data acquisition and accurate annotation. Multi-task learning is an approach that aims to improve the generalization performance of all tasks by leveraging useful information contained in multiple, related tasks [46]. It is assumed that all tasks, or at least a subset of them, are related to each other. Learning multiple tasks jointly has theoretically and empirically been found to achieve better performance than learning them independently [46]. Multi-task learning can allow access to more data overall, resulting in more robust and universal representations, and lower risk of overfitting for each task. Multi-task learning is related to transfer learning but has certain differences. In multi-task learning, all tasks are equal, and the aim is to improve the performance of all tasks [47]. However, transfer learning aims to improve the performance of a target task with the help of source tasks. Thus, the target task attracts more attention than the source tasks.

Multi-task DL, where each task is solved by its own deep network, can improve performance over single-task DL if the associated tasks share complementary information or act as regularizers for one another [47]. In addition, the inherent layer-sharing representation can reduce the memory footprint and avoid calculating features repeatedly in the shared layers. Thus, it can yield fast learning speed and increase data efficiency for related or downstream tasks. For example, Song *et al.* proposed an EEG classification method based on multi-task DL [11], as shown in Fig. 4. It consists of three modules, one for each of representation, classification, and reconstruction. The representation module learns shared features from EEG signals which are then sent to the classification module for prediction and then the reconstruction module to reconstruct the original EEG signal. The shared features work as a bridge to unite the classification and reconstruction tasks, and the two tasks are jointly optimized in an end-to-end manner. Through the interaction of the two tasks, the shared features maintain both classification and reconstruction abilities. Therefore, it can enhance the generalization ability of the deep model and improve classification performance with limited EEG data. Abdon *et al.* presented a multi-task cascaded deep neural network for joint prediction of people’s affective factors using EEG signals recorded from people while watching affective videos in either individual or group configuration [48]. The proposed network consists of two levels of prediction. The first level, affect network, is designed to predict the participant’s affective levels of valence and arousal expressed by the participants during single video segments. The second level, personal factors network, uses the prediction of affective levels of consecutive video segments

to perform multi-task prediction of personal factors. Liu *et al.* proposed a multiscale space-time-frequency feature-guided multi-task learning convolutional neural network architecture for EEG classification [49], which can fuse the complementary characteristics of different models. This method consists of four modules, *i.e.*, the space-time feature-based representation module, time-frequency feature-based representation module, multi-modal fused feature-guided generation module, and classification module. The four modules are trained using three tasks simultaneously and jointly optimized. Due to the interaction of the three tasks, it can improve the generalization ability and accuracy of subject-dependent and subject-independent methods with limited annotated data.

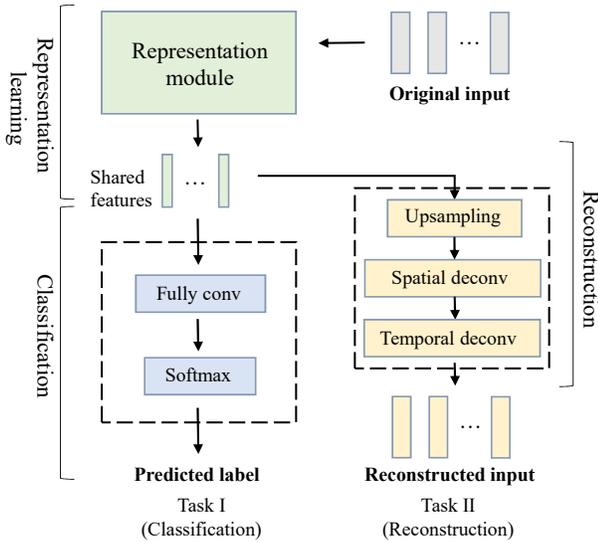


Fig. 4: EEG classification via multi-task DL (Figure adapted from [11]).

Multi-modal learning aims to utilize complementary or supplementary information from different modalities to complete a shared task or multiple related tasks [50]. The underlying motivation for using multi-modal data is that complementary or supplementary information can be extracted from different modalities for the shared task or multiple related tasks. It can obtain richer representation to achieve better performance than using only a single modality. Traditional multi-modal learning methods are shallow models that cannot learn the intrinsic representation of data. Thus, they cannot capture inter-modality representations and cross-modality complementary correlations of multi-modal data properly. However, DL models can learn a hierarchical representation of the data across hidden layers, and the learned representations of different modalities can be fused at various levels of abstraction.

Interestingly, multi-modal DL approaches appear to have some relevance for the visual pathway in the brain [51]. Palazzo *et al.* proposed a multi-modal DL method to learn a neural representation by classifying brain responses to natural images [51], and it can learn a joint brain-visual embedding and find similarities between brain representations and visual features. This embedding can be used to perform image classification, saliency detection, and visual scene analysis. The motivation is to learn reliable joint representations and

find correspondences between visual and brain features that can decode brain representations. In turn, these representations can also be used to build better DL models. Jia *et al.* proposed a multi-modal DL method for sleep stage classification by fusing EEG, EOG, and EMG signals [12]. Separate data representations were designed for each signal type. These representations were then input into a deep model to extract features from EEG, EOG, and EMG signals, respectively. Finally, a feature fusion module was used to fuse all extracted features for sleep stage classification. Cai *et al.* proposed a feature-level fusion method based on multi-modal EEG data for depression recognition [52]. The multi-modal EEG data were acquired under neutral, negative and positive audio stimulation to discriminate between depressed patients and normal controls. Then, a feature-level fusion method was used to fuse the EEG data of different modalities to construct a depression recognition model.

3) *Adversarial training for augmenting the training set with adversarial examples:* Although DL models have achieved outstanding performance in EEG feature extraction, they are vulnerable to adversarial attacks, where normal EEG samples are corrupted with small, seemingly innocuous perturbations [17]. DL models have been found to be easily fooled by adversarial examples, which are normal EEG examples with small perturbations. Figure 5 shows a normal EEG epoch and a corresponding adversarial example. The perturbations are usually too small for the human eyes to perceive. However, despite the slight changes, adversarial EEG samples can lead to a dramatic performance degradation with possible serious consequences. For example, adversarial attacks could lead to misdiagnosis of disorders of consciousness in patients in clinical applications. Adversarial perturbations can mislead the P300 and steady-state visual evoked potential brain-computer interface (BCI) spellers to spell anything the attacker wants [53]. Modern EEG monitoring systems designed to detect epileptic seizures could be vulnerable to an adversarial attack whereby an ictal (seizure) sample would be classified as inter-ictal (non-seizure) in an emergency situations, with possible dire implications [53].

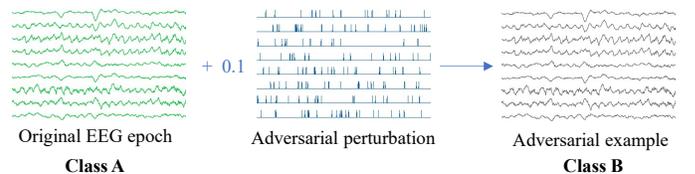


Fig. 5: A normal EEG epoch and its adversarial example.

