IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 September 2023, accepted 23 September 2023, date of publication 28 September 2023,
date of current version 5 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3320561

RESEARCH ARTICLE

Exploring the Potential of Attention Mechanism-Based Deep Learning for Robust Subject-Independent Motor-Imagery Based BCIs

AIGERIM KEUTAYEVA AND BERDAKH ABIBULLAEV, (Senior Member, IEEE)
Robotics Engineering Department, School of Engineering and Digital Sciences, 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 This study explores the use of attention mechanism-based deep learning models to construct subject-independent motor-imagery based brain-computer interfaces (MI-BCIs), which present unique and intricate challenges from a machine learning perspective. By comparing four attention mechanism-based models and employing nested LOSO methods for robust model selection, the study enhances the reliability of performance estimates and offers unique insights into the application of attention mechanisms in building subject-independent BCIs. The results indicate the potential of the Spatio-Temporal CNN + ViT model for practical BCI applications, as it outperforms other models on several datasets. Additionally, the study presents a realistic approach to building subject-independent BCIs by combining attention mechanisms and deep learning models to identify informative features common across subjects while filtering out noise and irrelevant data. While there are limitations and areas for future work to enhance the potential of these models, transformer-based models could become even more valuable in the BCI research field, leading to more robust and accurate subject-independent BCIs for various applications. The need for subject-independent MI-BCIs is amplified due to their potential in assisting individuals with severe neurological conditions, such as ALS and locked-in syndrome, which severely limit mobility and communication.

INDEX TERMS Attention mechanism, brain-computer interface (BCI), deep learning (DL), vision transformers (VT), motor imagery (MI), EEG, subject-independent BCIs, transformers.

I. INTRODUCTION

Brain-Computer interfaces (BCIs) have shown considerable promise in redefining the interaction between humans and external devices. Numerous applications have been explored, from aiding individuals with severe disabilities such as amyotrophic lateral sclerosis (ALS) and locked-in syndrome [1], [2], to the early identification of epileptic seizures [3], [4], [5]. Further applications include the use of advanced prosthetics [6], [7], [8], engagement in gaming and virtual reality [9], [10], as well as advancements in scientific research [6], [11], [12]. Among the methods employed in BCIs, Electroencephalography (EEG) stands as an essential technique. By recording the brain's electrical activity through non-invasive scalp electrodes, EEG provides invaluable real-time data on neural activity [13], [14]. Given its high temporal resolution, EEG is particularly useful for BCI systems that require rapid response times. Despite recent advances, the field still confronts a notable challenge: the variability of EEG data among individuals. Factors contributing to this variability include unique brain structures, differing mental states, and individual head shapes. This makes the task of developing universally applicable BCIs quite challenging. To address these challenges, research has generally split into two main approaches. The first involves subject-dependent models that require individualized

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

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

calibration. Although these models can be quite accurate, they require ongoing calibration for each new user, leading to resource-intensive implementations [15], [16]. The second approach focuses on subject-independent models, aiming to identify features that are consistent across different individuals to avoid the need for calibration [17], [18]. While this approach is less resource-intensive, it often results in compromised performance [19]. Consequently, resolving the issue of subject variability stands as a crucial task, not just as a research query but also as a practical imperative for realizing the comprehensive benefits of BCIs across a range of real-world applications.

"Attention is All You Need" by Vaswani et al. has been influential in the field of attention-based models, particularly due to its introduction of the Transformer architecture [20]. This framework has been successfully applied in a wide range of natural language processing tasks, including machine translation and language modeling. It has also shown promise in areas such as computer vision and EEG classification. Building upon this, a recent study by Sun et al. has made noteworthy contributions to EEG classification by proposing a transformer-based model [21]. Their work offers valuable insights into the role of attention mechanisms in enhancing the generalizability of BCIs. It highlights the merit of continued exploration in this direction, especially given the potential for such models to improve the lives of individuals with disabilities.

In the context of these contributions, this study seeks to examine the application of attention-based deep learning models to create more versatile, subject-independent BCIs. Our methodology involves utilizing attention mechanisms and deep learning to identify and learn the most relevant features that are uniformly present across multiple subjects while discarding noise and other irrelevant information. We assess the efficacy of this approach by analyzing its performance across four distinct Motor Imagery (MI)–based EEG datasets. The availability of such datasets opens up opportunities to further explore the role of attention mechanisms in the creation of more universally applicable BCIs.

Our ultimate goal is to leverage attention-based deep learning models to discern useful patterns in brain signals, with the aim of developing BCIs that are not just technically reliable but also user-centric. This work represents a constructive addition to ongoing efforts to make BCIs more accessible and beneficial, particularly for those with disabilities.

This paper provides several key contributions to the field of subject-independent BCIs, including:

1) Comparative Analysis: A comprehensive comparison of attention mechanism-based models - ViT, Spatial CNN + ViT, Temporal CNN + ViT, and Spatio-Temporal CNN + ViT - was provided. This analysis was in terms of classification accuracy, robustness, computational efficiency, and signal-to-noise ratio, furthering the understanding of the strengths and weaknesses of each model when applied to various BCI tasks and datasets.
2) Model Evaluation: The traditional Leave-one-subject-out (LOSO) and nested LOSO methods were employed for model selection, enhancing the reliability of performance estimates and promoting the development of more accurate and efficient BCIs.
3) Dataset Exploration: Four distinct Motor Imagery (MI) based EEG datasets were investigated. The analysis offered unique insights into the application of attention mechanisms in building subject-independent BCIs using these datasets.
4) Significance for BCI Development: The comparative analysis highlights the distinct performance characteristics of various models, notably highlighting the limitations of the Spatio-Temporal CNN + ViT model on the BCI IV 2a, 2b, and Weibo datasets. Such insights open the way for improving model selection criteria for practical BCI applications.
5) Innovative Approach: Attention mechanisms and deep learning models were combined to identify informative features common across subjects, effectively filtering out noise and irrelevant data.

These contributions presented advancements in subject-independent BCIs and furnished valuable insights for future research in this domain.

The paper is organized as follows: Section II presents a comprehensive literature review of previous work on BCIs and the application of attention mechanism-based deep learning models. Section III provides a detailed overview of the materials and methods used in the study, including the classification methods employed, the datasets used, the data preprocessing steps, the classification architectures developed, the CNN models implemented, and the performance evaluation techniques. Section IV discusses the results obtained from the study, delving into the performance and robustness of the models tested. Finally, Section V offers a summary of the study findings, their implications, and the scope for future work in the area of subject-independent BCIs using attention mechanisms. By breaking down the components of the study in this way, the paper aims to offer a thorough exploration of the potential of attention mechanisms in designing effective BCIs.

II. LITERATURE REVIEW

EEG serves as a useful tool for understanding the activities of neuronal groups within the central nervous system (CNS) and finds widespread application in both neurology and BCIs [22], [23]. Within the BCI ecosystem, the Motor Imagery (MI) paradigm is often utilized, where participants are encouraged to mentally simulate rather than physically perform specific movements [24], [25]. Accurate classification of EEG data derived from different MI tasks is fundamental for enabling BCIs to control various external

devices effectively, particularly in patient rehabilitation scenarios [25]. However, the task is far from straightforward. EEG signals are characterized by substantial variations between individuals, confined spatial resolution, and a high temporal resolution, compounded by a low signal-to-noise ratio [23]. These factors collectively pose challenges to the accurate classification of MI-EEG data, constraining the efficacy of BCI systems and their associated signal-processing techniques.

