IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 October 2023, accepted 27 October 2023, date of publication 2 November 2023, date of current version 16 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3329678

TOPICAL REVIEW

Deep Learning in EEG-Based BCIs: A Comprehensive Review of Transformer Models, Advantages, Challenges, and Applications

BERDAKH ABIBULLAEV[1], (Senior Member, IEEE), AIGERIM KEUTAYEVA[1], AND AMIN ZOLLANVARI[2], (Senior Member, IEEE)

[1]Department of Robotics Engineering, Nazarbayev University, 010000 Astana, Kazakhstan
[2]Department of Electrical and Computer Engineering, Nazarbayev University, 010000 Astana, Kazakhstan

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

This work was supported by Nazarbayev University under the Faculty Development Competitive Research Grant Program (FDCRGP) under Grant 021220FD2051.

ABSTRACT Brain-computer interfaces (BCIs) have undergone significant advancements in recent years. The integration of deep learning techniques, specifically transformers, has shown promising development in research and application domains. Transformers, which were originally designed for natural language processing, have now made notable inroads into BCIs, offering a unique self-attention mechanism that adeptly handles the temporal dynamics of brain signals. This comprehensive survey delves into the application of transformers in BCIs, providing readers with a lucid understanding of their foundational principles, inherent advantages, potential challenges, and diverse applications. In addition to discussing the benefits of transformers, we also address their limitations, such as computational overhead, interpretability concerns, and the data-intensive nature of these models, providing a well-rounded analysis. Furthermore, the paper sheds light on the myriad of BCI applications that have benefited from the incorporation of transformers. These applications span from motor imagery decoding, emotion recognition, and sleep stage analysis to novel ventures such as speech reconstruction. This review serves as a holistic guide for researchers and practitioners, offering a panoramic view of the transformative potential of transformers in the BCI landscape. With the inclusion of examples and references, readers will gain a deeper understanding of the topic and its significance in the field.

INDEX TERMS Deep learning, brain–computer interfaces, review, transformer architecture, EEG, emotion recognition, seizure detection, self-attention mechanism, neural networks, motor imagery, sleep stage analysis, transformer models, CNN, BCI.

I. INTRODUCTION

Brain-computer interfaces (BCI) enable communication between the human brain and external devices without the intervention of peripheral nerves and muscles. They provide a direct channel to translate mental processes into tangible actions, fundamentally reshaping how humans interact with technology. The concept of a BCI dates back to the early 1970s [1], [2], but it wasn't until the advent of sophisticated signal processing techniques and computational power in the late 20th century that significant progress was made [3], [4], [5], [6], [7], [8], [9], [10]. Early BCIs were primarily experimental, used in controlled laboratory settings, and often involved invasive procedures where electrodes were implanted directly into the brain [11], [12].

BCIs are generally categorized into three types, each differing in the degree to which they interface with the brain. Invasive BCIs necessitate the surgical implantation of electrodes directly into the brain [13], [14]. Although this type provides high-resolution neural signals, due to the inherent risks of surgery and the potential formation of scar tissue, it is seldom used in non-medical applications.

The associate editor coordinating the review of this manuscript and approving it for publication was Adnan Kavak.

Most non-invasive BCIs rely on Electroencephalography (EEG) to obtain real-time data on neural activity by placing electrodes on the scalp [15]. This technique enables immediate interfacing between the brain and external devices, which is invaluable in BCI applications. However, the spatial resolution is generally lower due to its non-invasive nature, as the signals have to pass through the skull and scalp. On the other hand, partially invasive BCIs are based on implanted electrodes within the skull, while the remaining electrodes are outside the brain [16]. These BCIs balance signal quality and medical risk, serving as a compromise between the invasive and non-invasive types. Nevertheless, these devices require neural surgery, worth considering the potential benefits they offer. There are also other types of BCIs that use different techniques, such as magnetoencephalography (MEG) [17], functional magnetic resonance imaging (fMRI) [18], and functional near-infrared spectroscopy (fNIRS) [19]. These methods allow for more precise measurements of brain activity, but they are often more expensive and less accessible than EEG-based BCIs.

BCIs find applicability in diverse domains such as medical rehabilitation [20], [21], entertainment [22], [23], and communication for those with motor or speech constraints [24]. For patients with paralysis or limb amputation, BCIs can help in controlling prosthetic limbs and wheelchairs or even restore speech [25], [26]. Gamers can use BCIs for a more immersive experience, where mental states or focus can control game elements. Moreover, BCIs are particularly useful for individuals with severe motor or verbal limitations, such as those with advanced ALS [27]. Nevertheless, BCI technology comes with challenges, including poor signal-to-noise ratio, user training, and hardware limitations. Accurately decoding brain signals can be difficult due to noise, and huge variability in data, especially in non-invasive BCIs [28], [29], [30]. Many BCIs require users to undergo training to use the system effectively [31]. Miniaturization and improvement in electrode technology are ongoing challenges [32], [33], [34].

A. EMERGENCE AND RELEVANCE OF DEEP LEARNING IN BCIS

BCIs have evolved significantly with the integration of advanced computational techniques. A prominent factor in this evolution is deep learning—a machine learning paradigm that utilizes multi-layered artificial neural networks to analyze data [35], [36]. Given its ability to handle extensive data sets and decode complex patterns, deep learning has become increasingly relevant to BCI applications [8], [37], [38], [39]. Deep learning, inherently inspired by the human brain's structure, leverages interconnected nodes in multiple layers to automatically learn and extract features from raw data. This capability becomes especially advantageous in the context of BCIs. Traditionally, BCIs relied on manual feature extraction and classical machine learning methods, which often required domain-specific expertise and were constrained by the limited capacity to process high-dimensional data [6], [15]. However, with deep learning, automated feature extraction from raw neural data became feasible, minimizing the need for manual intervention and domain-specific preprocessing [40], [41], showcasing enhanced performance in tasks such as motor imagery classification [42], [43]. Beyond enhancing accuracy in standard tasks, deep learning has expanded BCIs' scope. BCIs that utilize deep learning algorithms have improved the reaction time and accuracy of prosthetics and exoskeletons, particularly for individuals with mobility challenges [44]. Additionally, these systems can offer invaluable insights into a person's cognitive state during therapeutic scenarios [45]. The application of deep learning has also made BCIs more versatile and user-friendly, broadening their applicability to fields such as gaming and mindfulness practices [46], [47].

Nevertheless, deep learning models require significant amounts of data, which is problematic as brain data is often limited. The ''black-box'' nature of these models can also impede interpretation, raising concerns about the use of brain data in decision-making [48]. Additionally, deep learning models may overfit due to the high dimensionality of BCI data and the potential scarcity of samples.

B. INTRODUCTION TO TRANSFORMER MODELS

The transformer architecture, introduced in the groundbreaking paper ''Attention Is All You Need'' by Vaswani et al. in 2017 [49], has redefined the landscape of natural language processing. It has led to the development of models such as BERT, GPT, and many others, pushing the boundaries of machine learning tasks across various domains beyond just NLP [50], [51], [52], [53], [54].

Transformers are known for several unique and innovative components, most notably the self-attention mechanism, which allows the model to weigh the importance of different parts of the input data relative to each other, and the positional encoding, which gives the model a sense of order and ensures that it can account for the position of data in a sequence. Transformers also use multi-head attention, allowing the model to focus on different parts of the input simultaneously [55]. The benefits of transformer architectures include parallelization, scalability, and flexibility. Transformers process all data points in parallel, leading to faster training times. They are highly scalable, with large models capable of capturing intricate patterns in massive datasets, leading to state-of-the-art performance in various tasks. Although initially designed for NLP tasks, transformers have shown great potential in other domains, such as vision and BCIs [56], [57], [58], [59], [60]. The success of the initial transformer model led to the development of numerous variants tailored for different tasks, such as BERT, GPT, and Vision Transformers [52], [53], [54], [58], [61], [62], [63].

While the impact of the transformer architecture on machine learning and BCI applications are evident, it is also

important to understand its inherent challenges. Primarily, the intensive computational requirements of transformers can pose a barrier to individual researchers or smaller teams with limited computational resources. Additionally, the risk of overfitting, especially when working with the relatively smaller datasets frequently encountered in the BCI domain, is a pertinent concern. Nevertheless, the versatility of the transformer model and its capacity to address diverse problems underscore its significance in research. For optimal application in BCI decoding tasks, a clear understanding of both the strengths and limitations of transformer architecture is essential. With this balanced perspective, the academic community can harness the full capabilities of transformers, ensuring continued progress in machine learning research.

This work provides a comprehensive overview of the relevance and emergence of transformer models in non-invasive EEG-based BCI systems. We discuss several technical and practical considerations related to BCIs and deep learning, including signal accuracy, invasiveness, hardware limitations, and interpretation complexity. Overall, this study provides an overview of the current state of research into BCIs and deep learning, highlighting their potential and the challenges that must be overcome to realize their full potential.

The paper is structured into nine cohesive sections. Section I sets the stage with an overview of BCIs, emphasizing the role of deep learning architectures. In Section II, we delve into the foundational concepts of BCIs, tracing their historical evolution, identifying major applications, and emphasizing the pivotal role of deep learning in modern BCI research. Section III transitions to a deep dive into the Transformer Architecture, including mathematical formulations. This section elucidates its evolution from RNNs and LSTMs, shedding light on its distinct structure, the innovative self-attention mechanism, and its intrinsic benefits. Section IV reviews the real-world applications of transformers, from the decoding of Motor Imagery EEG to Emotion Recognition, underscoring their relevance and highlighting transformer model applications. The benefits of leveraging transformers in BCIs are systematically laid out in Section V. We discuss their properties for handling EEG's temporal nuances, capturing long-range dependencies, and ensuring scalability. Section VI critically assesses the challenges posed by transformer models, from their computational intensity and requirement for huge data to the interpretability issues in BCI contexts.

In Section VII, we explore potential avenues for future research in the field of BCIs. These include developing efficient transformer variants specifically designed for BCIs, integrating BCIs with other modalities using transformers, and investigating ways to achieve real-time BCI processing using these models. Sections VIII and IX summarize the key findings of the studies surveyed and distill the main insights gained from this exploration. Incorporating Transformer models into EEG-based BCIs is a complex process that requires careful attention to several key aspects to ensure effective and reliable system performance. To help with this, the Appendix provides a crucial checklist that outlines key questions to consider.

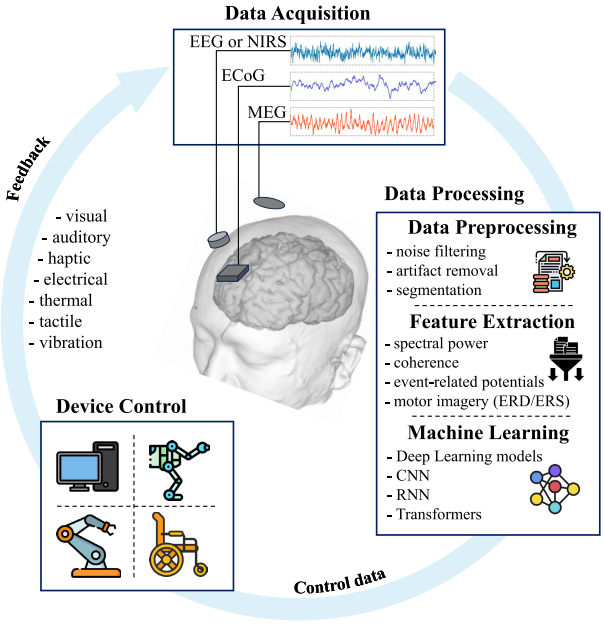

FIGURE 1. Illustration of the Brain-Computer Interface (BCI) Framework. The components represent the entirety of the BCI Framework, detailing the critical stages: Data Acquisition, Data Preprocessing, Feature Extraction, Machine Learning, and Feedback. Each stage is pivotal in determining the interpretive and responsive functionalities of BCI systems.

II. BCI FUNDAMENTALS

The brain-computer interface, alternatively recognized as a brain-machine interface or direct neural interface, represents a communication channel established directly between the human brain and external machinery. The premise of BCI technology centers on harnessing and interpreting neural signals, subsequently converting them into actionable commands that can drive external devices, ranging from computers to prosthetic devices [13], [64], [65].

Historical accounts pinpoint the genesis of BCI research to the early 1970s, marking the period when pioneering experiments involving animals were undertaken [66]. Progressing towards the close of the 20th century, the inaugural human-centric BCI experiments emerged, predominantly catering to medical interventions, especially for individuals grappling with neuromuscular impediments [67], [68]. Spanning several decades, the synergy of neuroscience, engineering, and computational disciplines has dynamically driven the development and democratization of BCI modalities [4].

BCIs are complex systems with multiple critical components, as detailed by Gerven et al. [69]. Primarily, data acquisition involves capturing neural signals from the brain using a variety of techniques. Invasive techniques might utilize micro-electrode arrays, positioned directly on the brain's surface [70], while non-invasive methods could employ EEG, where electrodes are placed on the scalp [71].

However, acquired neural data frequently contains noise and necessitates careful preprocessing. Essential preprocessing tasks include noise filtering, artifact removal, and segmentation [72]. Noise filtering seeks to eliminate undesired signals that might corrupt the true neural signals [73]. In contrast, artifact removal targets extraneous signals not originating from the brain, for instance, those originating from muscle contractions or eye movements [74]. Segmentation entails subdividing the continuous data into segments for further processing [75]. Subsequent to these preprocessing steps, the data undergoes further analysis to extract meaningful patterns or features, indicative of underlying neural activities. In EEG data, these patterns may include spectral power, coherence, event-related potentials, and motor imagery patterns such as event-related desynchronization (ERD) and event-related synchronization (ERS) [76]. Spectral power reflects the strength of the neural signals in different frequency bands [77]. Coherence indicates the degree of synchronization between various brain regions [78]. Event-related potentials represent the brain's response to a specific stimulus or task [79]. Meanwhile, ERD patterns indicate a decrease in power, usually related to motor preparation or movement initiation [80], whereas ERS patterns denote an increase in power, typically associated with motor termination or post-movement processes [81]. These patterns are then used in subsequent analysis within predictive models to classify them into different classes, such as motor imagery classification.

Machine learning algorithms play a crucial role in analyzing brain data [6] and decoding and interpreting specific brain states or intentions, enhancing the precision and effectiveness of BCI systems [38]. Once a mental intention is decoded, the BCI translates it into a specific action for the corresponding external device, with the user receiving real-time feedback. This enables an interactive loop, enhancing the usability of BCIs.

Figure 1 depicts the various components integral to the BCI framework, including data acquisition, data preprocessing, feature extraction, machine learning, and feedback. Each of these components is vital, collectively ensuring the efficacy and accuracy of the BCI's operations. While the BCI ecosystem spans from the initial data acquisition to the final real-time feedback, significant research gravitates towards the interpretation and prediction of brain data [35]. This concentrated focus underscores the critical importance of data analysis in harnessing the full capabilities of BCIs, enabling their effective deployment across a spectrum of applications [8].

A. MAIN APPLICATIONS

BCIs are useful systems that have had a significant impact in various domains. In the field of medical rehabilitation, they can provide a new level of independence for patients who are dealing with paralysis, neuromuscular disorders, or limb amputations. BCIs allow these individuals to control prosthetic limbs or wheelchairs in real-time [82]. This advancement has a profound effect on their quality of life, giving them a restored sense of autonomy [20], [21], [83]. Furthermore, BCIs have the potential to restore lost sensory feedback, especially for those who have lost the sense of touch or proprioception [84], [85], [86]. This allows patients to regain some of their lost motor skills, leading to a more independent daily life [31].

Simultaneously, in the domain of communication, BCIs can offer a practical means of communication [87], [88]. People with locked-in syndrome, advanced ALS, or other similar conditions are often unable to move or speak, making communication very challenging [82]. To date, most BCIs have been designed specifically for these individuals, to support them to communicate by translating their mental activities into commands. These systems can be used to type messages on a computer screen or even control external devices, such as wheelchairs or prosthetic limbs [89]. In summary, the development of BCI systems has helped to improve the quality of life for many individuals to communicate effectively.

A relatively novel application space for BCIs is emotion recognition. This area primarily involves the analysis of neural signals to understand and classify human emotional states [90], [91], [92]. By identifying distinct patterns in brain activity associated with various emotions, BCIs can offer a novel approach to detecting and analyzing these states. Potential applications include improved mental health diagnostics, where accurate emotion detection could provide clinicians with valuable insights. Moreover, in the domain of media, real-time emotion feedback can guide content adaptation, leading to user-specific experiences. Similarly, adaptive environments in educational or occupational settings can be developed based on emotional feedback, potentially enhancing learning and productivity. As research progresses, it is expected that the integration of BCIs in emotion recognition will open new avenues for further exploration and application.

Additionally, the gaming industry is actively investigating the use of BCI to create a more immersive gaming experience [93], [94]. BCIs allow players to navigate virtual environments using their thoughts and feelings. This means gamers might soon control game characters and execute actions just by thinking. Such an approach can increase the feeling of being ''in'' the game, making gameplay even more enjoyable. BCIs also offer a chance for people with disabilities, especially those who can not use regular game controllers or keyboards, to engage in gaming [22], [95], [96]. The integration of BCI in gaming suggests a future where games become more interactive and inclusive.

B. EMERGING APPLICATIONS

Beyond traditional applications, BCIs have expanded into cognitive enhancement, sleep analysis, seizure detection, and speech reconstruction.

Researchers are exploring the potential of BCIs in cognitive enhancement, specifically to improve concentration [97], [98]. Neurofeedback is one such technique that offers individuals real-time feedback on their neural activities, facilitating the enhancement of particular cognitive functions. This BCI application shows promise for individuals with attention disorders. As advancements in the field continue, neurofeedback and associated techniques might play a pivotal role in advancing our comprehension of cognitive enhancement and overall brain functionality [90], [91], [92], [99], [100], [101].

Sleep research can also benefit from BCI methods, particularly in the classification of sleep stages—a pivotal aspect of diagnosing sleep disorders [102]. Using EEG data, machine learning algorithms can discern sleep stages such as Wake, REM, and the non-REM stages (N1, N2, N3) with pronounced accuracy [103]. By decoding the EEG signatures of various sleep phases, machine learning methods used in BCIs can present diagnostic insights and therapeutic interventions. Moreover, combining BCIs with wearable technology holds promise for non-invasive, real-time sleep monitoring, warranting further investigation.

Another crucial application area is epileptic seizure detection. Given the characteristic irregular brain activities during seizures, BCIs, equipped with EEG monitoring, emerge as important tools for capturing these anomalies [104]. BCI-driven algorithms can pinpoint these atypical patterns, facilitating early seizure detection and intervention [105]. The prospective ability of BCIs to predict seizures before their onset can enhance the management of epilepsy, providing patients with preemptive alerts. The conjunction of BCIs with wearables emphasizes its importance in neurology and biomedical engineering research.

In speech reconstruction, BCIs are employed to decode neural activity related to auditory processes and produce intended speech [106]. This is particularly valuable for individuals who cannot communicate verbally due to specific conditions. By integrating machine learning models with BCIs, researchers are working to convert EEG-based neural patterns into understandable speech [107]. Although challenges, such as EEG noise and variations in individual neural patterns, persist, initial studies indicate the potential of BCIs in this area [108].

C. BCI CHALLENGES

Signal Accuracy: Acquiring accurate and consistent neural signals is fundamental for both neuroscience and BCI research. However, achieving consistent and precise recordings is challenging, particularly due to external interferences such as electronic devices [109]. Moreover, inherent variability in EEG data, attributed to individual brain differences [71], as well as inconsistencies across trials due to factors such as attention fluctuations [110], poses significant analytical challenges.