Adversarial attacks can be classified according to the degree of access the attacker has to the target model [54], *i.e.*, white-box attacks, black-box attacks, and gray-box attacks. White-box attacks [55] assume that the attacker can obtain all information of both the classifier and the defense mechanism. Black-box attacks assume that the attacker does not know the architecture or parameters of the target model. However, they can obtain the response to the input. Gray-box attacks assume that the attacker can obtain some information of the target model. Adversarial attack can also be classified according to

the stage that the attack is performed, including poisoning attacks and evasion attacks [53]. A poisoning attack takes place during the training time of the machine learning model [56]. An adversary tries to poison the training data by injecting carefully-designed samples to eventually compromise the whole learning process. An evasion attack, the most common type, tries to evade the system by adjusting malicious samples during the testing phase [56]. This setting does not assume any influence on the training data. According to the outcome, there are two types of adversarial attacks [53], namely, targeted attacks and non-targeted (indiscriminate) attacks. Targeted attacks force a model to classify either a particular subset of data samples or a particular region of feature space to a chosen (usually wrong) class. Non-targeted attacks force a model to misclassify certain data samples or regions of feature space, but do not specify which class they should be misclassified into.

In an open-world environment, obtained EEG signals are sent to a computer, a smart phone, or the cloud for further analysis. Goodfellow *et al.* proposed the fast gradient sign method (FGSM) [57], and soon it became a benchmark attack approach. Let g be the deep learning model, θ be its parameters, and J be the loss function for training g . The main idea of FGSM is to find an optimal perturbation η constrained by ε to maximize J . The perturbation can be calculated as follows

$$\eta = \varepsilon \cdot \text{sign}(\nabla_{x_i} J(\theta, x_i, y_i)), \quad (8)$$

where x_i and y_i represent the i th EEG trial and the corresponding label, respectively. It is not enough to know the architecture and parameters θ of the target model g , since it needs to know the true label y_i of x_i to generate the adversarial perturbation. Liu *et al.* propose an unsupervised FGSM (UFGSM) to deal with this problem [17]. UFGSM replaces the label y_i by $y'_i = g(x_i)$, *i.e.*, the estimated label from the deep model. The perturbation in UFGSM can be rewritten as follows

$$\eta = \varepsilon \cdot \text{sign}(\nabla_{x_i} J(\theta, x_i, y'_i)). \quad (9)$$

All EEG trials are needed to determine an adversarial perturbation for each trial. Here, EEG trial means the EEG signal corresponding to a trial. For example, during each trial in motor imagery (MI), the subject is required to perform either of the two (right hand and right foot) MI tasks for 3.5-s. However, it is inconvenient to compute the adversarial perturbation for each EEG trial. In addition, it requires all the EEG trials in advance to compute the adversarial perturbation [54]. It is impossible to attack as soon as an EEG trial starts. To address these issues, Liu *et al.* introduced a universal adversarial perturbation [58] method that could obtain the universal perturbation template offline and hence attack open-world EEG-based systems in real time. Therefore, it is critically important to pay attention to security concerns of open-world EEG-based systems. Meng *et al.* performed poisoning attack of EEG-based BCIs [59], and they proposed a practically realizable backdoor key, which can be inserted into original EEG signals during data acquisition.

To deal with adversarial attacks in EEG-based systems, many adversarial defense methods have been proposed [53].

The most representative method is to augment DL models with adversarial training [57]. Hussein *et al.* proposed a method to augment DL models with adversarial training for robust prediction of epilepsy seizures [60]. First, a DL classifier is constructed from available limited amount of labeled EEG data, and adversarial examples are obtained by performing white-box on the classifier. Then, the training set is augmented with adversarial examples. Finally, DL models are retrained with the augmented training set, which can improve the robustness of DL models in open-world EEG-based systems.

C. Classification

1) *Few-shot, zero-shot and semi-supervised learning for small sample size*: A variety of DL methods have shown superior performance compared to traditional methods in EEG classification [9]. For example, Tabar *et al.* proposed CNN and stacked autoencoders (SAE) to classify EEG MI signals [61], which can obtain better classification performance compared with other traditional methods. However, with inadequate training samples, DL models are prone to overfitting, which leads to a decrease in classification accuracy. Although each trial can be over-sampled to obtain a larger number of samples, these samples are highly dependent on each other, so that a larger number of EEG trials are still preferable to achieve reliable performance. BCI feedback applications typically require a tedious calibration process that can be challenging in some patient populations [62]. Clearly designing robust DL models for small sample sizes is important in EEG classification [63] [64].

To learn from a limited number of EEG training examples with supervised information, Cheng *et al.* proposed a deep forest model named multi-Grained Cascade Forest (gcForest) for multi-channel EEG-based emotion recognition task [66]. This method is insensitive to hyper-parameter setting, and greatly reduces the complexity of EEG emotion recognition. The model complexity of gcForest can be determined automatically for different size of training data, making it suitable for small-scale training data. In addition, a new machine learning paradigm called few-shot learning has been proposed [63]. The goal of few-shot learning is to classify unseen data instances (query data) into a set of classes, given just a small number of labeled instances (support examples) in each class. Typically, there are between 1 and 10 labeled support examples per class in the support set.

The EEG recording during MI (used to aid rehabilitation as well as autonomous driving) is a good scenario where these issues arise and has driven the development in this area. MI also allows users to generate the suppression of oscillatory neural activity in specific frequency bands over the motor cortex region without external stimuli [67]. The neurophysiological patterns of MI originate from changing brain areas' activations in the sensorimotor cortices similar to limb movements. Furthermore, a recent study has demonstrated MI-based BCI as an assistive tool in post-stroke motor rehabilitation. However, due to the scarcity of unseen subject data, complex dynamics of MI signals, inter-subject variability, and low signal-to-noise ratio, it is still challenging to improve the performance of MI-based classification tasks.

To solve the above mentioned problems, An *et al.* formulated an EEG-based MI classification task as a few-shot learning problem [65] (Fig. 6), and it could classify unseen subject data with a small number of MI EEG data. They also proposed a novel few-shot relation network consisting of a feature embedding module, attention module, and relation module in an end-to-end framework. The embedding module is used to extract semantic features from the support and query data. Given the extracted semantic features, the attention module is used to obtain the attention score for each support sample using both support and query features. Then, the representative vector for each class can be obtained using a weighted average of k support features with attention scores. Finally, the relation module is used to obtain the relation scores based on the distance metric between class-representative vectors and the query features. During training, the few-shot relation network is trained using pairs of support set and query data among different subjects in the training data. During testing, the label of a query datum can be taken as the class with the largest predicted relation score by using the k labeled support signals from an unseen subject. Therefore, the few-shot relation network enables good generalization ability to classify the query data of an unseen subject, even with a small amount of MI EEG data.

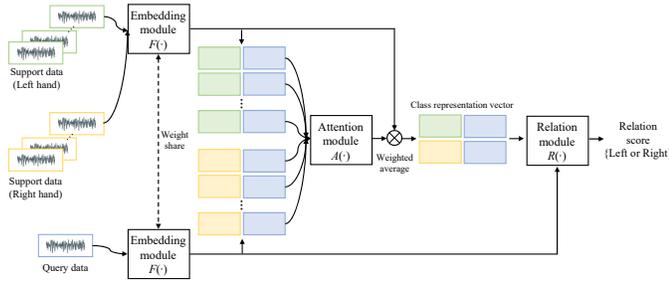


Fig. 6: Few-shot relation network (Figure adapted from [65]).

