IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 March 2024, accepted 16 April 2024, date of publication 29 April 2024, date of current version 8 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3394696

SURVEY

Data Constraints and Performance Optimization for Transformer-Based Models in EEG-Based Brain–Computer Interfaces: A Survey

AIGERIM KEUTAYEVA[1] AND BERDAKH ABIBULLAEV[2], (Senior Member, IEEE)
[1]Institute of Smart Systems and AI, Nazarbayev University, Astana 010000, Kazakhstan
[2]Department of Robotics and Mechatronics, Nazarbayev University, Astana 010000, Kazakhstan

Corresponding author: Berdakh Abibullaev (berdakh.abibullaev@nu.edu.kz)

This work was supported by the Institute of Smart Systems and Artificial Intelligence (ISSAI), Nazarbayev University, Astana, Kazakhstan.

ABSTRACT This work reviews the critical challenge of data scarcity in developing Transformer-based models for Electroencephalography (EEG)-based Brain-Computer Interfaces (BCIs), specifically focusing on Motor Imagery (MI) decoding. While EEG-BCIs hold immense promise for applications in communication, rehabilitation, and human-computer interaction, limited data availability hinders the use of advanced deep-learning models such as Transformers. In particular, this paper comprehensively analyzes three key strategies to address data scarcity: data augmentation, transfer learning, and the inherent attention mechanisms of Transformers. Data augmentation techniques artificially expand datasets, enhancing model generalizability by exposing them to a wider range of signal patterns. Transfer learning utilizes pre-trained models from related domains, leveraging their learned knowledge to overcome the limitations of small EEG datasets. By thoroughly reviewing current research and methodologies, this work underscores the importance of these strategies in overcoming data scarcity. It critically examines the limitations imposed by limited datasets and showcases potential solutions being developed to address these challenges. This comprehensive survey, focusing on the intersection of data scarcity and technological advancements, aims to provide a critical analysis of the current state-of-the-art in EEG-BCI development. By identifying research gaps and suggesting future directions, the paper encourages further exploration and innovation in this field. Ultimately, this work aims to contribute to the advancement of more accessible, efficient, and accurate EEG-BCI systems by addressing the fundamental challenge of data scarcity.

INDEX TERMS Brain-computer interfaces, data scarcity, survey, deep learning, transformer, EEG, self-attention mechanism, motor imagery, BCI.

I. INTRODUCTION

In the evolving landscape of Brain-Computer Interfaces (BCIs), the potential of Electroencephalography (EEG) to transform human interaction with external devices has become a focal point of both academic and practical interest. BCIs, leveraging the nuanced capabilities of EEG, promise unparalleled advancements across a spectrum of applications. These applications extend from providing lifelines of communication for individuals struggling with debilitating conditions such as amyotrophic lateral sclerosis [1], [2] and the identification and management of epileptic seizures [3]. Furthermore, the integration of EEG-based BCIs into cutting-edge prosthetics [4], immersive gaming environments [5], virtual reality platforms [6], and the broader domain of scientific research underscores the transformative impact of this technology [7].

Among the diverse methodologies deployed within BCIs, EEG distinguishes itself through the non-invasive monitoring of brain activity, offering a window into real-time neural dynamics with exceptional temporal resolution [8]. This characteristic of EEG, coupled with its accessibility,

The associate editor coordinating the review of this manuscript and approving it for publication was Md Kafiul Islam.

cost-effectiveness, and user-friendly nature, has driven EEG-based non-invasive BCIs to the forefront of neuroscience and technology discussions [9].

A critical area of focus within EEG-based BCIs is the analysis of Motor Imagery (MI) signals, which are generated by individuals imagining the movement of various body parts without actual movement. These signals are pivotal in interpreting intentions and facilitating interactions with BCI systems, especially in assistive healthcare and robotics [10], [11]. The MI signals, characterized by distinct patterns such as the modulation of α (8-12 Hz) and β (13-30 Hz) rhythms in the sensorimotor cortex, present a unique set of challenges and opportunities for BCI applications [8].

Despite the promising avenues EEG-based BCIs present, there are substantial hurdles in signal interpretation, primarily due to the complex, subtle, and often noisy nature of neural signals. The inherent challenges of weak signal-to-noise ratios, signal non-stationarity, lengthy calibration processes, and limited model generalization significantly complicate the decoding of MI EEG signals [7], [8]. These obstacles necessitate a shift towards more sophisticated analytical techniques capable of navigating the intricate landscape of EEG signals.

The introduction of deep learning methodologies into EEG signal analysis has opened new horizons for enhancing the accuracy and efficiency of MI signal decoding [12], [13], [14], [15], [16]. However, the transition towards leveraging deep learning for EEG interpretation is hindered by a significant challenge: the requirement for large, comprehensive datasets to train these complex models effectively [16], [17], [18], [19]. Given that MI EEG datasets are inherently limited in size, often comprising just a few hundred samples, the task of training robust, deep learning-based models becomes particularly challenging. This limitation is a direct consequence of the exhaustive and tiring nature of EEG data collection protocols, which can fatigue participants and thereby compromising the quality and quantity of usable data [10], [13].

Our survey focuses on addressing the data limitations and the need for efficient decoding techniques in EEG-based BCIs utilizing transformer-based models. We explore three key strategies to overcome challenges arising from limited datasets, as illustrated in Figure 1.

A. ATTENTION MECHANISM-BASED MODELS

Firstly, we delve into the significant impact of attention mechanisms in the deep learning domain, especially their ability to handle high-dimensional, non-stationary EEG data. By enabling models to dynamically prioritize the most relevant aspects of the input, even with small datasets, attention-based models have shown great promise in EEG-based BCIs. Recent studies have demonstrated their success in enhancing MI EEG signal decoding, significantly improving accuracy and computational efficiency [18], [20], [21], [22], [23].

B. DATA AUGMENTATION

Secondly, we investigate various data augmentation techniques, including geometric transformations, generative models, and feature transformations. These techniques aim to increase the diversity and volume of data available for model training, thereby improving model robustness and performance [24], [25], [26].

C. TRANSFER LEARNING

Lastly, we examine the role of transfer learning, which encompasses inductive, transductive, and unsupervised learning modalities. This approach leverages extensive pre-existing datasets to pretrain models, which are subsequently fine-tuned for specific MI decoding tasks. Transfer learning enables models to gain a deeper understanding of EEG signal characteristics, thereby enhancing decoding accuracy [19], [27], [28], [29].

Through an exhaustive review of these optimization strategies, this paper aims to provide a comprehensive overview of the current methodologies employed to navigate the challenges of data constraints in transformer-based MI EEG classification. By highlighting the advantages, limitations, and applicability of each approach, we aim to offer valuable insights into their potential impact on the future of EEG-based BCIs. Moreover, this discussion seeks to underline existing challenges and pave the way for future research endeavors. These efforts aim to improve transformer-based models for EEG-based BCIs.

II. OVERVIEW OF TRANSFORMER ARCHITECTURE

In the context of our survey on the impact of data scarcity in BCIs leveraging EEG signals, it is crucial to understand the role of advanced neural network architectures, particularly the Transformer model. Originally developed for natural language processing tasks, the Transformer architecture, introduced by Vaswani et al. in 2017, marked a significant departure from traditional recurrent neural network models such as RNNs and LSTMs [30]. These earlier models, despite their ability to handle sequential data, faced challenges such as difficulty in capturing long-term dependencies and limitations in parallel processing, which are crucial for efficient training on modern hardware.

The Transformer overcomes these challenges by employing a mechanism known as self-attention, enabling the model to process entire sequences of data in parallel. This not only improves training efficiency but also allows the model to dynamically concentrate on different sections of the input sequence as needed, increasing its ability to capture complex dependencies within the data. The architecture consists of an encoder and decoder, each made up of multiple layers that facilitate the construction of deep learning models. In the encoder architecture, every layer is equipped with a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. The decoder similarly includes these components but adds a cross-attention step

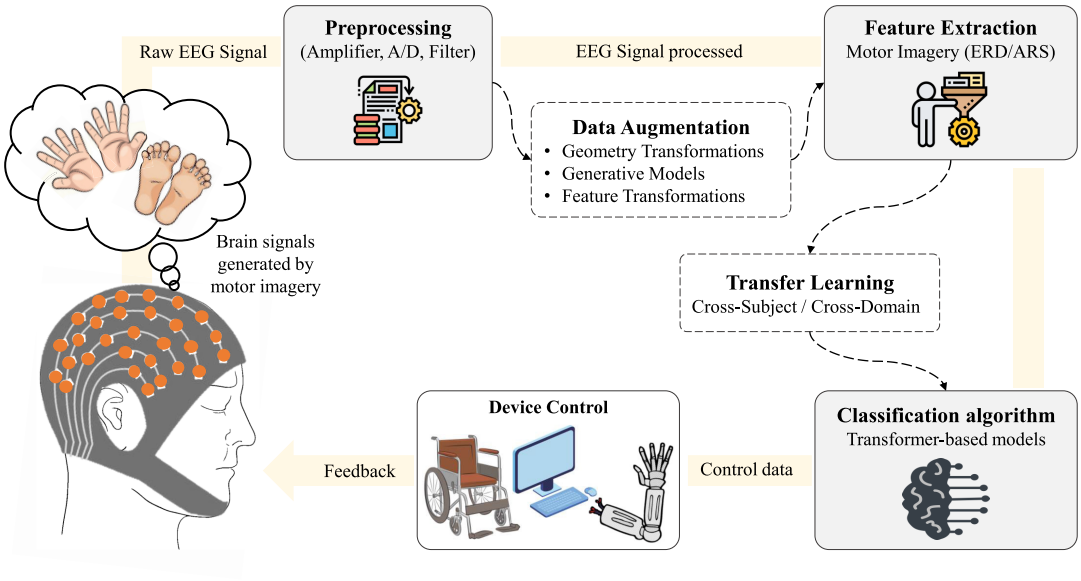

FIGURE 1. Schematic of EEG-based BCI framework, illustrating the stages from signal acquisition to device control using Transformer-based classification, highlighting preprocessing, data augmentation, and transfer learning steps.

between the self-attention and feed-forward layers, allowing it to selectively concentrate on pertinent segments of the input sequence based on the encoder's output. The overall architecture of the vanilla Transformer model is depicted in Figure 2.

For EEG classification tasks in BCIs, the Transformer architecture's ability to capture long-range dependencies and complex patterns within EEG signals makes it particularly suitable. It employs self-attention mechanisms to dynamically concentrate on relevant segments of the EEG sequence, improving the accuracy of classification tasks. Typically, the Transformer encoder is used to process EEG signals, with its output fed into a classification layer to predict mental states or tasks. This approach leverages the Transformer's strength in handling sequential data, making it a promising tool for advancing BCI technologies, especially in contexts where the availability of large-scale datasets is a challenge.

A. TRANSFORMER'S ARCHITECTURE FOR EEG CLASSIFICATION

To provide a comprehensive understanding of how a Transformer architecture can be applied to EEG data classification, we delve into the mathematical intricacies and highlight the relevance of data scarcity, particularly in the context of computing attention matrices and model training.

1) DATA STANDARDIZATION AND POSITIONAL ENCODING

Let's denote a dataset of EEG trials for training as $D_{\text{train}} = \{(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_n, y_n)\}$, where each $\mathbf{X}_i \in \mathbb{R}^{c \times p}$ represents the EEG data for trial i, with c channels and p time points. The class label y_i pertains to one of L possible classes. The raw data $\mathbf{X}_i$ undergoes channel-wise z-score normalization, resulting in $\hat{\mathbf{X}}_i$, to ensure homogeneity in data scale and distribution, which is critical for models sensitive to data variance.

Positional encoding is then added to $\hat{\mathbf{X}}_i$, yielding $\tilde{\mathbf{X}}_i$, to incorporate the sequential nature of EEG data into the model. This step is crucial since EEG signals are time-dependent, and capturing temporal relations enhances the model's predictive capability. The positional encoding employs sinusoidal functions to embed sequence position information, enabling the model to comprehend and utilize the sequential arrangement of EEG signal components effectively.

2) SELF-ATTENTION MECHANISM

The heart of the Transformer's ability to process sequences is its self-attention mechanism, which computes relevance-weighted aggregations of all elements in a sequence to generate context-aware representations. For a given trial $\tilde{\mathbf{X}}_i$, the scaled dot-product attention computes the output as:

$$\text{Attn}_{\mathbf{V},\mathbf{K},\mathbf{Q}}(\tilde{\mathbf{X}}_i) = \mathbf{V}\tilde{\mathbf{X}}_i \times \text{softmax}\left(\frac{\tilde{\mathbf{X}}_i^T \mathbf{K}^T \mathbf{Q} \tilde{\mathbf{X}}_i}{\sqrt{d_k}}\right), \quad (1)$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the queries, keys, and values matrices, respectively, obtained by projecting $\tilde{\mathbf{X}}_i$ with learned weights, and d_k is the scaling factor to prevent overly large dot products.