Feature extraction and classification have long been fundamental elements in traditional machine-learning approaches for classifying MI-EEG data. Conventional techniques for feature extraction often employ methods like Fast Fourier Transform [26] and Wavelet Transform [27]. For feature classification, supervised learning algorithms such as Support Vector Machines (SVM) [28] and Random Forest (RF) [29] are commonly used. However, these approaches risk omitting important EEG data during the feature extraction phase. To address this issue, deep learning models, which are capable of directly processing raw EEG data, have emerged as a promising alternative [23], [25]. In the realms of medicine and neuroscience, convolutional neural networks (CNN) [25], [30] and hybrid CNN architectures [31] have been particularly influential. These CNN models use spatial and temporal kernels to extract features from multiple EEG channels simultaneously or from a single channel across different time intervals [25], [30]. Some studies have even explored the synergy of CNNs and auto-encoders for EEG signal classification [32]. Although CNNs are frequently employed for motor imagery EEG classification, they do have limitations, such as constraints in detecting global dependencies [23], [30].

The attention mechanism serves as a valuable enhancement to the capabilities of deep learning models, drawing parallels to how attention functions in biological neural networks. Specifically, the multi-head attention mechanism, a variant of self-attention, establishes relationships between each pair of temporal points [33]. Unlike traditional recurrent or convolutional layers, the Transformer architecture capitalizes on multi-head attention for feature extraction, as evident in well-known models like BERT [34] and GPT-2 [35]. Initially gaining prominence in the realm of natural language processing (NLP), Transformer-based models have since broadened their utility to include image classification, video processing, speech recognition, and even music generation [36], [37], [38], [39].

This adaptability makes the Transformer particularly promising for addressing the challenges tied to EEG data classification, especially in capturing long-range dependencies [40]. It offers the unique advantage of achieving this without requiring convolutional neural networks (CNNs), thereby providing additional interpretability when compared to other deep learning architectures [21]. However, with this advantage comes the potential issue of prioritizing noisy segments in the EEG data. Additionally, the computational demands of attention mechanisms could present challenges for real-time applications or setups with constrained computational capabilities [41].

TABLE 1. Information on EEG datasets with "L" representing left-hand, and "R" representing right-hand motor imagery tasks.

Dataset	Participants	(trials × ch × time)	Sampling Rate (Hz)	Labels
BCI IV 2a	9	(2592 × 22 × 321)	250	L/R
BCI IV 2b	9	(6520 × 3 × 321)	250	L/R
Weibo	10	(1580 × 60 × 321)	1000	L/R
Physionet	109	(4683 × 64 × 201)	160	L/R

In the context of EEG-based applications, Transformer models have been explored for tasks such as imagined speech recognition, emotion detection, and sleep stage identification [42], [43], [44]. A limited number of studies have attempted to apply Transformer-based models specifically to motor imagery EEG (MI-EEG). Tao et al. [40], for instance, used a gated Transformer on the same PhysioNet dataset as the one examined in this study, although their focus was restricted to a single multi-class classification task. Other researchers such as Du et al. [33] and Kostas et al. [45] have used more complex pipelines involving spatial filtering and two-stage models, but their work often suffered from limitations such as a small number of subjects or lack of generalizability across different individuals.

This work seeks to build upon the efforts of these previous studies, specifically those of Du et al. and Song et al., by implementing the Vision Transformer both with and without CNNs as part of its architecture (see Section III for details). Our aim is to explore whether the Transformer architecture can offer an efficient and generalizable solution for EEG data classification, particularly in the context of subject-independent motor imagery BCIs.

III. MATERIALS AND METHODS

A. DATASET DESCRIPTION

This study utilized publicly available EEG data from multiple sources to analyze motor-imagery tasks performed by various subjects using chosen neural networks. The investigation focused on decoding motor imagery data for left-hand and right-hand tasks, and accordingly, datasets were selected. Additionally, the continuous EEG data was partitioned into separate 4-second trials for left-hand and right-hand imagery tasks, following the onset of mental imagery, as described in a paper by Abibullaev et al. [46]. Details on each dataset are provided in the following sections.

1) WEIBO2014

This dataset comprises EEG recordings from a sample of ten healthy right-handed participants, seven females and three males aged between 23 and 25 years. A 64-channel Neuroscan SynAmps2 amplifier was used to record EEG data at a sampling rate of 1000 Hz. The data was then band-pass filtered between 0.5 and 50 Hz and down-sampled to 200 Hz for analysis. The nose and grounded prefrontal

lobe were used as reference points for the international 10-20 electrode implantation method. Dataset was initially gathered to compare simple and complicated limb motor imagery's effects on EEG patterns. A red circle served as a visual indication for the participants during each eight-second trial, which was then followed by a cue suggesting either left- or right-hand imagery. Participants engaged in kinesthetic motor imagery for about four seconds during this time. Data were collected across nine sessions, each consisting of 60 trials of motor imagery data per task (for further details, refer to [47]).

2) PHYSIONET

The Physionet Motor/Mental Imagery database is a vast and comprehensive collection of motor-imagery EEG data. The data was obtained from a group of 109 participants, where they were asked to do motor/imagery tasks while the EEG signal was recorded by the BCI2000 system using the 10-10 electrode placement scheme. A total of 64 channels of EEG data were recorded per participant, with a sampling rate of 160 Hz. 14 experimental runs totaling two 1-min baseline runs and 12 2-min task runs were completed by each participant. In the task runs, participants were required to execute four motor/imaginary tasks, including performing the corresponding physical action when the target appeared on the computer, visualizing acting, and then resting when the target vanished. In this study, we concentrated on the classification of the motor imagery. We chose three task runs for the motor imagery of the left hand vs. the right hand, where participants had to visualize opening and clenching the matching hand while a target appeared on the left or right side of the computer screen. Each participant was assigned 46 trials for both right and left-hand tasks. Please see [48] and [49] for further details.

3) BCI 2A - DATASET IIA FROM BCI COMPETITION 4

This dataset was obtained using a cue-based brain-computer interface (BCI) paradigm from a sample of nine participants who performed four different types of imagery tasks involving left-hand, right-hand, both feet, and tongue movements. Data were gathered over the course of two sessions, each consisting of 288 trials for a specific visual task. Each session began with a fixation cross displayed on a blank screen, followed by a visual signal in the shape of an arrow pointing left, right, up, or down, shown for 1.25 seconds. Participants were instructed to carry out motor imagery exercises until six seconds had passed since the fixation cross last displayed on the screen. Monopolar EEG data was captured using 22 Ag/AgCl electrodes based on the 10-20 electrode positioning system. The right mastoid was used as the ground electrode, and the left mastoid served as the reference electrode. Data were bandpass filtered from 0.5 Hz to 100 Hz and sampled at 250 Hz. In this study, only left-hand and right-hand tasks were used. For further details, please refer to [50].