MI usually has only a few mental tasks that can be translated to corresponding commands (*e.g.*, imagine left-hand or right-hand movements). In a real-time BCI environment, a calibration procedure is particularly necessary for each user and each session. The aim of the calibration is to derive individualized parameterization of EEG signals. In fact, to construct the classifier, a separate parameterization is sought for discriminating each intentional control (IC) state, related to different MI, from the noncontrol (NC) state, and for discriminating among various IC states. The signal processing procedures of calibration include obtaining individualized parameters for classification, updating these parameters after the subsequent user training, and online signal processing and classification for BCI operation. The calibration consumes a significant amount of time that hinders the application of a BCI system in a real open-world scenario [62]. For example, the MI system requires a considerable amount of time to record sufficient EEG data for robust classifiers' training. Owing to these inevitable BCI environments, the MI-BCI system has considered the brain dynamics, reflecting each individual's EEG characteristics. In addition, the subjects tend toward a

state of inattention state in real-time experiments due to the long calibration times for offline experiments [68]. Despite the improved performance over the conventional methods, the deep learning methods often fail when training samples per subject are limited. Thus, a huge number of training samples need to be obtained from each target subject to train the robust model [65]. Moreover, the MI-based BCI system is often limited by the types of MI [13]. Usually, only a few mental tasks, such as the movements of left-hand, right-hand, and foot, can be recognized and translated to corresponding commands. Thus, it is important to simultaneously reduce the calibration time and increase the number of commands.

The learned classifier in supervised classification can only recognize the categories of target EEG signals that have been seen during training. However, it cannot deal with previously unseen classes. Seen or known classes refer to classes that are covered by the training dataset, while unseen or unknown classes refer to those that were not seen in the training dataset. Zero-shot learning, a powerful and promising learning paradigm [69], can recognize unknown categories of EEG signals [13], and has the potential to substantially reduce the calibration time. In zero-shot learning, there are some labeled training instances in the feature space. The classes covered by these training instances are referred to the known classes or seen classes. There are also some testing instances that do not belong to any known classes. These classes are referred to as the unknown classes or unseen classes. It is assumed that the classes covered by the training dataset and the classes of the testing dataset are disjoint.

However, the assumptions of zero-shot learning are so restrictive that it can only predict unknown classes. Thus, generalized zero-shot learning has been proposed to recognize both known and unknown categories of EEG signals. Duan *et al.* proposed a generalized zero-shot learning method for EEG classification in a MI-based BCI system [13], as shown in Fig. 7. It could recognize unknown task samples (*e.g.*, imagine left and right hands move simultaneously). The first step is to extract features from EEG signals using a common spatial pattern, and it aims to obtain discriminative spatial patterns by maximizing the variance ratios of filtered EEG signals for two classes. Then, the obtained EEG feature vectors are projected onto the target semantic space, which can help to recognize unknown task samples. The mean vector of each class is taken as semantic information, which can capture the distribution of different classes. Two fully connected layers with a \tanh activation function are used to map all the samples in each class to the corresponding mean vector. In this manner, testing samples of the known classes are clustered around the training samples from two classes (left-hand and right-hand MI). However, the unknown task samples are far away from the known classes. An outlier detection method is used to determine whether a mapped EEG feature belongs to known classes. If the mapped feature belongs to a known class, a classifier can be used to determine the class. Otherwise, it is assigned to a class based on the likelihood of being an unknown class. Therefore, the generalized zero-shot learning method can recognize not only unknown classes but also known classes. Hwang *et al.* proposed a new framework for

zero-shot EEG signal classification [70], which has three parts. The first part is an EEG encoder network that generates EEG features. The second part is a GAN that can recognize the unknown EEG labels with a knowledge base. The third part is a simple classification network to learn unseen EEG signals from the fake EEG features that are generated from the learned generator.

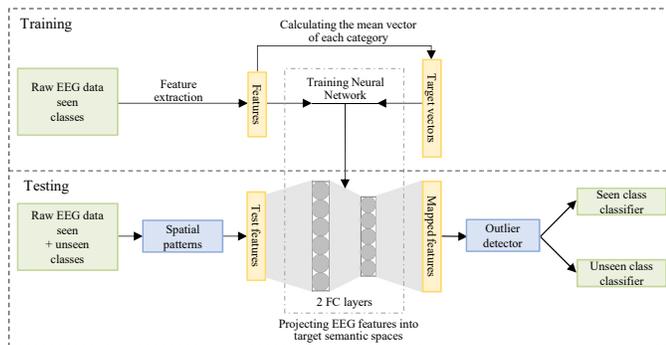


Fig. 7: Zero-shot learning for EEG based MI classification (Figure adapted from [13]).

To solve the small sample size problem in EEG classification, semi-supervised learning adopts a small number of labeled EEG data and a large amount of unlabeled EEG data simultaneously, which can be seen as a combination of supervised and unsupervised learning [71]. Since a large amount of unlabeled EEG data can help to capture the underlying distribution of data, semi-supervised learning methods can improve the performance of supervised learning by using unlabeled EEG data to learn more robust representations and hence alleviate the need for large amounts of labeled EEG data. For example, Jia *et al.* proposed a novel, semi-supervised DL framework for EEG emotion recognition [16]. They used label information in feature extraction and integrated the unlabeled information to regularize the supervised training. Specifically, they determined a sample’s potential contribution to the model training based on the uncertainty of the trained model over each unlabeled EEG sample. After the training stage, it can quickly predict the label for a test sample. With careful scrutinization, it is observed that the result is sometimes unreliable, that is, the value of conditional probability $P(y|X)$ for a test sample X with different label y can be very close. On the contrary, if the probability value for certain label y dominates the others, the model is quite confident with its decision. In this way, it can conclude that the sample in the former case contains more uncertainty [16], and it can result in a faster advance to the more accurate decision boundary. Thus, both supervised and unsupervised information can be jointly utilized in the entire training process to reduce model variance. Panwar *et al.* proposed a semi-supervised Wasserstein GAN with gradient penalty to classify driving fatigue from EEG signals [72]. This method is an extension of the Wasserstein GAN to include a classifier that predicts the class labels of the data, which enables the augmentation of limited training samples with generated EEG samples during training, and hence leads to improved classification performance.

2) *Self-supervised learning for discovering structure in unlabeled EEG data:* Most EEG-based DL models utilize supervised learning methods. However, supervised learning methods have several limitations. It is well known that DL requires a large amount of labeled EEG data to achieve satisfactory performance. However, most EEG studies are conducted under small labeled data regimes, and a few hundred subjects are often considered as big data. Large-scale EEG-based supervised learning is much rarer. Thus, it is difficult to achieve a desirable performance for most EEG-based DL models. In addition, supervised DL models must be trained from scratch for each task, and they require a large amount of computational resources and time. The learned representations are often very task-specific and are not expected to generalize well to other tasks. Furthermore, it is challenging to know exactly what the participants are thinking or doing in cognitive neuroscience experiments, and hence, it is difficult to obtain accurate labels [15]. Finally, supervised EEG-based DL models are prone to adversarial attacks.