Factors such as inter-subject differences in brain structures, intra-subject variability, inaccuracies in electrode placement [111], and device-specific biases [112] are notable contributors to data variability. External environmental conditions can further complicate the data collection process. However, by employing rigorous methodologies and leveraging advanced techniques, researchers can effectively mitigate these challenges, ensuring the integrity of neural data and thereby advancing the neuroscience domain.

Invasiveness: The invasiveness of BCIs poses one of the primary challenges in the realm of neural interfacing. While non-invasive methods such as EEG-based systems are widely adopted due to their relative safety and ease of application, they often compromise on signal quality and precision. In contrast, invasive methods, which involve the direct implantation of electrodes into or on the surface of the brain, can yield higher signal fidelity and specificity [113], [114]. However, these methods introduce increased medical risks, including potential complications from surgery and long-term biocompatibility concerns. The ethical considerations surrounding invasive procedures, especially in non-medical or elective contexts, further compound the challenges. Thus, determining the optimal balance between invasiveness and functionality remains a pivotal challenge in advancing BCI technology.

Hardware Limitations: While current BCI systems have made significant progress in helping individuals control devices with their brain activities, there is still much room for improvement. One area that particularly stands out is miniaturization [34], [115]. By reducing the size of BCI systems, they can become more portable and less obtrusive, allowing users to integrate them into their daily lives more easily. These improvements can lead to greater adoption and utilization of BCI technology in the future, ultimately benefiting individuals who rely on these systems for communication and independence.

Interpretation Complexity: The brain, which is the central organ of the nervous system, is incredibly complex in its structure and function. It is responsible for receiving and interpreting signals from various parts of the body, and it performs this task with remarkable efficiency. In fact, the brain works in tandem with the spinal cord to form the central nervous system, which controls all the functions of the body, including movement, sensation, and cognition [116], [117]. To accurately and consistently interpret the signals generated by the brain, scientists and researchers have developed sophisticated algorithms and models. These models are capable of processing vast amounts of data and identifying complex patterns that are simply impossible for humans to discern. However, the development of such models requires significant computational power, which can be a challenging task [118], [119]. Despite the complexity of the brain and the challenges associated with interpreting its signals, researchers are committed to unlocking its mysteries by developing new algorithms and models. This will help us better understand how the brain works and how we can use this knowledge to improve our lives and the world around us.

BCIs stand at the intersection of neuroscience and technology and hold the promise of fundamentally reshaping various sectors, particularly healthcare. As BCI technology continues to evolve, it is essential to approach its applications with a balance of optimism and caution, addressing challenges head-on.

D. DEEP LEARNING IN BCIS

Deep Learning is a subset of machine learning, which, in turn, falls under the broader category of artificial intelligence [15], [120], [121]. It's characterized by the use of deep neural networks – layered computational structures inspired by the biological neurons in the human brain. These networks, comprising of multiple layers, enable the algorithm to learn representations of data through multiple levels of abstraction automatically. Classic examples include Convolutional Neural Networks (CNNs) used in image recognition or recurrent neural networks used in sequence prediction tasks [122].

Where traditional machine learning techniques might require feature engineering – a manual process where the most relevant features of data are selected for model training – deep learning models are known for their ability to automatically extract and learn features directly from raw data. This makes them particularly powerful for tasks involving large and complex datasets, such as images, speech, and, notably for this context, brain signals [8], [37], [39], [40], [123], [124], [125], [126], [127].

1) IMPORTANCE AND IMPACT ON OF TRANSFORMERS IN BCI RESEARCH

- **Automated Feature Extraction**: The neural data from the brain is highly complex and multi-dimensional. Deep learning, with its ability for automatic feature extraction, has made it possible to interpret raw brain signals without the need for extensive manual feature engineering. This reduces the potential for human bias and error and simplifies the process of model development.
- **Enhanced Accuracy**: BCIs demand high accuracy to be practically useful, especially in medical or assistive contexts. Deep learning models, given their capability to handle vast datasets and complex structures, have consistently demonstrated superior accuracy in decoding brain signals compared to traditional methods.
- **Real-time Processing**: With advancements in hardware, such as GPUs, deep learning models can process and interpret brain signals in real time. This is crucial for BCIs where fast command output is key, such as in prosthetic limb control or communication aids.
- **Scalability**: Deep learning models are scalable. As more data becomes available – from a wider variety of subjects and conditions – these models can continue learning and refining their interpretations, improving the robustness and versatility of BCIs.
- **Addressing Subject Variability**: One of the core challenges in BCI is the variability of signals between different individuals. Deep learning, with architectures such as convolutional layers or attention mechanisms, has shown potential in capturing these latent temporal patterns, paving the way for subject-independent BCIs.

Deep learning has played an important role in advancing BCI research, enabling the automatic decoding of intricate neural patterns. This has resulted in the creation of more dependable, effective, and robust BCIs. With the continuous improvement of computational techniques and the expansion of BCI datasets, the integration of deep learning and BCIs has the potential to produce even more promising developments in the coming years.

III. TRANSFORMER ARCHITECTURE

Historically, sequence data processing in neural networks began with Recurrent Neural Networks (RNNs). These architectures were equipped with an inherent 'memory' mechanism, retaining hidden states across sequence steps. Yet, their efficacy weakened with longer sequences due to issues such as the vanishing and exploding gradient problems, which impeded successful training. Addressing these limitations, Long Short-Term Memory Networks (LSTMs) emerged as an advanced form of RNNs. By employing gated cells, LSTMs adeptly controlled information flow, enhancing the network's capacity to recognize long-term dependencies in sequential data [122]. Nevertheless, inherent to their design, LSTMs processed sequences serially, constraining parallel processing possibilities across sequence elements.

A paradigm shift occurred with the introduction of the Transformer architecture, as presented by Vaswani et al. in 2017 [49]. Eliminating the recurrent structure, Transformers capitalized on parallel processing capabilities, processing entire sequences concurrently. This adjustment significantly accelerated training phases, especially on contemporary hardware optimized for parallel computation. Central to Transformers is the attention mechanism, particularly self-attention, which intelligently assigns different weights to sequence elements based on their task-specific relevance. This mechanism enables the model to selectively emphasize parts of the input, optimizing comprehension and representation of data sequences.

A. TRANSFORMER'S ARCHITECTURE FOR EEG CLASSIFICATION

This section presents the ''standard'' approach for utilizing the Transformer encoder to classify EEG patterns for BCIs.

1) INPUT STANDARDIZATION AND POSITIONAL ENCODING

Let the set of pairs $D_{\text{train}} = \{(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_n, y_n)\}$ denote n trials of EEG recordings where y_i is the scaler class variable with L possible labels (e.g., RH and LH imagery in a binary classification) and $\mathbf{X}_i \in \mathbb{R}^{c \times p}$ is the collection of EEG observations in the i^{th} trial over c channels and p time points;

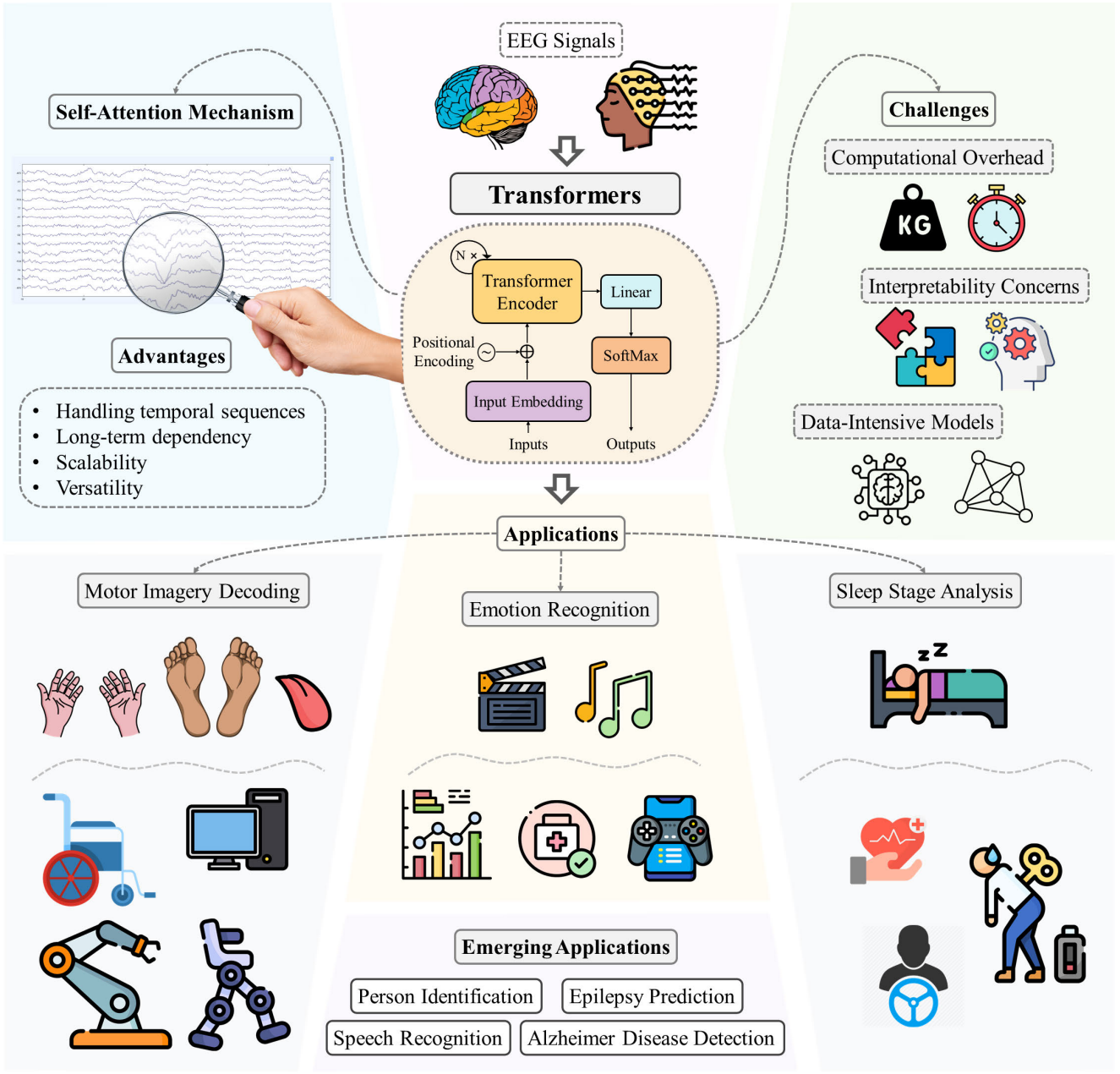

FIGURE 2. Visual summary underscores the transformative role of transformers models in Brain-Computer Interfaces (BCI), highlighting key application domains—including motor imagery decoding, emotion recognition, sleep stage analysis, as well as the emerging applications—and encapsulating both the advantages and inherent challenges in the BCI landscape.

that is to say,

$$\mathbf{X}_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{ic}]^T, \quad i = 1, \ldots, n, \tag{1}$$

with $\mathbf{x}_{ij} = [x_{ij1}, \ldots, x_{ijp}]^T \in \mathbb{R}^{p\times 1}, j = 1, \ldots, c$, where $x_{ijk}, k = 1, \ldots, p$ denotes the k^{th} element of vector $\mathbf{x}_{ij}$, and T denotes the transpose operator. The goal is to use D_{train} and train a classifier $\psi : \mathbb{R}^{c\times p} \rightarrow \{0, 1, \ldots, L-1\}$ that maps a given $\mathbf{X}$ to a possible value of the class variable.

It is common to apply standardization for each channel to make the sensory data across all channels comparable (see [128], [129], [130]). In this regard, each $\mathbf{X}_i$ is converted to $\hat{\mathbf{X}}_i$ where

$$\hat{\mathbf{X}}_i = [\hat{\mathbf{x}}_{i1}, \hat{\mathbf{x}}_{i2}, \ldots, \hat{\mathbf{x}}_{ic}]^T, \quad i = 1, \ldots, n, \tag{2}$$

and where $\hat{\mathbf{x}}_{ij} = [\hat{x}_{ij1}, \ldots, \hat{x}_{ijp}]^T$ such that

$$\hat{x}_{ijk} = \frac{x_{ijk} - m_{ij}}{s_{ij}}, \tag{3}$$

with m_{ij} and s_{ij} being the sample mean and sample standard deviation of vector $\mathbf{x}_{ij}$ given by

$$m_{ij} = \frac{1}{p}\sum_{k=1}^{p} x_{ijk}, \tag{4}$$

$$s_{ij} = \sqrt{\frac{1}{p}\sum_{k=1}^{p}(x_{ijk} - m_{ij})^2}, \tag{5}$$

respectively.

In order for the Transformer to make use of EEG recording orders, it is common to encode some information about the position of sequence elements in its input [49]. This *positional encoding* is generally realized by adding each $\hat{\mathbf{X}}_i$ to a matrix $\mathbf{P} \in \mathbb{R}^{c\times p}$ that is defined based on trigonometric functions with different frequencies for each channel [49]. As a result, we obtain

$$\tilde{\mathbf{X}}_i = \hat{\mathbf{X}}_i + \mathbf{P}, \quad i = 1, \ldots, n, \tag{6}$$

where the element on row (channel) $j = 1, \ldots, c$, and column (time index) $k = 1, \ldots, p$, of $\mathbf{P}$, denoted p_{jk} is given by

$$p_{jk} = \begin{cases} \sin\left(k/10000^{j/c}\right), & \text{for even } j, \\ \cos\left(k/10000^{(j-1)/c}\right), & \text{for odd } j. \end{cases} \tag{7}$$

2) SELF-ATTENTION MECHANISMS: CAPTURING CONTEXTS FOR EEG CLASSIFICATION

Capturing *contexts* is the essential concept that makes attention mechanism a promising operation for EEG classification. A context is simply another representation of an element of the input sequence (here, one column of each $\tilde{\mathbf{X}}_i$) based on its *compatbility* with other elements within the sequence. The most widely used attention operation for EEG classification is *scaled dot-product self-attention*, denoted $\mathrm{SA}^d_{\mathbf{V},\mathbf{K},\mathbf{Q}}(\tilde{\mathbf{X}}_i) : \mathbb{R}^{c\times p} \rightarrow \mathbb{R}^{d\times p}$, which was initially proposed and used for translation tasks [49]. In particular,

$$\mathrm{SA}^d_{\mathbf{V},\mathbf{K},\mathbf{Q}}(\tilde{\mathbf{X}}_i) = \mathbf{V}\tilde{\mathbf{X}}_i \times \mathrm{softmax}\left(\frac{\tilde{\mathbf{X}}_i^T\mathbf{K}^T\mathbf{Q}\tilde{\mathbf{X}}_i}{\sqrt{q}}\right), \tag{8}$$

where $\mathbf{V} \in \mathbb{R}^{d\times c}$, $\mathbf{K} \in \mathbb{R}^{q\times c}$, $\mathbf{Q} \in \mathbb{R}^{q\times c}$ are projection matrices that are learned in the training process, q is known as attention dimensionality, and d, which is generally a tuning parameter, denotes the dimensionality of the columns of the output matrix (*context vectors*). We use superscript d in $\mathrm{SA}^d_{\mathbf{V},\mathbf{K},\mathbf{Q}}(\tilde{\mathbf{X}}_i)$ to highlight the dimensionality of context vectors.

3) MULTI-HEAD SELF-ATTENTION

Rather than a single self-attention operation, it is generally beneficial to apply multiple self-attentions in parallel. Using this operation, we view the compatibility of sequence elements using different learned projections. In this context, it is also common to refer to the output matrix of each self-attention as a *head*. In particular, the multi-head self-attention, denoted $\mathrm{MHSA}(\tilde{\mathbf{X}}_i) : \mathbb{R}^{c\times p} \rightarrow \mathbb{R}^{d_h\times p}$, is defined as

$$\mathrm{MHSA}^{d_h}(\tilde{\mathbf{X}}_i) = \mathbf{W}[\mathrm{SA}^d_{\mathbf{V}_1,\mathbf{K}_1,\mathbf{Q}_1}(\tilde{\mathbf{X}}_i)^T, \ldots, \mathrm{SA}^d_{\mathbf{V}_m,\mathbf{K}_m,\mathbf{Q}_m}(\tilde{\mathbf{X}}_i)^T]^T \tag{9}$$

where $\mathbf{W} \in \mathbb{R}^{d_h\times md}$ is another learnable projection matrix, m is the number of self-attentions used in (9), which is also known as the number of heads, and d_h is the dimensionality of columns in the output of $\mathrm{MHSA}^{d_h}(\tilde{\mathbf{X}}_i)$ operation.

4) IDENTITY SKIP-CONNECTION AND LAYER NORMALIZATION

To ensure the stability and efficacy of the training process, especially with the complex nature of EEG data, the Transformer encoder utilizes identity skip-connections [68] followed by layer normalization [131]. Here, we define these operations. Let $\mathrm{SKP}\big(\mathrm{LAY}(\mathbf{Y})\big) : \mathbb{R}^{a\times b} \rightarrow \mathbb{R}^{a\times b}$ denote the identity skip-connection around a layer $\mathrm{LAY}(\mathbf{Y})$ (an operation) that operates on an input $\mathbf{Y} \in \mathbb{R}^{a\times b}$ to produce an output of *the same size* as the input. Then

$$\mathrm{SKP}\big(\mathrm{LAY}(\mathbf{Y})\big) = \mathbf{Y} + \mathrm{LAY}(\mathbf{Y}). \tag{10}$$

That is to say, we simply add the output of $\mathrm{LAY}(\mathbf{Y})$ to its input. Furthemore, let $\mathrm{LN}(\mathbf{Y}) : \mathbb{R}^{a\times b} \rightarrow \mathbb{R}^{a\times b}$ denote the layer normalization applied to an $(a > 1) \times b$ matrix $\mathbf{Y}$ with elements $y_{jk}, j = 1, \ldots, a, k = 1, \ldots, b$ where each row records measurements for a "features" (here, channel). Then, $\mathrm{LN}(\mathbf{Y})$ produces $\mathbf{Y}$, which is a matrix of the same size as $\mathbf{Y}$, with elements y_{jk} where

$$y_{jk} = \frac{y_{jk} - m_k}{s_k}, \tag{11}$$

and where

$$m_k = \frac{1}{a}\sum_{j=1}^{a} y_{jk}, \tag{12}$$

$$s_k = \sqrt{\frac{1}{a}\sum_{j=1}^{a}(y_{jk} - m_k)^2}. \tag{13}$$

In other words, $\mathbf{Y}$ is a type of standardization where the sample mean and sample standard deviation are computed for a column of $\mathbf{Y}$ (in the EEG context, means for each time point in the sequence) over all features. One place that these operations are used in the transformer encoder is to produce $\mathbf{X}_i$ as follows:

$$\mathbf{X}_i = \mathrm{LN}\Big(\mathrm{SKP}\big(\mathrm{MHSA}^c(\tilde{\mathbf{X}}_i)\big)\Big); \tag{14}$$

that is, the skip-connection is used around the multi-head self-attention, which is then followed by layer normalization. Note that the use of skip-connection in (14) enforces setting d_h defined in (9) to c, which is the number of channels.