3) MULTI-HEAD ATTENTION

To address the constraints inherent in single-head attention and to enhance the model's capacity to capture diverse facets of the data, the Transformer utilizes Multi-Head Self-Attention (MHSA). This approach involves running several attention mechanisms in parallel, each with its own set of learned projections, and then concatenating their outputs.

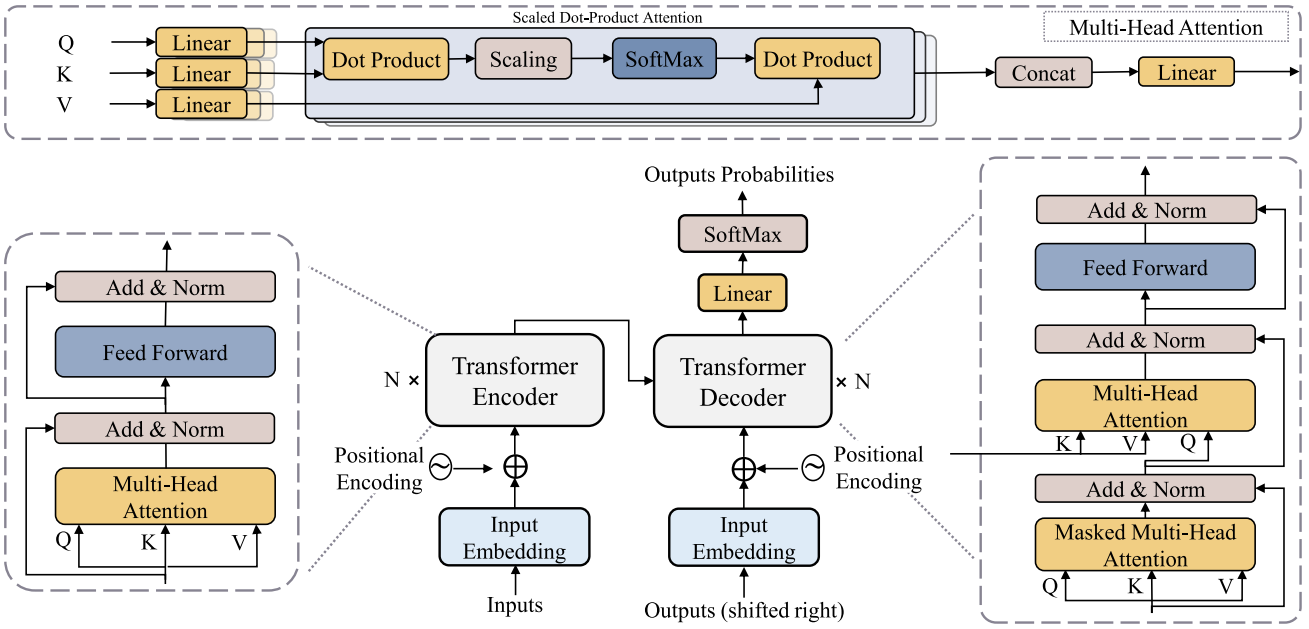

FIGURE 2. Architecture of the vanilla Transformer model. The model comprises an encoder and a decoder, each containing multiple identical layers. The layers within the encoder are equipped with multi-head self-attention mechanisms and feed-forward networks. In contrast, the layers in the decoder further integrate cross-attention mechanisms.

This can be mathematically represented as follows:

$$\text{MHSA}(\tilde{\mathbf{X}}_i) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O, \quad (2)$$

where $\text{head}_i = \text{Attn}_{\mathbf{V_i},\mathbf{K_i},\mathbf{Q_i}}(\tilde{\mathbf{X}}_i)$, and $\mathbf{W}^O$ is an output projection matrix. MHSA enables the model to concurrently attend to information from distinct representation subspaces, thereby augmenting its capacity to discern complex patterns.

4) SKIP CONNECTIONS AND LAYER NORMALIZATION

To further improve training stability and facilitate deeper model architectures, the Transformer architecture incorporates skip connections and layer normalization. Skip connections, also known as residual connections, facilitate the direct flow of gradients through the network, thereby alleviating issues associated with vanishing gradients. Following the attention and linear layers, the outputs are added to the original inputs (skip connection) and then normalized (layer normalization). This process is mathematically denoted as:

$$\mathbf{Y} = \text{LayerNorm}(\mathbf{X} + \text{Sublayer}(\mathbf{X})),$$

where $\mathbf{X}$ is the input to a sublayer, and $\text{Sublayer}(\mathbf{X})$ represents the output from either an attention or a linear layer. Layer normalization standardizes the outputs across the features for each data point, enhancing training efficiency and model performance.

III. DATA CONSTRAINTS IN EEG-BASED BCIS

Transformer-based models, with their capacity for handling sequential data and capturing long-range dependencies, hold significant promise for decoding and classifying MI EEG signals [7]. However, the effectiveness of these models heavily depends on the quantity and quality of the training data available [10], [13]. The scarcity of comprehensive, high-quality motor imagery EEG datasets constrains researchers' capacity to fully exploit the capabilities of Transformer-based deep learning techniques in MI EEG analysis. This scarcity is primarily attributed to the extensive experimental labor and the rigorous data collection processes that are required to gather MI EEG data [8]. Motor imagery tasks involve participants imagining the movement of their limbs without physically executing the movements. This process necessitates a careful experimental setup and strict adherence to protocol to guarantee the reliability and integrity of the collected EEG data [11], [12].

A. EXPERIMENTAL LABOR AND DATA COLLECTION CHALLENGES

The process of collecting motor imagery EEG data involves complex experimental setups that require precise calibration and configuration of EEG recording equipment [31]. Participants must be trained to perform motor imagery tasks consistently while minimizing artifacts that could contaminate the EEG signals, such as eye blinks or muscle movements unrelated to the imagined task [32], [33], [34].

An additional layer of complexity is introduced by the physical and mental demands placed on participants. The collection process is not only intricate and labor-intensive but also remarkably taxing for the subjects involved. They are required to sit still for extended periods, often stretching across several hours in a single session, and to repeat these sessions multiple times to gather sufficient data [35]. This

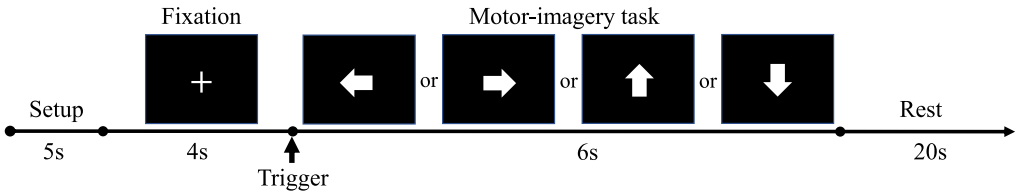

FIGURE 3. Experimental design for MI BCI paradigm.

requirement imposes significant physical and psychological strain on participants, influencing the feasibility and efficiency of data collection efforts [36].

- **Extended Sitting Periods**: The nature of EEG data collection necessitates minimal movement to avoid introducing noise into the data. For MI EEG tasks, this means participants must remain still for hours, focusing intensely on imagining specific movements without any physical execution [35]. This prolonged immobility can lead to physical discomfort, fatigue, and decreased concentration over time, impacting the quality of the collected data.
- **Multiple Sessions Requirement**: To achieve a dataset of sufficient size and variability, participants often need to undergo the data collection process multiple times. These repeated sessions, each requiring hours of intense concentration and stillness, can exacerbate the physical and mental fatigue experienced by participants [37]. The commitment required can also make it challenging to recruit and retain subjects, especially those who may find it difficult to visit the lab multiple times due to personal or logistical reasons. For instance, in motor imagery EEG experiments, estimating the number of trials that can be conducted within an hour involves considering several key factors: the duration of the motor imagery task, the rest period between tasks, and any initial setup or instruction time [31], [38]. Assuming a task duration of 10 seconds, a subsequent rest period of 20 seconds, and an additional 5 seconds allocated for setup or instructions before each task, the total duration for one trial is calculated at 35 seconds per trial (see Figure 3). This setup theoretically allows for around 102 trials to fit into a one-hour window, based on dividing the total number of seconds in an hour by the duration of each trial. However, the practical execution may result in a lower number of trials, as this estimate does not account for potential delays such as equipment calibration, extended breaks for participant comfort, or task repetitions due to data quality issues or artifacts [38], [39]. The balance between achieving a high number of trials and maintaining data quality and participant well-being is crucial, with specific experimental designs tailored to the study's goals and the participants' capacity.
- **Psychological Fatigue**: Beyond the physical discomfort of prolonged stillness, the mental effort involved in consistently generating vivid motor imagery can be mentally exhausting [40]. This cognitive strain, compounded over several sessions, can lead to diminished performance in the tasks, potentially affecting the reliability and quality of the EEG data collected.

The physical and psychological demands of participating in Motor Imagery EEG studies profoundly impact both data collection processes and the subsequent development of Transformer-based deep learning models [35], [41]. Ensuring the comfort and sustained engagement of participants is crucial for gathering high-quality, reliable data. The fatigue and discomfort experienced by subjects, necessitated by hours of stillness and intense concentration across multiple sessions, can lead to significant variability in data quality [42], [43]. This variability, stemming from the difficulty in consistently performing motor imagery tasks, directly contributes to the broader challenge of data scarcity, complicating the development and refinement of advanced deep learning methodologies.

The challenge of recruiting and retaining participants, amplified by the tiring nature of EEG data collection, further increases the scarcity of MI EEG datasets. This scarcity critically limits the training, testing, and validation of Transformer models and undermines the models' generalizability.

B. VARIABILITY AND NOISE IN EEG SIGNALS

Another challenge arises from the inherent variability in participants' ability to generate clear and consistent motor imagery patterns, which can significantly affect the quality and usability of the collected data. This variability necessitates the inclusion of a larger number of participants to ensure that the datasets are representative and robust enough for training deep learning models. There are other significant challenges that come with analyzing EEG signals due to various constraints. For example, these constraints include high variability between and within individuals, the presence of noise and artifacts, and the high dimensionality of the EEG data. Addressing these challenges is essential for improving the accuracy and reliability of the analysis [17], [44], [45], [46]. For instance, EEG signals exhibit considerable variability across different individuals (inter-subject variability) and within the same individual across different sessions (intra-subject variability), which stems from differences in brain anatomy, neural processing, and physiological and environmental factors [42], [47]. This variability presents

a significant challenge in developing robust BCI systems that exhibit strong generalization capabilities across different subjects and sessions.

Inter-subject variability is influenced by various factors, including lifestyle, gender, age, and psychological states such as motivation and frustration, which affect BCI performance [43], [48], [49], [50], [51]. For instance, it has been reported that females and individuals skilled in playing a musical instrument tend to be better BCI performers [43]. Furthermore, studies have identified a correlation between age and the duration of daily hand/arm movements with the modulation of α rhythms, which is pivotal for the performance of brain-computer interfaces [48].

Intra-subject variability can be affected by changes in physiological and psychological states over time. Factors such as motivation and fear have been reported to influence BCI performance positively and negatively, respectively [40]. Designing engaging and immersive BCI systems, such as those incorporating virtual reality, has been shown to improve user performance by enhancing attention and motivation [44], [45], [52], [53], [54].

EEG signals are also known for their low signal-to-noise ratio (SNR), making them susceptible to various types of noise and artifacts. These unwanted signals can originate from external sources, such as electrical equipment, or physiological sources such as eye movements (EOG), muscle activity (EMG), and heartbeats (ECG) [15], [42], [55], [56].

These physiological sources can include:

- **Electrooculographic Activity (EOG):** Eye movements, such as blinking and saccades, generate electrical activity that can be picked up by EEG electrodes placed on the scalp. EOG artifacts can obscure genuine brain signals, particularly in experiments involving eye-related tasks [57], [58], [59], [60], [61].
- **Electrocardiographic Activity (ECG):** The electrical activity of the heart, represented by the ECG, can introduce interference into EEG recordings. Heart-related artifacts can be challenging to distinguish from neural signals, particularly in studies where precise timing is crucial [36], [62].
- **Electromyographic Activity (EMG):** Muscle contractions and movements, especially in facial muscles, can contaminate EEG signals with electromyographic artifacts. These artifacts can be more pronounced during tasks that involve facial expressions or speech [58], [59], [61].
- **Ballistocardiographic Activity:** The pulsatile motion of the heart can generate ballistocardiographic artifacts in EEG recordings. These artifacts can vary with body position and may require specialized techniques for removal [33], [36].
- **Respiration:** Variations in breathing patterns can introduce fluctuations in EEG signals, particularly when electrodes are placed near the nose or mouth. These respiratory artifacts can affect the accuracy of BCI decoding, especially in tasks that require precise timing [34].