4) BCI 2B - DATASET IIB FROM BCI COMPETITION 4

This dataset contains EEG data collected from a sample of nine healthy, right-handed participants. EEG data from three channels (C3, Cz, and C4) were captured at a sampling rate of 250 Hz and bandpass filtered between 0.5 Hz and 100 Hz. Five data collection sessions were conducted on the participants, with the first two sessions occurring without feedback and the final three with feedback. The Fz electrode was used as an EEG ground. During the study, participants engaged in two-class motor imagery tasks involving left- and right-hand motions, using a cue-based paradigm. Each motor imagery task consisted of 120 trials per session. Each trial began with a fixation cross and a brief warning tone, followed by a 1.25-second visual cue arrow. The motor imagery tasks lasted four seconds, with a brief intermission of 1.5 seconds between each set of trials. In sessions with feedback, participants were instructed to perform the motor imagery tasks after a cheerful face appeared on the screen for 3 to 7.5 seconds. For further details, refer to [51].

B. DATA PREPROCESSING

Deep learning research has demonstrated an increased efficacy in directly learning from raw EEG data, reducing the necessity for elaborate preprocessing or feature engineering [23], [52], [53]. In alignment with this, we employed a basic preprocessing strategy on the four datasets detailed in the preceding section. For more insights into the preprocessing methods and the effectiveness of the deep learning models used, we direct readers to consult [52], [54].

Specifically, we applied a high-pass filter to the EEG waveforms with a cut-off frequency of 4 Hz, using a fourth-order Butterworth IIR filter. This filtering was performed to minimize electro-oculographic artifacts caused by eye movements, which are typically dominant in the 0.1 to 4 Hz frequency range within EEG recordings. Apart from this, and in line with the recommendations from [52], we refrained from applying low-pass filtering to preserve the integrity of the raw EEG data. Additionally, the continuous EEG recordings were divided into left-hand and right-hand motor imagery trials, each lasting four seconds and commencing at the onset of the motor imagery. These segmented EEG data trials were then subjected to artifact correction through a statistical thresholding technique, which filtered out: (i) trials with prominent movement-related noise and (ii) channels with potential poor scalp connectivity that resulted in noise. To identify these, the mean absolute value per trial was computed, and any trials with values exceeding three standard deviations above the mean were excluded. All the methodologies outlined here were implemented in the MNE Python framework [55]. Our approach to minimal preprocessing allows the deep learning models in this study to automatically learn meaningful features from EEG data that is nearly raw, thereby eliminating the need for prior assumptions or extensive preprocessing.

Addressing the common issue of limited trials in motor imagery (MI) datasets, a range of data augmentation strategies were employed. Traditional methods, such as cropping or adding Gaussian noise, can degrade the signal-to-noise ratio or risk compromising the inherent coherence.

To generate new augmented data for MI datasets, which typically have fewer trials, this work utilizes a segmentation and reconstruction (S&R) technique in the time domain. The method presented in [41] partitions training data into Ns segments, which are then randomly concatenated while preserving the original time sequence. Here, Ns was set to 3, as time samples from EEG data, such as 321 and 201, are divisible by 3. When combined with the original data, the augmented data effectively doubles the dataset size because it is generated with each iteration to match the batch size. Additionally, Z-score normalization was applied to reduce the non-stationarity of the EEG data before data augmentation, as shown in the following equation:

$$x_i = \frac{x_0 - \mu}{\sigma} \tag{1}$$

where x_0 is the original EEG data, and x_i is the data after standardization. μ and σ represent the mean and standard deviation.

C. DEEP LEARNING MODELS

In this section, we will discuss the use of multiple deep-learning models.[1]

1) CONVOLUTIONAL NEURAL NETWORKS FOR EEG-BASED MI TASKS

Convolutional Neural Networks (CNNs) have gained recognition for their effectiveness in analyzing EEG data, particularly in the context of Motor Imagery (MI) tasks [46]. Unlike traditional Artificial Neural Networks (ANNs), CNNs incorporate specialized layers–namely convolutional and pooling layers–alongside fully-connected layers which serve as classifiers [25]. The convolutional layers utilize a sliding window approach, using filters or kernels, to identify intricate patterns within the data. These filters are designed to recognize progressively more complex features as we delve deeper into the network [56]. Pooling layers further simplify the data by employing kernels to perform operations like max-pooling or average-pooling, thereby reducing computational demands [46], [56].

In this study, we employ CNNs due to their ability to automatically learn high-level features from EEG data, which are crucial for accurate MI classification.

a: MATHEMATICAL FORMULATION

Let's consider a 3D input tensor $\mathbf{X}$ of dimensions $c_{\mathbf{X}} \times h_{\mathbf{X}} \times w_{\mathbf{X}}$, where $c_{\mathbf{X}} = 1$ is analogous to a grayscale image and $h_{\mathbf{X}}$ corresponds to the number of EEG channels. The convolutional layer employs a 4D kernel tensor $\mathbf{K}$ of dimensions $c_{\mathbf{Y}} \times c_{\mathbf{X}} \times h_{\mathbf{K}} \times w_{\mathbf{K}}$ to compute a 3D output tensor $\mathbf{Y}$, following the equation:

$$\mathbf{Y}_{l,m,n} = \sum_{i,j,k} \mathbf{X}_{i,m+j-1,n+k-1} \mathbf{K}_{l,i,j,k} \tag{2}$$

In a multi-layer CNN tailored for EEG-based MI classification, the output tensor from one layer serves as the input tensor for the subsequent layer. Each convolutional layer consists of $c_{\mathbf{Y}}(c_{\mathbf{X}} \times h_{\mathbf{K}} \times w_{\mathbf{K}} + 1)$ learnable parameters, enabling the network to adapt to the unique characteristics of EEG data [57].

2) TRANSFORMERS MODELS

When using RNN and LSTM models for sequence-to-sequence NLP tasks, issues with long-range dependencies need to be addressed [20]. Unlike RNNs, Transformers do not have recurrent connections to record distant relationships. Instead, they use self-attention to process complete sequences simultaneously and deal with the memory problem. One potential solution to the variability problem in BCIs is to use attention mechanisms in transformer models. These mechanisms allow deep learning models to focus on important input data features, thus reducing noise and variability effects. The combined methods used to create the original transformer are discussed below, followed by an explanation of the transformer's overall architecture.