5) POSITION-WISE FEED-FORWARD NETWORKS

The Transformer encoder utilizes a fully connected feed-forward network that transforms each element of a given sequence individually. Let $\mathbf{Y} \in \mathbb{R}^{a\times b}$ be the generic matrix defined before. The effect of this position-wise feed-forward network operated on an input $\mathbf{Y}$, denoted $\mathrm{FFN}^s(\mathbf{Y})$, is:

$$\mathrm{FFN}^s(\mathbf{Y}) = [g(\mathbf{y}_1), \ldots, g(\mathbf{y}_b)], \tag{15}$$

where $\mathbf{y}_k, k = 1, \ldots, b$, are columns of $\mathbf{Y}$ and

$$g(\mathbf{y}_k) = \mathbf{W}_2 \times f(\mathbf{W}_1\mathbf{y}_k + \mathbf{b}_1) + \mathbf{b}_2, \tag{16}$$

where $f(.)$ denotes an element-wise nonlinear activation function (e.g., ReLU), and $\mathbf{W}_1 \in \mathbb{R}^{r\times a}$, $\mathbf{W}_2 \in \mathbb{R}^{s\times r}$, $\mathbf{b}_1 \in \mathbb{R}^{r\times 1}$, and $\mathbf{b}_2 \in \mathbb{R}^{s\times 1}$ are learnable matrices and vectors—r is generally a tuning parameter. We use superscript s in $\mathrm{FFN}^s(\mathbf{Y})$ to highlight the dimensionality of output vectors in (15). In the Transformer encoder, a position-wise feed-forward network is used to produce an output $\mathbf{O}_i$ from $\mathbf{X}_i$ obtained in (14), which is then added to its input through the skip-connection, followed by layer normalization. This operation is characterized as follows:

$$\mathbf{O}_i = \mathrm{LN}\Big(\mathrm{SKP}\big(\mathrm{FFN}^c(\mathbf{X}_i)\big)\Big). \tag{17}$$

Note that the use of skip-connection in (17) enforces setting s defined in (15) to c. The classification can be performed by vectorizing $\mathbf{O}_i$ and using that as the input to a fully connected layer with a softmax activation function.[1]

B. BENEFITS OF THE TRANSFORMER MODELS

Transformers offer a significant advantage in parallelization. Unlike traditional RNNs that process sequences one element at a time, transformers can handle all sequence elements simultaneously due to their non-recurrent nature. This parallel processing approach is highly optimized for parallel hardware such as GPUs, leading to noticeably reduced training times [156]. Additionally, transformers are adept at handling long-range dependencies within sequences, due to their attention mechanisms. These mechanisms allow the model to effectively associate distant elements in sequences, offering a notable improvement over conventional models such as LSTMs in capturing the context within extended texts [157]. On the scalability front, transformer-based models, including BERT [63] and GPT [50], have demonstrated the ability to scale up to billions of parameters. This scalability has played a key role in setting new performance benchmarks across a wide array of tasks. Moreover, the flexibility of transformer models is prominent. Even though they were initially designed for sequence-to-sequence tasks, they have been successfully adapted for a variety of applications, ranging from classification to image processing [52]. Their capability to cater to diverse data types underscores their potential as a versatile tool in machine learning research. The combination of parallel processing, efficient handling of long dependencies, scalability, and broad adaptability make transformer models a popular choice for various tasks.

[1]Supplementary Material provides a sample code in PyTorch that complements the mathematical formulation discussed in this section

IV. APPLICATIONS OF TRANSFORMERS IN BCIS

With the advent of the Transformer architecture, there has been an increasing interest in leveraging these state-of-the-art machine-learning models to advance BCIs. In the subsequent sections, we will explore how transformer models have significantly advanced the BCI domain. Figure 2 provides an overview of the fundamental principles, areas of application, benefits, and challenges of transformers in BCIs. It highlights the potential of transformers in tasks such as motor imagery decoding and sleep stage analysis, where they can effectively handle brain signal dynamics. We will discuss the various ways in which transformers can be utilized in BCIs, potentially leading to beneficial developments in both research and practical applications.

Figure 3 presents diverse architectural configurations that integrate the Convolutional Neural Network (CNN) and Transformer Encoder components, as discussed in this study, tailored for EEG data analysis for BCIs. These configurations differ in their approach to feature extraction from EEG signals: while some models emphasize temporal characteristics, others concentrate on spatial patterns. Notably, there are also hybrid models designed to capture both spatial and temporal features concurrently. By fusing the capabilities of CNNs with Transformer techniques, these architectures underscore the potential and adaptability of advanced transformer approaches in EEG data processing. Furthermore, in Table 1, we list abbreviations used in the work. It includes full terms in the first and third columns and their corresponding abbreviations in the second and fourth columns. The terms cover relevant topics in BCI research and are crucial for understanding the discussed components and techniques. For example, it includes abbreviations for movements such as "Left Hand" (LH) and "Both Feet" (BF), metrics such as "Average" (AVG) and "Precision" (PREC), and technical terms such as "Multi-Head Self-Attention" (MHSA) and "Batch Normalization" (BN).

A. MOTOR IMAGERY EEG DECODING

Motor Imagery (MI) refers to the cognitive process where individuals mentally visualize and rehearse specific motor tasks without physically moving. This mental representation is important in BCIs, particularly in creating effective communication and control methods for people with severe motor disabilities. MI-BCIs aim to convert these imagined motor tasks into control signals that can be used to operate external devices, such as computer cursors, robotic prosthetics, and electric wheelchairs [7], [158], [159], [160], [161].

However, there are several challenges associated with the decoding process of MI:

- **Intra and Inter-Subject Variability**: EEG patterns that correspond with MI manifest considerable variability. This variability is discernible not only within individual sessions but also across different individuals. Such

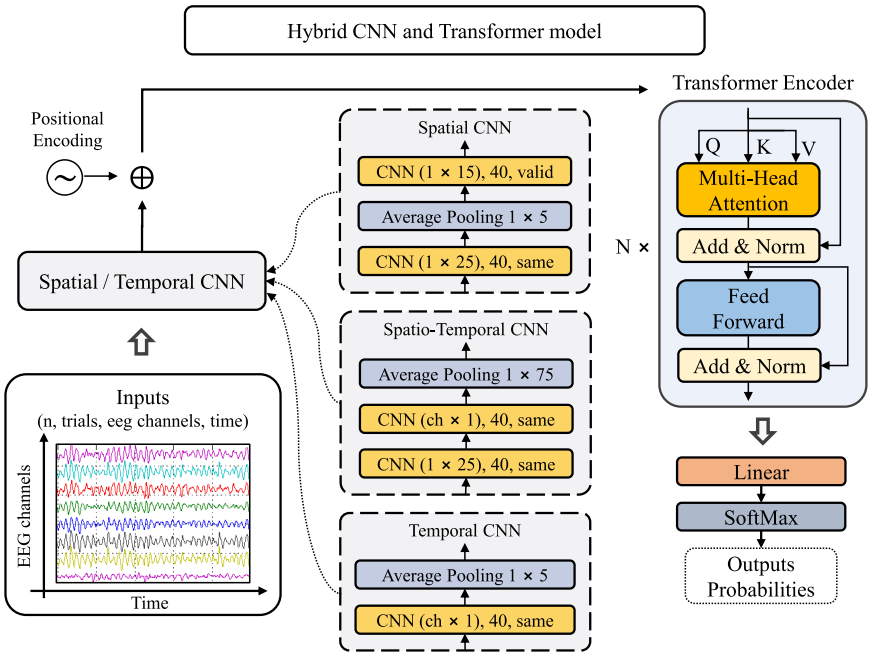

FIGURE 3. Architectural Configurations of Convolutional Neural Network (CNN) and Transformer Encoder Components for EEG Analysis. The diagram illustrates the diverse possible configurations combining CNNs and Transformers, tailored to extract temporal, spatial, or integrated spatiotemporal features using advanced transformer methodologies.

TABLE 1. List of abbreviations used in this work.

Full Term	Abbreviation	Full Term	Abbreviation
Left hand	LH	Subject independent	SI
Right hand	RH	Subject dependent	SD
Both feet	BF	Eyes open	EO
Tongue	T	Eyes closed	EC
Average	AVG	Physical action	PHY
Precision	PREC	Imagined action	IMA
Accuracy	ACC	Cross-subject experiment	CSE
Sensitivity	SENS	Multi-Head Self-Attention	MHSA
Specificity	SPEC	Multi-Head Attention	MHA
Fully Connected	FC	Feedforward	FF
Rectified Linear Unit	ReLU	Batch Normalization	BN
Exponential Linear Units	ELU	Global Average Pooling	GAP
Position Encoding	PE	Virtual Reality	VR
Motor Imagery	MI	Brain-Computer Interface	BCI
Long Short-Term Memory	LSTM	Convolutional Neural Network	CNN
Particle Swarm optimization	PSO	EEG channel-attention	ECA
Layer Normalization	LN	Matthews Correlation Coefficient	MCC
Pearson Correlation Coefficient	PCC		

inconsistencies render the formulation of universally applicable models a complex undertaking [129], [132], [133], [135], [136], [146], [148].

- **Susceptibility to Noise**: Inherent noise within EEG signals, compounded by external artifacts originating from muscle twitches, eye movements, or other external electromagnetic interferences, can convolute the accurate decoding of MI patterns [129], [130], [135], [142], [151].
- **Non-stationarity**: Overextended durations, EEG patterns linked with MI may undergo alterations attributable to factors such as fatigue, learning adaptations,

TABLE 2. Summary of datasets used in MI EEG-based studies.

Dataset Name	Detailed Info	Channels	Subjects	Targets	Studies Used
BCI Competition IV 2a [10]	9 individuals, 22 channels, 288 trials, 500/1000 time samples	22	9	LH, RH, BF, T	[129], [130], [132]–[140]
BCI Competition IV 2b [10]	9 individuals, 3 channels, 5 sessions, 120-160 trials/session	3	9	LH, RH	[129], [133]–[135]
BCI Comp. 2003 Data III [141]	1 subject (25 y.o. female), 7 sessions, 40 trials/session	3	1	LH, RH	[142]
BCI competition IV dataset 1 [143]	7 subjects, 59 channels, and 200 trials.	59	7	LH, RH, BF	[144]
OpenBMI [145]	54 subjects, 62 channels, 400 trials	62	54	LH, RH	[133], [139], [146]
Physionet [147]	109 subjects, 64 channels, 11354 samples, 656-time steps	64	109	LH, RH, EO, BF	[60], [128], [140], [148]
Private Datasets					
By Huashan Hospital	108 trials, 7 stroke patients, 31 channels	31	7	Motor, Reset	[149]
Private Dataset 1	25 subjects, 60 channels, 320 trials	60	25	RH, BF	[144]
Private Dataset 2	5 subjects, 8 channels, 300 trials (AO + MI)	8	5	LH, RH, No Move	[150]
Private Dataset 3	20 subjects, 59 channels, 26400 epochs, 6 skeleton points	59	20	LH, RH	[151]
Private Dataset 4	40 subjects, 64 channels, 50 trials/class	64	40	LH, RH	[130]
Other Datasets					
Brain-Visual [152]	6 subjects, 128 channels, 11964 samples, 500-time steps	128	6	40 Classes	[148]
ASU (Speech Imagery) [153]	6 subjects, 60 channels, 2 classes, 100 trials/class	60	6	In, Cooperate	[130]

or other dynamic neural processes, leading to potential decrements in decoding efficacy [132], [135], [142].

- **High Dimensionality**: Given the time-series nature of EEG recordings, the data encapsulates high dimensionality. This necessitates the deployment of refined, computationally intensive algorithms to sift through, process, and efficaciously decode the embedded information [142], [146], [151].

It is crucial to take a comprehensive and diverse approach to the development and optimization of MI-BCI systems to ensure their effectiveness in various real-world situations, given the aforementioned challenges.

B. TRANSFORMER-BASED MODELS IN MI-BCI

The transformer architecture has been widely adopted in the research community to enhance Motor Imagery (MI) decoding in BCIs. A number of EEG analysis models have been devised and assessed in this context, with technical details of certain models summarized in Table 3, Table 4, and Table 5. These tables outline distinct approaches, datasets used for testing, and their corresponding performance outcomes. In addition to the technical details of the studies, we include Table 2 to provide a better understanding of the data sources used. This table outlines each dataset's specific details and is organized with various columns that offer a clear overview of each dataset. The ''Private Datasets'' section is proprietary and absent from public access, while the ''Other Datasets'' section covers datasets with applications beyond Motor Imagery, such as speech imagery.

A study conducted by Ma et al. [138] demonstrates significant progress in this field. Their hybrid CNN-Transformer model, which includes spatial, spectral, and temporal transformers, achieved an impressive MI-EEG decoding performance improvement with an accuracy of 83.91%. Following a similar path, Wu et al. [144] introduced the TransEEG model. This model combines a CNN encoder with transformer blocks and enhances it further with graph embedding. It achieved accuracies of 89.5% and 77.4% in their private data set and the BCI IV-1 dataset, respectively. Ma et al. presented important work [139], where their CNN model equipped with an attention mechanism decoded temporal EEG features. Their work performed well on the BCI Competition IV 2a and OpenBMI datasets, achieving session-dependent accuracies of 82.32% and 77.52% and session-independent accuracies of 79.48% and 70.43%, respectively. Kostas et al. [140] presented the BENDR methodology, a unique approach that uses unlabeled EEG data. Their results were impressive across several MI datasets, notably achieving 86.7% accuracy on the PhysioNet dataset. Moving forward, Hameed et al. [129] introduced the Temporal-Spatial Transformer model. This model incorporates an ICA filter and attention mechanisms, achieving remarkable accuracies of 96.11% and 84.89% in the BCI IV 2a and 2b datasets, respectively. In a novel merger of BCI and VR, Lee et al. [150] employed the TSTN model with continuous learning. Through VR-aided MI tasks, they achieved progressive accuracy increments in AO+MI sessions, showcasing the transformative power of VR in MI EEG research. Wang et al. [151] embarked on a unique endeavor, using MI EEG signals to decode sign language. With their Motion Imagery Trajectory Reconstruction Transformer model, they achieved an impressive accuracy of 0.975 in reconstructing motion trajectories. Ahn et al. [130] further contributed to the field with their multiscale convolutional transformer, which combines various imagery tasks. This model achieved notable accuracies of 0.62, 0.70, and 0.70 on their private dataset, the BCI IV 2a dataset, and the Arizona State University (ASU) dataset, respectively. The aforementioned studies show the remarkable potential of transformer models in MI-EEG decoding.

Several other research groups have proposed novel transformer models for MI-BCI systems. One notable innovation is the hierarchical transformer introduced by

TABLE 3. Overview of transformer models in motor imagery brain-computer interfaces (BCIs) (Continued on next page).

Ref	Protocol	EEG Data	Model Description	Performance	Highlights
[146]	LH vs. RH; SI / SD tests.	OpenBMI dataset: 400 trials x 62 ch	Two-level hierarchical transformer: - LLT: Extracts short-term features - HLT: Focuses on relevant features	81.3% (SI) (Acc), 82.1% (SD) (Acc)	Hierarchical Attention: Introduced a two-level transformer that separately focuses on short-term and relevant features. Outperformed conventional CNN models in the subject-independent test.
[132]	LH vs. RH vs. BF vs. T; CSE	BCI Competition IV 2a dataset: 288 trials x 22 ch x 500 samples	VAT-TransEEGNet: - VAT regularization and PSO - Adds self-attention to EEGNet	63.56% (CSE) (Avg Acc)	Regularization and Particle Swarm: Integrated Variational Adversarial Training with Particle Swarm Optimization in EEGNet, enhancing robustness and outperforming multiple baseline models.
[149]	Motor vs. Reset state	Dataset by Huashan Hospital: 108 trials x 31 ch	ST with ECA: - Powerful ML block - Shifted window-based MSA module	87.67% (Avg Prec)	Utilized the Swin Transformer's shifted window-based multi-head self-attention (MHSA) module for better EEG representation. Outperformed other standard models such as CSP and CNN.
[133]	LH vs. RH; SI tests.	BCI Competition IV 2a/2b, OpenBMI	SMT: - CNN, MSA block - Mirror network for data augmentation	2a - 67.28% (Avg Prec), 2b - 76.41% (Avg Prec), OpenBMI - 79.76% (Avg Prec)	Proposed a shallow transformer model combined with a mirror network for EEG data augmentation. Highlighted the potential of multi-head self-attention in subject-independent (SI) motor imagery BCI tasks.
[135]	LH vs. RH vs. BF vs. T; CSE	BCI Competition IV 2a/2b	GAT: - Feature extractor, global adaptor - Domain discriminator, classifier	2a - 76.58% (Avg Acc), 2b - 84.44% (Avg Acc)	Domain Adaptation with Attention: Proposed a global adaptive transformer emphasizing domain adaptation using attention mechanisms.
[134]	Multiple Protocols	Multiple Datasets	EEG Conformer: - CNN, AvgPooling - Self-attention, classifier	2a - 78.66% (Avg Acc) 2b - 84.63% (Avg Acc)	Achieved impressive results across both motor imagery and emotion recognition tasks using the EEG Conformer model, showcasing versatility.
[136]	LH vs. RH vs. BF vs. T; SI tests.	BCI Competition IV 2a dataset	CRAM: - Two stacked recurrent networks, a self-attention module - Classification block.	59.10% (Avg Acc)	Outperformed a series of models as CNN, RNN, and Attention-based models by integrating stacked recurrent networks with a self-attention mechanism.
[148]	Multiple Protocols, SI / SD tests.	MI Physionet dataset: 64 ch x 11354 samples Brain-visual dataset: 128 ch x 11964 samples	Gated Transformer: - PreLN Transformer and Post-LN Transformer - Input Embedding, PE, Encoder Block, and a Classifier - Encoder Block consists of a LN, MHA, a Gating layer, and a FF layer	Physionet - 55.40% (Avg Acc) Brain-visual - 61.11% (Avg Acc)	Temporal-Spatial Decomposition: Used a two-encoder transformer strategy focusing separately on temporal and spatial EEG features, achieving outstanding results across different EEG states.
[128]	Four states: EO, EC, PHY, IMA.	Physionet dataset	ETST: - TTE: A temporal transformer encoder - STE: A spatial transformer encoder	Single state (Avg Acc): EO - 100% EC - 99.96% PHY - 99.97% IMA - 100% Two states (Avg Acc): PHY - 97.29%, IMA - 97.45% All states (Avg Acc): 99.90%	Provided three sub-experiments, and outperformed many baseline models and shows strong generalization ability.
[137]	LH vs. RH vs. BF vs. T;	BCI Competition IV 2a dataset: 288 trials x 22 ch x 1000 samples	CNN-Transformer: - Temporal 1D-CNN, MaxPooling, ReLU activation, BN layer - PE, Transformer Encoder, Transformer Decoder, Classifier	99.29% (Avg Acc)	Integrated the entire vanilla transformer architecture with convolutional layers, achieving top-tier performance against other models such as SNN-LSTM.

TABLE 4. (Continued from previous page) Overview of transformer models in motor imagery brain-computer interfaces (BCIs).