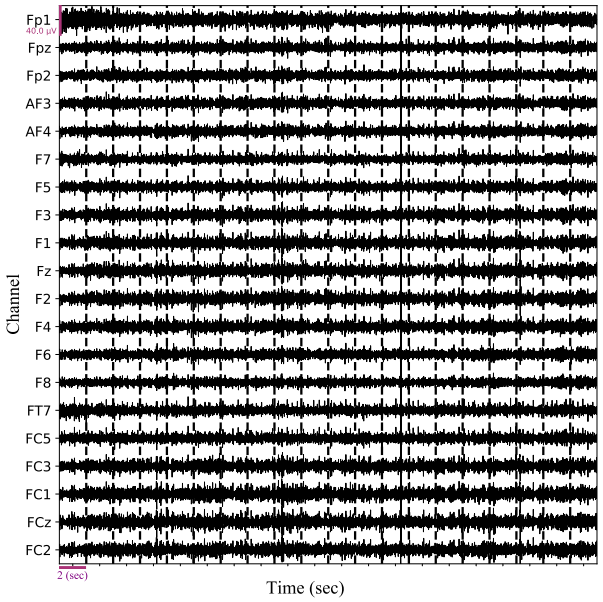

FIGURE 4. High dimensional representation of MI EEG data structure from Weibo2014 dataset (selected 20 channels) [66].

Effective preprocessing and denoising techniques are essential to improve the quality of EEG data and ensure accurate feature extraction and classification, addressing both environmental and physiological sources of noise [33], [35], [63], [64].

Recent approaches to mitigating the impact of variability and noise in EEG signals include the development of adaptive models that can adjust to individual characteristics and the use of transfer learning techniques to leverage data from multiple subjects. For example, the Temporal-Spatial Transformer (TST) model aims to enhance robustness and accuracy in classifying EEG signals across diverse subjects by focusing on both temporal and spatial aspects of the data, representing a promising direction in overcoming these challenges in BCI development.

Moreover, advanced denoising methods, such as the GAN-guided parallel CNN and transformer network (GCTNet) proposed by Yin et al. [65], show promise in effectively removing artifacts while preserving important EEG information. These approaches aim to strike a balance between eliminating noise and retaining the integrity of the underlying neural signals.

Another critical consideration in EEG data is its high dimensionality. EEG recordings often consist of a multitude of channels, each measuring electrical activity from a specific scalp location. High-density EEG systems can include dozens or even hundreds of electrodes, leading to a substantial number of channels. Addressing the high dimensionality of EEG data is crucial for enhancing the efficiency and accuracy of BCI systems. This aspect is further discussed in subsection III-C, and a high-dimensional representation of EEG data can be seen in Figure 4.

C. HIGH DIMENSIONALITY OF EEG DATA

The high dimensionality of EEG data poses several challenges in the context of BCI research:

- **Curse of Dimensionality:** High dimensionality can lead to the "curse of dimensionality," which refers to the exponential increase in data volume as the number of dimensions (channels) grows. This can result in increased computational complexity and memory requirements for processing and storing EEG data [24], [25].
- **Data Sparsity:** Another challenge in EEG-based BCI research is data sparsity. Limited data availability makes it challenging to apply transformer models reliably to EEG datasets. Researchers have been exploring data augmentation methods to address this issue and expand the dataset effectively [65].
- **Feature Extraction and Selection:** The large number of EEG channels necessitates careful feature selection and extraction to reduce dimensionality and focus on the most informative aspects of the data. Researchers often employ techniques like Common Spatial Patterns (CSP) to identify spatial patterns that discriminate between different cognitive states while reducing the dimensionality of the feature space [20], [25].

Recently, approaches such as the use of convolutional neural networks (CNNs) and transformer-based models have been explored to reduce dimensionality while capturing relevant spatial and temporal features. For example, the EEG Conformer model proposed by Song et al. [25] combines CNN elements with self-attention mechanisms to efficiently handle both local and global EEG features. Additionally, the development of models like the Swin Transformer for EEG-based BCIs by Wang et al. [20] showcases the potential of transformers in managing high-dimensional data while maintaining high accuracy. These advancements highlight the critical role of innovative model architectures in the future of EEG-based BCI research.

IV. CHALLENGES OF APPLYING TRANSFORMERS TO LIMITED EEG DATA

The application of transformer models to EEG data in BCIs presents several challenges. These include computational overheads, overfitting due to data sparsity, and the need to integrate domain-specific knowledge. Addressing these challenges is crucial for harnessing the potential of transformers in EEG-based BCIs.

A. COMPUTATIONAL OVERHEADS IN APPLYING TRANSFORMERS TO EEG DATA

Transformers bring a sophisticated level of data processing to BCIs, thanks to their self-attention mechanisms. However, their integration is not without significant computational demands, which are particularly acute when working with the continuous and high-dimensional nature of EEG data streams.

- **Computational Complexity:** Transformers' complex architecture, while powerful in identifying intricate patterns within data, requires substantial computational power, often resulting in increased processing times and resource demands, especially in settings with limited computational capacity [21], [25].
- **Continuous EEG Data Processing:** The real-time requirement of BCIs to process ongoing EEG data increases these computational challenges. Efficiently managing memory and computational load becomes essential to maintain the responsiveness of BCI systems [19].
- **Scalability and Resource Constraints:** In resource-constrained environments, the scalability of large transformer models is particularly affected, leading to limitations in their practical deployment for real-time EEG analysis [65].

Models such as the Gated Transformer [21] introduce gating mechanisms to manage the high temporal resolution of EEG signals, while multiscale convolutional transformers [15] blend convolutional layers with transformers to better capture spatial-temporal dynamics. Despite their effectiveness, the computational intensity of these models poses a barrier to their application in BCIs, where processing efficiency and low latency are of the essence.

The challenges of applying transformers to limited EEG data, characterized by data sparsity and the need for integrating domain-specific knowledge, underscore the necessity for optimized model architectures and advanced data augmentation techniques. Employing strategies such as data augmentation through generative models and transfer learning can address data scarcity and enhance the training of these computationally demanding models. These solutions, when effectively implemented, promise to overcome the current limitations and facilitate the broader adoption of transformer models in EEG-based BCI applications.

B. OVERFITTING CHALLENGES WITH TRANSFORMERS IN EEG DATA

The limited availability of EEG data elevates the risk of overfitting, a common challenge when applying deep learning models to small datasets. Transformers, with their large number of parameters, are particularly prone to overfitting when trained on limited data [18], [29].

Several strategies have been developed to mitigate overfitting in transformer-based models for EEG data. For example, Luo et al.'s shallow mirror transformer (SMT) introduces a novel network structure that augments data within the training phase, thereby enhancing model generalization on extensive EEG trial data. However, the variability in SMT's performance with shorter trial lengths underscores the challenge of adapting transformer architectures to the diverse sequence lengths intrinsic to EEG recordings [24].

The EEG Conformer, proposed by Song et al., combines CNN layers with self-attention mechanisms, facilitating the model's learning of both local and global temporal

EEG features. This integration helps reduce overfitting by leveraging the comprehensive temporal characteristics of EEG signals, although the scarcity of specialized calibration data for this approach remains a bottleneck for EEG decoding efficiency [25].

Tan et al.'s VAT-TransEEGNet model employs virtual adversarial training (VAT), introducing regularization that aids the model in optimizing the extraction of global features. This model is further refined through a particle swarm optimization (PSO) algorithm, enhancing its robustness in decoding the intricate neural activity patterns associated with motor imagery. Despite its advancements, the model's dependency on virtual noise and optimization underscores the persistent challenge of developing robust classifiers capable of navigating the complexities of EEG-based motor imagery classification [18].

These approaches demonstrate the concerted effort in the field to develop transformer models that can effectively circumvent the overfitting challenge, aiming to create more reliable and generalizable systems for EEG-based BCI applications. The continuous evolution of these strategies is indicative of the dynamic nature of BCI research, wherein model robustness and adaptability are in constant development to meet the demanding requirements of EEG signal interpretation.

V. PERFORMANCE OPTIMIZATION TECHNIQUES

Advanced performance optimization techniques are crucial for enhancing the functionality of EEG-based BCIs, particularly when addressing the challenges posed by the limited data availability of MI EEG signals. This section delves into attention mechanism-based models, which have shown promise in improving the performance of transformer-based models under these constraints. It is followed by an exploration of two additional strategies that can further enhance these models.

A. ATTENTION MECHANISM-BASED MODELS

Attention mechanisms are key components of transformer-based models, which have significantly advanced BCI systems. These methods are particularly adept at managing high-dimensional, non-stationary EEG data, enabling the model to dynamically prioritize the most pertinent aspects of the input. This ability to focus on crucial information allows transformer-based models with attention mechanisms to sometimes outperform state-of-the-art models, even when working with small datasets. Tables 1 and 2 provide an overview of transformer-based models in MI BCIs with attention methods applied, showcasing the diversity and effectiveness of these models in the BCI domain.

One notable example is the Temporal-Spatial Convolution and Transformer (TSCT) model developed by Shi et al. [67], which features a spatiotemporal convolution layer designed to autonomously focus on essential brain regions during the training phase. By combining convolutional blocks with partial Transformer encoders, the model significantly improves feature recognition and analysis, highlighting the effectiveness of attention mechanisms in optimizing EEG-based BCI classification.

Expanding on transformer configurations, Tao et al. [21] explored various models, including the Post-LN Transformer, Pre-LN Transformer, and GRUGate Transformer, for decoding human brain EEG signals. Their research demonstrates the versatility of attention mechanisms in managing the unique attributes of EEG data and offers insights into customizing transformer models for specific BCI applications.

Tan et al.'s [18] VAT-TransEEGNet algorithm, which employs Virtual Adversarial Training (VAT) for model regularization, showcases another innovative application of attention mechanisms. By addressing overfitting in limited EEG datasets, VAT enhances the model's stability and generalization ability, emphasizing the need for computational efficiency and refined hyperparameter tuning in such models.

Wang et al. [20] introduced the Swin Transformer for EEG-based BCIs, leveraging its shifted windows design to increase modeling capacity. This model's approach to integrating diverse data modalities, including visual, textual, and EEG signals, underlines the potential of attention mechanisms in broadening the capabilities of BCI systems.

Ahn et al. [15] utilized regularization techniques and multi-modal tasks to enhance the robustness of EEG data interpretation. By introducing diversity in learning patterns, these techniques have improved model performance, highlighting the importance of selecting appropriate regularization methods and managing multi-modal task complexities.

Hameed et al.'s [68] Temporal-Spatial Transformer (TST) employs attention mechanisms to adeptly navigate the temporal and spatial dynamics of EEG data, offering promising results in continuous EEG signal analysis with low signal-to-noise ratios. Challenges such as data scarcity and model tuning intricacies remain, with data augmentation and comprehensive hyperparameter optimization poised as potential solutions.

Deny et al.'s [17] hierarchical transformer architecture, which dissects long MI trials into short-term intervals, exemplifies how finely tuned attention mechanisms can significantly enhance BCI application accuracy. Despite high accuracy achievements, enhancing generalization across subjects remains an area ripe for improvement.

Lastly, Ma et al. [14] and Xie et al. [69] highlight the value of incorporating attention mechanisms for decoding MI-EEG and classifying raw EEG data, respectively. These studies underscore the benefits of attention-based models in improving the specificity and generalizability of BCI systems, indicating a fruitful direction for future research in making BCIs more personalized and effective.

Together, these advancements highlight the significant potential of attention mechanism-based models in EEG-based BCIs. While these models have shown considerable promise in enhancing the accuracy and reliability of BCI systems, opportunities for further enhancement remain.

Addressing the challenges outlined in Section IV and further exploring the performance optimization methods discussed in subsections B and C of Section V is crucial. Additionally, the selected studies provided in Section VI, underscore the ongoing need for refining these models to overcome current limitations and fully realize their potential in EEG BCI applications.

B. DATA AUGMENTATION TECHNIQUES

In EEG signal classification, data augmentation is a pivotal strategy to address the challenge of overfitting, often arising from the sparse nature of the data. By using a range of techniques to artificially expand the dataset, data augmentation plays a crucial role in improving model training and performance. This subsection delves into three key data augmentation methods: geometric transformations, generative models, and feature transformations, as illustrated in Figure 5. Table 3 provides a comprehensive overview of how these data augmentation (DA) methods have been applied in the context of MI-based BCI research.

1) GEOMETRY TRANSFORMATIONS

Geometric transformations have emerged as a crucial technique for enhancing the performance of EEG-based BCIs through data augmentation. By altering the spatial or temporal structure of EEG data, these methods introduce essential variability into training datasets, enhancing model generalizability. Recent studies have demonstrated the effectiveness of geometric transformations in improving BCI performance.

The work of Wang et al. [71] exemplifies this approach by employing repeated trial augmentation techniques such as random cropping and erasing. These methods modify the EEG data's structure in a controlled manner, simulating potential variability within EEG recordings and preventing the model from learning noise-specific patterns that could lead to overfitting.

Building on this, Song et al. [25] have explored the temporal domain through segmentation and reconstruction (S&R) techniques. This approach disassembles and reconstructs EEG sequences to generate an augmented dataset with a diverse range of temporal patterns, enabling models to better handle the temporal irregularities encountered in EEG signal analysis.

Further contributions to this field include the work of Wang et al. [41], who adopted an overlapped time slice strategy to overlay slices of EEG data, creating composite samples that enhance temporal diversity. Similarly, Luo et al. [24] have introduced a novel spatial transformation technique, the mirror EEG trials, which simulates different spatial perspectives and brain activity patterns, beneficial for ensemble learning approaches.