1) **Tokenization:** A conversion into numbers is necessary for a computer to comprehend natural language. An extremely basic illustration of such a transformation from the NLP field entails giving each word a unique number and then altering input phrases employing these numbers [20].
2) **Text embeddings:** While computers can process natural language by turning it into numbers, whether they can learn from it is another matter. The vectors of numbers sent to learning algorithms must be meaningful for them to be able to gain knowledge from natural language. Text embeddings transform the meaning of a word from dense fixed-length real-valued vectors to variable-length token vectors, where words that are close together in the vector space are expected to have similar meanings. The original transformer uses an embedding layer that trains the entire model while also learning word embeddings, although text embeddings can be produced using external pre-trained models. The embedding layer consists of a comprehensive lookup table that contains an embedding for each possible token. These embeddings are modified during the model's training [20].
3) **Positional embeddings:** As mentioned earlier, the transformer uses self-attention to make up for the lack of recurrence. Transformers can be trained significantly faster than RNNs because they do not require repetition and can process whole sentences at once. However, self-attention does not entirely compensate for the loss of recurrence. Transformers receive whole phrases as

[1]Note that the supplementary materials present the detailed outcomes of the models during the selection process

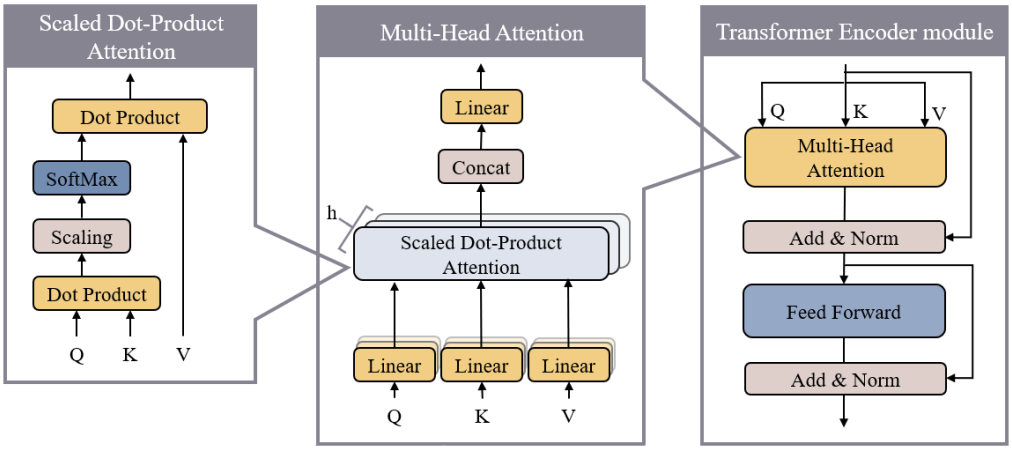

FIGURE 1. Comprehensive architecture of the transformer encoder module, featuring scaled Dot-Product attention and multi-head attention mechanisms.

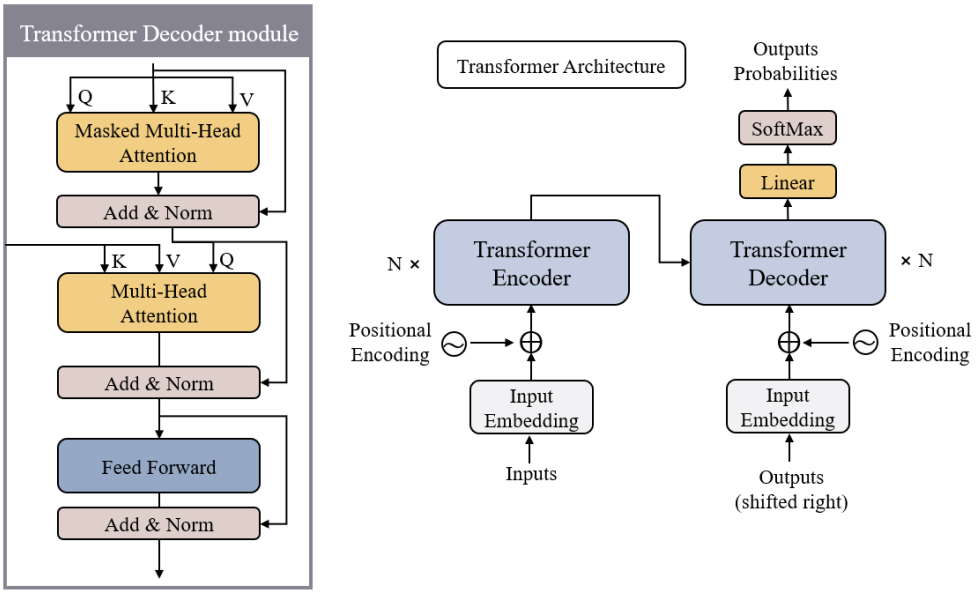

FIGURE 2. Illustration of the complete transformer architecture, highlighting the transformer decoder module. The architecture is adapted from the seminal work "Attention is all you need" by Vaswani et al. [20].

input, but each word is given its attention. Therefore, the significance of the word's placement in the phrase needs to be recovered. The solution is to supplement each text embedding vector with a positional embedding vector that records the positions of the input sequence's words. Both fixed and learnable positional embeddings are possible. The original transformer uses fixed sinusoidal positional embeddings [20] (see Equation 3).

$$PE(pos, 2i) = sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right);$$
$$PE(pos, 2i+1) = cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right), \quad (3)$$

in which pos is the position of the word in a phrase.

4) **Attention:** In its layers, the transformer replaces repetition with layers of self-attention. The aforementioned layers attempt to capture the knowledge that inputs have about one another. Self-attention operates in the following manner in its most basic form. The query W^q, key W^k, and value W^v matrices are among the three sizable matrices that are learned during training [20]. Each component of an embedded input sequence that enters a self-attention layer is multiplied by these matrices to produce three new vectors for each input, namely the query Q, key K, and value V vectors:

$$Q_i = \overrightarrow{x}_i \cdot W^q \quad (4)$$

$$K_i = \overrightarrow{x}_i \cdot W^k \quad (5)$$
$$V_i = \overrightarrow{x}_i \cdot W^v \quad (6)$$

The "Scaled Dot-Product Attention" shown in Fig. 1 was obtained with the formula 7.

$$\text{Attention}(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

The model can concentrate simultaneously on data from distinct representational subdomains at different locations thanks to the "Scaled Dot-Product Attention" layers that made up multi-head attention [20]. Figure 1 and formula (8) both depict the "Multi-Head Attention":

$$\text{MultiHead}(Q, K, V) = [head_1, \ldots, head_h],$$
$$head_i = \text{Attention}(Q_i, K_i, V_i) \quad (8)$$

3) SPATIAL TEMPORAL ATTENTION

Attention mechanisms have been increasingly used in deep learning models for analyzing EEG data, and here are some common attention mechanisms that can be used:

- **Spatial Attention:** Spatial attention is a mechanism that learns the weight the contributions of different spatial locations of the EEG electrodes. This can be particularly useful when the location of the electrodes is relevant to the task.
- **Temporal Attention:** Temporal attention is a mechanism that allows the model to focus on specific time intervals in the EEG signal. They are useful when the time course of EEG features is important for the task.

These attention mechanisms can be used in various deep learning models, such as Convolutional Neural Networks (CNNs), and Transformers to improve the performance of the model in analyzing EEG data. In this work we combine CNNs with Vision Transformers (ViT).

5) **Architecture:** Vaswani et al. [20] introduced the original transformer (see Figure 2), featuring an encoder-decoder architecture. The encoder processes tokenized input sequences through multiple blocks, each with a multi-headed attention layer followed by a feed-forward layer. The decoder is autoregressive, generating each token iteratively. Decoder blocks consist of a feed-forward neural network, a multi-headed attention layer, and a masked multi-headed attention layer, followed by residual connections and layer normalization. After the decoder blocks, a linear and softmax layer is added to compute probabilities for each possible output token, with the highest probability token being the model's output.

a: THE VISION TRANSFORMER (VIT)

is a neural network architecture primarily designed for image classification tasks [36]. Instead of using convolutional layers, it relies on the transformer encoder from the original transformer architecture initially developed for NLP tasks [20].