Ref	Protocol	EEG Data	Model Description	Performance	Highlights
[142]	LH vs. RH	BCI Competition 2003 Data III: 280 trials.	Transformers with Auto-Encoders: - FBCSP Feature Extraction - AE Dimensionality Reduction - Vanilla Transformer, Classifier	91.30% (Avg Acc)	Transformer with Auto-Encoders: Improved classification results by combining transformers with auto-encoders, outperforming KNN, LDA+KNN, and standalone transformer models.
[60]	Two-class: left fist/ right fist (L/R), Three-class: left fist/ right fist/eyes open (L/R/O), Four-class: left fist/ right fist/ eyes open/ feet (L/R/O/F) SI tests.	MI Physionet dataset: 3s of data 480 samples, 6s of data 960 samples.	Five different models: - s-Trans: spatial-Transformer - t-Trans: temporal-Transformer - s-CTrans: spatial-CNN, Transformer - t-CTrans: temporal-CNN, Transformer - f-CTrans: fusion-CNN, Transformer	Highest Acc: 3s data (f-CTrans): L/R: 83.31% (s-CTrans), L/R/O: 74.44%, L/R/O/F: 64.22% 6s data (t-CTrans): L/R: 87.80%, L/R/O: 78.98%, L/R/O/F: 68.54%	Presented five transformer-based models, focusing on spatial and temporal EEG aspects. Achieved the highest performance with fusion-based CNN and Transformer models on subject-independent tasks.
[138]	2a: LH vs. RH vs. BF vs. T; Within-subject classification	BCI Competition IV 2a: 576 trials x 22 ch x 1000 samples	Hybrid CNN-Transformer: - Spatial Transformer, - Spectral Transformer (FBCSP), - CNNs, Temporal Transformer (FF, MHA), - Classifier. Early stopping was applied.	83.91% (Acc)	Outperforming Benchmark: Achieved superior results on the BCI IV dataset compared to previous state-of-the-art methods, reinforcing the model's effectiveness. [154]
[144]	Private dataset: RH vs. BF; BCI IV-1: LH vs. RH vs. BF	Private dataset: 320 trials x 60 ch BCI competition IV dataset 1: 200 trials x 59 ch	TransEEG: - CNN encoder: 2D-Conv layers, BN layer, ELU activation, Max-Pooling, Dropout - Three transformer blocks with graph embedding	Private - 89.5% (Avg Acc) BCI IV-1 - 77.4% (Avg Acc)	Graph Embedding for EEG: Introduced graph embedding to represent multichannel EEG data, achieving robust and precise results by capturing spatial relationships between channels.
[139]	2a: LH vs. RH vs. BF vs. T; OpenBMI: LH vs. RH; session-dependent / session-independent	BCI Competition IV 2a dataset: 576 trials x 22 ch OpenBMI dataset: 400 trials x 20 ch	CNN with an attention mechanism: - Spatial convolutional layers, BN, and ELU. - Temporal segmentation and feature extraction, - Temporal attention module, depth-wise separable convolution and MHA. - Classifier Early stopping was applied.	2a (Avg Acc): Session-dependent: 82.32% Session-independent: 79.48% OpenBMI (Avg Acc): Session-dependent: 77.52% Session-independent: 70.43%	Temporal Dependencies: By leveraging attention mechanisms and a 2D map for feature extraction, the model successfully exploited temporal dependencies in MI-EEG, enhancing performance across multiple datasets.
[140]	Multiple Protocols	TUEG dataset; MI PhysioNet dataset; MI BCI IV 2a dataset; ERN dataset: 56 ch P300 dataset: 64 ch SSC dataset: 2 ch.	BENDR: - A series of 1D convolutions with short-receptive fields and a transformer encoder.	PhysioNet (Acc): 86.7% BCI IV 2a (Acc): 42.6% ERN (AUROC): 0.65% P300 (Acc): 0.72% SSC (AUROC): 0.72%	Universal EEG Learning: Developed a strategy to learn from a wide range of EEG data without requiring labels. The approach promises more generalized EEG models by tapping into broader EEG data distributions.
[129]	2a: LH vs. RH vs. BF vs. T; 2b: LH vs. RH; SI / SD tests.	BCI competition IV 2a and 2b datasets	TST: - ICA filter - Temporal and Spatial transformations using an attention mechanism - Classifier: FC layer and GAP	SD tests (Avg Acc): 2a: TST: 96.11% TST-ICA: 97.77% 2b: TST: 84.89% TST-ICA: 85.90% 5-fold CV (TST-ICA) (Avg Acc): 2a: 88.75% 2b: 84.20% LOSO (TST-ICA) (Avg Acc): 2a: 93.94% 2b: 87.29%	Dual Approach with ICA: Demonstrated the prowess of using both temporal and spatial transformers, with enhanced results when integrating Independent Component Analysis (ICA).

TABLE 5. (Continued from previous page) Overview of transformer models in motor imagery brain-computer interfaces (BCIs).

Ref	Protocol	EEG Data	Model Description	Performance	Highlights
[150]	AO+MI session: LH vs. RH; MI-FB sessions: LH vs. RH vs. No Movement	Private dataset: AO + MI session: 300 trials MI-FB sessions: 900 trials	The study incorporates TSTN, proposed by Song et al. (2021) [154], it focuses on: - AO+MI: Action Observation and MI - MI-FB: MI with Feedback - Continual Learning strategy.	AO+MI: 0.63 (Avg Acc) 1st MI-FB: 0.68 (Avg Acc) 2nd MI-FB: 0.75 (Avg Acc) 3rd MI-FB: 0.77 (Avg Acc)	Virtual Reality & Motor Imagery: Gathered EEG data using Virtual Reality (VR), exploring the potential of combining action observation with motor imagery tasks for richer EEG interpretations.
[151]	Multiple Protocols	Private dataset: 59 ch x 26400 epochs	MITRT: - Transformer encoder: pre-processed EEG input - Transformer decoder: input corrected joint points location input	0.975 (PCC)	Motion Trajectory in 3D: Successfully reconstructed upper limb motion trajectories from MI EEG signals, showcasing potential applications in understanding Chinese sign language in a 3D space.
[130]	Multiple Protocols	Private dataset: 64 ch (MI and VI) 200 trials (SI) BCI competition IV 2a, Arizona State University (ASU) dataset: 200 trials	Multiscale convolutional transformer: - Temporal convolutional blocks based on TSception, - A temporal transformer encoder, - Parallel spatial convolutional blocks, - A spatial transformer encoder, - A fusion convolutional block.	Private: 0.62 (Acc) BCI IV 2a: 0.70 (Acc) ASU: 0.70 (Acc)	Multiscale Imagery Paradigm: Designed a unique experimental paradigm encompassing motor, visual, and speech imagery tasks, presenting a comprehensive approach for EEG data collection and interpretation.
[155]	LH, RH SI tests.	BCI competition IV 2a: 2592 trials x 22 ch x 321 samples BCI competition IV 2b: 6520 trials x 3 ch x 321 samples MI Physionet dataset: 4683 trials x 64 ch x 201 samples Weibo dataset: 1580 trials x 60 ch x 321 samples	CNN with Vision Transformers: - s-CViT: Spatial CNN, Vision Transformer, - t-CViT: Temporal CNN, Vision Transformer, - st-CViT: Spatio-Temporal CNN, Vision Transformer, - Classifier	2a: 80.44% (Avg Acc) 2b: 74.73% (Avg Acc) Physionet: 83.08% (Avg Acc) Weibo: 73.88% (Avg Acc)	Presented a realistic approach to building subject-independent BCIs using nested LOSO method and combination of CNN and Vision Transformers.

Deny et al. [146], which partitions attention across two levels: a low-level transformer for feature extraction and a high-level transformer for highlighting crucial features. Tan et al. [132] combined VAT regularization with Particle Swarm Optimization to enhance EEGNet with self-attention. Wang et al. [149] fused EEG channel attention with the Swin Transformer, diverging from traditional CSP and CNN models. Additionally, Luo et al. [133] ventured into data augmentation and probability ensembling with a shallow mirror transformer. Song et al. [135] developed an attention-based domain adaptation model to enhance decoding across subjects. Tailoring transformer models to BCI's unique requirements, Tao et al. [148] demonstrated the efficacy of gated Transformer models. In contrast, Jiang et al. [142] combined Auto Encoders with FBCSP for efficient feature extraction. In addition to these contributions, Xie et al. [60] combined spatial and temporal CNN transformers, achieving good results in subject-independent scenarios. Ma et al. [138] refined the spectral transformer, while Wu et al. [144] explored graph embeddings for dynamic extraction. Reflecting the versatility of transformer models, Kostas et al. [140] delivered a methodology that spans multiple EEG datasets. In an imaginative blend, Lee et al. [150] integrated their model within the immersive realm of virtual reality, creating a data set rooted in VR-based motor imagery tasks. Other novel applications include Wang et al.'s [151] effort to interpret 3D motion from Chinese sign language, and Ahn et al.'s [130] multiscale convolutional transformer for diverse mental imagery decoding.

When considering the specific uses of transformer models in the context of MI-BCIs, we can identify several important utilities:

- **Feature Extraction**: One of the prominent advantages of transformers is their ability to automatically extract features from raw EEG signals. This largely eliminates the need for laborious and complex manual feature engineering, which has traditionally been a significant part of BCI research [128], [129], [132].
- **Robustness to Noise**: Transformers utilize multi-head attention mechanisms to focus on the relevant portions of

an EEG sequence selectively. This feature contributes to substantial reductions in the influence of noise and other artifacts, leading to more reliable BCI outputs [129], [132], [135], [151].

- **Temporal Dynamics**: The Transformer architecture is particularly well-suited for handling the sequential nature of EEG data. This enables it to capture the temporal dynamics of Motor Imagery (MI) with greater efficacy than traditional models such as CNNs or RNNs. This has been shown to improve decoding accuracy across multiple studies [129], [134], [136], [137], [146], [148], [151].
- **Transfer Learning**: The adaptability of Transformer models also allows them to benefit from pre-training on large datasets. This enables these models to start at a more advanced point when tailored for specific BCI tasks, potentially mitigating challenges related to inter-subject variability [135].
- **Hybrid Models**: Some research efforts have explored the combination of Transformers with other machine learning architectures, such as CNNs. These hybrid models aim to capture spatial and temporal features more effectively for enhanced MI decoding [128], [132], [133], [134], [137], [162].

The research landscape in MI-EEG decoding shows a growing interest in transformer models. Various models have shown promising results in different datasets and methodologies. These advancements highlight the significant impact of these models and the wide potential of BCI research.

C. EMOTION RECOGNITION

BCIs, originally designed to aid communication and control for those with motor impairments, are now expanding in their potential applications [163]. Recent advancements in BCIs have shown promise in recognizing and decoding emotions directly from the brain, with potential uses in entertainment and medical therapy [164], [165]. By accurately detecting human emotions, machines can respond in a more empathetic manner, resulting in more natural and personalized interactions [166]. Furthermore, continuous emotional monitoring can facilitate the early detection of mood disorders and severe conditions such as post-traumatic stress disorder (PTSD) [161]. The entertainment industry, particularly gaming and virtual reality, can capitalize on this technology to customize user experiences based on their emotions, leading to heightened engagement and satisfaction [167]. BCIs can also offer real-time emotional feedback for patients undergoing therapy for trauma or emotional disorders, enabling therapists to devise more effective and tailored treatment plans [168]. Emotion recognition BCIs utilize EEG data for its high temporal resolution [167]. With advancements in signal processing and machine learning, including the use of Transformers [168], emotion decoding accuracy has significantly improved.

D. TRANSFORMER MODELS IN EMOTION RECOGNITION

The application of Transformers in EEG data for emotion recognition has generated significant interest in the literature. Transformer models automatically extract relevant features from EEG signals associated with emotional states, eliminating the need for laborious manual feature engineering.

Several EEG datasets related to emotions were examined as summarized in Table 6. These datasets vary in the number of subjects, trials per session, EEG channels used, emotional states examined, and relevant academic literature. The DEAP and SEED datasets have gained popularity for EEG-based emotion recognition, as indicated by insights from these studies. A number of studies on the applications of transformers in BCI-based emotion recognition have been devised and assessed in this context, with technical details of certain models summarized in Table 7, Table 8, and Table 9. These tables outline distinct approaches, datasets used for testing, and their corresponding performance outcomes. Studies presented in the tables have empirically demonstrated the effectiveness of Transformers in capturing long-term dependencies in EEG data for emotion recognition tasks. They also highlight the challenges and potential for further optimization in this domain. Transformers, when combined with other techniques such as convolution or tailored for specific spatial-temporal features, can achieve impressive results. However, a common challenge across these studies is the decrease in performance in subject-independent classification tasks, underscoring the need for models that generalize well across different individuals.

For instance, Li et al. [61] achieved superior classification accuracies by utilizing the DEAP and DREAMER datasets. Their proposed Transformer Neural Architecture Search model, which integrated a Supernet with a Multi-Objective Evolutionary Algorithm (MOEA) and a classifier, achieved the highest average accuracy in emotion classification for the DREAMER dataset. Koorathota et al. [181] introduced the Multimodal Neurophysiological Transformer (MNT) on the DEAP dataset. By integrating raw time series with extracted features, their model demonstrated the potential for sequential modeling of EEG data, achieving notable results for valence and arousal. Xiao et al. [170] worked with the DEAP, SEED, and SEED-IV datasets. Their four-dimensional attention-based neural network, which integrated spectral, spatial, and temporal attention mechanisms, achieved commendable performance across multiple datasets. Wang et al. [182] addressed binary and four-class emotion classifications using the DEAP and MAHNOB-HCI datasets. Their Hierarchical Spatial Learning Transformer, focusing on electrodes and brain-region-level spatial learning, highlighted the contribution of brain regions in capturing enhanced spatial dependencies. Sun et al. [62] leveraged the DEAP, SEED, and SEED-IV datasets and proposed a Dual-Branch Dynamic Graph Convolution with Adaptive Transformer Feature Fusion. Their proposed model achieved impressive results on SEED and SEED-IV, showcasing its potential. Arjun et al. [183] reported exceptionally high

TABLE 6. Overview of EEG datasets utilized for emotion recognition with transformers.

Dataset Name	Subjects	Channels	Emotions	Studies Used
SEED [169]	15	62	Positive, Negative, Neutral	[62], [134], [170]–[177]
SEED-IV [178]	15	62	Neutral, Sad, Fear, Happy	[62], [170]–[173], [177]
DREAMER [179]	23	14	Potency, Arousal, Dominance	[61], [173], [176]
DEAP [180]	32	32	Valence, Arousal, Liking, Dominance	[61], [62], [170], [181], [182] [174], [176], [177], [183]
MAHNOB-HCI [184]	27	32	Arousal, Valence, Dominance, Predictability, Emotional Keywords	[182]
Private Datasets				
Private dataset 1 (HIED)	30	64	Happiness, Inspiration, Neutral, Anger, Fear, Sadness	[174]
Private dataset 2	32	32	Positive, Negative, Neutral	[185]

accuracies using Continuous Wavelet Transform (CWT) and raw EEG signal for emotion classification. Although specific model details were not provided, their study demonstrated the effectiveness of these techniques. In addition to these studies, Song et al. [134] presented an EEG Conformer model that effectively generalized across both Motor Imagery (MI) and emotion recognition tasks. Liu et al. [186] proposed the EEG Emotion Recognition Transformer (EeT) and achieved noteworthy results, particularly on the SEED dataset.

The aforementioned studies highlight the effectiveness of Transformers in utilizing a multi-head attention mechanism to focus on relevant parts of an EEG sequence. This improves the model's ability to filter out noise, leading to enhanced accuracy in emotion recognition. Additionally, studies demonstrate that Transformer models are scalable, benefiting from pre-training on extensive datasets and fine-tuning on specific BCI emotion datasets. This allows the models to leverage prior knowledge and achieve more precise emotion decoding. These models are also flexible and can be customized to recognize a wide range of emotions, including subtle shifts in emotional states. This capability enables a detailed emotional spectrum instead of simple or binary classifications.

The use of BCIs for emotion recognition holds transformative potential across various sectors. As our understanding of the neural basis of emotions and Transformer architectures deepens, we can expect continuous improvements in the accuracy, reliability, and adaptability of emotion recognition BCIs. With these advancements, significant progress is anticipated in this field in the coming years.

E. OTHER EMERGING APPLICATIONS

The potential of transformer models in the field of BCIs extends beyond motor imagery decoding and emotion recognition. Several lesser-known but promising areas of EEG research are utilizing transformer models, hinting at a future where our understanding and interaction with the brain will be transformed. In this section, we will outline some emerging applications that harness the power of transformers.

While motor imagery is already extensively explored, another promising avenue is language reconstruction from neural data. The goal is to reconstruct perceived or imagined speech directly from neural signals. Achieving this would have profound implications, enabling locked-in patients to communicate or even translating thoughts into understandable speech. Additionally, sleep stage classification can benefit from transformer intervention. Traditionally, EEG data has been used to classify different sleep phases. Transformers, with their ability to process and assign significance to temporal data points, offer improved accuracy in discerning sleep stages.

In Tables 10 and 11, we present studies covering emerging application areas of transformer models including person identification, sleep stage classification, speech reconstruction, epilepsy prediction, Alzheimer's disease detection, and seizure detection. A study by Du et al. [128] focused on using an EEG temporal-spatial transformer for person identification. They used an EEG temporal-spatial transformer, which consisted of a temporal transformer encoder (TTE) and a spatial transformer encoder (STE) and achieved impressive results, with accuracies ranging from 97.29% to 100% for different states of EEG signals. Transformers have also shown promise in sleep stage classification. Dai et al. [56] proposed a multi-channel sleep network that combined transformer encoders with other feature extraction techniques. Their approach achieved accuracies of 85.0% to 87.5% on different datasets, demonstrating a strong correlation with the physiological features of sleep stages. Kostas et al. [140] presented a methodology called BENDR that utilized Transformers for analyzing multiple domains of EEG data, including motor imagery, event-related potentials, and sleep staging. Their approach achieved competitive performance on different datasets, showcasing the ability to learn from diverse EEG tasks and generalize well. In another study, Lee et al. [188] explored the classification of imagined speech and overt speech using EEG signals. They proposed a classification framework that incorporated convolution layers, separable convolution layers, self-attention mechanisms, and feed-forward networks. The results indicated that overt speech recognition outperformed imagined speech, although the difference was not as significant as initially anticipated.

Several other studies have investigated the use of Transformers for seizure detection and prediction. Hussein et al. [189] introduced MViT, a multi-channel vision Transformer, for epileptic seizure prediction. Their model achieved high prediction sensitivity across different public

TABLE 7. Summary of reviewed studies on the applications of transformers in BCI-based emotion recognition (continued on next page).