Expanding upon these spatial and temporal strategies, Ozelbas et al. [72] have employed a combination of techniques, including random time shifting and Gaussian noise addition, further broadening the data augmentation spectrum. Random time shifting, especially, is critical for temporal transformation, fortifying the model's resilience to variations in signal timing, a common occurrence in EEG data collection.

2) GENERATIVE MODELS

Generative models, specifically Generative Adversarial Networks (GANs), are transforming the field of EEG data augmentation. These models can generate synthetic datasets that are realistic in nature. By leveraging GANs to create synthetic EEG data, researchers are effectively addressing the issue of data scarcity, which often hampers the development of robust EEG-based BCI systems.

Penava and Buettner [74] have utilized GANs to generate synthetic EEG data, demonstrating the potential of GANs to expand the breadth of available EEG datasets for model training. Similarly, Habashi et al. [75] have employed GANs to produce synthetic images of EEG spectra, which can be particularly useful for tasks involving spectral analysis.

FBGAN, a specialized variant of GAN, was adopted by Zhang et al. [76] for EEG data generation. This adaptation of GAN technology is tailored specifically to the unique properties of EEG signals, reflecting ongoing innovation in generative model applications.

Beyond traditional GANs, Liang et al. [77] have proposed an auxiliary synthesis framework that uses generative models to synthesize supplementary data, thereby providing additional resources for model training without the need for extensive data collection.

The target-centered subject transfer framework presented by Yin et al. [78] represents another step forward, utilizing generative models to transfer a subset of source data to the target domain. This approach not only enhances the explainability of the target domain data but also ensures the augmentation process adds real, relevant data as opposed to mere noise, thus improving the overall performance of data-driven models.

3) FEATURE TRANSFORMATIONS

Feature transformation techniques are emerging as pivotal methods to enhance the robustness and generalizability of deep learning models. These techniques directly manipulate the extracted features from EEG signals, offering innovative ways to augment data and improve model performance.

Arı and Taçgın [73] highlight the importance of jittering and scaling operations as basic yet effective feature modifications. Jittering introduces slight variations to the data, simulating natural fluctuations in EEG signals, while scaling adjusts their amplitude, enabling models to recognize and adapt to different signal intensities.

Building on these concepts, Xie et al. [26] incorporate signal scaling, signal mixing, and the addition of Gaussian noise. Signal mixing merges features from different EEG recordings, enhancing the model's exposure to diverse brain activities. The addition of Gaussian noise mimics real-world artifacts and irregularities in EEG signals, training the model

TABLE 1. Overview of transformer-based models in motor imagery-based brain-computer interfaces with attention methods applied (SI: subject-independent, SD: subject-dependent) (continued on next page).

Ref.	Published	Model	Performance	Challenges	Future direction
[65]	2023	GCTNet (GAN-guided parallel CNN and transformer network) comprises: - Generator: Utilizes parallel CNN blocks and transformer blocks to capture local and global temporal dependencies, respectively. - Discriminator: Engaged to identify and rectify holistic discrepancies between clean and denoised EEG signals. - Evaluation: The network has been assessed using both semi-simulated and real data.	12.53% reduction in RMSE, 9.81% improvement in SNR over other methods.	The authors underscored the limitations of current deep learning-based EEG denoising techniques, notably their insufficient consideration of the temporal characteristics of artifacts and their neglect of the holistic consistency between denoised and clean EEG signals.	The authors suggest further exploration of the proposed method in diverse practical scenarios and the potential of integrating different temporal characteristics for artifact separation.
[14]	2023	CNN with an Attention Mechanism: - It is engineered to assimilate both spatial and spectral information, employing spatial convolutional layers. - The network includes temporal segmentation and feature extraction, succeeded by a temporal attention module. - Classification: Flatten operation and fully connected (FC) layers. - Early stopping.	SD: 2a: 82.32% OpenBMI: 77.52% SI: 2a: 79.48% OpenBMI: 70.43%	The paper addresses the difficulties in thoroughly examining temporal dependencies among MI-related patterns at various stages of MI tasks, which culminates in constrained decoding performance of MI-EEG.	The paper suggests further investigation into the proposed method across different practical scenarios and EEG datasets, emphasizing the potential of temporal dependency learning in improving EEG decoding performance.
[69]	2022	Each of the five models integrates a Transformer component, either combined with a spatial or temporal CNN or as independent spatial-Transformer and temporal-Transformer models.	3s data (f-CTrans): L/R/O: 74.44% L/R/O/F: 64.22% 6s data (t-CTrans): L/R/O: 78.98% L/R/O/F: 68.54%	The paper examines the challenges associated with EEG signal classification, such as limited spatial resolution, high temporal resolution, low signal-to-noise ratio, and substantial inter-individual variability.	The authors suggest further optimization of the Transformer models by exploring multi-scale attention models and reducing computation load by removing less contributive attention heads. Also, authors think the performance might be improved with more EEG data included.
[67]	2023	TSCT (Temporal-Spatial Convolution and Transformer): - It incorporates a spatiotemporal convolution layer with multiple trainable kernels serving as a feature extractor, automatically focusing on key brain regions during training. - A multi-branch Convolution block is employed to sustain the downsampling process. - A Partial Transformer encoder is integrated with a MLP to discern distinctive features from EEG signals.	Subject-specific testing: 2a: 83.3%	The authors address hurdles such as the considerable variability observed in EEG signals among individual subjects and recording sessions, alongside the constraints of conventional machine learning techniques in EEG decoding.	The authors suggest further exploration of data enhancement or transfer learning methods to improve the model's performance and plan to apply the proposed model to an online BCI system to verify its effectiveness and robustness.
[21]	2021	Gated Transformer (Post-LN Transformer, Pre-LN Transformer, GRUGate Transformer): - The architecture includes Input Embedding and Positional Encoding, which is followed by N times Encoder Block. - Each Encoder Block consists of a Layer Norm, MHA, and a Gating layer, followed by another Layer Norm, a FF layer, and another Gating layer. - The Gating layer is an extension of the Gated Recurrent Unit (GRU) approach.	SI: 55.40%	The authors delve into the difficulties associated with decoding EEG signals, stemming from their elevated temporal resolution and extensive sequences.	The paper suggests further evaluation of gated Transformers in other applications and incorporating them into end-to-end BCI development. Additionally, exploring factors that impact performance, such as band-pass filtering frequencies and electrode placement, is mentioned as future work.
[17]	2023	The two-level hierarchical transformer architecture: - The low-level transformer (LLT) is responsible for extracting features from short-term intervals. - The high-level transformer (HLT) pays more attention to the features from more relevant short-term intervals.	SD: BCI IV 2a: 90% PhysioNet: 83.5% Cho: 84.6% Lee: 82.1% SI: BCI IV 2a: 70.30% PhysioNet: 80.20% Cho:83.40% Lee: 81.30%	The paper discusses the challenges in designing accurate MI classification algorithms due to the variability in environmental and physiological conditions affecting EEG signal characteristics.	The authors suggest further exploration of the hierarchical transformer architecture in different practical scenarios and datasets to assess its generalization capabilities.
[68]	2023	TST (Temporal-Spatial Transformer): - After removing artifacts with an ICA filter, a TST is used to enhance EEG data quality. - This involves constructing temporal and spatial transformations using an attention mechanism. - A simple classifier is then built with a FC layer and GAP.	SD: 2a: 97.77% 2b: 85.90% LOSO: 2a: 93.94% 2b: 87.29%	The paper discusses challenges such as the complexity, variability, and low signal-to-noise ratio of EEG data, particularly in subject-independent classification.	The authors suggest further exploration of the TST model's performance in different practical scenarios and EEG datasets.

TABLE 2. (Continued from previous page) Overview of transformer-based models in motor imagery-based brain-computer interfaces with attention methods applied (CSE: cross-subject experiment).

Ref.	Published	Model	Performance	Challenges	Future direction
[18]	2023	VAT-TransEEGNet (EEGNet Combined with Transformer Architecture): - VAT is employed to regularize the neural network, introducing virtual noise at the input level while ensuring consistent outputs from the encoder. - The PSO algorithm is utilized to enhance the optimization of global features and parameters within the model.	CSE: 63.56%	The paper addresses challenges related to the suboptimal performance of current computer-aided classification frameworks, which are largely attributed to the complexity of the brain's neural electric field activity.	The authors suggest further exploration of the proposed method in different practical scenarios and EEG datasets to assess its generalization capabilities.
[20]	2023	Swin Transformer (ST) with EEG channel-attention (ECA): - ST vombined with ECA creates a powerful block for machine learning models. - The ST block consists of a shifted window-based MSA module, which is followed by a 2-layer MLP with GELU non-linearity at the center of the structure.	87.67%	The paper discusses challenges related to the high dimensionality of EEG datasets and the need for models that can capture the intricate temporal relationships in EEG signals.	The authors suggest further exploration of the proposed method in different practical scenarios and EEG datasets. They also plan to improve the model and reduce training time by incorporating few-shot learning techniques.
[47]	2019	CRAM (Convolutional Recurrent Attention model): - Temporal segments are partitioned and encoded through spatio-temporal mechanisms prior to the extraction of attentive temporal dynamics, which is facilitated by the integration of two sequentially arranged recurrent networks alongside a self-attention module. - This process is succeeded by a classification block.	59.10%	The paper discusses the challenges of subject-independent EEG signal analysis, including the high variability of EEG signals across different subjects.	The authors suggest further exploration of the proposed model in different practical scenarios and EEG datasets.
[23]	2022	ETST (EEG Temporal-Spatial Transformer): - Incorporates a Temporal Transformer Encoder (TTE), - Accompanied by a Spatial Transformer Encoder (STE).	Two states: PHY: 97.29%, IMA: 97.45% All states: 99.90%	The paper discusses challenges in EEG-based person identification, such as variability in brain signals and the need for robust models.	The authors suggest further exploration of feature extraction methods for EEG signals and the potential of applying their model to different states and conditions.
[22]	2022	CNN-Transformer: - A 1D-CNN based on Transformer architecture, consists of several components. - It starts with a convolutional layer with a ReLU activation function. - Followed by a temporal 1D-CNN with Max-Pooling, a convolutional layer with ReLU activation, and a BN layer. - A position encoder is then applied, followed by a Transformer with an Encoder and a Decoder. - The final component is a FC layer and a Soft-Max activation function.	99.29%	The paper discusses challenges related to the non-linearity and low signal-to-noise ratio of motor imagery EEG signals and the difficulty in effectively combining features from different domains.	The paper proposes that the model offers a conceptual framework for enhancing the accuracy of motor imagery EEG classification and recognition. Furthermore, it establishes a foundational basis for the broad implementation of motor imagery-based brain-computer interfaces.
[55]	2022	Transformers with Auto-Encoders: - Employ Filter Bank Common Spatial Pattern (FBCSP) for feature extraction, - Utilize Auto-Encoders (AE) for dimensionality reduction, - Implement a Vanilla Transformer for sequence modeling, - Conclude with a classification module.	91.30%	The paper discusses challenges related to the high-dimensional nature of EEG signals and the need for effective dimensionality reduction and classification techniques.	Not mentioned.
[56]	2022	TransEEG model: - CNN encoder, which consists of 2 2DConv layers followed by BN layer, ELU activation, Maxpooling, and Dropout; - Followed by three transformer blocks with graph embedding.	Private: 89.5% BCI IV 1: 77.4%	Challenges related to the high dimensionality of EEG datasets and the need for models that can capture the intricate temporal relationships in EEG signals are discussed.	Not mentioned.
[70]	2023	MITRT (motion imagery trajectory reconstruction Transformer) utilizes: - Pre-processed EEG as input for its Transformer encoder, - Corrected joint points location for its Transformer decoder. Its structure is similar to that of a vanilla Transformer.	97.5%	Challenges include the difficulty of collecting ground truth for imagery sign language motion and the high cost of EEG data acquisition.	The paper suggests exploring the reconstruction of real motion trajectory data using transfer learning, developing a cross-subject model for patients with limb disorders, and increasing data collection to enhance model training.
[15]	2023	The multiscale convolutional transformer incorporates: - Multi-kernel temporal convolutional blocks inspired by TSception, - A temporal transformer encoder for temporal sequence encoding, - Concurrent spatial convolutional blocks, - A spatial transformer encoder for spatial feature encoding, - A fusion convolutional block to integrate the diverse feature sets.	Private: 62% BCI IV 2a: 70% ASU: 70%	Challenges include the low spatial resolution and low signal-to-noise ratio of EEG, and the difficulty in extracting discriminative features as the number of classes increases.	Not mentioned.

TABLE 3. Overview of data augmentation (DA) methods in motor imagery-based brain-computer interfaces.