Self-supervised learning, an unsupervised learning paradigm, can learn the underlying features from large-scale unlabeled data without using any labeled data, thereby avoiding the extensive cost of collecting and annotating large-scale datasets. Self-supervised learning usually reformulates the unsupervised learning problem as a supervised learning problem and is composed of a pretext task and downstream task [15]. Pretext tasks are pre-designed tasks that are helpful for downstream tasks. The supervision signal is generated from the data itself by leveraging the structure, instead of manual annotation, and the features can be learned by training the objective function of pretext tasks. After training the pretext task, the features learned by the pretext task can be transferred to the downstream task. By training the model to solve well-designed pretext tasks, self-supervised learning can help the model to learn more generalized representations from unlabeled data. Thus, it can achieve better performance and generalization on downstream tasks. In general, annotated labels are required to solve downstream tasks. However, in several applications, the downstream task can be the same as the pretext task without using any annotated labels.

Self-supervised learning has several advantages over supervised learning for EEGs. For example, self-supervised learning can exploit the structure of unlabeled data to provide supervision. The learned representations are often more general than task-specific supervised learning, and more robust to inter-class variation and intra-class variation. Thus, it can be reused for different tasks and save computation time compared to training a model from scratch for each task. In addition, self-supervised learning can make full use of large amounts of unlabeled data, which in turn can help to train much deeper and more sophisticated networks. Self-supervised learning can also improve the performance of DL when there are limited or no labeled data. Banville *et al.* investigated self-supervised learning to determine the structure in unlabeled data to learn useful representations of EEG signals [15]. Three pretext tasks were designed for EEG, *i.e.*, relative positioning, temporal shuffling, and contrastive predictive coding. Relative positioning aims to discriminate pairs of EEG samples based

on their relative positions. This is based on the assumption that an appropriate representation of the data should evolve slowly over time and EEG samples close in time should share the same label. Pairs of windows are sampled from time series S (EEG recording) so that the two windows of a pair are either close in time ('positive pairs') or farther away ('negative pairs'), and an end-to-end feature extractor h_{Θ} is trained to predict whether a pair is positive or negative. For temporal shuffling, triplets of windows are sampled from S , and a triplet is given a positive label if its windows are ordered or a negative label if they are shuffled. For contrastive predictive coding, sequences of $N_c + N_p$ consecutive windows are sampled from S along with random distractor windows ('negative samples'). Given the first N_c windows of a sequence ('context'), a neural network is trained to identify which window out of a set of distractor windows actually follows the context. Downstream tasks were performed on two EEG-based clinical applications, sleep staging and pathology detection. Experiments demonstrated that linear classifiers trained on self-supervised learned features can outperform supervised deep models under small labeled data regimes [15] and achieve competitive performance when all labels are available. In addition, the learned representation can reveal the latent structures related to physiological and clinical phenomena. Furthermore, all self-supervised learning tasks systematically outperformed or matched the compared methods in low-to-medium labeled data regimes, and remained competitive in a high labeled data regime.

Recently, contrastive learning has been successful in computer vision for representation learning. Chen *et al.* introduced a simple framework for contrastive learning of visual representations (SimCLR) [73], which can learn representations that are invariant under a set of augmentations through a contrastive loss. To learn EEG representations, Mohsenvand *et al.* modify the SimCLR framework to work with time-series EEG data [74]. The core idea of contrastive self-supervised learning is to train the networks while maximizing similarity for augmented instances of the same data point and minimizing the similarity between different data points. In contrast to images where the set of augmentations are intuitive and easily verifiable by the human eye, it is not clear what augmentations could be beneficial for EEG. They trained a channel-wise feature extractor by extending the contrastive learning framework to time-series data and introduced a set of augmentations for EEG. Mohsenvand *et al.* consulted four neurologists and two post-doctoral researchers at anonymized hospital research group specializing in clinical interpretation of EEG to identify a set of augmentations that do not change the interpretation of EEG data [74]. They chose the transformations that were easy to randomize programmatically, and ran preliminary experiments to choose a minimal effective set. Experiments demonstrated that the learned features can improve the accuracy of EEG classification and significantly reduce the amount of labeled data required for three EEG tasks, *i.e.*, emotion recognition, normal/abnormal EEG classification, and sleep-stage scoring.

3) *Robust EEG classification in the presence of noisy label:* In EEG experiments, each trial is associated with a stimulus and response, *i.e.*, the stimulus to the participant and the

behavioral response of the participant to it. The behavioral response can be a label, and it is assumed to be in accordance with the stimulus [14]. However, it requires a large amount of well-labeled EEG data for deep supervised learning to achieve desirable performance, and this may not always be available for real open-world applications. It is difficult to interpret and annotate EEG signals due to noise in the data and the complexity of brain processes, which can lead to high inter-rater variability, *i.e.*, label noise [15]. In addition, poisoning attacks can poison the training data by modifying their labels [75]. Label noise: the sample is valid but the label is wrong due to mislabeling. Data noise: the data is noisy but the label is valid, for example, samples caused by corruption, occlusion, distortion, and so on. Also, it is challenging to know exactly what the participants are thinking or doing in cognitive neuroscience experiments. In imagery tasks, for instance, the subjects might not be following instructions or the process under study might be difficult to quantify objectively (*e.g.* meditation, emotions). For example, participants may not always generate the intended emotions when watching emotion-eliciting stimuli [76]. Moreover, if participants become sleepy, bored, or distracted [14], it would lead to a significant increase in mislabeled trials. DL models would overfit to noisy labels due to the capability of learning any complex function, and the parameters obtained after training would deviate from the true optimal value, leading to a decline in the classification accuracy during testing. Therefore, it is necessary to consider noisy labels in real open-world applications. For example, in the field of EEG-based emotion recognition, humans have natural bias and inconsistencies in their judgments, which creates noise in their ratings. It is generally acknowledged that emotions are subjective, and studies have indicated that humans understand and perceive emotions varyingly. Moreover, participants may not always generate the intended emotion when watching emotion-eliciting stimuli, and the emotion label may be noisy and inconsistent with the actual elicited emotions. Zhong *et al.* proposed a regularized graph neural network for EEG emotion recognition using node-wise domain adversarial training and emotion-aware distribution learning to deal with noisy labels [76]. Instead of learning the traditional single-label classification, the emotion-aware distribution learning method learns a distribution of labels of the training data and thus acts as a regularizer to improve the robustness of our model against noisy labels. Up to now, few works have been proposed for EEG classification in the presence of noisy label. A number of methods have been proposed for image classification with DL in the presence of noisy labels [77], which can provide inspiration for noisy label EEG classification. For example, Patrini *et al.* proposed a loss correction method to make deep neural networks robust to label noise [78], and the minimizer of the corrected loss under the noisy distribution was the same as the minimizer of the original loss under the clean distribution. Huang *et al.* proposed a simple but effective noisy label detection method for deep neural networks without human annotations [79], and it only required adjusting the hyper-parameters of the deep neural network to make it transfer from overfitting to underfitting cyclically. Ren *et al.* proposed a novel method

that learns to assign weights to training examples based on their gradient directions [80], and it adopted a meta gradient descent step on the current mini-batch example weights to minimize the loss on a clean unbiased validation set.