Ref	Protocol	EEG Data	Model Description	Performance	Highlights
[61]	arousal, valence, dominance	DEAP: 40 videos x 32 ch x 8064 data DREAMER: 18 videos x 14 ch x 25472 data	TNAS model: - Supernet with a Multi-Objective Evolutionary Algorithm (MOEA) - Incorporates an Emotion Classifier	DEAP (Avg Acc): A(98.66%) V(98.68%) D(98.67%) DREAMER (Avg Acc): A(96.95%) V(96.41%) D(96.90%)	The proposed model set a new benchmark in emotion classification using the DREAMER dataset, showcasing its potential for real-world applications.
[181]	valence, arousal	DEAP: 40 videos x 32 ch	MNT: - Adaptable Conv1D - Crossmodal Transformer: PPG, EEG, GSR, Freq. - Self-attention, Classifier	V(58.0%) (Acc) A(69.4%) (Acc)	Multimodal Integration: Introduced a novel transformer designed specifically for multimodal neurophysiological data, achieving competitive results by leveraging EEG, PPG, and GSR data streams.
[170]	DEAP: valence, arousal SEED: positive, neutral, negative SEED-IV: neutral, sad, fear, happy Intra-subject splitting	DEAP; SEED; SEED-IV.	4D-aNN: - Spectral and spatial attention mechanisms - CNN for spectral and spatial information of the 4D representations. - Temporal attention mechanism is integrated into a bidirectional LSTM	DEAP (Acc): V(96.90%), A(97.39%) SEED (Acc): 96.25% SEED-IV (Acc): 86.77%	Attention-Driven Adaptability: The 4D attention-based neural network adeptly harnessed attention mechanisms to recognize discriminative EEG patterns, optimizing model performance across several datasets.
[182]	DEAP: arousal, valence, dominance MAHNOB-HCI: arousal, valence, dominance, predictability, emotional keywords SI tests.	DEAP: 40 trials x 32 ch MAHNOB-HCI: 32 ch Binary classification (samples): 34599 (DEAP) 29952 (MAHNOB-HCI) Four-class classification (samples): 12635 (DEAP) 12338 (MAHNOB-HCI)	HSLT: - EEG feature extraction (PSD features) - Division of the electrode patches (pre-frontal, frontal, and so on, 9 clusters), - Electrode-level spatial learning (Linear embedding + Transformer encoder) - Brain-region-level spatial learning (Linear embedding + Transformer encoder) - Inspired from Vision Transformers.	Binary (Acc): DEAP: A(65.75%), V(66.51%) MAHNOB-HCI: A(66.20%), V(66.63%)	Spatial Hierarchies in EEG: The hierarchical spatial learning transformer effectively delineated brain region contributions and inter-region dependencies, enhancing the model's understanding of spatial EEG data and achieving commendable binary classification results.
[62]	SEED: negative, positive, neutral SEED-IV: neutral, sad, fear, happy valence, arousal, liking, dominance	DEAP; SEED; SEED-IV.	DBGC-ATFFNet-AFTL: - Dual-branch dynamic graph convolution - Adaptive transformer feature fusion	SEED (Acc): 97.31% SEED-IV (Acc): 89.97% DEAP (Acc): V(95.91%), A(94.61%)	The DBGC-ATFFNet-AFTL model combined dual-branch dynamic graph convolution with adaptive transformer feature fusion, enhancing the model's ability to capture EEG channel connections and efficiently fuse features.
[183]	DEAP: valence, arousal, liking, dominance	DEAP	ViT for CWT images and the raw EEG signal: - Patch Embeddings, PE - Transformer Encoder, MLP, Classifier	CWT (Avg Acc): V(97%), A(95.75%) Raw EEG (Avg Acc): V(99.4%), A(99.1%)	Utilizing both CWT image representations and raw EEG signals, the proposed ViT model achieved exceptionally high accuracies for valence and arousal classifications, showcasing the versatility of the approach.
[134]	MI: LH vs. RH vs. BF vs. T SEED (emotions): positive, neutral, negative SD tests	BCI competition IV 2a and 2b datasets (MI); SEED: 62 ch x 3394 samples	EEG Conformer: - Convolution module, average pooling - Self-attention module, classifier	2a: 78.66% (Avg) 2b: 84.63% (Avg Acc) SEED: 95.30% (Avg Acc)	Demonstrated high accuracy across both motor imagery (MI) and emotion recognition tasks, showcasing the model's generalizability across paradigms.

TABLE 8. (Continued from previous page) Summary of reviewed studies on the applications of transformers in BCI-based emotion recognition.

Ref	Protocol	EEG Data	Model Description	Performance	Highlights
[186]	DEAP: valence, arousal SEED: positive, neutral, negative SEED-IV: neutral, sad, fear, happy	DEAP; SEED; SEED-IV.	EeT: - S: Spatial attention - T: Temporal attention - S-T: Sequential spatial-temporal attention - S+T: Simultaneous spatial-temporal attention - DNN, Classifier	DEAP (Avg Acc): A(93.34%), V(92.86%) SEED (Avg Acc): 96.28% SEED-IV (Avg Acc): 83.27%	Reported that the simultaneous spatio-temporal attention gets the best results among the four designed structures, the result is also better than most state-of-the-art methods.
[171]	SEED: negative, positive, neutral SEED-IV: neutral, sad, fear, happy	SEED: 45 trials x 62 ch x 5076 samples SEED-IV: 72 trials x 62 ch x 5043 samples	ACTNN: - Feature extration, Spatial projection - Spatial and spectral attention branch - Spatial-spectral convolution part - Temporal encoding, Classifier	SEED (Avg Acc): 98.47% SEED-IV (Avg Acc): 91.90%	The ACTNN achieved the best classification rates, particularly excelling in distinguishing positive and neutral states in SEED and identifying sadness in SEED-IV.
[187]	valence, arousal, liking, dominance CSE.	DEAP: 32 ch x 2400 samples	ADDA-TF: - Domain adaptation, attention mechanism - Feature-Channel Transformer - Global Temporal Transformer	V(0.61), A(0.64) (Avg Acc)	The ADDA-TF reached the highest performance than single TF or ADDA, proving the advantage of their combination.
[172]	SEED: negative, positive, neutral SEED-IV: neutral, sad, fear, happy	SEED: 75 trials x 62 ch SEED-IV: 72 trials x 62 ch	Bi-ViTNet: - Spatial-frequency feature extraction branch - Spatial-temporal feature extraction branch - Each branch: Linear Embedding and Transformer Encoder - Classifier	SEED (Avg Acc): 97.55% SEED-IV (Avg Acc): 88.08%	Dual-branch Efficiency: The Bi-ViTNet's spatial-frequency branch outperformed its spatial-temporal counterpart, underscoring the significance of spatial-frequency features in emotion recognition.
[173]	SEED: negative, positive, neutral SEED-IV: neutral, sad, fear, and happy / valence and arousal ratings. DREAMER: potency, arousal, dominance CSE	SEED; SEED-IV; DREAMER.	STGATE: - TLB: Transformer learning block, utilizes 2D-CNN and Transformer Encoder - STGAT: Spatial-temporal Graph Attention (STGAT) mechanism, to learn temporal information	SEED (Avg Acc): 90.37% SEED-IV (Avg Acc): 76.43% DREAMER: 76.35%	Topological Learning: STGATE's approach tackled non-Euclidean data with a topological graph structure, addressing the limitations inherent in CNNs and boosting performance across multiple datasets.
[174]	HIED: happiness, inspiration, neutral, anger, fear, sadness SEED: negative, positive, neutral DEAP: valence, arousal SI / SD tests.	Private dataset (HIED): 54000 samples SEED: 13500 samples DEAP: 51200 samples	SECT: - CT and S-CT: LN, MLP - Classifier	HIED (Avg Acc): PSD: 82.51% (SD) DE: 84.76% (SD) PSD(50.12%) (SI) DE(52.94%) (SI) SEED (Avg Acc): 85.43% (SI) DEAP (Avg Acc): V(66.83%), A(65.31%) (SI)	Superior Decoding Efficiency: Demonstrated leading performance across various datasets and states with notable standard deviation improvements, particularly for hearing-impaired subjects using differential entropy (DE).

datasets. Similarly, Hu et al. [190] proposed a hybrid Transformer model for epilepsy prediction, which demonstrated excellent performance compared to CNN-based structures. The application of Transformers in Alzheimer's disease detection has also been explored. Ravikanti [191] developed EEGAlzheimer'sNet, a transformer-based attention LSTM network for detecting Alzheimer's disease using EEG signals. The model achieved high accuracy and demonstrated potential for early diagnosis. Transformers have been applied to speech recognition and motor action recognition tasks as well. Murphy et al. [106] successfully decoded unigram and bigram parts-of-speech tags from single-trial EEG

TABLE 9. (Continued from previous page) Summary of reviewed studies on the applications of transformers in BCI-based emotion recognition.

Ref	Protocol	EEG Data	Model Description	Performance	Highlights
[176]	DEAP: arousal, valence, liking, dominance Dreamer: arousal, valence SEED: positive, negative, neutral	DEAP: 40 trials x 32 ch x 2400 samples/subject Dreamer; SEED.	TSFFN: - Transformer, 3D-CNN - Temporal and spatial feature extraction modules - Temporal-spatial feature fusion module.	DEAP (Acc): TSDFN: A(96.14%), V(95.76%) TSFFN: A(98.53%), V(98.27%) Dreamer (Acc): TSFFN: A(97.74%), V(96.80%) SEED (Acc): TSFFN: 97.64%	Proven Model Adequacy: With the TSFFN, strong performance was consistently observed across DEAP, Dreamer, and SEED datasets, validating the proposed model's utility.
[175]	SEED: positive, negative, neutral SI / SD tests.	SEED: 45 trials	DCoT: - DW-CONV: Depthwise convolutionlayer, PE, learnable embeddings, Transformer encoders, linear layers. - Transformer encoder: LN, MHSA, FF	93.83% (SD) (Avg Acc) 83.03% (SI) (Avg Acc)	Interpretable Feature Extraction: The DCoT model offers not just high performance but also deeper insights into significant brain areas during emotional activities.
[185]	Ternary classification: positive, negatve, neutral Binary classification: Positive, negative	Private dataset: 32 ch x 512 samples/epoch 240 epochs (binary), 360 epochs (ternary)	Spatial-temporal transformer: - Two channels, spatial and temporal EEG epochs - Linear Projection, Transformer Encoder - Weighted sum of the outputs from the two channels.	Binary (Avg Acc): 97.3% Ternary (Avg Acc): 97.1%	Hybrid Model Enhancement: Demonstrated that combining spatial and temporal EEG data with transformers can significantly improve emotion recognition accuracy.
[177]	DEAP: valence, arousal, liking, dominance. SEED: negative, positive, neutral. SEED-IV: happy, sad, fear, neutral.	DEAP: 40 videos x 32 ch x 60 features x 5 freq bands. SEED: 45 videos x 62 ch x 185 features x 5 freq bands. SEED-IV: session1: 45 videos x 62 ch x 10 features x 5 freq bands, session2: 45 videos x 62 ch x 12 features x 5 freq bands, session3: 45 videos x 62 ch x 14 features x 5 freq bands.	MSDTTs: - MST: multi-domain spatial transformer module - DTT: dynamic temporal transformer module	DEAP (Avg Acc): V(98.91%), A(98.89%), β frequency band max SEED (Avg Acc): 97.52%, γ frequency band max SEED-IV (Avg Acc): 96.70%, γ	Positive Emotion Recognition Bias: In multi-domain analysis, positive emotions consistently showed easier recognition and higher accuracies than their negative counterparts.

data using Transformers. Kaushik et al. [192] proposed an ensemble of BLSTM-LSTM and EEG-Transformer models for motor action recognition, achieving superior performance compared to existing methods. These studies represent just a fraction of the emerging applications of Transformers in EEG research. The versatility and effectiveness of Transformers in handling EEG data offer exciting possibilities for advancing our understanding of brain dynamics and developing innovative EEG-based applications in various domains.

V. ADVANTAGES OF TRANSFORMER-BASED MODELS FOR BCIS

Based on the earlier reviewed studies, the integration of transformer-based models in BCIs is a remarkable achievement. This is due to their proven effectiveness in processing sequential data. In the subsequent section, we will outline the essential advantages of utilizing transformer architectures in BCI research.

A. HANDLING OF TEMPORAL SEQUENCES IN EEG

EEG data, intrinsic to many BCI applications, is inherently temporal. It captures the dynamic changes in the brain's electrical activity over time. Traditional models such as RNNs and LSTMs were initially favored for such sequence data. However, transformers, through their self-attention mechanisms, offer a more adaptive way to weigh different time points based on their importance, ensuring that recent data points do not outweigh crucial temporal patterns.

B. ABILITY TO CAPTURE LONG-TERM DEPENDENCIES

Transformers excel at recognizing long-term dependencies in data due to their parallel processing and dynamic weighting through attention mechanisms. In the context of BCIs, brain

TABLE 10. Overview of emerging applications using transformers across a spectrum of eeg-driven applications (Continued on next page).

Ref	Application area	Protocol	EEG Data	Model Description	Performance	Highlights
[128]	Person identification	Four states: EO, EC, PHY, IMA	MI Physionet dataset	ETST: - TTE: A temporal transformer encoder - STE: A spatial transformer encoder	Single state (Avg Acc): EO - 100% EC - 99.96% PHY - 99.97% IMA - 100% Two states (Avg Acc): PHY - 97.29%, IMA - 97.45% All states (Avg Acc): 99.90%	Exceptional Sensitivity: Demonstrated robust performance across various states, especially excelling in the MI Physionet dataset.
[56]	Sleep Stage classification	W, N1, N2, N3, REM.	SleepEDF-20: 43141 samples SleepEDF-78: 196350 samples SHHS: 324854 samples	MultiChannelSleepNet: - Single-channel feature extraction (PE + Transformer Encoder); - Multichannel feature fusion (LN, PE, Transformer Encoder); - Classifier	SleepEDF-20 (Acc): 87.2% SleepEDF-78 (Acc): 85.0% SHHS (Acc): 87.5%	Physiological Correlation: Achieved a correlation with physiological sleep stage features using MultiChannelSleepNet.
[140]	Multiple domains: MI, ERN, ERP, sleep staging	Multiple Protocols	TUEG dataset; MI PhysioNet dataset; MI BCI IV 2a dataset; ERN dataset: 56 ch P300 dataset: 64 ch SSC dataset: 2 ch.	BENDR: - A series of 1D convolutions with short-receptive fields and a transformer encoder.	PhysioNet (Acc): 86.7% BCI IV 2a (Acc): 42.6% ERN (AUROC): 0.65% P300 (Acc): 0.72% SSC (AUROC): 0.72%	Versatility in Unlabeled Data: BENDR showcases the ability to learn across varied EEG data distributions, spanning multiple people, sessions, and tasks without labeled data.
[188]	Imagined speech and overt speech	12 words (ambulance, clock, hello, help me, light, pain, stop, thank you, toilet, TV, water, and yes) and resting state.	Private dataset: 9 subjects, 300 trials/condition, 25 experiments/every 12 words.	- Convolution layers, separable convolution layers; - Self-attention, FF, Dropout; - Residual connection with subsequent LN.	Imagined speech: 35.07% (Avg Acc) Overt speech: 49.5% (Avg Acc)	Comparative Analysis: Overt speech EEG showed marginally better performance compared to imagined speech, contrary to significant differences expected.
[102]	Sleep Stage classification	W, REM, N1, N2, N3	Private dataset: 6 EEG channels, only 1 used. Train (1590 patients), Val (341 patients), Test (343 patients)	- Transformer - Transformer + RNN (CNN + Transformer Encoder) - Inner + Outer Transformer (Transformer Encoders, where the output of first is the input of the second)	Single-Epoch (Acc): Transformer (89.50%) Multi-Epoch (Acc): Transformer + RNN (91.38%) Inner + Outer Transformer (91.45%)	Advanced Hybrid Approaches: The multi-epoch model variants, especially the Inner + Outer Transformer, exhibited superior performance in sleep stage classification.
[193]	Sleep Stage classification	W, N1, N2, N3, REM	Montreal Archive of Sleep Studies (MASS), Sleep-EDF dataset	Residual based attention model: - Feature extractor: CNN, Maxpooling - Encoder: residual blocks, GAP layer - Decoder: MHA, temporal context	MASS: 86.5% (Acc) Sleep-EDF: 80.7% (Acc)	Accelerated Processing: The residual-based attention model provides both training and inference speeds that are over ten times faster than other methods.
[194]	Sleep Stage classification	Sleep-EDF: N1, N2, N3, Wake, REM Subject-specific training	Sleep-EDF: 148471 samples, 197 whole-night PSG recordings.	CNN-Transformer DL model: - Sequential Conv layers (to extract time-invariant data) - Transformer (MHA, Dense layers) (to learn time-variant data) - Classifier.	Basic training (Acc): 77.5% Subject-specific training (Acc): 79.5%	Efficiency and Portability: Comparable performance to state-of-the-art methods but at significantly lower computational costs, with successful testing on a low-cost Arduino board.
[190]	Epilepsy Prediction	Preictal/Interictal	CHB-MIT scalp EEG database: 24 cases from 23 pediatric patients, 18 EEG channels	Hybrid Transformer model: - Rhythm embedding block (8 Conv layers, Avg Pool, 2 FC, Conv layer), - PE, Self-attention block (MHA, FFN) - Classifier block.	91.7% (SENS), 0.00/h (FPR)	Outperforming CNNs: The hybrid transformer model in epilepsy prediction demonstrated superior results over pure CNN structures.
[106]	Speech Recognition	Binary classification: open class vs. closed class words	University of Birmingham private dataset: 4479 sentences, 75 sessions	Linear SVMs and Transformers: - Four encoder blocks and a dense layer - Pretraining used.	68% (Avg Acc)	Decoding Word Types: Successfully decoded unigram and bigram PoS tags from single-trial EEG data, highlighting the efficiency of transformers over SVMs.

signals often carry patterns where an earlier signal might influence a much later one, and recognizing this relationship can be important for accurate decoding. While LSTMs were designed to mitigate the vanishing gradient problem of RNNs

TABLE 11. (Continued from previous page) Overview of emerging applications using transformers across a spectrum of EEG-driven applications.

Ref	Application area	Protocol	EEG Data	Model Description	Performance	Highlights
[189]	Epileptic Seizure Prediction	Preictal vs. Interictal	CHB–MIT Scalp EEG Dataset: 22 patients, 198 seizure events Kaggle/American Epilepsy Society (AES) Invasive EEG Dataset: 2 adult human and 5 canine subjects. Kaggle/Melbourne University (MU) Invasive EEG Dataset: 3 patients with epilepsy	MViT: - Stack of N transformer encoders - Each encoder processes image tokens from an individual EEG channel - MLP for classification.	Surface EEG - 99.80% (Avg SENS) Invasive EEG data - 90.28–91.15% (Avg SENS) CHB–MIT Scalp EEG Dataset: 99.8% (Acc), 99.8% (SENS), 0.004/h (FPR) AES - 90.28% (SENS) MU - 91.15% (SENS)	Achieved a high sensitivity of 90.28–99.80% across three independent datasets.
[151]	Decoding the Continuous Motion Imagery Trajectories of Upper Limb Skeleton Points	left and right shoulders, elbow, wrist skeletons 30 Chinese sign language sentences	Private dataset 20 subjects, 26400 epochs, 59 ch, 6 key skeleton points.	MITRT: - Transformer encoder - Corrected joint points location for Transformer decoder - Similar to Vanilla Transformer.	0.975 (PCC)	Introduced a novel approach using MI EEG signals to reconstruct 3D motion trajectories of upper limb based on Chinese sign language.
[191]	Alzheimer disease detection	Normal vs. Abnormal	Private dataset from [195]	EEGAlzheimer'sNet: - Transformer-based attention, - LSTM network.	96% (Acc) 98% (MCC)	According to the results, the accuracy of the suggested model is 4% greater than CNN, 2.6% greater than RNN, 2.5% greater than SVM, and 0.2% greater than A-LSTM.
[196]	Seizure Detection	non-seizure vs. seizure	2023 ICASSP Signal Processing Grand Challenge dataset: bhe-EEG for training, and a subset of the Temple University Hospital (TUH) Seizure Corpus (469 seizure events across 43 patients) to pretrain.	A mixed Transformer and CNN: - several transformer blocks (MHA, FFN) - Avg pooling, Classifier	100% (SENS), 1.78/h (FPR)	Achieved high sensitivity and low false alarm rates, highlighting the power of pre-trained transformer models on seizure classification.
[197]	Epilepsy Detection	CHB-MIT Scalp EEG dataset: non-seizure vs. seizure	CHB-MIT Scalp EEG dataset:	EEGformer: - Raw EEG, Input embedding - PE, Transformer Encoder, Output	65.5% (SENS), 99.9% (SPEC), 0.8/h (FPR)	Presented an EEGformer that offers shorter detection latency in epilepsy detection and aligns well with state-of-the-art, especially when considering an artifact-removal stage.
[105]	Automated detection of epilepsy	Normal vs. Epilepsy	Kaggle/Turkish Epilepsy EEG dataset: 71 healthy subjects, 50 epileptic patients. The dataset is provided in Kaggle.	EpilepsyNet: - Pearson Product-moment Correlation Coefficients, - Correlation Coefficients Embedding, - PE, Transformer Encoder, Classifier	85% (Acc) 82% (SENS) 87% (SPEC) 82% (Positive Pred)	Introduced EpilepsyNet, a less computationally intensive solution, as an effective tool for the classification of AD patients vs. control subjects.
[198]	Alzheimer Detection	AD/CN FTD/CN	Private dataset: 36 AD, 23 Frontotemporal dementia (FTD), and 29 healthy individuals (CN). 19 scalp electrodes EEG	DICE-net: - Convolution, Transformer Encoder, and FF layers	AD/CN: 83.28% (Acc) FTD/CN: 74.96% (Acc)	Reported the proposed model can effectively capture the complex features of EEG signals for the classification of AD patients vs. control subjects.
[199]	Sleep Staging	Wake, N1, N2, N3, REM	SHHS: 5736 subjects Wake (2371496 samples), N1 (166619 samples), N2 (809155 samples), N3 (732389 samples), REM (214985 samples)	A Transformer-Based Spatial-Temporal Sleep Staging Model: - Inception Module, 30s Sequence patches, Linear projection - Transformer model (PE, Transformer Encoder), SoftMax Classifier	F1-score: Wake (0.92), N1 (0.34), N2 (0.85), N3 (0.84), REM (0.76)	Proposed model outperforms state-of-the-art in classifying Wake, N2, and N3 sleep stages and offers a fully automatic system from feature extraction to staging.
[200]	Seizure prediction	Seizure vs. Non-seizure	CHB-MIT: 22 patients with epilepsy, 198 seizures	A personalized seizure prediction model: - Vision Transformer: EEG, STFT, Vision Transformer	chb21 patient: 94.6% (Acc) 98.6% (Recall) 89.8% (SPEC) 90.5% (PREC) 0.989 (AUC)	Personalized seizure prediction model showcased potential for early epilepsy prediction and holds promise for the diagnosis of seizures in epileptic patients.
[192]	Motor Action Recognition	newline walking, sitting, chewing, blinking, boxing, fist closing, fist opening, drop, EC, EO, hand on, hand out, lift, pull, push, squats, standing.	Private dataset: 20 subjects, 17 day-to-day motor activities	Ensemble of BLSTM-LSTM and EEG-Transformer	BLSTM-LSTM: 97.9% (Acc) EEG-Transformer: 96.7% (Acc) Ensemble: 98.5% (Acc)	Ensemble of BLSTM-LSTM and EEG-Transformer achieved 98.5% accuracy in motor action recognition, marking a significant advancement over current methods.

and capture such dependencies, transformers excel in this domain due to their parallel processing of sequences and the dynamic weighting through attention mechanisms.