Ref.	Year	Dataset	Data Augmentation	Model	Performance	Pros and Cons
[72]	2024	BCI Competition IV 2a	Adaptive Cross-Subject Segment Replacement (ACSSR), along with other techniques like random time shifting, Gaussian noise addition, and cross-trial segment replacement for comparison.	Parallel two-branch convolutional neural network model with a spatiotemporal branch and a temporal branch.	80.47% - with DA 77.63% - no DA	- **Pros:** Addresses challenges in motor imagery signal classification, improves classification performance, and showcases the effectiveness of the ACSSR method. - **Cons:** Not provided.
[73]	2024	BCI Competition IV 2a and 2b	Jittering and scaling operations were applied to the data.	A generalized CNN model called No-Filter EEG (NF-EEG) that performs automatic feature extraction using raw EEG data.	2a: 93.56% - 2-class 2b: 88.40% - 2-class 2a: 81.05% - 4-class	- **Pros:** High classification accuracy without the need for signal preprocessing; Reduced time and effort required for data preparation; Effective input reshaping and data augmentation techniques. - **Cons:** The complexity of the model may require significant computational resources; Performance may vary depending on the quality and quantity of the available data.
[26]	2024	BCI Competition IV 2a and 2b	Signal mixing, Gaussian noise addition, and signal scaling.	The proposed BFATCNet model is structured around four principal components: a temporal feature block, an attention (AT) block, a temporal convolution (TC) block, and a fully connected block. It integrates advanced mechanisms such as a multi-head self-attention mechanism, a Convolutional Block Attention Module (CBAM), and a Bidirectional Feature Pyramid Network (Bi-FPN).	2a: 87.5% 2b: 86.3%	- **Pros:** The paper highlights the effectiveness of the BFATCNet model in EEG-based motor imagery classification, with high accuracy and stability across subjects. - **Cons:** Not provided.
[74]	2023	EEG dataset from 62 adult subjects	Generative Adversarial Networks (GANs) were used to generate synthetic EEG data.	Conditional Tabular GAN (CT-GAN) was used for synthetic data generation.	SD: 95.72% SI: 83.51% Predictive gains of 17.53% (SD) and 7.51% (SI) were observed with DA	- **Pros:** The paper highlights the potential of GANs to improve classification accuracy and simplify data acquisition for EEG-based BCIs. - **Cons:** It acknowledges limitations such as the need for external validation and further testing on different datasets.
[75]	2023	BCI Competition IV	Generative Adversarial Networks (GANs) were used to generate synthetic EEG spectrum images.	Two different Convolutional Neural Network (CNN) architectures were examined in the context of Motor Imagery (MI) classification.	76.71% - 2-class Enhancement of 2.5% with DA	- **Pros:** The paper highlights the effectiveness of GANs in improving MI BCI systems with limited data. - **Cons:** Not provided.
[76]	2023	BCI Competition IV 2a	Filter Bank Generative Adversarial Network (FBGAN) for generating high-quality EEG data.	A hybrid neural network combining a filter bank approach, sparse Common Spatial Pattern (CSP) features, and a Convolutional Recurrent Neural Network with Discriminative Features (CRNN-DF).	72.74% - 4-class	- **Pros:** The paper highlights the effectiveness of the hybrid neural network in subject-independent EEG classification. - **Cons:** Not provided.
[71]	2023	BCI competition IV 2a OpenBMI	Repeated trial augmentation consisting of random cropping and random erasing.	Interactive Frequency Convolutional Neural Network (IFNet), which explores cross-frequency interactions for enhancing representation of motor imagery (MI) characteristics.	2a: 11% improvement	- **Pros:** The paper highlights the effectiveness and superiority of the proposed IFNet for MI decoding. - **Cons:** Not provided.
[77]	2023	BCI Competition IV 2a	Auxiliary synthesis framework using a pre-trained auxiliary decoding model and a generative model to synthesize artificial data.	The framework consists of an auxiliary decoding model and a generative model. The auxiliary decoding model extracts temporal-spatial features of EEG signals, and the generative model synthesizes new data based on Gaussian noise.	2a: 4.72% improvement	- **Pros:** The paper highlights the effectiveness of the auxiliary synthesis framework in enhancing EEG-based classification with limited data. - **Cons:** Not provided.
[78]	2023	OpenBMI	Proposed a target-centered subject transfer framework for EEG data augmentation.	The framework includes a source-target relevance maximization strategy and self-adjustable generative learning.	Cycle-GAN: 82.34%	- **Pros:** The paper highlights the importance of using real data for augmentation and minimizing inter-subject variability through distribution transfer. - **Cons:** Not provided.

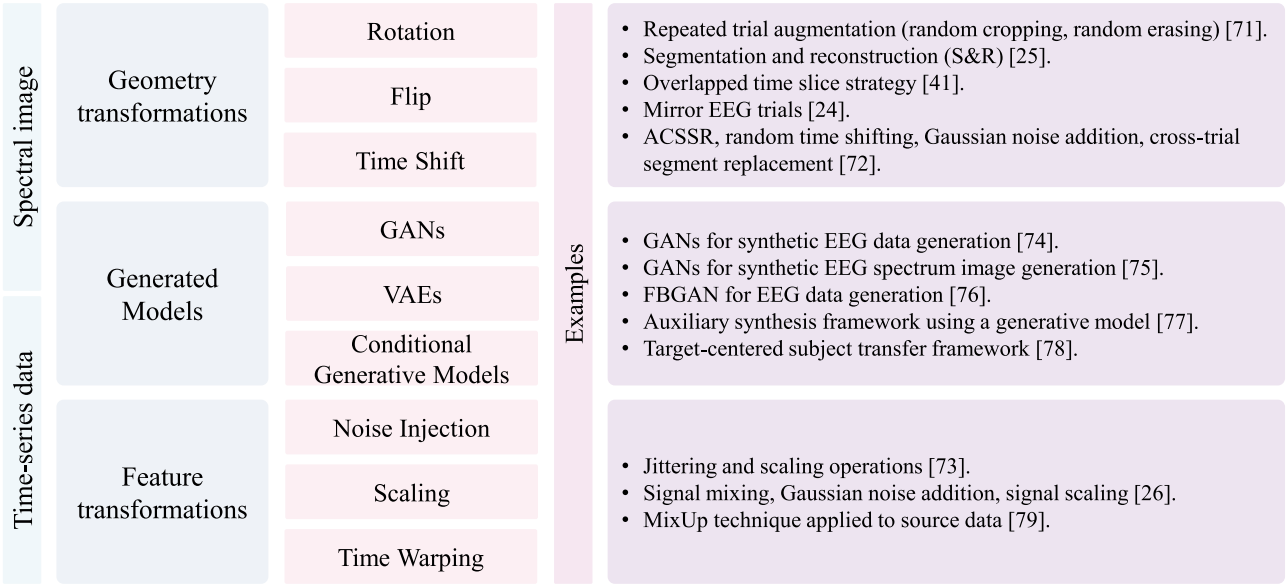

FIGURE 5. Categorization of data augmentation techniques for EEG-based brain-computer interfaces.

to filter out irrelevant information and focus on underlying patterns.

Yin et al. [79] introduce a technique called MixUp, which creates convex combinations of pairs of examples and their labels. MixUp promotes smoother model generalization across various states and conditions of EEG data, increasing the diversity of the training dataset and encouraging the model to interpolate between different signal types, thereby improving predictive accuracy and robustness.

Together, these feature transformation techniques provide a comprehensive toolkit for overcoming challenges associated with EEG data scarcity and variability. By modifying EEG signal features through jittering, scaling, signal mixing, noise addition, and the MixUp technique, researchers can significantly enhance their datasets. This not only addresses the limitations of available data but also equips deep learning models to handle the complex and dynamic nature of brain activity signals, paving the way for more effective and reliable BCIs.

In summary, the use of data augmentation techniques is critical in addressing the challenges of insufficient EEG data and its variability, which is fundamental for developing accurate deep-learning models such as transformers that can accurately interpret motor imagery EEG signals. Incorporating generative models plays a significant role in overcoming the constraints of EEG datasets that are both complex and limited in size. By generating new data and diversifying the training samples, these techniques are crucial for developing robust deep learning systems towards greater EEG decoding accuracy and the ability to adapt to various scenarios.

C. TRANSFER LEARNING

Transfer learning has emerged as an important strategy in refining deep neural networks for EEG-based BCI systems, particularly addressing the challenge of data scarcity and the requirement for transformer models to access extensive datasets for generalization. This approach, which leverages knowledge from previously learned tasks or domains, can significantly enhance model performance in new, related tasks or domains, even when labeled data are limited. This section delves into the various transfer learning methodologies applied in MI-BCI, as summarized in Table 4 and illustrated in Figure 6.

Inductive transfer learning focuses on leveraging knowledge from a source task to enhance performance on a related but distinct target task. For instance, in motor imagery BCI, knowledge gained from analyzing hand movement imagery tasks can be applied to improve models designed for foot movement imagery tasks. This category of transfer learning is invaluable when labeled data for the target task are available but perhaps not in sufficient volume required for training from scratch [19], [80].

Transductive transfer learning, or domain adaptation, aims to adapt a model from a source domain to a target domain where the tasks remain the same, but the data distributions differ. This approach is particularly relevant in MI-BCI systems where a model trained on EEG data from one group of subjects is adapted to perform well on data from another subject with unique brain signal characteristics. This approach addresses the variability in EEG signals across individuals, enhancing the model's applicability across diverse user groups [79], [81].

Unsupervised transfer learning is especially useful when no labeled data are available for the target task in the target domain. It allows for the utilization of unlabeled data from new subjects to improve or adapt the performance of pre-existing models. This method is crucial for MI-BCI applications where collecting labeled data can be labor-intensive and time-consuming. By leveraging unlabeled data, unsupervised transfer learning can significantly reduce the dependency on

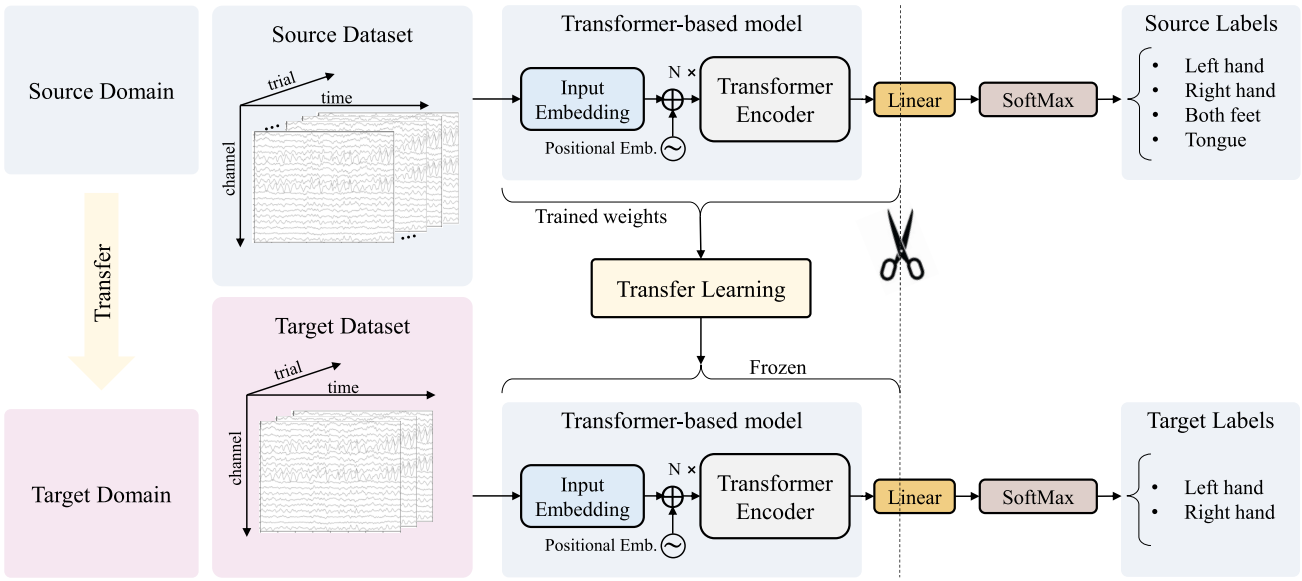

FIGURE 6. Schematic representation of transfer learning mechanism in MI-EEG BCI using transformer-based models, highlighting the process from source domain adaptation through transfer learning to target domain application.

extensive labeled datasets, making it easier to extend BCI systems to new tasks or subject groups without starting from zero [27], [29], [82], [83].

These transfer learning methods offer a suite of solutions to overcome the challenges posed by the need for large datasets for training transformer models in the BCI domain. By intelligently leveraging knowledge from related tasks or domains, transfer learning not only addresses data scarcity but also facilitates the development of more versatile and effective BCI systems capable of adapting to a wide range of tasks and individual differences in EEG signals.