III. CONCLUSION AND FUTURE DIRECTIONS

EEG decoding has made significant progress in the past few decades. In this article, we comprehensively surveyed existing EEG decoding methods using DL in the open-world setting. The challenges facing open-world EEG decoding were described for each step involved, *i.e.*, preprocessing, feature extraction and classification. In addition, approaches for solving open-world EEG decoding using DL were briefly introduced and summarized according to the core concepts, theory, progress, and examples. Despite the considerable progress that has been achieved in open-world EEG decoding using DL, there are several problems to be solved in future work, as illustrated in Table II.

The data noise in real applications is much more complex and with a different distribution than simulated EEG data noise. The characteristics of real-world data noise can differ with regard to different EEG collection settings and conditions. Thus, the problem of domain shift between real-world noise and simulated data noise should be considered in open-world EEG artifact removal. Traditional DL-based artifact removal methods aim to learn an end-to-end nonlinear function to map a noisy EEG signal to a clean EEG signal with known statistics. However, they tend to lack flexibility for blind and real open-world data noise. For example, a traditional deep network is trained under a specific level of SNR, and the trained network would fail for an unseen noise level. Thus, a more universal network should be trained with the entire expected range of SNR for blind denoising of contaminated EEG signals. The denoising performance of most CNN-based methods largely relies on supervised learning with a large amount of paired clean-noisy EEG signals. However, it would be difficult to collect true clean EEG signals for several open-world EEG decoding applications, and new DL-based artifact removal methods should be designed without access to clean EEG training examples or to paired clean-noisy EEG training examples. Unpaired DL methods can be used to solve this problem. In addition, it is important to train DL-based artifact removal methods using only noisy EEG signals. In sum, EEG artifact removal in open-world EEG decoding is a highly challenging task. Data noise can be natural or added by an adversarial attack. Similarly, label noise can be natural, or added by a poisoning attack. When there are both nature noise and adversarial noise for data and label, this is the worst-case for data noise and label noise in open-world EEG decoding, which needs to delve deep research in future work.

Most traditional domain adaptation methods in EEG signal analysis are assumed to have access to the source data during training. In several open-world EEG decoding applications, the requirements for accessing source domain data are restrictive. Sharing data can be a concern due to privacy and security issues. In addition, it is difficult to store, transmit, and process a large amount of source EEG data. Thus, it is necessary to

conduct source-free unsupervised domain adaptation, where the pre-trained model of the source domain is expected to adapt to unlabeled target data. Moreover, unsupervised domain adaptation models are usually applied to a single source domain and a single target domain, while multi-source and multi-target domain adaptation are typically encountered in open-world EEG decoding. Thus, single-domain adaptation may be suboptimal as it ignores the knowledge shared across multiple domains. Furthermore, when applying the classifier trained on one dataset to other datasets, the performance will be degraded significantly, and it is important to improve both cross-subject and cross-dataset classification performance.

In terms of other solutions for solving open-world EEG decoding using DL, only a few studies have started to focus on few-shot learning, zero-shot learning, semi-supervised learning, self-supervised learning, and noisy labels for EEG analysis. Thus, there is considerable scope for conducting in-depth studies on the above-mentioned approaches. Moreover, class imbalance occurs when the minority classes contain significantly fewer EEG samples than the majority classes. When a class imbalance exists in the training data, the learned classifier overclassifies the majority classes owing to their increased prior probability. Thus, the samples belonging to minority classes are more prone to misclassification than those belonging to the majority classes. The class imbalance problem is common in several EEG applications, such as EEG seizure detection tasks. The duration of seizure events is typically much shorter than that of non-seizure periods in long-term continuous EEG data. Thus, the classifiers will be biased toward non-seizure EEG signals if the class imbalance problem is not considered. Methods for addressing class imbalance problem based on deep learning can be divided into two main categories [81]. The first category is data level methods that operate on training set and change its class distribution [81]. They aim to alter dataset in order to make standard training algorithms work. For example, random minority oversampling simply replicates randomly selected samples from minority classes. However, as opposed to oversampling, undersampling removes randomly from majority classes until all classes have the same number of examples. The other category covers classifier (algorithmic) level methods. These methods keep the training dataset unchanged and adjust training or inference algorithms [81]. For example, threshold moving adjusts the decision threshold of a classifier. It is applied in the test phase and involves changing the output class probabilities.

There are also some new problems to be solved for open-world EEG decoding. Hybrid problems are a combination of existing problems. To solve multiple issues simultaneously, several new problems will arise in open-world EEG decoding, such as few-shot transfer learning, transfer learning in the presence of noisy labels, multi-task and multi-modal transfer learning, adversarial transfer learning, adversarial training with noisy labels, self-supervised transfer learning, and multi-task zero-shot learning.

Most DL-based classification methods in EEG analysis are offline learning algorithms. However, the classifier must be retrained using all training data together with newly-arriving EEG samples, making these methods inefficient and unscalable

TABLE II: Challenges and future directions.

	Challenges	Possible solutions
Preprocessing	How to solve the domain shift between real-world EEG noise and simulated data noise	Designing a more universal deep network
	How to design deep models without access to paired clean-noisy EEG samples	Unpaired deep models via GAN
Feature extraction	How to perform domain adaptation without access to the source EEG data	Source-free unsupervised domain adaptation
	How to exploit multiple source and target EEG data in transfer learning	Multiple source and target domain adaptation
Classification	How to solve the class imbalance problem in EEG signal analysis	Imbalanced learning
	How to train a deep model incrementally from a stream of EEG samples	Online learning
	How to design the architecture of a deep network automatically for EEG recognition	Neural architecture search
	How to deploy deep models on portable EEG devices with limited resources	Model compression
	How to achieve a higher level of automation in solving diverse EEG-based tasks	Automatic machine learning
	How to exploit the interpretability of deep models for EEG analysis	Interpretable DL
	How to design EEG models that are comparable or superior to human intelligence	Strong artificial intelligence methods

for real-time EEG data stream analysis. Online learning can train a deep prediction model incrementally from a stream of EEG samples without requiring re-analysis of previous data, and hence it has high efficiency, strong adaptability, and excellent scalability to dynamical environments. Therefore, it is necessary to apply online learning to open-world EEG decoding, which can learn new knowledge from incoming EEG samples incrementally.