C. SCALABILITY AND PERFORMANCE BENEFITS

Transformer models offer numerous advantages for analyzing large EEG datasets. With sufficient computational infrastructure, these models can be trained effectively on extensive EEG recordings, leading to the extraction of complex neural patterns that are often overlooked by less capable models. Unlike sequential models such as RNNs, transformer models can be trained faster due to their capacity for parallel processing. Additionally, streamlined models such as distilled transformers help mitigate computational restrictions. Hence, transformer-based architectures can be favored in BCIs, especially when EEG data exhibit high variability and noise.

D. VERSATILITY ACROSS DIVERSE BCI TASKS

The architectural design of transformers makes them adaptable. Whether it is motor imagery tasks, emotion recognition, or sleep stage classification, the underlying principles of transformers remain consistent. This versatility ensures that researchers do not have to reinvent the wheel for every unique BCI challenge but can adapt and fine-tune established transformer models.

Transformers are pivotal in BCI research for handling temporal sequences and discerning long-term dependencies. Their versatility spans numerous tasks. As these models evolve and integrate, they further broaden the horizons for BCI potential.

VI. CHALLENGES AND LIMITATIONS

While Transformer architectures undeniably augment the performance of BCIs, their successful integration into the BCIs is not without challenges. Key among these challenges are computational efficiency, data variability, and model interpretability. Addressing these concerns is essential to harnessing the full potential of Transformers in BCI development and application.

A. NEED FOR LARGE DATASETS

Transformers, given their huge parameters, often require vast datasets for effective training. In the domain of BCIs, obtaining large and high-quality datasets is challenging due to several reasons. Firstly, there is a trade-off between invasive and non-invasive methods. Invasive methods, using micro-electrode arrays, provide fine-grained neural data but involve surgical procedures and are typically limited to animal models or specific clinical cases. On the other hand, non-invasive methods such as EEG are more common but provide lower spatial resolution data. Collecting brain data also requires adherence to strict ethical guidelines, adding an additional challenge. Moreover, conducting experiments to collect BCI data is time and cost-intensive, requiring specialized equipment, trained personnel, and often lengthy sessions with participants. Technical challenges include ensuring consistent electrode placement, handling equipment-related issues, and ensuring participant comfort and safety. Finally, the complexity of brain signals adds to the challenge. The brain's activity is multidimensional and complex, and capturing all relevant information, especially in real-world scenarios outside controlled lab environments, is difficult.

Potential Solutions: Addressing the challenges in acquiring extensive, high-quality datasets for BCIs requires a multifaceted approach. Adopting a hybrid approach, combining non-invasive methods, e.g., EEG with fNIRS or MEG, can enhance spatial and temporal resolutions [201]. Centralized, open-source repositories can facilitate data sharing [145], [202], [203], [204], while advanced data augmentation techniques, such as Generative Adversarial Networks (GANs), can artificially enlarge datasets [205], [206]. Another approach would be using crowdsourcing to gather BCI data from a broader population. This approach can help in obtaining diverse datasets, capturing a wide range of neural activities and conditions [207]. Additionally, transfer learning allows models to adapt using smaller, task-related datasets [208], [209]. Together, these solutions can lead the BCI community towards better data practices, setting the stage for advanced, reliable future applications.

B. COMPUTATIONAL OVERHEAD

The complex architecture of Transformers, especially in their advanced configurations, places substantial demands on computational capabilities and memory allocation. When these models are tasked with training or fine-tuning on complex and high-dimensional EEG data, the computational burden becomes especially pronounced. This demand can establish significant limitations for research groups with limited resources from fully utilizing the capabilities of these models. Such computational requirements could potentially limit the democratization of transformer-based BCIs.

Mitigating Strategies: Several strategies can be employed to counteract these challenges. One such approach is model distillation, where a smaller, more manageable model is trained to mimic the behavior and performance of its larger counterpart, allowing for efficient deployment without a drastic decline in performance [210]. Additionally, there has been a surge in research focusing on creating optimized versions of transformer architectures that are specifically designed to maintain performance while being more computationally efficient [57], [211]. Techniques that focus on effective training strategies, sparse activations, and model pruning are also being explored to reduce the computational overhead associated with these models [212]. By adopting these methods, the broader research community can harness the potential of transformers in BCIs without being burdened by computational constraints.

C. MODEL INTERPRETABILITY

The interpretability of transformer models is a crucial aspect of deep learning that has been extensively researched [213]. Despite their ability to process vast amounts of data and generate exceptional results across a range of tasks, these models can be incredibly complex, making it difficult to understand how they arrive at their predictions. This lack of transparency in decision-making processes can have serious consequences in fields such as healthcare, where EEG data plays a significant role, and errors in interpretation can result in incorrect medical interventions. Consequently, it is imperative to improve the transparency and interpretability of these models, ensuring that the rationale behind their decisions is accessible and understandable.

Addressing the Challenge: As the deep learning community acknowledges the importance of interpretability, several methods are emerging to understand the inner workings of transformers and other deep learning models: [57], [214], [215], [216]:

- **Attention Visualization**: Given that transformers utilize attention mechanisms, visualizing attention weights can offer insights into which parts of the input data the model deems significant during predictions.
- **Saliency Maps**: These graphical representations highlight input features that are most influential for a given prediction, providing a visual guide to a model's focus areas.
- **Feature Attribution**: By determining the contribution of individual features to the final output, we can gain a clearer understanding of what drives model decisions.

Collaborative efforts between neuroscientists and machine learning researchers can also be beneficial in bridging the interpretability gap.

D. TRANSFER LEARNING AND DOMAIN ADAPTATION

The consistent performance of transformer models across various subjects and devices can be a significant challenge in the BCI field. Training a model on data from a specific group of subjects or a particular device may result in reduced efficiency when applied to a different cohort or another device. This inconsistency can be attributed to several factors.

- **Inter-Subject Variability**: Every individual's brain has unique characteristics and patterns. Differences in anatomy, functional organization, and neural plasticity can lead to distinct EEG signal patterns, even for similar tasks.
- **Inter-Trial Variablity**: A single individual's brain signals can exhibit variations over different trials and even sessions, due to factors such as fatigue, attention levels, or even the time of day.
- **Electrode Placement Inconsistencies**: Minor discrepancies in electrode placement across sessions or individuals can introduce variability. This can arise due to human error, differences in head shape, or hair density.
- **Device-Specific Biases**: Different EEG devices might have unique calibration settings, sampling rates, or signal-to-noise ratios, which can introduce discrepancies in the recorded data.
- **Environmental Noise**: External factors, such as ambient light, noise, or even the room's temperature, can influence an individual's brain signals and further complicate inter-subject comparisons.

Addressing the Challenge: Addressing this challenge requires sophisticated techniques that can normalize and adapt to the inherent variabilities present, ensuring that BCI models are robust, generalizable, and not just narrowly tailored to a specific dataset. Domain adaptation techniques, which adjust a model trained on one domain to perform well on a different but related domain, can be explored [217], [218], [219]. Fine-tuning on smaller, target-specific datasets or employing strategies such as meta-learning [220], where models are trained to quickly adapt to new tasks, can also be beneficial.

While Transformers show promise for BCI research, it is important to be aware of their limitations and work to address them. Collaboration among researchers is key to finding innovative solutions and advancing BCIs to greater efficiency and applicability as the field evolves. For researchers looking to leverage the benefits provided by Transformer models for BCIs, in Appendix IX, we provide a checklist that can serve as a practical guide.

VII. FUTURE DIRECTIONS

The application of transformer architectures to BCIs remains a developing domain with significant potential. With advancements in technology and research methodologies, there are several exploration avenues that promise enhanced integration of transformers in BCIs.

Efficient Transformer Variants: Researchers are currently focusing on enhancing the effectiveness of transformer architectures while maintaining their performance. This area of exploration is rapidly advancing, with new adaptations seeking to minimize computational burdens. This is a crucial stage in enabling their application in real-time BCI situations. The emergence of these variants, primarily targeted at natural language processing, opens up possibilities for customizations tailored for BCI endeavors. Methods such as pruning (purging redundant model parameters) or quantization (reducing parameter precision) could be recalibrated considering the nuances of BCI data.

A. FUSION WITH OTHER MODALITIES

The technological development in data acquisition has sparked interest in combining EEG data with other modalities, such as functional near-infrared spectroscopy (fNIRS), magnetoencephalography (MEG), or even peripheral physiological metrics such as heart rate or skin conductance. Such integrative efforts can amplify the richness of data, potentially catalyzing superior model outputs. Transformers, with their inherent ability to handle sequential data from diverse sources, can play a foundational role in processing

multi-modal data. By synchronizing and processing features from various modalities, transformers can offer a more comprehensive perspective, enabling the discernment of complex patterns in the synergized data.

B. REAL-TIME PROCESSING AND FEEDBACK

Instantaneous processing of brain signals is essential for various applications, particularly those related to assistive technologies or neurofeedback systems. Such processing provides immediate feedback, making BCIs more interactive and intuitive. Nevertheless, achieving real-time processing requires precise, fast, and efficient models. However, computational intensity makes it challenging to meet this level of performance. To tackle this issue, efficient transformer derivatives, combined with hardware optimizations, could offer a way forward for real-time BCI applications. Furthermore, exploring sparse transformers or those designed explicitly for streaming data processing could prove invaluable in this regard.

In summary, the potential for transformers in BCIs is promising, offering opportunities to redefine boundaries. By navigating constraints and exploring new trajectories, transformers have the potential to impact the BCI field.

VIII. DISCUSSION

BCI research has transitioned through various phases of development. From its early reliance on basic signal processing techniques for EEG analysis, it has evolved to embrace advanced machine learning algorithms, particularly deep learning models such as Transformers. The integration of Transformer architectures into BCIs represents a significant milestone. However, there are still challenges that need to be addressed to fully leverage the potential of Transformers in BCIs. These challenges include efficient training techniques, reducing computational overhead, ensuring interpretability, addressing the EEG data deficiency, and exploring transfer learning.

The examined studies underscore the adaptability and efficacy of Transformer architectures in diverse BCI applications. They have demonstrated enhanced performance in capturing temporal dependencies in tasks as motor imagery decoding. Notable models, such as VAT-TransEEGNet [132] and Swin Transformer with ECA [149], have outperformed traditional methods by implementing sophisticated techniques such as VAT regularization and particle swarm optimization. Different protocols and tests are also used to evaluate the generalizability and robustness of these models across various datasets and tasks.

The review extends to applications of transformers in emotion recognition, showing high accuracies in classification tasks with models such as the Transformer Neural Architecture Search model [61] and Multimodal Neurophysiological Transformer [181]. It emphasizes the capability of transformers to handle complex cognitive tasks, capture spatial dependencies, and improve feature extraction processes effectively.

Furthermore, Transformers have proven useful in medical diagnostics and sleep stage classification, with models such as EEG temporal-spatial transformer [128] and MultiChannelSleepNet [56] paving the way for innovative applications in EEG, ranging from refined speech differentiation to comprehensive sleep analysis.

While promising, the incorporation of transformers in BCI is not devoid of challenges, primarily due to their extensive computational requirements and the complexity of the models. These challenges are augmented by the necessity for large datasets and the intricate balance needed to avoid overfitting, especially crucial for real-time and cross-subject applications.

As BCI research integrates more deeply into human-machine interaction frameworks, the question is no longer if Transformers can be used but rather how they should be implemented to maximize their advantages while mitigating their limitations. For researchers and practitioners contemplating the incorporation of Transformers into BCIs, several key considerations come into play:

1) Is the computational trade-off justified by the enhanced performance?
2) How can we efficiently collect and preprocess large, high-quality BCI datasets that can feed into these demanding models?
3) How can the model's complexity be managed to suit real-time and cross-subject applications?
4) What strategies can be employed to make these models more interpretable, given that interpretability is often crucial for clinical applications?

Answering these questions is crucial for determining the feasibility, efficacy, and broader implications of integrating complex Transformer architectures with the ever-advancing BCI domain.

IX. CONCLUSION

This review has thoroughly examined the applications of transformer models for EEG classification tasks. The unique self-attention mechanism of transformers allows them to handle long-term dependencies in EEG sequences, improving the accuracy of BCIs in various applications such as motor imagery decoding, emotional state recognition, sleep stage classification, and epilepsy prediction.

While integrating transformers into BCIs has many advantages, there are challenges to consider. These include computational intensity, the need for large datasets, and the complexity that may limit real-time applications or cross-subject compatibility. However, ongoing research and development offer reasons for optimism. More resource-efficient transformer variants, opportunities for multi-modal data fusion, and improvements in real-time processing algorithms are emerging. The combination of Transformers with BCIs has the potential to bring significant advancements in sectors such as healthcare, entertainment, and assistive technologies. It demonstrates promising applications in emotion recognition, sleep stage classification, epilepsy prediction,

and other areas within the BCI field through detailed EEG analysis.

APPENDIX
CHECKLIST FOR RESEARCHERS ADOPTING TRANSFORMER MODELS IN BCIS

The incorporation of Transformer models into EEG-based BCIs is not a straightforward task and requires careful consideration of several key aspects to ensure effective and reliable system performance. For researchers interested in leveraging the benefits of Transformers for BCIs, the following questions may serve as a critical checklist:

- **Data Availability**: How much EEG data is available for training the Transformer model?
  - Depending on the complexity of the task, you may need hundreds to thousands of labeled EEG samples. Insufficient data can lead to overfitting and poor generalization to new, unseen data, making this a critical first step.
- **Feature Engineering**: Will the Transformer model handle feature extraction from raw EEG data, or will some form of feature engineering be necessary?
  - Transformers can handle raw EEG data, but preprocessing steps such as filtering and normalization often improve performance. Your choice between manual feature engineering and automated extraction will significantly impact model complexity and interpretability.
- **Noise Handling**: How well can the Transformer model adapt to noisy EEG signals and artifacts?
  - Transformers can be sensitive to noise; consider preprocessing techniques or noise-reduction layers. Robust handling of noise is essential for the model to be applicable in real-world, noisy conditions typically encountered in EEG data.
- **Inter-Subject Variability**: How does the model perform across different subjects, and is there a need for domain adaptation or transfer learning?
  - Transfer learning or domain adaptation techniques may be necessary for better performance across subjects. The ability to generalize across subjects is crucial for building models that are more widely applicable and cost-effective.
- **Real-Time Processing**: Can the Transformer model process EEG data in real-time, considering its complexity and computational requirements?
  - Real-time processing is possible but may require hardware acceleration due to the Transformer's computational complexity. Real-time processing is vital for interactive applications such as neuroprosthetics or live emotional feedback systems.
- **Temporal Dynamics**: How effectively can the Transformer model capture temporal dependencies in EEG signals?
  - Transformers excel in capturing long-term dependencies, but attention mechanisms should be appropriately configured. Temporal information is often key in EEG-based tasks, such as sleep stage classification or motor imagery tasks.
- **Model Complexity**: Given the Transformer architecture's complexity, how will it impact computational efficiency and deployment feasibility?
  - Deploying the model on low-resource devices might require lightweight versions of the Transformer model. Computational efficiency is paramount for real-world applications where resources may be limited.
- **Generalization**: Can the model generalize well to new, unseen data or different BCI tasks?
  - Regularization techniques and data augmentation can improve the model's ability to generalize to new tasks or data. The utility of a model increases significantly if it can adapt to new, unseen conditions.
- **Hyperparameter Tuning**: What hyperparameters are most crucial for this application, and how should they be selected?
  - Attention heads, the number of layers, and learning rates are critical hyperparameters. Their optimal settings can significantly impact the model's performance, making this an important aspect of model tuning.
- **Evaluation Metrics**: What metrics will be used to evaluate the model's performance, such as accuracy, precision, recall, and computational time?
  - Consider using a combination of accuracy, precision, recall, and F1-score, but also explore domain-specific measures when needed. The choice of metrics provides a nuanced understanding of both the model's strengths and weaknesses.
- **Comparison with Existing Models**: How does the Transformer-based approach compare with existing methods such as CNNs, RNNs, or traditional machine learning algorithms in terms of performance, interpretability, and usability?
  - Performance comparisons with state-of-the-art models in terms of accuracy, computational time, and interpretability are essential. This ensures that the Transformer model either outperforms or offers specific advantages over existing methods.
- **Interpretability**: How do transformer models ensure interpretability in EEG and BCIs, and why is it essential?
  - To enhance the interpretability of transformer models, researchers can employ attention visualization and feature attribution methods, crucial for validating model decisions in critical applications.
- **User Experience**: How user-friendly is the BCI system with the Transformer model, and how much calibration is required for a new user?
  - The system should require minimal calibration and provide intuitive feedback to users. A user-friendly system is more likely to gain acceptance and be adopted for practical applications.
- **Scalability**: How scalable is the model in terms of adding more EEG channels or dealing with longer time series data?

- While Transformers scale well with more data, they may require proportionally more computational resources. The model's scalability is vital when extending the application to more complex tasks or larger datasets.

Considering these questions can provide valuable insights into the applicability and limitations of using Transformer models in BCIs.