Building on the foundation of transfer learning in MI EEG-based BCI, it is evident that categorizing these methods based on the type of transfer and domain is crucial for maximizing their effectiveness. Here are some common categories of transfer learning that can be applied to MI BCI systems (see Figure 7):

- **Feature Transfer** is an approach where features extracted from one subject or task are utilized to enhance the model's performance on another. This method is particularly effective in managing the inherent variability found in EEG signals, which can vary widely across individuals and tasks. By adapting these feature representations, Feature Transfer aims to make the model more versatile and capable of understanding and interpreting EEG data from different domains with greater accuracy [81].
- **Model Transfer** adopts a more holistic approach by applying the principles of transfer learning at the model level. In this method, a neural network initially trained on a specific set of EEG data is fine-tuned to accommodate new subjects or tasks. This process leverages the neural network's learned patterns from its original training to improve its performance on related but previously unseen data [84].
- **Domain Adaptation** focuses on minimizing the differences in data distribution between the source and target domains. Techniques employed in this approach, such as domain-invariant feature extraction, strive to make the model's performance more consistent across varied subject data. This is crucial for ensuring that a model trained on one group of subjects can adapt and perform effectively on data from another group, despite differences in EEG signal characteristics [28], [41].
- **Instance Transfer** is the method where specific data instances from the source domain are selectively used in the training process for the target domain. The selection of these instances is critically dependent on their relevance or similarity to the target domain, thus optimizing the knowledge transfer and making the training process more efficient and targeted [81].
- **Hybrid Transfer** leverages the strengths of two or more transfer learning strategies to achieve optimal results. By integrating approaches such as Feature Transfer and Instance Transfer, it is possible to significantly improve the model's classification accuracy. This method ensures that both the most relevant features and data instances contribute to the model's learning process, enhancing its ability to adapt and perform across diverse tasks and subject data [85].

In addition to the transfer learning frameworks mentioned above, exploring generative models such as GANs introduces a novel dimension to enhancing EEG signal analysis. The use of GANs, particularly in conjunction with CNNs and transformer networks for tasks such as artifact removal, exemplifies the innovative integration of pre-trained models for refining EEG signal quality. Jin Yin et al.'s [65] work

TABLE 4. Overview of transfer learning (TL) methods applied in the context of motor imagery-based brain-computer interfaces.

Ref.	Year	Dataset	Data Augmentation	Model	Performance	Transfer Learning	Pros and Cons
[41]	2023	BCI competition IV 2a	Overlapped time slice strategy.	Domain Adversarial Training of Neural Network (DANN) with feature extractors.	75.96%	The study implements calibration-free transfer learning through the utilization of Riemannian and Euclidean alignment, Multiple Kernel-Maximum Mean Discrepancy (MK-MMD), and Domain Adversarial Training of Neural Networks (DANN).	- **Pros:** Reduces calibration time, maintains classification accuracy, improves practicality, and can be extended to large-scale applications without additional trials for collecting labeled data. - **Cons:** The dataset used in the experiment is relatively small, and the impact of excluding different numbers of subjects in larger datasets is not considered. The CNN model used in feature extraction is relatively simple, and more advanced deep learning methods could be applied to improve performance.
[84]	2023	BCI Competition IV 2a	Not mentioned.	Ensembles of pretrained CNNs (AlexNet, GoogLeNet, and SqueezeNet) with a novel signal-to-image conversion approach.	84.18%	Uses pretrained CNNs (AlexNet, GoogLeNet, and SqueezeNet).	- **Pros:** Improved classification performance with limited labeled data, lower computational cost compared to time-frequency domain methods, potential for real-time MI-EEG applications. - **Cons:** Not provided.
[80]	2023	High-Gamma dataset (HGD), OpenBMI, GIST dataset	Not mentioned.	Lightweight CNN for cross-task adaptation between motor execution (ME) and motor imagery (MI) EEG data.	OpenBMI: 80% GIST: 72.73%	Uses a novel explainable cross-task adaptive transfer learning method for MI EEG decoding, leveraging ME EEG data for pre-training and fine-tuning with partial MI EEG data.	- **Pros:** Demonstrates the feasibility of transferring knowledge from ME tasks to MI tasks for EEG decoding, provides a complete pipeline for ME/MI cross-task adaptive transfer learning, and includes explainability analysis to understand discriminative EEG features. - **Cons:** Not provided.
[79]	2024	OpenBMI BCI Competition IV 2a	MixUp technique applied to source data.	A generative adversarial network (GAN) based framework with an encoder-generator structure for feature extraction and domain adaptation.	OpenBMI: 84% 2a+: 82.9% 2a: 69.2%	Utilizes unsupervised domain adaptation (UDA) for transferring knowledge between subjects.	- **Pros:** High performance in cross-subject EEG motor imagery classification; Effective integration of outlier removal and data augmentation techniques; Improved interpretability with Layer-Wise Relevance Propagation (LRP) analysis. - **Cons:** The complexity of the model may require significant computational resources; Performance may vary depending on the quality and quantity of the available data.

demonstrates the effectiveness of GANs in generating cleaner EEG data, a crucial step for improving the reliability and accuracy of BCI systems.

Further advancing the transfer learning landscape, the application of domain discriminators offers a strategic method for feature alignment across varied domains. Inspired by the principles behind GANs, this technique employs a domain discriminator to challenge and thus refine the model's ability to generalize features. By obfuscating the differences between source and target domain features, this approach significantly enhances the model's ability to adapt across diverse subject data, boosting its generalizability and performance [28].

The introduction of the Domain-Free Transformer (DFformer) by Kim et al. [29] represents a forward leap in creating adaptable and versatile models for BCI applications. This pre-trained model, designed to operate seamlessly across different EEG datasets, embodies the principles of transfer learning by offering a foundational model that can be tailored to a wide array of BCI tasks without the need for extensive retraining.

Collectively, these advancements underscore the dynamic potential of transfer learning strategies in surmounting the challenges faced in EEG-based BCI development. By leveraging the adaptability of transfer learning, alongside the power of generative models and domain adaptation techniques, the path is set toward developing BCI systems that are not only more effective and precise but also accessible and user-centric. This holistic approach marks a significant stride in realizing the full potential of BCIs, ensuring they can be tailored to meet the needs of diverse applications and user groups.

VI. EMPIRICAL STUDIES ON THE IMPACT OF DATA AUGMENTATION AND TRANSFER LEARNING ON TRANSFORMER-BASED MODELS FOR MI EEG CLASSIFICATION

In this section, we explore empirical studies on the impact of data augmentation and transfer learning on transformer-based models for Motor Imagery EEG classification, focusing on their applications, benefits, challenges, and potential improvements (see Table 5 for detailed overview).

Song et al. [25] developed the EEG Conformer, which combines CNNs with self-attention mechanisms. This model achieved promising accuracy of up to 92.96% across various datasets, indicating its potential for event-related potential (ERP) analysis and subject-independent models.

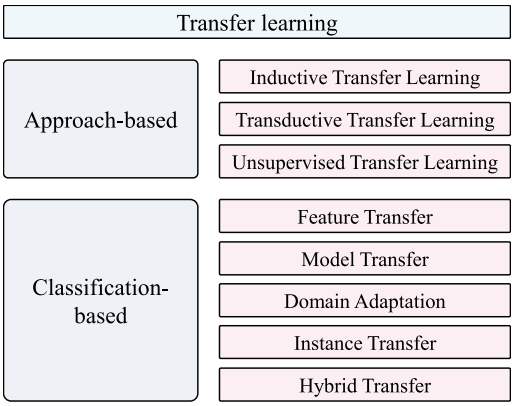

FIGURE 7. A Taxonomy of transfer learning methods in EEG-based motor imagery BCI, delineating the approaches based on the learning paradigm (Inductive, Transductive, and Unsupervised), and further classified by the techniques applied (Feature, Model, Domain Adaptation, Instance, and Hybrid Transfer).

Lee et al. [19] introduced the transformer-based spatial-temporal network (TSTN) for decoding MI intentions. Despite moderate accuracy ranging from 63% to 77% and progressive improvement, the TSTN faces challenges such as a small sample size and high computational demand. Future research could address these limitations by expanding datasets and optimizing model efficiency.

Kim et al. [29] proposed the domain-free transformer (DFformer), designed for dataset-agnostic pre-training. While demonstrating pre-training capability, challenges remain in optimizing a class token for single-trial data and achieving robust performance, particularly when fine-tuned on specific datasets like the Sleep Heart Health Study (SHHS).

Yin et al. [65] presented GCTNet, a model combining a GAN-guided structure with CNN and transformer networks for EEG denoising. Although it shows promise in artifact removal, GCTNet requires further exploration for direct application to motor imagery tasks, and the study lacks performance metrics for quantitative evaluation.

Finally, Song et al. [28] introduced the Global Adaptive Transformer (GAT), which employs domain adaptation to enhance cross-subject classification, achieving improved performance in this domain. However, further validation on diverse paradigms and a deeper understanding of the attention mechanism's role in decoding are recommended for further development.

These case studies collectively highlight the transformative potential of transformer-based models in EEG-BCIs. However, they also emphasize the importance of continued research to address limitations and challenges faced by these models, such as data scarcity, computational efficiency, and generalizability across different subjects and tasks. By refining these models and their associated techniques, there is a clear path towards developing more robust, accurate, and user-friendly BCIs, holding significant promise for individuals with neurological disabilities and the broader field of neurotechnology.

VII. FUTURE DIRECTIONS AND OPEN CHALLENGES

Building on the insights gained from our review, this section outlines potential future directions and open challenges in the field of EEG-based Brain-Computer Interfaces, with a focus on advancing transformer-based models.

1) ENHANCING DATA EFFICIENCY

One of the primary challenges in EEG-based BCIs is the scarcity of large, labeled datasets. Future research should explore innovative data augmentation techniques to mitigate this issue. For example, Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have shown promise in generating synthetic EEG data, which can be used to expand training datasets and improve model generalization [26], [65].

2) IMPROVING MODEL GENERALIZATION

Despite the success of transformer-based models in EEG classification, there is still room for improvement in their generalization capabilities across different subjects and tasks. Future work could focus on developing more robust models that can adapt to individual variations in EEG signals, potentially through the use of transfer learning techniques [28], [29].

3) ADDRESSING COMPUTATIONAL CHALLENGES

Transformer models are often computationally intensive, which can be a limitation for real-time BCI applications. Research should aim to optimize these models for efficiency, possibly by exploring lightweight architectures or hardware acceleration techniques. Additionally, further investigation is needed to understand and reduce the impact of factors such as electrode placement and signal preprocessing on model performance [20], [21].

4) ADVANCING INTERPRETABILITY AND PERSONALIZATION

Interpretable models are crucial for gaining insights into the neural mechanisms underlying BCI control. Future research should focus on developing interpretable transformer-based models that can provide meaningful explanations of their decisions. Moreover, personalizing models to individual users' neural signatures remains a critical challenge, and approaches such as hierarchical architectures may offer a promising direction for enhancing personalization in BCIs [17], [68].

5) LEVERAGING MULTIMODAL DATA

Integrating multimodal data, such as combining EEG with other physiological signals, can potentially enhance the performance and robustness of BCIs. Future research should explore how transformer-based models can effectively fuse and analyze data from multiple sources to improve BCI accuracy and user experience [70].

In general, while transformer-based models have shown great promise in EEG-based BCIs, there are still several open

TABLE 5. Overview of transformer-based models in motor imagery-based brain-computer interfaces with different optimization methods applied.

Ref.	Year	Datasets	Data Augmentation	Model	Performance	Transfer learning	Challenges
[27]	2021	TUEG dataset PhysioNet BCI IV 2a ERN dataset P300 dataset SSC dataset	Not mentioned.	BENDR: - A series of 1D convolutions with short-receptive fields and a transformer encoder. - The stack of convolutions and encoder allows for the representation of data in a more efficient and effective manner.	PhysioNet: 86.7% 2a: 42.6%	The study focuses on self-supervised learning and transfer learning from pre-trained models to specific EEG classification tasks.	The paper discusses challenges related to the limited labeled data in EEG and the need for models that can learn from large amounts of unlabeled data.
[24]	2023	BCI IV 2a and 2b OpenBMI dataset	Mirror EEG trials for data augmentation and ensemble learning. The electrode configuration in the mirror EEG trial mirrors that of the original EEG trial, reflecting an analogous spatial arrangement.	SMT (A shallow mirror transformer): - CNN with temporal convolution and spatial filtering, - MSA block consisting of position embedding, MSA layer, and position-wise FF layer, - Classification block. - At the training stage, a mirror network structure is used to implement data augmentation by creating mirror EEGs and an ensemble of probabilities is generated at the predicting stage. 29842 parameters	74.48% - new subjects 76.10% - existing subjects 2a: 67.28%, 2b: 76.41%, OpenBMI: 79.76% SI evaluation.	Not mentioned.	The proposed SMT model exhibits superior performance in extended EEG trials, with its efficacy diminishing as the duration of the EEG trials decreases. Consequently, this model is more suitable for analyzing long-sequence EEG data.
[25]	2023	BCI IV 2a and 2b SEED dataset (emotion)	Segmentation and Reconstruction (S&R) in the time domain are utilized for data generation. Training samples within the same category are uniformly segmented into Ns divisions. Subsequently, these segments are randomly concatenated while preserving their inherent temporal sequence.	EEG Conformer: - Employs convolution techniques to extract local temporal and spatial features; - Integrates self-attention mechanisms to capture global temporal characteristics.	2a: 78.66% 2b: 84.63% SEED: 92.96% SD experiments.	Not mentioned.	The paper discusses the challenges in EEG decoding, such as the need for learning both local and global temporal features and the limited data available for calibration.
[28]	2023	BCI IV 2a and 2b	Not mentioned.	GAT (Global Adaptive Transformer): - Feature extractor: parallel convolution to capture temporal and spatial features first. - Global adaptor: a novel attention-based adaptor that implicitly transfers source features to the target domain, includes Linear, MHA, and FC layers. - Domain discriminator inspired by GAN. - A classifier that uses two FC layers and a SoftMax activation.	2a: 76.58% 2b: 84.44% SI, cross-subject evaluation.	The GAT model utilizes a domain adaptation approach to enhance cross-subject EEG classification.	The authors discuss the challenges of poor feature representation and neglect of long-range dependencies in existing transfer learning methods for EEG classification.
[19]	2023	MI-VR private dataset	Not mentioned.	TSTN (transformer-based spatial-temporal network), proposed by Song et al. [87]: - Extracts spatial and temporal features, and employs attention mechanisms across both spatial and temporal dimensions to discern global dependencies. - AO+MI (Action Observation + Motor Imagery) technique; - MI-FB (Motor Imagery with Feedback) technique. The study also employs a Continual Learning strategy.	AO+MI: 63% 1st MI-FB: 68% 2nd MI-FB: 75% 3rd MI-FB: 77%	The study focuses on continual learning, where an initial classifier trained on action observation (AO) and MI data was continually improved with MI-FB data.	The paper explores the challenges associated with developing a universal pattern for initial BCI training and the complexities involved in transitioning from offline calibration to online feedback. Furthermore, the authors highlight the limitation of small data sizes, noting that data collection for a single subject required four weeks to complete.
[29]	2024	BCI IV 2a and 2b Sleep-EDF Sleep Heart Health Study (SHHS)	Not mentioned.	DFformer (domain-free transformer): - Tokenizer compresses information from the high-frequency raw EEG signals into the patch. - Biaxial information embedding block enriches the compressed EEG data with auxiliary information, including positional encoding and both intra- and inter-channel class tokens. - DFformer blocks; - Classification head.	2a: 58.41% 2b: 76.57%	The paper focuses on pre-training the model using an autoencoder-based reconstruction task to improve generalization across different EEG datasets.	The paper discusses challenges related to the unique characteristics of EEG datasets, such as different configurations and the need for a unified pre-trained model that can be applied across datasets without distortion.