DL methods have been widely used in EEG signal analysis. The network architecture design has a significant impact on the final EEG decoding performance. Various network architectures have been designed to achieve good performance. However, network architecture design relies heavily on prior knowledge and experience. Therefore, neural architecture search [82] aims to design the architecture of a network automatically to reduce human intervention as much as possible. Neural architecture search methods can be categorized according to three dimensions [82]: search space, search strategy, and performance estimation strategy. The search space defines which architectures can be represented in principle. The search strategy details how to explore the search space, which is often exponentially large or even unbounded. The objective of neural architecture search is to find architectures that achieve high performance on unseen data. Performance estimation refers to the process of estimating the performance, the simplest way is to perform a standard training and validation of the architecture on data. For example, Li *et al.* proposed a novel neural architecture search framework based on reinforcement learning for EEG-based emotion recognition [83], which can automatically design network architectures.

The good performance of deep models for EEG decoding is at the cost of huge memory consumption and high computational complexity. There are growing interests in deploying deep models on edge devices (*e.g.*, wearable device, mobile phone, medical equipment, etc.) that have a stringent budget on the resource and energy, and expect real-time processing. As a result, reducing the cost of memory and computational complexity in deep models, that is, model compression [84] of deep models without significantly decreasing the model performance for EEG analysis becomes an urgent and promising topic. For example, Wang *et al.* adopted the knowledge distillation to extract the distribution of training data from the complex network (teacher network) to a simple network (student network) for EEG emotion recognition [85].

Current artificial intelligence for EEG decoding mainly focuses on bridging the performance gap between machines

and human beings. However, general artificial intelligence replaces task-specific models with general artificial intelligence algorithmic systems, which can achieve a higher level of automation in solving diverse tasks. Automatic machine learning is a general artificial intelligence algorithm approach that can be applied to a wide range of tasks, including vastly different ones. The hyper-parameter settings of DL models have a significant impact on the final EEG decoding performance in the open world. Manual testing is a traditional and prevalent approach for tuning hyper-parameters, and it requires a deep understanding of the DL algorithms and their hyper-parameter value settings. However, manual tuning is ineffective for several problems in open-world EEG decoding. This has inspired studies on the automatic optimization of hyper-parameters. Hyper-parameter optimization aims to automate the hyper-parameter tuning process [86], which makes it possible for users to apply DL models to open-world EEG decoding problems effectively. In addition, when facing complex decision-making tasks in open-world EEG decoding, it is necessary to design automatic machine learning methods to solve complex tasks adaptively. Reinforcement learning is a branch of machine learning in which an agent can learn from interacting with an environment. Reinforcement learning does not require extensive engineering and heuristic design. In addition, reinforcement learning updates the parameters through trial and error, does not require the expected reward to be differentiable, and can deal with the search problem in a discrete space directly. Thus, deep reinforcement learning can combine the advantages of DL and reinforcement learning for open-world EEG decoding and hence can enable the agent to solve complex decision-making tasks.

Most deep models for EEG decoding are over-parameterized black-box models, and they can obtain high classification accuracy without interpretable knowledge representations. Therefore, it is often difficult to understand the prediction logic of deep models hidden inside the network. Thus, it is important to exploit the interpretability of deep models [87] for EEG decoding in both theory and practice.

Most current EEG decoding methods are not comparable with human intelligence, especially in changing, dynamic, and complex open-world environments. Strong artificial intelligence aims to design algorithms that are comparable or superior to human intelligence. Therefore, it is important to develop strong artificial intelligence methods to solve the challenging problems associated with open-world EEG decoding. For example, deep models usually require a large amount of

training data to achieve good performance. However, their ability to quickly learn new concepts is relatively limited. Meta-learning is known as “learning to learn models” [88]. It treats tasks as training examples and aims to train a model to adapt to all such training tasks. Thus, meta-learning can improve the ability of model generalization for open-world EEG decoding, and it can potentially design general methods applicable to both in-distribution and out-of-distribution tasks.

IV. ACKNOWLEDGMENTS

Xun Chen, Chang Li, and Aiping Liu were partially supported by the National Natural Science Foundation of China (grants 61922075, 41901350, and 61701158), University of Science and Technology of China research funds from the Double First-Class Initiative (grant YD2100002004), and Anhui Huami Information Technology Co., Ltd. Martin J. McKeown was supported by John Nichol Chair in Parkinson’s Research. All future correspondence should be sent to the corresponding author, Chang Li (changli@hfut.edu.cn).

V. AUTHORS

Xun Chen (xunchen@ustc.edu.cn) received his Ph.D. degree in biomedical engineering from the University of British Columbia. He is a distinguished professor with the Department of Neurosurgery, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, 230001, China. In addition, he heads the Department of Electrical Engineering and Information Science and directs the Institute of Advanced Technology-Huami Joint Laboratory for Brain-Machine Intelligence at the University of Science and Technology of China. His research interests include signal processing and machine learning in biomedical applications. He is a Senior Member of IEEE.

Chang Li (changli@hfut.edu.cn) received the B.S. degree in information and computing science from Wuhan Institute of Technology in 2012, and the Ph.D. degree in electronic information and communications from Huazhong University of Science and Technology in 2018. He is currently a Lecturer with the Department of Biomedical Engineering, Hefei University of Technology, Hefei, China. His current research interests include the areas of biomedical signal processing, hyperspectral image analysis and information fusion.

Aiping Liu (aipingl@ustc.edu.cn) received her Ph.D. degree in electrical and computer engineering from the University of British Columbia. She is an associate professor in the School of Information Science and Technology, University of Science and Technology of China. Her research interests include biomedical signal processing, neuroimaging analysis, and noninvasive brain stimulation. She is a Member of IEEE.

Martin J. McKeown (martin.mckeown@ubc.ca) received his M.D. degree from the University of Toronto. He completed a three-year research fellowship in the Computational Neurobiology Laboratory at the Salk Institute for Biological Studies in San Diego, California. He was an assistant professor of medicine at Duke University from 1998 to 2003, and he is currently the John Nichol Chair in Parkinson’s Research, a professor of medicine, and the director of the Pacific Parkinson’s Research Center, Vancouver, British Columbia, V6E 2M6, Canada. His research interests include examining novel treatments for Parkinson’s disease.

Ruobing Qian (qianruobing@fsyy.ustc.edu.cn) is the Chief Physician with the Department of Neurosurgery at The First Affiliated Hospital of University of Science and Technology of China (Anhui Provincial Hospital), the Executive Director of the Epilepsy Center, the President of the Anhui Association Against Epilepsy (AAAE). His current research interests include multi-modal neuroimaging, neuroelectrophysiology, surgical treatment of refractory epilepsy, and cognitive science. He has published over 50 scientific articles in prestigious journals and conferences.

Z. Jane Wang (zjanew@ece.ubc.ca) received her Ph.D. degree in electrical engineering from the University of Connecticut. She has been a research associate at the University of Maryland, College Park from 2002 to 2004. She is a full professor in the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada. Her research interests include statistical signal processing and machine learning, with applications in digital media and biomedical data analytics. She is a Fellow of IEEE and the Canadian Academy of Engineers.