REFERENCES

[1] J. J. Vidal, "Toward direct brain–computer communication," *Annu. Rev. Biophys. Bioeng.*, vol. 2, no. 1, pp. 157–180, Jun. 1973.
[2] M. A. Lebedev and M. A. L. Nicolelis, "Brain–machine interfaces: Past, present and future," *Trends Neurosci.*, vol. 29, no. 9, pp. 536–546, Sep. 2006.
[3] S. Makeig, C. Kothe, T. Mullen, N. Bigdely-Shamlo, Z. Zhang, and K. Kreutz-Delgado, "Evolving signal processing for brain–computer interfaces," *Proc. IEEE*, vol. 100, no. Special Centennial Issue, pp. 1567–1584, May 2012.
[4] J. L. Collinger and D. J. Krusienski, "The 8th international brain–computer interface meeting, BCIs: The next frontier," *Brain-Comput. Interfaces*, vol. 9, no. 2, pp. 67–68, Apr. 2022.
[5] A. Bashashati, M. Fatourechi, R. K. Ward, and G. E. Birch, "A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals," *J. Neural Eng.*, vol. 4, no. 2, pp. R32–R57, Jun. 2007.
[6] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update," *J. Neural Eng.*, vol. 15, no. 3, Jun. 2018, Art. no. 031005.
[7] 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.
[8] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: A systematic review," *J. Neural Eng.*, vol. 16, no. 5, Aug. 2019, Art. no. 051001.
[9] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019.
[10] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz, "Review of the BCI competition IV," *Frontiers Neurosci.*, vol. 6, p. 55, 2012.
[11] R. Schwartz, J. Dodge, N. Smith, and O. Etzioni, "Green AI," *Commun. ACM*, vol. 63, no. 12, pp. 54–63, Dec. 2020.
[12] J. L. Collinger, B. Wodlinger, J. E. Downey, W. Wang, E. C. Tyler-Kabara, D. J. Weber, A. J. Mcmorland, M. Velliste, M. L. Boninger, and A. B. Schwartz, "High-performance neuroprosthetic control by an individual with tetraplegia," *Lancet*, vol. 381, no. 9866, pp. 557–564, Feb. 2013.
[13] L. R. Hochberg, D. Bacher, B. Jarosiewicz, N. Y. Masse, J. D. Simeral, J. Vogel, S. Haddadin, J. Liu, S. S. Cash, P. van der Smagt, and J. P. Donoghue, "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm," *Nature*, vol. 485, no. 7398, pp. 372–375, May 2012.
[14] E. C. Leuthardt, G. Schalk, J. R. Wolpaw, J. G. Ojemann, and D. W. Moran, "A brain–computer interface using electrocorticographic signals in humans," *J. Neural Eng.*, vol. 1, no. 2, pp. 63–71, Jun. 2004.
[15] B. He, B. Baxter, B. J. Edelman, C. C. Cline, and W. W. Ye, "Noninvasive brain–computer interfaces based on sensorimotor rhythms," *Proc. IEEE*, vol. 103, no. 6, pp. 907–925, Jun. 2015, doi: 10.1109/JPROC.2015.2407272.
[16] A. Gunduz and G. Schalk, "Ecog-based BCIs," in *Brain-Computer Interfaces Handbook*. Boca Raton, FL, USA: CRC Press, 2018, pp. 297–322.
[17] J. Mellinger, G. Schalk, C. Braun, H. Preissl, W. Rosenstiel, N. Birbaumer, and A. Kübler, "An MEG-based brain–computer interface (BCI)," *NeuroImage*, vol. 36, no. 3, pp. 581–593, Jul. 2007.
[18] B. Sorger, J. Reithler, B. Dahmen, and R. Goebel, "A real-time fMRI-based spelling device immediately enabling robust motor-independent communication," *Current Biol.*, vol. 22, no. 14, pp. 1333–1338, Jul. 2012.
[19] N. Naseer and K.-S. Hong, "fNIRS-based brain–computer interfaces: A review," *Frontiers Hum. Neurosci.*, vol. 9, p. 3, Jan. 2015.
[20] N. A. Bhagat, A. Venkatakrishnan, B. Abibullaev, E. J. Artz, N. Yozbatiran, A. A. Blank, J. French, C. Karmonik, R. G. Grossman, M. K. O'Malley, G. E. Francisco, and J. L. Contreras-Vidal, "Design and optimization of an EEG-based brain machine interface (BMI) to an upper-limb exoskeleton for stroke survivors," *Frontiers Neurosci.*, vol. 10, p. 122, Mar. 2016.
[21] A. Venkatakrishnan, G. E. Francisco, and J. L. Contreras-Vidal, "Applications of brain-machine interface systems in stroke recovery and rehabilitation," *Current Phys. Med. Rehabil. Rep.*, vol. 2, pp. 93–105, 2014.
[22] D. Marshall, D. Coyle, S. Wilson, and M. Callaghan, "Games, gameplay, and BCI: The state of the art," *IEEE Trans. Comput. Intell. AI Games*, vol. 5, no. 2, pp. 82–99, Jun. 2013.
[23] B. Kerous, F. Skola, and F. Liarokapis, "EEG-based BCI and video games: A progress report," *Virtual Reality*, vol. 22, no. 2, pp. 119–135, Jun. 2018.
[24] J. Pan, X. Chen, N. Ban, J. He, J. Chen, and H. Huang, "Advances in P300 brain–computer interface spellers: Toward paradigm design and performance evaluation," *Frontiers Hum. Neurosci.*, vol. 16, Dec. 2022, Art. no. 1077717.
[25] M. Vilela and L. R. Hochberg, "Applications of brain–computer interfaces to the control of robotic and prosthetic arms," in *Handbook of Clinical Neurology*, vol. 168. Amsterdam, The Netherlands: Elsevier, 2020, pp. 87–99.
[26] C. Guger, B. Z. Allison, and A. Gunduz, "Brain–computer interface research: A state-of-the-art summary 10," in *Brain-Computer Interface Research*. Berlin, Germany: Springer, 2021.
[27] 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, Jul. 2010.
[28] 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.
[29] L. M. M. Roijendijk, "Variability and nonstationarity in brain computer interfaces," M.S. thesis, Radboud Univ., Nijmegen, The Netherlands, 2009.
[30] B. Abibullaev and A. Zollanvari, "Learning discriminative spatiospectral features of ERPs for accurate brain–computer interfaces," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 5, pp. 2009–2020, Sep. 2019.
[31] X. Gao, Y. Wang, X. Chen, and S. Gao, "Interface, interaction, and intelligence in generalized brain–computer interfaces," *Trends Cognit. Sci.*, vol. 25, no. 8, pp. 671–684, Aug. 2021.
[32] M. Kim, S. Yoo, and C. Kim, "Miniaturization for wearable EEG systems: Recording hardware and data processing," *Biomed. Eng. Lett.*, vol. 12, no. 3, pp. 239–250, Aug. 2022.
[33] G. Niso, E. Romero, J. T. Moreau, A. Araujo, and L. R. Krol, "Wireless EEG: A survey of systems and studies," *NeuroImage*, vol. 269, Apr. 2023, Art. no. 119774.
[34] W. Byun, M. Je, and J.-H. Kim, "Advances in wearable brain–computer interfaces from an algorithm-hardware co-design perspective," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 7, pp. 3071–3077, Jul. 2022.
[35] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.
[36] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Computer*, vol. 53, no. 9, p. 17, Sep. 2020.
[37] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.
[38] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," 2015, *arXiv:1511.06448*.
[39] B. Abibullaev, K. Kunanbayev, and A. Zollanvari, "Subject-independent classification of P300 event-related potentials using a small number of training subjects," *IEEE Trans. Human-Mach. Syst.*, vol. 52, no. 5, pp. 843–854, Oct. 2022.
[40] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.

[41] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," *J. Neural Eng.*, vol. 14, no. 1, Feb. 2017, Art. no. 016003.

[42] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *J. Neural Eng.*, vol. 16, no. 3, Jun. 2019, Art. no. 031001.

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

[44] N. Robinson, R. Mane, T. Chouhan, and C. Guan, "Emerging trends in BCI-robotics for motor control and rehabilitation," *Current Opinion Biomed. Eng.*, vol. 20, Dec. 2021, Art. no. 100354.

[45] T. Karácsony, J. P. Hansen, H. K. Iversen, and S. Puthusserypady, "Brain computer interface for neuro-rehabilitation with deep learning classification and virtual reality feedback," in *Proc. 10th Augmented Hum. Int. Conf.*, Mar. 2019, pp. 1–8.

[46] V. More, M. A. Khalil, and K. George, "Using motor imagery and deeping learning for brain–computer interface in video games," in *Proc. IEEE World AI IoT Congr. (AIIoT)*, Jun. 2023, pp. 0711–0716.

[47] B. Liu, "Deep learning for meditation's impact on brain–computer interface performance," in *Proc. Int. Commun. Eng. Cloud Comput. Conf. (CECCC)*, Oct. 2022, pp. 64–69.

[48] F.-L. Fan, J. Xiong, M. Li, and G. Wang, "On interpretability of artificial neural networks: A survey," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 5, no. 6, pp. 741–760, Nov. 2021.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017.

[50] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[51] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019, *arXiv:1910.10683*.

[52] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[53] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[54] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh, "Large batch optimization for deep learning: Training BERT in 76 minutes," 2019, *arXiv:1904.00962*.

[55] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into deep learning," 2021, *arXiv:2106.11342*.

[56] Y. Dai, X. Li, S. Liang, L. Wang, Q. Duan, H. Yang, C. Zhang, X. Chen, L. Li, X. Li, and X. Liao, "MultiChannelSleepNet: A transformer-based model for automatic sleep stage classification with PSG," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 9, pp. 4204–4215, Sep. 2023.

[57] Z. Miao, M. Zhao, X. Zhang, and D. Ming, "LMDA-Net: A lightweight multi-dimensional attention network for general EEG-based brain–computer interfaces and interpretability," *NeuroImage*, vol. 276, Aug. 2023, Art. no. 120209.

[58] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, Oct. 2022.

[59] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.

[60] 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.

[61] C. Li, Z. Zhang, X. Zhang, G. Huang, Y. Liu, and X. Chen, "EEG-based emotion recognition via transformer neural architecture search," *IEEE Trans. Ind. Informat.*, vol. 19, no. 4, pp. 6016–6025, Apr. 2023.

[62] M. Sun, W. Cui, S. Yu, H. Han, B. Hu, and Y. Li, "A dual-branch dynamic graph convolution based adaptive TransFormer feature fusion network for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2218–2228, Oct. 2022.

[63] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[64] J. R. Wolpaw and D. J. McFarland, "Control of a two-dimensional movement signal by a noninvasive brain–computer interface in humans," *Proc. Nat. Acad. Sci.*, vol. 101, no. 51, pp. 17849–17854, Dec. 2004.

[65] J. Wolpaw and E. W. Wolpaw, *Brain–Computer Interfaces: Principles and Practice*. Oxford, U.K.: OUP, 2012.

[66] E. E. Fetz, "Operant conditioning of cortical unit activity," *Science*, vol. 163, no. 3870, pp. 955–958, Feb. 1969.

[67] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor, "A spelling device for the paralysed," *Nature*, vol. 398, no. 6725, pp. 297–298, Mar. 1999.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Dec. 2015, *arXiv:1512.03385*.

[69] M. Gerven, J. Farquhar, R. Schaefer, R. Vlek, J. Geuze, A. Nijholt, N. Ramsey, P. Haselager, L. Vuurpijl, and S. Gielen, "The brain–computer interface cycle," *J. Neural. Eng.*, vol. 6, no. 4, pp. 1–9, 2009.

[70] G. Buzsáki, C. A. Anastassiou, and C. Koch, "The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes," *Nature Rev. Neurosci.*, vol. 13, no. 6, pp. 407–420, 2012.

[71] E. Niedermeyer and F. L. da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2005.

[72] W. Zhang, C. Tan, F. Sun, H. Wu, and B. Zhang, "A review of EEG-based brain–computer interface systems design," *Brain Sci. Adv.*, vol. 4, no. 2, pp. 156–167, Dec. 2018.

[73] A. Widmann, E. Schröger, and B. Maess, "Digital filter design for electrophysiological data—A practical approach," *J. Neurosci. Methods*, vol. 250, pp. 34–46, Jul. 2015.

[74] M. K. Islam, A. Rastegarnia, and Z. Yang, "Methods for artifact detection and removal from scalp EEG: A review," *Neurophysiologie Clinique/Clinical Neurophysiol.*, vol. 46, nos. 4–5, pp. 287–305, Nov. 2016.

[75] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[76] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," *Clin. Neurophysiol.*, vol. 110, no. 11, pp. 1842–1857, Nov. 1999.

[77] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoustics*, vol. 15, no. 2, pp. 70–73, Jun. 1967.

[78] P. L. Nunez and R. Srinivasan, *Electric Fields of the Brain: The Neurophysics of EEG*. London, U.K.: Oxford Univ. Press, 2006.

[79] S. J. Luck, *An Introduction to the Event-Related Potential Technique*. Cambridge, MA, USA: MIT Press, 2014.

[80] H. Gastaut, R. Naquet, and Y. Gastaut, "A study of mu rhythm in subjects lacking one or more limbs," in *Electroencephalogr. Clin. Neurophysiol.*, vol. 18, no. 7, p. 720, 1965.

[81] C. Neuper, A. Schlögl, and G. Pfurtscheller, "Enhancement of left-right sensorimotor EEG differences during feedback-regulated motor imagery," *J. Clin. Neurophysiol.*, vol. 16, no. 4, pp. 373–382, Jul. 1999.

[82] U. Chaudhary, N. Birbaumer, and A. Ramos-Murguialday, "Brain–computer interfaces in the completely locked-in state and chronic stroke," *Prog. Brain Res.*, vol. 228, pp. 131–161, Jan. 2016.

[83] M. J. Young, D. J. Lin, and L. R. Hochberg, "Brain–computer interfaces in neurorecovery and neurorehabilitation," *Seminars Neurol.*, vol. 41, no. 2, pp. 206–216, Apr. 2021.

[84] A. J. Suminski, D. C. Tkach, A. H. Fagg, and N. G. Hatsopoulos, "Incorporating feedback from multiple sensory modalities enhances brain–machine interface control," *J. Neurosci.*, vol. 30, no. 50, pp. 16777–16787, Dec. 2010.

[85] M. M. Shanechi, "Brain–machine interfaces from motor to mood," *Nature Neurosci.*, vol. 22, no. 10, pp. 1554–1564, Oct. 2019.

[86] P. D. Ganzer, S. C. Colachis, M. A. Schwemmer, D. A. Friedenberg, C. F. Dunlap, C. E. Swiftney, A. F. Jacobowitz, D. J. Weber, M. A. Bockbrader, and G. Sharma, "Restoring the sense of touch using a sensorimotor demultiplexing neural interface," *Cell*, vol. 181, no. 4, pp. 763.e12–773.e12, May 2020.

[87] N. Birbaumer, "Breaking the silence: Brain–computer interfaces (BCI) for communication and motor control," *Psychophysiology*, vol. 43, no. 6, pp. 517–532, Nov. 2006.
[88] D. J. McFarland, "Brain–computer interfaces for amyotrophic lateral sclerosis," *Muscle Nerve*, vol. 61, no. 6, pp. 702–707, 2020.
[89] E. Sellers and E. Donchin, "A P300-based brain–computer interface: Initial tests by ALS patients," *Clin. Neurophysiol., Off. J. Int. Fed. Clin. Neurophysiol.*, vol. 117, pp. 538–548, Apr. 2006.
[90] J. Marín-Morales, C. Llinares, J. Guixeres, and M. Alcañiz, "Emotion recognition in immersive virtual reality: From statistics to affective computing," *Sensors*, vol. 20, no. 18, p. 5163, Sep. 2020.
[91] F. Dehais, A. Lafont, R. Roy, and S. Fairclough, "A neuroergonomics approach to mental workload, engagement and human performance," *Frontiers Neurosci.*, vol. 14, p. 268, Apr. 2020.
[92] M.-P. Deiber, R. Hasler, J. Colin, A. Dayer, J.-M. Aubry, S. Baggio, N. Perroud, and T. Ros, "Linking alpha oscillations, attention and inhibitory control in adult ADHD with EEG neurofeedback," *NeuroImage, Clin.*, vol. 25, Jan. 2020, Art. no. 102145.
[93] Z. Wang, Y. Yu, M. Xu, Y. Liu, E. Yin, and Z. Zhou, "Towards a hybrid BCI gaming paradigm based on motor imagery and SSVEP," *Int. J. Hum.-Comput. Interact.*, vol. 35, no. 3, pp. 197–205, Feb. 2019.
[94] G. A. M. Vasiljevic and L. C. de Miranda, "Brain–computer interface games based on consumer-grade EEG devices: A systematic literature review," *Int. J. Hum.-Comput. Interact.*, vol. 36, no. 2, pp. 105–142, Jan. 2020.
[95] E. M. Holz, J. Höhne, P. Staiger-Sälzer, M. Tangermann, and A. Kübler, "Brain–computer interface controlled gaming: Evaluation of usability by severely motor restricted end-users," *Artif. Intell. Med.*, vol. 59, no. 2, pp. 111–120, Oct. 2013.
[96] A. Nijholt, D. P.-O. Bos, and B. Reuderink, "Turning shortcomings into challenges: Brain–computer interfaces for games," *Entertainment Comput.*, vol. 1, no. 2, pp. 85–94, Apr. 2009.
[97] J. H. Gruzelier, "EEG-neurofeedback for optimising performance. I: A review of cognitive and affective outcome in healthy participants," *Neurosci. Biobehav. Rev.*, vol. 44, pp. 124–141, Jul. 2014.
[98] R. Sitaram, T. Ros, L. Stoeckel, S. Haller, F. Scharnowski, J. Lewis-Peacock, N. Weiskopf, M. L. Blefari, M. Rana, E. Oblak, N. Birbaumer, and J. Sulzer, "Closed-loop brain training: The science of neurofeedback," *Nature Rev. Neurosci.*, vol. 18, no. 2, pp. 86–100, Feb. 2017.
[99] G. Papanastasiou, A. Drigas, C. Skianis, and M. Lytras, "Brain computer interface based applications for training and rehabilitation of students with neurodevelopmental disorders. A literature review," *Heliyon*, vol. 6, no. 9, Sep. 2020, Art. no. e04250.
[100] G. Viviani and A. Vallesi, "EEG-neurofeedback and executive function enhancement in healthy adults: A systematic review," *Psychophysiology*, vol. 58, no. 9, p. e13874, Sep. 2021.
[101] A. U. Patil, D. Madathil, Y.-T. Fan, O. J. L. Tzeng, C.-M. Huang, and H.-W. Huang, "Neurofeedback for the education of children with ADHD and specific learning disorders: A review," *Brain Sci.*, vol. 12, no. 9, p. 1238, Sep. 2022.
[102] D. Kim, J. Lee, Y. Woo, J. Jeong, C. Kim, and D.-K. Kim, "Deep learning application to clinical decision support system in sleep stage classification," *J. Personalized Med.*, vol. 12, no. 2, p. 136, Jan. 2022.
[103] D. Jiang, Y.-N. Lu, Y. Ma, and Y. Wang, "Robust sleep stage classification with single-channel EEG signals using multimodal decomposition and HMM-based refinement," *Expert Syst. Appl.*, vol. 121, pp. 188–203, May 2019.
[104] M.-P. Hosseini, D. Pompili, K. Elisevich, and H. Soltanian-Zadeh, "Optimized deep learning for EEG big data and seizure prediction BCI via Internet of Things," *IEEE Trans. Big Data*, vol. 3, no. 4, pp. 392–404, Dec. 2017.
[105] O. S. Lih, V. Jahmunah, E. E. Palmer, P. D. Barua, S. Dogan, T. Tuncer, S. García, F. Molinari, and U. R. Acharya, "EpilepsyNet: Novel automated detection of epilepsy using transformer model with EEG signals from 121 patient population," *Comput. Biol. Med.*, vol. 164, Sep. 2023, Art. no. 107312.
[106] A. Murphy, B. Bohnet, R. McDonald, and U. Noppeney, "Decoding part-of-speech from human EEG signals," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2201–2210.
[107] G. Krishna, C. Tran, M. Carnahan, and A. H. Tewfik, "EEG based continuous speech recognition using transformers," May 2020, *arXiv:2001.00501*.
[108] A. Kamble, P. H. Ghare, V. Kumar, A. Kothari, and A. G. Keskar, "Spectral analysis of EEG signals for automatic imagined speech recognition," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–9, 2023.
[109] R. Oostenveld and P. Praamstra, "The five percent electrode system for high-resolution EEG and ERP measurements," *Clin. Neurophysiol.*, vol. 112, no. 4, pp. 713–719, Apr. 2001.
[110] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis," *Brain Res. Rev.*, vol. 29, nos. 2–3, pp. 169–195, Apr. 1999.
[111] E. S. Kappenman and S. J. Luck, "The effects of electrode impedance on data quality and statistical significance in ERP recordings," *Psychophysiology*, vol. 47, no. 5, pp. 888–904, 2010.
[112] M. De Vos, K. Gandras, and S. Debener, "Towards a truly mobile auditory brain–computer interface: Exploring the P300 to take away," *Int. J. Psychophysiol.*, vol. 91, no. 1, pp. 46–53, Jan. 2014.
[113] J. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller, and T. Vaughan, "Brain–computer interfaces for communication and control," *Clin. Neurophys.*, vol. 113, no. 6, pp. 767–791, Jun. 2002.
[114] A. Jackson and E. E. Fetz, "Compact movable microwire array for long-term chronic unit recording in cerebral cortex of primates," *J. Neurophysiol.*, vol. 98, no. 5, pp. 3109–3118, Nov. 2007.
[115] A. B. Rapeaux and T. G. Constandinou, "Implantable brain machine interfaces: First-in-human studies, technology challenges and trends," *Current Opinion Biotechnol.*, vol. 72, pp. 102–111, Dec. 2021.
[116] G. Székely, "An approach to the complexity of the brain," *Brain Res. Bull.*, vol. 55, no. 1, pp. 11–28, May 2001.
[117] T. Takahashi, "Complexity of spontaneous brain activity in mental disorders," *Prog. Neuro-Psychopharmacol. Biol. Psychiatry*, vol. 45, pp. 258–266, Aug. 2013.
[118] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *Science*, vol. 320, no. 5880, pp. 1191–1195, May 2008.
[119] O. Sporns, *Networks of the Brain*. Cambridge, MA, USA: MIT Press, 2016.
[120] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Shelter Island, NY, USA: MIT Press, 2016.
[121] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," Apr. 2014, *arXiv:1206.5538*.
[122] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[123] R. Ma, T. Yu, X. Zhong, Z. L. Yu, Y. Li, and Z. Gu, "Capsule network for ERP detection in brain–computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 718–730, 2021.
[124] J. León, J. J. Escobar, A. Ortiz, J. Ortega, J. González, P. Martín-Smith, J. Q. Gan, and M. Damas, "Deep learning for EEG-based motor imagery classification: Accuracy-cost trade-off," *PLoS ONE*, vol. 15, no. 6, Jun. 2020, Art. no. e0234178.
[125] M. Zabcikova, Z. Koudelkova, R. Jasek, and J. J. L. Navarro, "Recent advances and current trends in brain–computer interface research and their applications," *Int. J. Develop. Neurosci.*, vol. 82, no. 2, pp. 107–123, Apr. 2022.
[126] B. Abibullaev and A. Zollanvari, "A systematic deep learning model selection for P300-based brain–computer interfaces," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 52, no. 5, pp. 2744–2756, May 2022.
[127] Y. Zhang, P. Tino, A. Leonardis, and K. Tang, "A survey on neural network interpretability," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 5, pp. 726–742, Oct. 2021.
[128] Y. Du, Y. Xu, X. Wang, L. Liu, and P. Ma, "EEG temporal–spatial transformer for person identification," *Sci. Rep.*, vol. 12, no. 1, Aug. 2022, Art. no. 14378.
[129] 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.
[130] 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.