The first step in applying ViT to MI data is to structure the data appropriately. One strategy is considering the data composed of a series of "tokens" or "patches." The entire sequence represents the whole signal, and each token can be considered a time frame in the EEG data. After processing this string of tokens, the transformer architecture can identify temporal or spatial correlations and discover the necessary attributes for categorization.

The transformer architecture requires a lot of processing, particularly for lengthy sequences. To balance accuracy and computational demands, it is crucial to carefully choose the token size, sequence length, and model parameters.

This study implements Vision Transformers, where inputs are embedded and directly fed into a Transformer Encoder. Also, this study experiments with CNNs with ViT inspired by [36] and [41].

D. EXPLORED MODELS

1) CNN MODELS

As illustrated in Figure 3, shows an example of a CNN architecture used in this study with varying layers of convolutional kernels, specifically of sizes [64, 32, 16, 8] and kernel dimensions of 3×8. We denote CNN models by C[Chan 1, …, Chan N]_K(height, width). For the output channels, both increasing and decreasing configurations for $c_{\mathbf{Y},j}$ in different layers were examined. These configurations were defined by $c_{\mathbf{Y},j} = 2^{2+j}$ and $c_{\mathbf{Y},j} = 2^{L+3-j}$, as described in [46] and [56]. Regarding kernel dimensions, a consistent rectangular kernel shape was adopted across all layers. This kernel is defined by $h_{\mathbf{K},j} = h_{\mathbf{K}} = 3$ and $w_{\mathbf{K},j} = w_{\mathbf{K}} = 8 \times \tau$, where τ was varied within $[1, 3]$ to capture a diverse range of EEG temporal features. For instance, considering $fs = 80$ Hz as a sampling rate of EEG data, the temporal window of $t = 100$ ms would be covered by a kernel width of $w_{\mathbf{K}} = 8$, and $t = 300$ ms is covered by kernel widths of $w_{\mathbf{K}} = 24$. We evaluated a variety of convolutional layers and kernel dimensions, resulting in 12 distinct models for baseline comparison. These models are compared alongside EEGNet [56] and other machine learning architectures. Our design closely follows the methodology presented in [46], as it employs the same MI datasets and preprocessing approaches, albeit tailored to the unique constraints of subject-specific analysis.

2) EEGNET

The EEGNet architecture consists of three convolutional layers. The first layer employs temporal convolution to learn frequency filter parameters. The second uses depth-wise convolution to capture spatial filters specific to frequency characteristics. The third layer employs separable convolution to generate temporal summaries for each feature map. Figure 4, depicts the EEGNet architecture in detail.

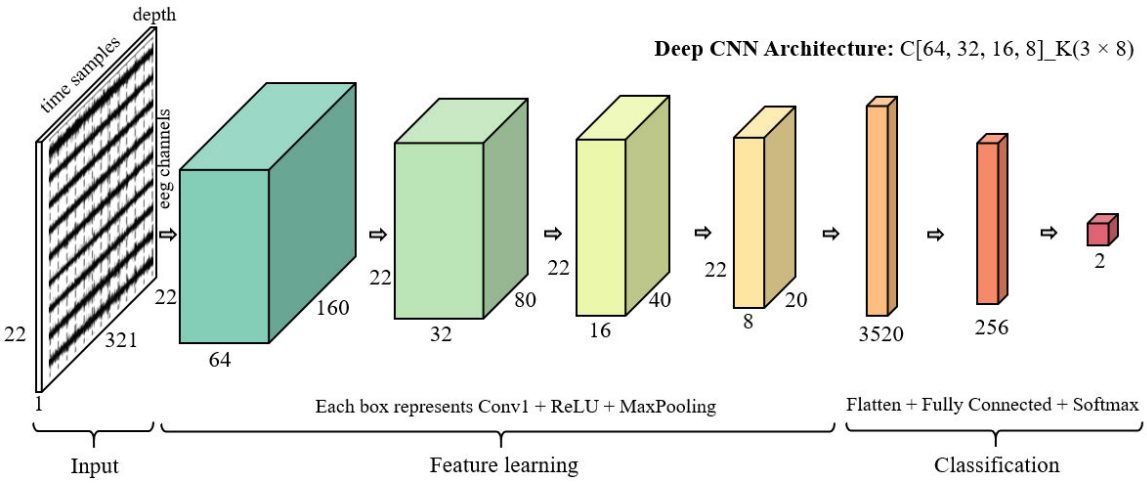

FIGURE 3. In-depth architecture of convolutional neural networks for EEG signal decoding in BCIs applications.

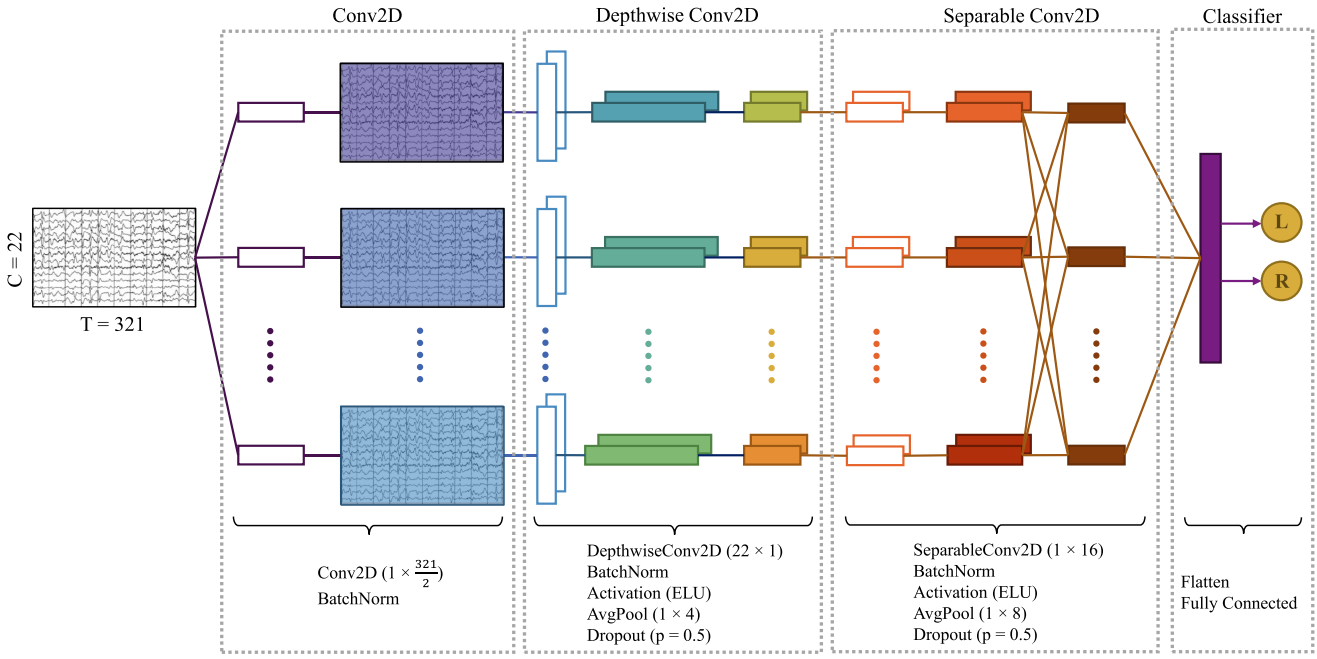

FIGURE 4. Illustration of the EEGNet architecture, inspired from [56]. Employed for EEG signal decoding in the present study.