REFERENCES

- [1] M. Hassan and F. Wendling, "Electroencephalography source connectivity: aiming for high resolution of brain networks in time and space," *IEEE Signal Processing Magazine*, vol. 35, no. 3, pp. 81–96, 2018.
- [2] J. Hubbard, A. Kikumoto, and U. Mayr, "Eeg decoding reveals the strength and temporal dynamics of goal-relevant representations," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.
- [3] W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, F. Wan, and X. Chen, "Eeg-based emotion recognition via channel-wise attention and self attention," *IEEE Transactions on Affective Computing*, 2020, doi: 10.1109/TAFFC.2020.3025777.
- [4] X. Chen, Z. J. Wang, and M. McKeown, "Joint blind source separation for neurophysiological data analysis: Multiset and multimodal methods," *IEEE Signal Processing Magazine*, vol. 33, no. 3, pp. 86–107, 2016.
- [5] R. T. Schirmer, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggersperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.
- [6] H. Zhang, R. Chavarriaga, Z. Khalilardali, L. Gheorghe, I. Iturrate, and J. d R Millán, "Eeg-based decoding of error-related brain activity in a real-world driving task," *Journal of neural engineering*, vol. 12, no. 6, p. 066028, 2015.
- [7] J. Minguillon, M. A. Lopez-Gordo, and F. Pelayo, "Trends in eeg-bci for daily-life: Requirements for artifact removal," *Biomedical Signal Processing and Control*, vol. 31, pp. 407–418, 2017.
- [8] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A review on transfer learning in eeg signal analysis," *Neurocomputing*, vol. 421, pp. 1–14, 2021.
- [9] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of neural engineering*, vol. 16, no. 5, p. 051001, 2019.
- [10] H. Lu, H.-L. Eng, C. Guan, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized common spatial pattern with aggregation for eeg classification in small-sample setting," *IEEE transactions on Biomedical Engineering*, vol. 57, no. 12, pp. 2936–2946, 2010.
- [11] Y. Song, D. Wang, K. Yue, N. Zheng, and Z.-J. M. Shen, "Eeg-based motor imagery classification with deep multi-task learning," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.
- [12] Z. Jia, X. Cai, G. Zheng, J. Wang, and Y. Lin, "Sleepprintnet: A multivariate multimodal neural network based on physiological time-series for automatic sleep staging," *IEEE Transactions on Artificial Intelligence*, 2021, doi: 10.1109/TAI.2021.3060350.
- [13] L. Duan, J. Li, H. Ji, Z. Pang, X. Zheng, R. Lu, M. Li, and J. Zhuang, "Zero-shot learning for eeg classification in motor imagery-based bci system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 11, pp. 2411–2419, 2020.
- [14] A. K. Porbadnigk, N. Görnitz, C. Sannelli, A. Binder, M. Braun, M. Kloft, and K.-R. Müller, "When brain and behavior disagree: Tackling systematic label noise in eeg data with machine learning," in *2014 International Winter Workshop on Brain-Computer Interface (BCI)*. IEEE, 2014, pp. 1–4.
- [15] H. Banville, O. Chehab, A. Hyvarinen, D. Engemann, and A. Gramfort, "Uncovering the structure of clinical eeg signals with self-supervised learning," *Journal of Neural Engineering*, 2020.
- [16] X. Jia, K. Li, X. Li, and A. Zhang, "A novel semi-supervised deep learning framework for affective state recognition on eeg signals," in *2014 IEEE international conference on bioinformatics and bioengineering*. IEEE, 2014, pp. 30–37.
- [17] X. Zhang and D. Wu, "On the vulnerability of cnn classifiers in eeg-based bcis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 5, pp. 814–825, 2019.
- [18] X. Chen, X. Xu, A. Liu, S. Lee, X. Chen, X. Zhang, M. J. McKeown, and Z. J. Wang, "Removal of muscle artifacts from the eeg: a review and recommendations," *IEEE Sensors Journal*, vol. 19, no. 14, pp. 5353–5368, 2019.
- [19] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. McKeown, V. Iragui, and T. J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, no. 2, pp. 163–178, 2000.
- [20] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, vol. 454, no. 1971, pp. 903–995, 1998.
- [21] R. R. Vázquez, H. Velez-Perez, R. Ranta, V. L. Dorr, D. Maquin, and L. Maillard, "Blind source separation, wavelet denoising and discriminant analysis for eeg artefacts and noise cancelling," *Biomedical signal processing and control*, vol. 7, no. 4, pp. 389–400, 2012.
- [22] H. Zhang, M. Zhao, C. Wei, D. Mantini, Z. Li, and Q. Liu, "Eeg-denoiset: A benchmark dataset for deep learning solutions of eeg denoising," *Journal of Neural Engineering*, vol. 18, no. 5, p. 056057, 2021.
- [23] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "Iclabel: An automated electroencephalographic independent component classifier, dataset, and website," *NeuroImage*, vol. 198, pp. 181–197, 2019.
- [24] R. Ghosh, N. Sinha, and S. K. Biswas, "Automated eye blink artefact removal from eeg using support vector machine and autoencoder," *IET Signal Processing*, vol. 13, no. 2, pp. 141–148, 2019.
- [25] G. Lin, J. Zhang, and Y. Liu, "Single shot reversible gan for bcg artifact removal in simultaneous eeg-fmri," *arXiv preprint arXiv:2011.01710*, 2020.
- [26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [27] N. K. Manjunath, H. Paneliya, M. Hosseini, W. D. Hairston, T. Mohsenin *et al.*, "A low-power lstm processor for multi-channel brain eeg artifact detection," in *2020 21st International Symposium on Quality Electronic Design (ISQED)*. IEEE, 2020, pp. 105–110.
- [28] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.
- [29] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [30] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.
- [31] A. Gupta, S. Parameswaran, and C.-H. Lee, "Classification of electroencephalography (eeg) signals for different mental activities using kullback leibler (kl) divergence," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 1697–1700.
- [32] J. Giles, K. K. Ang, L. S. Mihaylova, and M. Arvaneh, "A subject-to-subject transfer learning framework based on jensen-shannon divergence for improving brain-computer interface," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3087–3091.
- [33] C. W. Anderson, J. N. Knight, T. O'Connor, M. J. Kirby, and A. Sokolov, "Geometric subspace methods and time-delay embedding for eeg artifact removal and classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 142–146, 2006.
- [34] H. He and D. Wu, "Transfer learning for brain-computer interfaces: A euclidean space data alignment approach," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 2, pp. 399–410, 2019.
- [35] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.
- [36] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *arXiv preprint arXiv:1411.1792*, 2014.
- [37] S. Raghu, N. Sriram, Y. Temel, S. V. Rao, and P. L. Kubben, "Eeg based multi-class seizure type classification using convolutional neural network and transfer learning," *Neural Networks*, vol. 124, pp. 202–212, 2020.
- [38] W. Zhang, F. Wang, Y. Jiang, Z. Xu, S. Wu, and Y. Zhang, "Cross-subject eeg-based emotion recognition with deep domain confusion," in *International conference on intelligent robotics and applications*. Springer, 2019, pp. 558–570.
- [39] Y. Luo, S.-Y. Zhang, W.-L. Zheng, and B.-L. Lu, "Wgan domain adaptation for eeg-based emotion recognition," in *International Conference on Neural Information Processing*. Springer, 2018, pp. 275–286.
- [40] H. He and D. Wu, "Different set domain adaptation for brain-computer interfaces: A label alignment approach," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 5, pp. 1091–1108, 2020.