challenges and opportunities for improvement. Addressing these challenges will require interdisciplinary collaboration and continued innovation to unlock the full potential of BCIs for enhancing human-computer interaction and improving the lives of individuals with neurological disabilities.

VIII. DISCUSSION AND CONCLUSION

This survey examined the application of Transformer architectures for electroencephalography (EEG)-based Brain-Computer Interfaces (BCIs), focusing on the challenge of data scarcity. We explored how effective Transformers are in classifying EEG signals, particularly noting their emphasis on long-range dependencies (multi-head attention) and stable training (layer normalization).

Attention mechanism-based models have emerged as a significant advancement, effectively handling high-dimensional, non-stationary EEG data by dynamically prioritizing relevant input aspects. These models have shown promise in enhancing MI EEG signal decoding, even with small datasets. However, there are opportunities for improvement, especially in addressing computational challenges and enhancing model generalization across different subjects and tasks.

Our investigation centered on how data augmentation and transfer learning were utilized as key strategies to mitigate the challenges arising from limited EEG datasets. Data augmentation expanded datasets, providing the model with a broader learning base and mitigating overfitting. By infusing the training data with a wider range of signal features,

data augmentation allowed the Transformer model to better capture the complex temporal patterns inherent in EEG data.

Transfer learning further strengthened the Transformer's performance by leveraging pre-trained models from domains such as vision and applying their learned knowledge to the EEG classification task. This cross-domain knowledge sharing allowed the Transformer model to perform better with limited training data, ultimately enhancing its accuracy in EEG signal classification.

Throughout this survey, several open challenges and future directions have been identified. Key areas for future research include addressing data scarcity through innovative augmentation techniques, improving model generalization, optimizing computational efficiency, enhancing interpretability and personalization, and leveraging multimodal data integration.

In conclusion, this survey positions Transformer-based models as a promising avenue for advancing EEG-based BCI technology, potentially improving the lives of individuals with neurological disabilities and broadening BCI applications. The continuous evolution of this field underscores the importance of interdisciplinary collaboration, particularly involving fields such as neuroscience, data science, and engineering, and innovative solutions to fully leverage the potential of Transformer-based models in EEG-BCI systems.

REFERENCES

[1] S. B. Borgheai, J. McLinden, A. H. Zisk, S. I. Hosni, R. J. Deligani, M. Abtahi, K. Mankodiya, and Y. Shahriari, "Enhancing communication for people in late-stage ALS using an fNIRS-based BCI system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1198–1207, May 2020.

[2] D. Lesenfants, D. Habbal, Z. Lugo, M. Lebeau, P. Horki, E. Amico, C. Pokorny, F. Gómez, A. Soddu, G. Müller-Putz, S. Laureys, and Q. Noirhomme, "An independent SSVEP-based brain–computer interface in locked-in syndrome," *J. Neural Eng.*, vol. 11, no. 3, May 2014, Art. no. 035002.

[3] Y. Zhang, Y. Guo, P. Yang, W. Chen, and B. Lo, "Epilepsy seizure prediction on EEG using common spatial pattern and convolutional neural network," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 465–474, Feb. 2020.

[4] G. R. M'uller-Putz and G. Pfurtscheller, "Control of an electrical prosthesis with an SSVEP-based BCI," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 361–364, Jan. 2008.

[5] B. van de Laar, H. Gürkök, D. Plass-Oude Bos, M. Poel, and A. Nijholt, "Experiencing BCI control in a popular computer game," *IEEE Trans. Comput. Intell. AI Games*, vol. 5, no. 2, pp. 176–184, Jun. 2013.

[6] F. Lotte, J. Faller, C. Guger, Y. Renard, G. Pfurtscheller, A. Lécuyer, and R. Leeb, *Combining BCI With Virtual Reality: Towards New Applications and Improved BCI*. Berlin, Germany: Springer, 2013, pp. 197–220.

[7] B. Abibullaev, A. Keutayeva, and A. Zollanvari, "Deep learning in EEG-based BCIs: A comprehensive review of transformer models, advantages, challenges, and applications," *IEEE Access*, vol. 11, pp. 127271–127301, 2023.

[8] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of EEG-based brain–computer interface paradigms," *J. Neural Eng.*, vol. 16, no. 1, Jan. 2019, Art. no. 011001.

[9] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.

[10] T. Mulder, "Motor imagery and action observation: Cognitive tools for rehabilitation," *J. Neural Transmiss.*, vol. 114, no. 10, pp. 1265–1278, Oct. 2007.

[11] H. Yuan and B. He, "Brain–computer interfaces using sensorimotor rhythms: Current state and future perspectives," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1425–1435, May 2014.

[12] I. Dolzhikova, B. Abibullaev, R. Sameni, and A. Zollanvari, "Subject-independent classification of motor imagery tasks in EEG using multisubject ensemble CNN," *IEEE Access*, vol. 10, pp. 81355–81363, 2022.

[13] A. Al-Saegh, S. A. Dawwd, and J. M. Abdul-Jabbar, "Deep learning for motor imagery EEG-based classification: A review," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102172.

[14] X. Ma, W. Chen, Z. Pei, J. Liu, B. Huang, and J. Chen, "A temporal dependency learning CNN with attention mechanism for MI-EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 3188–3200, 2023.

[15] H.-J. Ahn, D.-H. Lee, J.-H. Jeong, and S.-W. Lee, "Multiscale convolutional transformer for EEG classification of mental imagery in different modalities," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 646–656, 2023.

[16] A. Keutayeva and B. Abibullaev, "Exploring the potential of attention mechanism-based deep learning for robust subject-independent motor-imagery based BCIs," *IEEE Access*, vol. 11, pp. 107562–107580, 2023.

[17] P. Deny, S. Cheon, H. Son, and K. Won Choi, "Hierarchical transformer for motor imagery-based brain computer interface," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 11, pp. 5459–5470, Nov. 2023.

[18] X. Tan, D. Wang, J. Chen, and M. Xu, "Transformer-based network with optimization for cross-subject motor imagery identification," *Bioengineering*, vol. 10, no. 5, p. 609, May 2023.

[19] P.-L. Lee, S.-H. Chen, T.-C. Chang, W.-K. Lee, H.-T. Hsu, and H.-H. Chang, "Continual learning of a transformer-based deep learning classifier using an initial model from action observation EEG data to online motor imagery classification," *Bioengineering*, vol. 10, no. 2, p. 186, Feb. 2023.

[20] H. Wang, L. Cao, C. Huang, J. Jia, Y. Dong, C. Fan, and V. H. C. de Albuquerque, "A novel algorithmic structure of EEG channel attention combined with Swin Transformer for motor patterns classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 3132–3141, 2023.

[21] Y. Tao, T. Sun, A. Muhamed, S. Genc, D. Jackson, A. Arsanjani, S. Yaddanapudi, L. Li, and P. Kumar, "Gated transformer for decoding human brain EEG signals," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 125–130.

[22] H. Liu, Y. Liu, Y. Wang, B. Liu, and X. Bao, "EEG classification algorithm of motor imagery based on CNN-transformer fusion network," in *Proc. IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Dec. 2022, pp. 1302–1309.

[23] Y. Du, Y. Xu, X. Wang, L. Liu, and P. Ma, "EEG temporal–spatial transformer for person identification," *Sci. Rep.*, vol. 12, no. 1, p. 14378, Aug. 2022.

[24] J. Luo, Y. Wang, S. Xia, N. Lu, X. Ren, Z. Shi, and X. Hei, "A shallow mirror transformer for subject-independent motor imagery BCI," *Comput. Biol. Med.*, vol. 164, Sep. 2023, Art. no. 107254.

[25] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG conformer: Convolutional transformer for EEG decoding and visualization," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 710–719, 2023.

[26] X. Xie, L. Chen, S. Qin, F. Zha, and X. Fan, "Bidirectional feature pyramid attention-based temporal convolutional network model for motor imagery electroencephalogram classification," *Frontiers Neurorobot.*, vol. 18, Jan. 2024, Art. no. 1375309, doi: 10.3389/fnbot.2024.1375309.

[27] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data," *Frontiers Hum. Neurosci.*, vol. 15, Jun. 2021, Art. no. 653659, doi: 10.3389/fnhum.2021.653659.

[28] Y. Song, Q. Zheng, Q. Wang, X. Gao, and P.-A. Heng, "Global adaptive transformer for cross-subject enhanced EEG classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2767–2777, 2023.

[29] S.-J. Kim, D.-H. Lee, H.-G. Kwak, and S.-W. Lee, "Toward domain-free transformer for generalized EEG pre-training," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 482–492, 2024.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017, pp. 5998–6008.

[31] M.-H. Lee, O.-Y. Kwon, Y.-J. Kim, H.-K. Kim, Y.-E. Lee, J. Williamson, S. Fazli, and S.-W. Lee, "EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy," *GigaScience*, vol. 8, no. 5, May 2019, Art. no. giz002.

[32] W. Xiong and Q. Wei, "Reducing calibration time in motor imagery-based BCIs by data alignment and empirical mode decomposition," *PLoS One*, vol. 17, no. 2, Feb. 2022, Art. no. e0263641.