3) VIT MODELS

The architecture of the Transformer Encoder module requires the input data to be organized in a specific manner, treating EEG channels and time samples as analogous to the height and width dimensions of an image. However, the dimensions of EEG data are unique and often larger than typical image dimensions. To address this, Average Pooling was strategically employed to reduce the effective width and height of the input, as illustrated in Figure 5. Following the dimensional reduction, positional encoding was applied to the data to imbue the Transformer Encoder with a sense of sequence or temporal order. The technique for positional encoding was adopted from [41] and was implemented through a one-dimensional convolutional layer. This prepared input was then processed by the Transformer Encoder module. For the final EEG classification, features extracted by the Transformer Encoder were forwarded to a sequence of fully connected layers, culminating in a Softmax activation function to produce class probabilities. In terms of model selection for the ViT, particular emphasis was placed on examining the number of Transformer modules in the architecture. This was a key factor influencing both classification performance and computational load on the CUDA hardware. We conducted experiments with a range of 4 to 10 Transformer modules and found that using either 3 or 6 modules yielded optimal results for

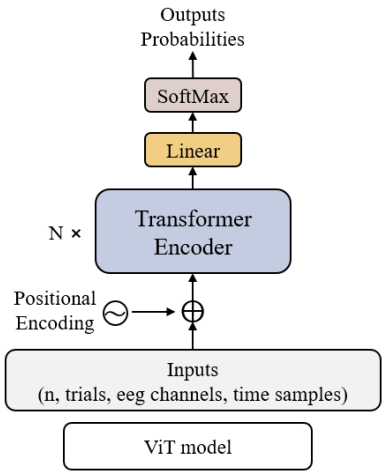

FIGURE 5. The detailed architecture of the vision transformer (ViT) model applied to EEG classification tasks.

different datasets, contingent on the shape of the input data.

4) SPATIAL/TEMPORAL CNN + VIT

Figure 6 presents the frameworks of Spatial CNN combined with ViT (s-CViT) and Temporal CNN combined with ViT (t-CViT) for EEG classification. These frameworks bear a close resemblance to the Spatio-Temporal CNN + ViT architecture, with the primary differences lying in the convolutional modules.

- In s-CViT, the convolutional module consists of two layers and an average pooling layer. The first convolutional layer employs 40 kernels of dimensions 1×16 and utilizes SAME padding to capture temporal characteristics of EEG data. This is followed by an average pooling layer with a pooling size of 1×5. The second convolutional layer, utilizing VALID padding, features 40 kernels with dimensions 1×15.
- The t-CViT model has a simpler CNN module, including a single convolutional layer and an average pooling layer. The convolutional layer adopts 40 kernels of size ch $\times$ 1 with SAME padding to focus on extracting spatial information from EEG signals. The subsequent average pooling layer has a size of 1×5. After pooling, the features are transposed for further processing.

For both s-CViT and t-CViT models, features extracted by the CNN modules are enriched with position encoding and then processed through Transformer modules. Finally, they are classified using fully connected layers, as illustrated in Figure 6.

5) SPATIO-TEMPORAL CNN + VIT

As depicted in Figure 6, the st-CViT model comprises three main components: spatial and temporal convolutional layers, a transformer encoder module, and a fully connected classification layer. The input consists of a group of pre-processed EEG trials with channel and sample dimensions, as well as one additional dimension for the convolution channel. The output is the probability of various EEG categories.

1) **Spatial and Temporal Convolution layers:** We constructed two specialized one-dimensional convolutional layers designed to process temporal and spatial dimensions of EEG data. The first layer focuses on the temporal axis and consists of k kernels, each with dimensions of 1×25 and a stride of 1×1. Subsequently, a second layer is utilized as a spatial filter and encompasses 40 kernels. These kernels are tailored to the count of EEG channels, denoted as ch, with dimensions of $ch \times 1$ and a stride of 1×1 (e.g., $ch = 22$ for the BCI IV 2a dataset). To maintain training stability and improve generalization, we integrated batch normalization and Exponential Linear Units (ELUs) for normalization and activation, respectively. For the dual purpose of diminishing overfitting risks and optimizing computational efficiency, an average pooling layer was introduced in the temporal domain. This layer features a stride of 1×15 and kernel dimensions of 1×75. In the concluding stages, the architecture undergoes a transformation that condenses the electrode channel dimension, simultaneously transposing the convolution channel with the temporal dimension. Additionally, the feature maps outputted from the convolutional module are restructured. As a result, the feature channels corresponding to each temporal instance are tokenized, setting the stage for subsequent modules.
2) **Transformer Encoder Module:** To investigate the global temporal relationships among EEG features, self-attention mechanisms are employed. Sequential tokens from the convolutional module are converted into Query, Key, and Value (QKV) matrices via linear transformations. Dot products of these matrices allow us to analyze associations between tokens, a process further described in Figure 1. A scaling factor and the application of the Softmax function ensure stability during training. The attention score, or weighting matrix, is derived by applying the Softmax function to the resultant output. The attention score is then multiplied by the Value (V) using a dot product, as demonstrated in Equation 7. To enhance the model's adaptability, a pair of fully connected feed-forward layers are appended consecutively. This procedure's input and output dimensions remain invariant, and the module is reiterated three or six times, contingent upon the model's complexity. Moreover, the multi-head approach is employed to augment the diversity of representations. The tokens are partitioned into three or four equidistant segments, each independently processed through the self-attention module. The final output is obtained by concatenating the outcomes from each segment [58].
3) **Fully-Connected Classifier:** The last component is a straightforward classification layer. It consists of two