[131] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[132] 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.

[133] 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.

[134] 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.

[135] 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.

[136] 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.

[137] 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.

[138] Y. Ma, Y. Song, and F. Gao, "A novel hybrid CNN-transformer model for EEG motor imagery classification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8.

[139] 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.

[140] 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.

[141] *BCI Competition II Data Set III*. Accessed: Oct. 10, 2023. [Online]. Available: https://www.bbci.de/competition/ii/

[142] 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.

[143] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The non-invasive Berlin brain–computer interface: Fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, pp. 539–550, Sep. 2007.

[144] 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)*, May 2022, pp. 404–408.

[145] 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, p. giz002, May 2019.

[146] P. Deny and K. W. Choi, "Hierarchical transformer for brain computer interface," in *Proc. 11th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2023, pp. 1–5.

[147] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[148] 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.

[149] 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.

[150] 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.

[151] 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.

[152] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, J. Schmidt, and M. Shah, "Decoding brain representations by multimodal learning of neural activity and visual features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3833–3849, Nov. 2021.

[153] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: A new approach using Riemannian manifold features," *J. Neural Eng.*, vol. 15, no. 1, Feb. 2018, Art. no. 016002.

[154] Y. Song, X. Jia, L. Yang, and L. Xie, "Transformer-based spatial-temporal feature learning for EEG decoding," 2021, *arXiv:2106.11170*.

[155] 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.

[156] S. M. Jain, "Introduction to transformers," in *Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems*. Berlin, Germany: Springer, 2022, pp. 19–36.

[157] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," Jun. 2019, *arXiv:1901.02860*.

[158] R. Alazrai, M. Abuhijleh, H. Alwanni, and M. I. Daoud, "A deep learning framework for decoding motor imagery tasks of the same hand using EEG signals," *IEEE Access*, vol. 7, pp. 109612–109627, 2019.

[159] H. Altaheri, G. Muhammad, M. Alsulaiman, S. U. Amin, G. A. Altuwaijri, W. Abdul, M. A. Bencherif, and M. Faisal, "Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review," *Neural Comput. Appl.*, vol. 35, no. 20, pp. 14681–14722, Jul. 2023.

[160] S. Gong, K. Xing, A. Cichocki, and J. Li, "Deep learning in EEG: Advance of the last ten-year critical period," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 2, pp. 348–365, Jun. 2022.

[161] C. Mühl, B. Allison, A. Nijholt, and G. Chanel, "A survey of affective brain computer interfaces: Principles, state-of-the-art, and challenges," *Brain-Comput. Interfaces*, vol. 1, no. 2, pp. 66–84, Apr. 2014.

[162] J. Sun, J. Xie, and H. Zhou, "EEG classification with transformer-based models," in *Proc. IEEE 3rd Global Conf. Life Sci. Technol. (LifeTech)*, Mar. 2021, pp. 92–93.

[163] S. Koelstra, A. Yazdani, M. Soleymani, C. Mühl, J.-S. Lee, A. Nijholt, T. Pun, T. Ebrahimi, and I. Patras, "Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos," in *Brain Informatics* (Lecture Notes in Computer Science), Y. Yao, R. Sun, T. Poggio, J. Liu, N. Zhong, and J. Huang, Eds. Berlin, Germany: Springer, 2010, pp. 89–100.

[164] J. LeDoux, "The emotional brain, fear, and the amygdala," *Cellular Mol. Neurobiol.*, vol. 23, nos. 4–5, pp. 727–738, Oct. 2003.

[165] B. Chakravarthi, S.-C. Ng, M. R. Ezilarasan, and M.-F. Leung, "EEG-based emotion recognition using hybrid CNN and LSTM classification," *Frontiers Comput. Neurosci.*, vol. 16, Oct. 2022, Art. no. 1019776.

[166] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, Apr. 2012.

[167] D. Wu, B.-L. Lu, B. Hu, and Z. Zeng, "Affective brain–computer interfaces (aBCIs): A tutorial," *Proc. IEEE*, vol. 111, no. 10, pp. 1314–1332, Oct. 2023.

[168] H. Liu, Y. Zhang, Y. Li, and X. Kong, "Review on emotion recognition based on electroencephalography," *Frontiers Comput. Neurosci.*, vol. 15, p. 84, Oct. 2021.

[169] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[170] G. Xiao, M. Shi, M. Ye, B. Xu, Z. Chen, and Q. Ren, "4D attention-based neural network for EEG emotion recognition," *Cognit. Neurodynamics*, vol. 16, no. 4, pp. 805–818, Aug. 2022.

[171] L. Gong, M. Li, T. Zhang, and W. Chen, "EEG emotion recognition using attention-based convolutional transformer neural network," *Biomed. Signal Process. Control*, vol. 84, Jul. 2023, Art. no. 104835.

[172] W. Lu, T.-P. Tan, and H. Ma, "Bi-branch vision transformer network for EEG emotion recognition," *IEEE Access*, vol. 11, pp. 36233–36243, 2023.

[173] J. Li, W. Pan, H. Huang, J. Pan, and F. Wang, "STGATE: Spatial-temporal graph attention network with a transformer encoder for EEG-based emotion recognition," *Frontiers Hum. Neurosci.*, vol. 17, Apr. 2023, Art. no. 1169949.

[174] Z. Bai, F. Hou, K. Sun, Q. Wu, M. Zhu, Z. Mao, Y. Song, and Q. Gao, "SECT: A method of shifted EEG channel transformer for emotion recognition," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 10, pp. 4758–4767, Oct. 2023.

[175] J.-Y. Guo, Q. Cai, J.-P. An, P.-Y. Chen, C. Ma, J.-H. Wan, and Z.-K. Gao, "A transformer based neural network for emotion recognition and visualizations of crucial EEG channels," *Phys. A, Stat. Mech. Appl.*, vol. 603, Oct. 2022, Art. no. 127700.

[176] J. Sun, X. Wang, K. Zhao, S. Hao, and T. Wang, "Multi-channel EEG emotion recognition based on parallel transformer and 3D-convolutional neural network," *Mathematics*, vol. 10, no. 17, p. 3131, Sep. 2022.

[177] C. Cheng, Y. Zhang, L. Liu, W. Liu, and L. Feng, "Multi-domain encoding of spatiotemporal dynamics in EEG for emotion recognition," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 3, pp. 1342–1353, Mar. 2023.

[178] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.

[179] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 98–107, Jan. 2018.

[180] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis ;Using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[181] S. Koorathota, Z. Khan, P. Lapborisuth, and P. Sajda, "Multimodal neurophysiological transformer for emotion recognition," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 3563–3567.

[182] Z. Wang, Y. Wang, C. Hu, Z. Yin, and Y. Song, "Transformers for EEG-based emotion recognition: A hierarchical spatial information learning model," *IEEE Sensors J.*, vol. 22, no. 5, pp. 4359–4368, Mar. 2022.

[183] A. Arjun, A. S. Rajpoot, and M. Raveendranatha Panicker, "Introducing attention mechanism for EEG signals: Emotion recognition with vision transformers," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 5723–5726.

[184] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[185] Y. Zhou and J. Lian, "Identification of emotions evoked by music via spatial-temporal transformer in multi-channel EEG signals," *Frontiers Neurosci.*, vol. 17, Jul. 2023, Art. no. 1188696.

[186] J. Liu, H. Wu, L. Zhang, and Y. Zhao, "Spatial–temporal transformers for EEG emotion recognition," in *Proc. 6th Int. Conf. Adv. Artif. Intell. (ICAAI)*. New York, NY, USA: Association for Computing Machinery, 2023, pp. 116–120.

[187] S. Sartipi and M. Cetin, "Adversarial discriminative domain adaptation and transformers for EEG-based cross-subject emotion recognition," in *Proc. 11th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Apr. 2023, pp. 1–4.

[188] Y.-E. Lee and S.-H. Lee, "EEG-transformer: Self-attention from transformer architecture for decoding EEG of imagined speech," in *Proc. 10th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2022, pp. 1–4.

[189] R. Hussein, S. Lee, and R. Ward, "Multi-channel vision transformer for epileptic seizure prediction," *Biomedicines*, vol. 10, no. 7, p. 1551, Jun. 2022.

[190] S. Hu, J. Liu, R. Yang, Y. Wang, A. Wang, K. Li, W. Liu, and C. Yang, "Exploring the applicability of transfer learning and feature engineering in epilepsy prediction using hybrid transformer model," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1321–1332, 2023.

[191] D. K. Ravikanti and S. Saravanan, "EEGAlzheimer'sNet: Development of transformer-based attention long short term memory network for detecting Alzheimer disease using EEG signal," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105318.

[192] P. Kaushik, I. Tripathi, and P. P. Roy, "Motor activity recognition using eeg data and ensemble of stacked BLSTM-LSTM network and transformer model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[193] W. Qu, Z. Wang, H. Hong, Z. Chi, D. D. Feng, R. Grunstein, and C. Gordon, "A residual based attention model for EEG based sleep staging," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2833–2843, Oct. 2020.

[194] Z. Yao and X. Liu, "A CNN-transformer deep learning model for real-time sleep stage classification in an energy-constrained wireless device," in *Proc. 11th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Apr. 2023, pp. 1–4.

[195] G. Fiscon, E. Weitschek, A. Cialini, G. Felici, P. Bertolazzi, S. De Salvo, A. Bramanti, P. Bramanti, and M. C. De Cola, "Combining EEG signal processing with supervised methods for Alzheimer's patients classification," *BMC Med. Informat. Decis. Making*, vol. 18, no. 1, Dec. 2018.

[196] S. Panchavati, S. V. Dussen, H. Semwal, A. Ali, J. Chen, H. Li, C. Arnold, and W. Speier, "Pretrained transformers for seizure detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–2.

[197] P. Busia, A. Cossettini, T. M. Ingolfsson, S. Benatti, A. Burrello, M. Scherer, M. A. Scrugli, P. Meloni, and L. Benini, "EEGformer: Transformer-based epilepsy detection on raw EEG traces for low-channel-count wearable continuous monitoring devices," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Burrel, Benin, Oct. 2022, pp. 640–644.

[198] A. Miltiadous, E. Gionanidis, K. D. Tzimourta, N. Giannakeas, and A. T. Tzallas, "DICE-Net: A novel convolution-transformer architecture for Alzheimer detection in EEG signals," *IEEE Access*, vol. 11, pp. 71840–71858, 2023.

[199] G. Shi, Z. Chen, and R. Zhang, "A transformer-based spatial-temporal sleep staging model through raw EEG," in *Proc. Int. Conf. High Perform. Big Data Intell. Syst. (HPBD&IS)*, Dec. 2021, pp. 110–115.

[200] X. Zhang and H. Li, "Patient-specific seizure prediction from scalp EEG using vision transformer," in *Proc. IEEE 6th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, vol. 6, Mar. 2022, pp. 1663–1667.

[201] Y. Li, X. Zhang, and D. Ming, "Early-stage fusion of EEG and fNIRS improves classification of motor imagery," *Frontiers Neurosci.*, vol. 16, Jan. 2023, Art. no. 1062889.

[202] C. J. Markiewicz, K. J. Gorgolewski, F. Feingold, R. Blair, Y. O. Halchenko, E. Miller, N. Hardcastle, J. Wexler, O. Esteban, M. Goncavles, A. Jwa, and R. Poldrack, "The OpenNeuro resource for sharing of neuroscience data," *eLife*, vol. 10, Oct. 2021, Art. no. e71774.

[203] V. Jayaram and A. Barachant, "MOABB: Trustworthy algorithm benchmarking for BCIs," *J. Neural Eng.*, vol. 15, no. 6, Dec. 2018, Art. no. 066011.

[204] P. Dreyer, A. Roc, L. Pillette, S. Rimbert, and F. Lotte, "A large EEG database with users' profile information for motor imagery brain–computer interface research," *Sci. Data*, vol. 10, no. 1, p. 580, Sep. 2023.

[205] F. Fahimi, S. Dosen, K. K. Ang, N. Mrachacz-Kersting, and C. Guan, "Generative adversarial networks-based data augmentation for brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 4039–4051, Sep. 2021.

[206] K. Kunanbayev, B. Abibullaev, and A. Zollanvari, "Data augmentation for P300-based brain–computer interfaces using generative adversarial networks," in *Proc. 9th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2021, pp. 1–7.

[207] S. Dikker, G. Michalareas, M. Oostrik, A. Serafimaki, H. M. Kahraman, M. E. Struiksma, and D. Poeppel, "Crowdsourcing neuroscience: Inter-brain coupling during face-to-face interactions outside the laboratory," *NeuroImage*, vol. 227, Feb. 2021, Art. no. 117436.

[208] D. Wu, Y. Xu, and B.-L. Lu, "Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 1, pp. 4–19, Mar. 2022.

[209] W. Li, W. Huan, B. Hou, Y. Tian, Z. Zhang, and A. Song, "Can emotion be transferred?—A review on transfer learning for EEG-based emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 3, pp. 833–846, Sep. 2022.

[210] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[211] Y. Liu, E. Sangineto, W. Bi, N. Sebe, B. Lepri, and M. Nadai, "Efficient training of visual transformers with small datasets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23818–23830.

[212] C. Zhu, W. Ping, C. Xiao, M. Shoeybi, T. Goldstein, A. Anandkumar, and B. Catanzaro, "Long-short transformer: Efficient transformers for language and vision," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 17723–17736.

[213] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—Explainable artificial intelligence," *Sci. Robot.*, vol. 4, no. 37, p. eaay7120, 2019.

[214] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 782–791.

[215] P. Komorowski, H. Baniecki, and P. Biecek, "Towards evaluating explanations of vision transformers for medical imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 3725–3731.

[216] V. Mun and B. Abibullaev, "Explainable deep learning for brain–computer interfaces through layerwise relevance propagation," in *Proc. 11th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2023, pp. 1–5.

[217] D. Zhang, H. Li, and J. Xie, "MI-CAT: A transformer-based domain adaptation network for motor imagery classification," *Neural Netw.*, vol. 165, pp. 451–462, Aug. 2023.

[218] D. Zhang, H. Li, J. Xie, and D. Li, "MI-DAGSC: A domain adaptation approach incorporating comprehensive information from MI-EEG signals," *Neural Netw.*, vol. 167, pp. 183–198, Oct. 2023.

[219] J. J. Bird, J. Kobylarz, D. R. Faria, A. Ekárt, and E. P. Ribeiro, "Cross-domain MLP and CNN transfer learning for biological signal processing: EEG and EMG," *IEEE Access*, vol. 8, pp. 54789–54801, 2020.

[220] J. Li, F. Wang, H. Huang, F. Qi, and J. Pan, "A novel semi-supervised meta learning method for subject-transfer brain–computer interface," *Neural Netw.*, vol. 163, pp. 195–204, Jun. 2023.

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 was with the Daegu Gyeongbuk Institute of Science and Technology (2010–2013) and the Samsung Medical Center (2013–2014). In 2014, he joined the University of Houston, TX, USA, as a Postdoctoral Research Fellow II, supported by the National Institute of Health. He is currently an Associate Professor with the Robotics Department, Nazarbayev University, Kazakhstan. His research is centered around developing machine learning techniques to solve inference problems in brain–computer interfaces. He is an Associate Editor of IEEE Access.

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. Since 2019, she has been a Research Assistant with the School of Engineering and Digital Sciences, Nazarbayev University, and a member of the Young Researchers Alliance, Astana. Her current research interests include machine learning, brain–computer interfaces, pattern recognition, and digital twins.

AMIN ZOLLANVARI (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Shiraz University, Iran, and the Ph.D. degree in electrical engineering from Texas A&M University, College Station, TX, USA, in 2010. He was a Postdoctoral Researcher with the Harvard Medical School, and Brigham and Women's Hospital, Boston, MA, USA, from 2010 to 2012, and then joined the Department of Statistics, Texas A&M University, as an Assistant Research Scientist, from 2012 to 2014. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Nazarbayev University, Kazakhstan. His research interests include machine learning, signal processing, and biomedical informatics.