[33] N. Elsayed, Z. Saad, and M. Bayoumi, "Brain computer interface: EEG signal preprocessing issues and solutions," *Int. J. Comput. Appl.*, vol. 169, no. 3, pp. 12–16, Jul. 2017.
[34] S. Vaid, P. Singh, and C. Kaur, "EEG signal analysis for BCI interface: A review," in *Proc. 5th Int. Conf. Adv. Comput. Commun. Technol.*, Feb. 2015, pp. 143–147.
[35] M. Rashid, N. Sulaiman, A. P. P. A. Majeed, R. M. Musa, A. F. A. Nasir, B. S. Bari, and S. Khatun, "Current status, challenges, and possible solutions of EEG-based brain–computer interface: A comprehensive review," *Frontiers Neurorobot.*, vol. 14, Jun. 2020, Art. no. 25, doi: 10.3389/fnbot.2020.00025.
[36] S. R. Benbadis and K. Lin, "Errors in EEG interpretation and misdiagnosis of epilepsy: Which EEG patterns are overread?" *Eur. Neurol.*, vol. 59, no. 5, pp. 267–271, Feb. 2008.
[37] W. Ko, E. Jeon, S. Jeong, J. Phyo, and H.-I. Suk, "A survey on deep learning-based short/zero-calibration approaches for EEG-based brain–computer interfaces," *Frontiers Hum. Neurosci.*, vol. 15, May 2021, Art. no. 643386, doi: 10.3389/fnhum.2021.643386.
[38] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI competition 2008—Graz data set A," Inst. Knowl. Discovery, Lab. Brain-Comput. Interfaces, Graz Univ. Technol., 2008, vol. 16, pp. 1–6.
[39] R. Leeb, C. Brunner, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI competition 2008—Graz data set B," Graz Univ. Technol., Graz, Austria, 2008, pp. 1–6.
[40] F. Nijboer, N. Birbaumer, and A. Kübler, "The influence of psychological state and motivation on brain–computer interface performance in patients with amyotrophic lateral sclerosis—A longitudinal study," *Frontiers Neurosci.*, vol. 4, p. 55, Jun. 2010.
[41] Y. Wang, J. Wang, W. Wang, J. Su, and Z.-G. Hou, "Calibration-free transfer learning for EEG-based cross-subject motor imagery classification," in *Proc. IEEE 19th Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2023, pp. 1–6.
[42] M. Ahn and S. C. Jun, "Performance variation in motor imagery brain–computer interface: A brief review," *J. Neurosci. Methods*, vol. 243, pp. 103–110, Mar. 2015.
[43] A. B. Randolph, "Not all created equal: Individual-technology fit of brain–computer interfaces," in *Proc. 45th Hawaii Int. Conf. Syst. Sci.*, Jan. 2012, pp. 572–578.
[44] R. Leeb, F. Lee, C. Keinrath, R. Scherer, H. Bischof, and G. Pfurtscheller, "Brain–computer communication: Motivation, aim, and impact of exploring a virtual apartment," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 4, pp. 473–482, Dec. 2007.
[45] M. González-Franco, P. Yuan, D. Zhang, B. Hong, and S. Gao, "Motor imagery based brain–computer interface: A study of the effect of positive and negative feedback," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 6323–6326.
[46] F. Lotte, F. Larrue, and C. Mühl, "Flaws in current human training protocols for spontaneous brain–computer interfaces: Lessons learned from instructional design," *Frontiers Hum. Neurosci.*, vol. 7, Sep. 2013, Art. no. 568, doi: 10.3389/fnhum.2013.00568.
[47] D. Zhang, L. Yao, K. Chen, and J. Monaghan, "A convolutional recurrent attention model for subject-independent EEG signal analysis," *IEEE Signal Process. Lett.*, vol. 26, no. 5, pp. 715–719, May 2019.
[48] A. B. Randolph, M. M. Jackson, and S. Karmakar, "Individual characteristics and their effect on predicting mu rhythm modulation," *Int. J. Hum.-Comput. Interact.*, vol. 27, no. 1, pp. 24–37, Dec. 2010.
[49] C. Guger, G. Edlinger, W. Harkam, I. Niedermayer, and G. Pfurtscheller, "How many people are able to operate an EEG-based brain–computer interface (BCI)?" *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 145–147, Jun. 2003.
[50] W. Burde and B. Blankertz, "Is the locus of control of reinforcement a predictor of brain-computer interface performance?" in *Proc. 3rd Int. Brain-Comput. Interface Workshop Training Course*. Graz, Austria: TU Graz, 2006.
[51] A. Vuckovic and B. A. Osuagwu, "Using a motor imagery questionnaire to estimate the performance of a brain–computer interface based on object oriented motor imagery," *Clin. Neurophysiol.*, vol. 124, no. 8, pp. 1586–1595, Aug. 2013.
[52] A. Chatterjee, V. Aggarwal, A. Ramos, S. Acharya, and N. V. Thakor, "A brain–computer interface with vibrotactile biofeedback for haptic information," *J. NeuroEng. Rehabil.*, vol. 4, no. 1, pp. 1–12, Dec. 2007.
[53] M. Gomez-Rodriguez, J. Peters, J. Hill, B. Schölkopf, A. Gharabaghi, and M. Grosse-Wentrup, "Closing the sensorimotor loop: Haptic feedback facilitates decoding of motor imagery," *J. Neural Eng.*, vol. 8, no. 3, Apr. 2011, Art. no. 036005.
[54] F. Lotte, J. Faller, C. Guger, Y. Renard, G. Pfurtscheller, A. Lécuyer, and R. Leeb, "Combining BCI with virtual reality: Towards new applications and improved BCI," in *Towards Practical Brain-Computer Interfaces: Bridging the Gap From Research to Real-World Applications*, 2013, pp. 197–220.
[55] R. Jiang, L. Sun, X. Wang, and Y. Xu, "Application of transformer with auto-encoder in motor imagery EEG signals," in *Proc. 14th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Nov. 2022, pp. 1–7.
[56] Z. Wu, B. Sun, and X. Zhu, "Coupling convolution, transformer and graph embedding for motor imagery brain–computer interfaces," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Jun. 2022, pp. 404–408.
[57] C. I. Salis, A. E. Malissovas, P. A. Bizopoulos, A. T. Tzallas, P. A. Angelidis, and D. G. Tsalikakis, "Denoising simulated EEG signals: A comparative study of EMD, wavelet transform and Kalman filter," in *Proc. 13th IEEE Int. Conf. Bioinf. BioEng.*, Mali, Nov. 2013, pp. 1–4.
[58] H. Zhang, C. Wei, M. Zhao, Q. Liu, and H. Wu, "A novel convolutional neural network model to remove muscle artifacts from EEG," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1265–1269.
[59] E. Brophy, P. Redmond, A. Fleury, M. De Vos, G. Boylan, and T. Ward, "Denoising EEG signals for real-world BCI applications using GANs," *Frontiers Neuroergonom.*, vol. 2, Jan. 2022, Art. no. 805573, doi: 10.3389/fnrgo.2021.805573.
[60] T. W. Picton, P. van Roon, M. L. Armilio, P. Berg, N. Ille, and M. Scherg, "The correction of ocular artifacts: A topographic perspective," *Clin. Neurophysiol.*, vol. 111, no. 1, pp. 53–65, Jan. 2000.
[61] D. Moretti, F. Babiloni, F. Carducci, F. Cincotti, E. Remondini, P. Rossini, S. Salinari, and C. Babiloni, "Computerized processing of EEG-EOG-EMG artifacts for multi-centric studies in EEG oscillations and event-related potentials," *Int. J. Psychophysiol.*, vol. 47, no. 3, pp. 199–216, Mar. 2003.
[62] D. Craven, B. McGinley, L. Kilmartin, M. Glavin, and E. Jones, "Adaptive dictionary reconstruction for compressed sensing of ECG signals," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 3, pp. 645–654, May 2017.
[63] J. Minguillon, M. A. Lopez-Gordo, and F. Pelayo, "Trends in EEG-BCI for daily-life: Requirements for artifact removal," *Biomed. Signal Process. Control*, vol. 31, pp. 407–418, Jan. 2017.
[64] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.
[65] J. Yin, A. Liu, C. Li, R. Qian, and X. Chen, "A GAN guided parallel CNN and transformer network for EEG denoising," *IEEE J. Biomed. Health Inform.*, early access, pp. 1–12, May 23, 2023, doi: 10.1109/JBHI.2023.3277596.
[66] W. Yi, S. Qiu, K. Wang, H. Qi, L. Zhang, P. Zhou, F. He, and D. Ming, "Evaluation of EEG oscillatory patterns and cognitive process during simple and compound limb motor imagery," *PLoS One*, vol. 9, no. 12, Dec. 2014, Art. no. e114853.
[67] X. Shi, B. Li, W. Wang, Y. Qin, H. Wang, and X. Wang, "Classification algorithm for electroencephalogram-based motor imagery using hybrid neural network with spatio-temporal convolution and multi-head attention mechanism," *Neuroscience*, vol. 527, pp. 64–73, Sep. 2023.
[68] A. Hameed, R. Fourati, B. Ammar, A. Ksibi, A. S. Alluhaidan, M. B. Ayed, and H. K. Khleaf, "Temporal–spatial transformer based motor imagery classification for BCI using independent component analysis," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105359.
[69] J. Xie, J. Zhang, J. Sun, Z. Ma, L. Qin, G. Li, H. Zhou, and Y. Zhan, "A transformer-based approach combining deep learning network and spatial–temporal information for raw EEG classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2126–2136, 2022.
[70] P. Wang, P. Gong, Y. Zhou, X. Wen, and D. Zhang, "Decoding the continuous motion imagery trajectories of upper limb skeleton points for EEG-based brain–computer interface," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.
[71] J. Wang, L. Yao, and Y. Wang, "IFNet: An interactive frequency convolutional neural network for enhancing motor imagery decoding from EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1900–1911, 2023.
[72] E. Ozelbas, E. E. Tülay, and S. Ozekes, "Improving cross-subject classification performance of motor imagery signals: A data augmentation-focused deep learning framework," *Mach. Learning: Sci. Technol.*, vol. 5, no. 1, Mar. 2024, Art. no. 015021.
[73] E. Arı and E. Taçgın, "NF-EEG: A generalized CNN model for multi class EEG motor imagery classification without signal preprocessing for brain computer interfaces," *Biomed. Signal Process. Control*, vol. 92, Jun. 2024, Art. no. 106081.

[74] P. Penava and R. Buettner, "A novel small-data based approach for decoding yes/no-decisions of locked-in patients using generative adversarial networks," *IEEE Access*, vol. 11, pp. 118849–118864, 2023.
[75] A. G. Habashi, A. M. Azab, S. Eldawlatly, and G. M. Aly, "Motor imagery classification enhancement using generative adversarial networks for EEG spectrum image generation," in *Proc. IEEE 36th Int. Symp. Computer-Based Med. Syst. (CBMS)*, Jun. 2023, pp. 354–359.
[76] H. Zhang, H. Ji, J. Yu, J. Li, L. Jin, L. Liu, Z. Bai, and C. Ye, "Subject-independent EEG classification based on a hybrid neural network," *Frontiers Neurosci.*, vol. 17, Jun. 2023, Art. no. 1124089, doi: 10.3389/fnins.2023.1124089.
[77] S. Liang, S. Kuang, D. Wang, Z. Yuan, H. Zhang, and L. Sun, "An auxiliary synthesis framework for enhancing EEG-based classification with limited data," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2120–2131, 2023.
[78] K. Yin, B.-H. Lee, B.-H. Kwon, and J.-H. Cho, "Target-centered subject transfer framework for EEG data augmentation," in *Proc. 11th Int. Winter Conf. Brain-Computer Interface (BCI)*, Feb. 2023, pp. 1–4.
[79] K. Yin, E. Y. Lim, and S.-W. Lee, "GITGAN: Generative inter-subject transfer for EEG motor imagery analysis," *Pattern Recognit.*, vol. 146, Feb. 2024, Art. no. 110015.
[80] M. Miao, Z. Yang, H. Zeng, W. Zhang, B. Xu, and W. Hu, "Explainable cross-task adaptive transfer learning for motor imagery EEG classification," *J. Neural Eng.*, vol. 20, no. 6, Dec. 2023, Art. no. 066021.
[81] M. S. Aldayel, M. Ykhlef, and A. N. Al-Nafjan, "Electroencephalogram-based preference prediction using deep transfer learning," *IEEE Access*, vol. 8, pp. 176818–176829, 2020.
[82] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. Learn.* New York, NY, USA: Association for Computing Machinery, Jun. 2007, pp. 759–766.
[83] H. Chang, J. Han, C. Zhong, A. M. Snijders, and J.-H. Mao, "Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1182–1194, May 2018.
[84] C. M. Yilmaz, B. H. Yilmaz, and C. Kose, "A novel signal-to-image conversion approach with ensembles of pretrained CNNs for motor imagery EEG signals," in *Proc. 10th Int. Conf. Electr. Eng., Comput. Sci. Informat. (EECSI)*, Sep. 2023, pp. 49–53.
[85] A. M. Azab, H. Ahmadi, L. Mihaylova, and M. Arvaneh, "Dynamic time warping-based transfer learning for improving common spatial patterns in brain–computer interface," *J. Neural Eng.*, vol. 17, no. 1, Feb. 2020, Art. no. 016061.
[86] Y. Song, X. Jia, L. Yang, and L. Xie, "Transformer-based spatial–temporal feature learning for EEG decoding," 2021, *arXiv:2106.11170*.

AIGERIM KEUTAYEVA received the B.S. degree in robotics and mechatronics and the M.S. degree in robotics from Nazarbayev University, Astana, Kazakhstan, in 2021 and 2023, respectively. From 2019 to 2023, she was a Research Assistant with the School of Engineering and Digital Sciences (SEDS), Nazarbayev University, and a member with the Young Researchers Alliance, Astana. Since January 2024, she has been a Research Assistant with the Institute of Smart Systems and Artificial Intelligence (ISSAI). Her current research interests include machine learning, brain–computer interfaces, signal processing, computer vision, and digital twin.

BERDAKH ABIBULLAEV (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in electronic engineering from Yeungnam University, South Korea, in 2006 and 2010, respectively.

He held research scientist positions with Daegu Gyeongbuk Institute of Science and Technology, from 2010 to 2013, and Samsung Medical Center, Seoul, South Korea, from 2013 to 2014. In 2014, he received the National Institute of Health Postdoctoral Research Fellowship II to join a multi-institutional research project between the University of Houston Brain-Machine Interface Systems Team and Texas Medical Center in developing neural interfaces for rehabilitation in post-stroke patients. He is currently an Associate Professor with the Robotics Department, Nazarbayev University, Kazakhstan. His current research interests include signal processing and machine learning algorithms for the inference problems of brain–computer interfaces. He is an Associate Editor of IEEE Access and *PeerJ Computer Science*.