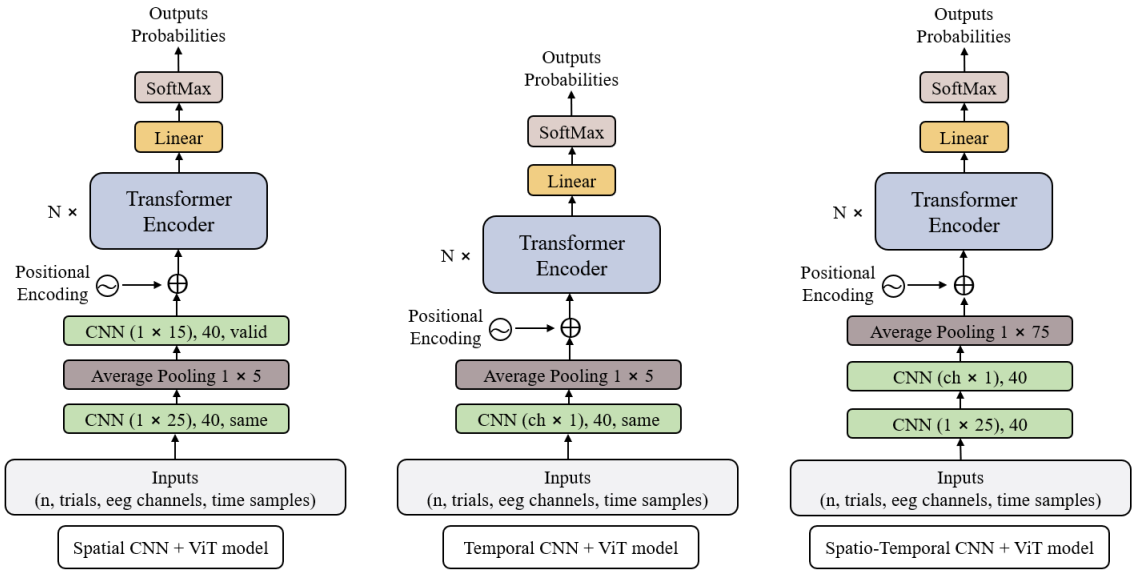

FIGURE 6. Comprehensive architectural schematics of spatial CNN + ViT (s-CViT), temporal CNN + ViT (t-CViT), and Spatio-Temporal CNN + ViT (st-CViT) models, specifically designed for EEG classification tasks.

fully connected layers followed by a Softmax activation to produce an M-dimensional probability vector. The cross-entropy loss function, defined as

$$L = -\frac{1}{N_b}\sum_{i=1}^{N_b}\sum_{c=1}^{M} y\log(\hat{y}), \tag{9}$$

guides the training process. Here $M = 2$ (for EEG classes), y and $\hat{y}$ are the true and predicted labels, respectively, and N_b represents the number of trials in each batch.

E. PERFORMANCE EVALUATION

1) LEAVE-ONE-SUBJECT-OUT EVALUATION

The Leave-One-Subject-Out (LOSO) cross-validation technique is a specialized form of k-fold cross-validation tailored for datasets with multiple samples from individual subjects. Nested cross-validation offers a systematic framework for hyperparameter tuning and model selection. When combined with LOSO, it ensures unbiased evaluation, especially in datasets with distinct samples from individual subjects. The outer loop is responsible for model evaluation, while the inner loop fine-tunes the model hyperparameters (see Fig. 7).

2) PROCEDURE FOR NESTED LOSO-BASED MODEL SELECTION

Given a set of candidate BCI models $\mathcal{M} = \{M_1, M_2, \ldots, M_n\}$, each with a potential set of hyperparameters:

1) For each model $M_k \in \mathcal{M}$:
   a) For each subject s_i in the dataset S:
      i) Partition samples from all subjects excluding s_i into inner training and validation sets.
      ii) Conduct hyperparameter tuning on M_k using inner LOSO cross-validation.
      iii) With the optimal hyperparameters determined, train M_k on the complete inner training set.
      iv) Validate M_k using only samples from s_i.
      v) Document the performance metric for this outer iteration, represented as $P_{k,i}$.
   b) Compute the mean performance for M_k across all subjects:

$$P_{\text{LOSO},k} = \frac{1}{|S|}\sum_{i=1}^{|S|} P_{k,i}$$

2) Identify the model with the top mean performance:

$$M^* = \arg\max_{M_k \in \mathcal{M}} P_{\text{LOSO},k}$$

In BCI contexts, employing nested LOSO cross-validation is important to ensure that models remain resilient against overfitting to EEG patterns distinct to individual subjects. Moreover, it ensures the models' capability to generalize effectively over heterogeneous user data. Despite its computational intensity, the nested procedure provides both unbiased estimates, confirming the adaptability of BCIs to new, unseen users. Additionally, the methodology is designed to minimize information leakage, thus offering a robust representation of a model's ability to decode variance across subjects.

F. HYPERPARAMETER SELECTION

This section outlines the hyperparameter search space for attention mechanism-based models, specifically ViT, Spatial CNN + ViT, Temporal CNN + ViT, and Spatio-Temporal

FIGURE 7. Illustration of Leave-One-Subject-Out (LOSO) and nested LOSO Cross-Validation techniques for model evaluation.

CNN + ViT. Selecting the right hyperparameters is crucial for training and evaluating Transformer models effectively. To ensure a fair comparison and rigorous evaluation, we have standardized the hyperparameter search space for CNN and Transformer models based on the following key assumptions:

1) **Epochs**: Model training was confined to a maximum of 150 epochs, incorporating early stopping criteria.
2) **Batch Size**: For transformer models mini-batch size of 72 was utilized to strike a balance between computational efficiency and model performance after trying different range of batch sizes. In the case of CNN models, and EEGNet batch size of 32 was employed.
3) **Optimization Algorithm**: We employed mini-batch gradient descent using the Adam optimizer. The learning rate and decay parameters were set at 0.001 and 0.0001, respectively [59].
4) **Loss Function**: Cross-entropy was selected as the loss function for evaluating model performance [60].
5) **Transformer Layers**: Transformer modules from 4 to 10 were tested, and 6 was chosen depending on computational efficiency and model performance.
6) **Attention Heads**: The multi-head self-attention mechanism was configured to have 8 to 10 heads.
7) **Embedding Size**: The dimensionality of the input embeddings ranged between 40 and 64, depending on computational efficiency and model performance.
8) **CNN Architecture**: Our convolutional neural networks were structured with up to four convolutional layers, succeeded by a fully connected layer.
9) **CNN Kernel Dimensions**: A consistent rectangular kernel shape was employed across all layers, defined by $h_{\mathbf{K},j} = h_{\mathbf{K}} = 3$ and $w_{\mathbf{K},j} = w_{\mathbf{K}} = 8 \times \tau$, where τ was varied within $[1, 3]$ to capture a diverse range of EEG temporal features.
10) **CNN Output Channels**: We examined both increasing and decreasing configurations for $c_{\mathbf{Y},j}$ in different layers, with specific configurations given by $c_{\mathbf{Y},j} = 2^{2+j}$ and $c_{\mathbf{Y},j} = 2^{L+3-j}$ [46], [56].

TABLE 2. Algorithmic hyperparameters for CNNs and EEGNet models: Key terms include BS (Batch Size), LR (Learning Rate), and WD (Weight Decay).

Models	BS	LR	WD	Dropout	Pooling layer	Activation
CNN	32	1e-3	1e-4	0.3	Max	ReLU
EEGNet	32	1e-3	1e-4	0.5	Average	ELU

TABLE 3. Hyperparameter configuration for attention-based models across diverse datasets: Depth values are specified for BCI 2a/2b, Weibo, and physionet datasets.

Models	Heads	Embedding size	Depth	Batch size
ViT	8	64	2 (All)	72
s-CViT	10	60	6 / 3 / 4	72
t-CViT	8	64	2 (All)	72
st-CViT	10	40	6 (All)	72

11) **Dropout**: A dropout layer with a retention probability $p = 0.3$ and $p = 0.5$ were used prior to the fully connected layer to minimize overfitting [61].

These hyperparameters were selected based on a combination of empirical testing and best practices from the existing literature. The aim is to allow for meaningful comparisons of models, as well as to facilitate the replication of our experiments.