- [41] Z. Cao, L. Ma, M. Long, and J. Wang, "Partial adversarial domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 135–150.
- [42] P. Panareda Busto and J. Gall, "Open set domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 754–763.
- [43] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada, "Open set domain adaptation by backpropagation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 153–168.
- [44] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Universal domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2720–2729.
- [45] B.-Q. Ma, H. Li, W.-L. Zheng, and B.-L. Lu, "Reducing the subject variability of eeg signals with adversarial domain generalization," in *International Conference on Neural Information Processing*. Springer, 2019, pp. 30–42.
- [46] Y. Zhang and Q. Yang, "An overview of multi-task learning," *National Science Review*, vol. 5, no. 1, pp. 30–43, 2018.
- [47] X. Ma, S. Qiu, Y. Zhang, X. Lian, and H. He, "Predicting epileptic seizures from intracranial eeg using lstm-based multi-task learning," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2018, pp. 157–167.
- [48] J. A. Miranda-Correa and I. Patras, "A multi-task cascaded network for prediction of affect, personality, mood and social context using eeg signals," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 373–380.
- [49] X. Liu, L. Lv, Y. Shen, P. Xiong, J. Yang, and J. Liu, "Multiscale space-time-frequency feature-guided multitask learning cnn for motor imagery eeg classification," *Journal of Neural Engineering*, vol. 18, no. 2, p. 026003, 2021.
- [50] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.
- [51] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, J. Schmidt, and M. Shah, "Decoding brain representations by multimodal learning of neural activity and visual features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [52] H. Cai, Z. Qu, Z. Li, Y. Zhang, X. Hu, and B. Hu, "Feature-level fusion approaches based on multimodal eeg data for depression recognition," *Information Fusion*, vol. 59, pp. 127–138, 2020.
- [53] D. Wu, W. Fang, Y. Zhang, L. Yang, X. Xu, H. Luo, and X. Yu, "Adversarial attacks and defenses in physiological computing: A systematic review," *arXiv preprint arXiv:2102.02729*, 2021.
- [54] Z. Liu, L. Meng, X. Zhang, W. Fang, and D. Wu, "Universal adversarial perturbations for cnn classifiers in eeg-based bcis," *Journal of Neural Engineering*, vol. 18, no. 4, p. 0460a4, 2021.
- [55] D. J. Miller, Z. Xiang, and G. Kesidis, "Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks," *Proceedings of the IEEE*, vol. 108, no. 3, pp. 402–433, 2020.
- [56] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *arXiv preprint arXiv:1810.00069*, 2018.
- [57] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
- [58] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.
- [59] L. Meng, J. Huang, Z. Zeng, X. Jiang, S. Yu, T.-P. Jung, C.-T. Lin, R. Chavarriga, and D. Wu, "Eeg-based brain-computer interfaces are vulnerable to backdoor attacks," *arXiv preprint arXiv:2011.00101*, 2020.
- [60] A. Hussein, M. Djandji, R. A. Mahmoud, M. Dhaybi, and H. Hajj, "Augmenting dl with adversarial training for robust prediction of epilepsy seizures," *ACM Transactions on Computing for Healthcare*, vol. 1, no. 3, pp. 1–18, 2020.
- [61] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of eeg motor imagery signals," *Journal of neural engineering*, vol. 14, no. 1, p. 016003, 2016.
- [62] K.-H. Shim, J.-H. Jeong, and S.-W. Lee, "Gradual relation network: Decoding intuitive upper extremity movement imaginations based on few-shot eeg learning," *arXiv preprint arXiv:2005.02602*, 2020.
- [63] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.
- [64] R. Keshari, S. Ghosh, S. Chhabra, M. Vatsa, and R. Singh, "Unravelling small sample size problems in the deep learning world," in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*. IEEE, 2020, pp. 134–143.
- [65] S. An, S. Kim, P. Chikontwe, and S. H. Park, "Few-shot relation learning with attention for eeg-based motor imagery classification," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10933–10938.
- [66] J. Cheng, M. Chen, C. Li, Y. Liu, R. Song, A. Liu, and X. Chen, "Emotion recognition from multi-channel eeg via deep forest," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 453–464, 2020.
- [67] G. Pfurtscheller and F. L. Da Silva, "Event-related eeg/meg synchronization and desynchronization: basic principles," *Clinical neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.
- [68] Y. Jiao, Y. Zhang, X. Chen, E. Yin, J. Jin, X. Wang, and A. Cichocki, "Sparse group representation model for motor imagery eeg classification," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 631–641, 2018.
- [69] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–37, 2019.
- [70] S. Hwang, K. Hong, G. Son, and H. Byun, "Ezsl-gan: Eeg-based zero-shot learning approach using a generative adversarial network," in *2019 7th International Winter Conference on Brain-Computer Interface (BCI)*. IEEE, 2019, pp. 1–4.
- [71] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.
- [72] S. Panwar, P. Rad, J. Quarles, E. Golob, and Y. Huang, "A semi-supervised wasserstein generative adversarial network for classifying driving fatigue from eeg signals," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 3943–3948.
- [73] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [74] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Contrastive representation learning for electroencephalogram classification," in *Machine Learning for Health*. PMLR, 2020, pp. 238–253.
- [75] B. Biggio, B. Nelson, and P. Laskov, "Support vector machines under adversarial label noise," in *Asian conference on machine learning*. PMLR, 2011, pp. 97–112.
- [76] P. Zhong, D. Wang, and C. Miao, "Eeg-based emotion recognition using regularized graph neural networks," *IEEE Transactions on Affective Computing*, 2020, doi: 10.1109/TAFFC.2020.2994159.
- [77] G. Algan and I. Ulusoy, "Image classification with deep learning in the presence of noisy labels: A survey," *Knowledge-Based Systems*, vol. 215, p. 106771, 2021.
- [78] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1944–1952.
- [79] J. Huang, L. Qu, R. Jia, and B. Zhao, "O2u-net: A simple noisy label detection approach for deep neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3326–3334.
- [80] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4334–4343.
- [81] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.
- [82] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 1997–2017, 2019.
- [83] C. Li, Z. Zhang, R. Song, J. Cheng, Y. Liu, and X. Chen, "Eeg-based emotion recognition via neural architecture search," *IEEE Transactions on Affective Computing*, 2021, doi: 10.1109/TAFFC.2021.3130387.
- [84] L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model compression and hardware acceleration for neural networks: A comprehensive survey," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 485–532, 2020.
- [85] Z. Wang, T. Gu, Y. Zhu, D. Li, H. Yang, and W. Du, "Fldnet: Frame level distilling neural network for eeg emotion recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2533–2544, 2021.
- [86] T. Yu and H. Zhu, "Hyper-parameter optimization: A review of algorithms and applications," *arXiv preprint arXiv:2003.05689*, 2020.

- [87] Q.-s. Zhang and S.-c. Zhu, "Visual interpretability for deep learning: a survey," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 27–39, 2018.
- [88] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, "Meta-learning in neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, doi: 10.1109/TPAMI.2021.3079209.