The following tables 2 and 3 list the hyperparameters of the attention model for models.

All models were implemented using a Windows workstation with AMD Ryzen 7 5800H (3.20 GHz) processor, 32GB of RAM, and Nvidia GeForce RTX 3060 (RAM=6GB, CUDA Cores: 3840). The entire model selection workflow was implemented in the PyTorch deep learning environment.

IV. RESULTS AND DISCUSSION

A. PERFORMANCE ANALYSIS OF BENCHMARK MODELS

In this section, the performance of various deep learning models on BCI datasets (BCI IV 2a, BCI IV 2b, Weibo, and Physionet datasets) is examined. These datasets have previously been studied using different methods, but differences in their setup and testing methods were observed in those studies. Models such as CNN, EEGNet, ViT, Spatial CNN + ViT, Temporal CNN + ViT, and Spatio-Temporal CNN + ViT were tested. The efficacy of each method for each subject was determined using the Leave-One-Subject-Out (LOSO) method.

To provide a comparison with previous studies, changes were made to all models. These modifications are detailed in Sections III-B, III-D, and III-E. In Tables 4, 5, 6, and 7, the performance of different models on the four MI datasets using the top LOSO results is presented. Detailed results for each subject can be found in Appendix A. The performance of EEGNet and Spatio-Temporal CNN + ViT on the BCI IV 2b dataset using the nested LOSO evaluation accuracies is presented in Tables 8 and 9, respectively.

CNN models, represented as C[Chan 1, …, Chan N]_K(height, width), were defined by their output channels and kernel dimensions. Output channels vary in increasing or decreasing configurations across layers. Consistently, a rectangular kernel shape was adopted across layers, with adjustments in width to capture diverse EEG temporal features.

B. THE EFFECT OF CONVOLUTIONAL LAYERS AND KERNEL SIZE ON THE PERFORMANCE OF CNN MODELS

The convolutional layers and kernel size can significantly affect the performance of the CNNs as reflected in the four datasets. These parameters are critical for the model's capability to extract relevant features from the given MI-based BCI datasets.

On the BCI IV 2a dataset (see Table 4 and Figure 8), we observed an interesting pattern. As the depth of the network increased (i.e., more convolutional layers), the average accuracy tended to improve. Specifically, the model with the configuration C[64, 32, 16, 8]_K(3, 8) yielded the highest average accuracy of 68.25%, despite having the lowest number of parameters (1.01M). This suggests that deeper networks could potentially capture more complex patterns in the data, leading to better overall performance. As for the kernel size, networks with kernel size (3, 24) generally outperformed those with (3, 8), indicating that larger kernels might have contributed to a more effective feature extraction.

The performance trend continued in the BCI IV 2b dataset (see Table 5). Again, the CNN model with configuration C[64, 32, 16, 8]_K(3, 24) showed the best average accuracy (73.54%) with a moderate amount of parameters (328.06K). The increased kernel size seemed to have a positive effect on the model's performance, further demonstrating the importance of a suitable kernel size in achieving high classification accuracy.

In the Weibo dataset (see Table 6), the CNN model with the configuration C[64, 32, 16, 8]_K(3, 8) performed the best, with an average accuracy of 66.38% and 2.65M parameters. However, as the kernel size increased, the best average accuracy (64.80%) was attained by the configuration C[32, 16, 8]_K(3, 24). This suggests that while deeper networks may yield better results, the impact of kernel size may be more nuanced and potentially dataset-dependent.

In the Physionet dataset (see Table 7), the CNN model C[64, 32, 16, 8]_K(3, 8) demonstrated the best performance, with an average accuracy of 74.42% and 1.77M parameters. As the kernel size increased, the highest average accuracy (74.00%) was achieved by the configuration C[32, 16, 8]_K(3, 24), aligning with the trends observed in previous datasets.

To summarize, the number of convolutional layers and the kernel size play substantial roles in the performance of CNNs on BCI datasets. It appears that deeper networks tend to produce superior results, possibly due to their ability to

TABLE 4. Performance comparison on unseen BCI IV 2a test subjects: Classification accuracy of CNN, EEGNet, and transformer models evaluated via LOSO methodology. Best performing models for each subject are highlighted in bold (in %). Different configurations of CNN models are denoted as C[16, 8]_K(3, 8).

Models / Subjects	S1	S2	S3	S4	S5	S6	S7	S8	S9	Average	Params
C[16, 8]_K(3, 8)	67.71	60.07	75.69	65.63	57.64	65.97	61.46	72.57	73.61	66.71	3.65M
C[32, 16, 8]_K(3, 8)	70.14	50.69	78.82	60.76	55.90	74.31	59.03	75.69	72.92	66.47	1.86M
C[64, 32, 16, 8]_K(3, 8)	71.88	59.03	77.78	65.28	56.94	61.46	57.99	85.76	78.13	68.25	1.01M
C[16, 8]_K(3, 24)	61.81	59.38	71.88	60.42	57.64	64.24	55.56	81.94	72.92	65.09	3.66M
C[32, 16, 8]_K(3, 24)	68.40	54.86	82.64	61.11	55.90	65.28	58.33	77.08	76.39	66.67	1.90M
C[64, 32, 16, 8]_K(3, 24)	74.31	57.99	80.56	64.58	55.56	62.85	63.89	76.74	75.69	68.02	1.15M
C[8, 16]_K(3, 8)	73.26	57.99	68.75	61.46	56.60	69.10	55.21	73.96	67.71	64.89	7.30M
C[8, 16, 32]_K(3, 8)	67.36	55.90	75.69	61.46	54.51	67.36	53.82	80.56	69.79	65.16	7.41M
C[8, 16, 32, 64]_K(3, 8)	69.79	50.69	80.21	62.50	52.08	68.06	54.86	76.74	74.65	65.51	7.64M
C[8, 16]_K(3, 24)	64.93	60.76	77.43	58.68	54.86	65.97	56.94	74.65	72.57	65.20	7.31M
C[8, 16, 32]_K(3, 24)	65.97	48.26	80.56	61.81	52.78	66.32	57.99	76.39	75.00	65.01	7.44M
C[8, 16, 32, 64]_K(3, 24)	65.97	51.74	74.31	57.64	53.13	69.44	55.56	73.61	71.88	63.70	7.76M
EEGNet	79.51	53.13	86.11	54.51	57.99	62.50	57.99	86.81	67.01	67.28	19.95k
ViT	53.13	55.56	55.21	55.90	54.86	56.94	54.86	55.21	54.17	55.09	2.73M
s-CViT	58.68	58.68	58.33	55.56	54.86	58.68	60.07	54.17	61.46	57.83	989.43k
t-CViT	72.92	58.68	**92.01**	69.79	59.03	72.92	55.56	95.49	77.08	72.61	769.83k
st-CViT	**89.58**	**63.89**	90.63	**78.13**	**70.14**	**79.17**	**74.31**	**96.53**	**81.60**	**80.44**	136k

TABLE 5. Performance comparison on unseen BCI IV 2b Test Subjects: Classification accuracy of CNN, EEGNet, and transformer models evaluated via LOSO methodology. Best performing models for each subject are highlighted in bold (in %). Different configurations of CNN models are denoted as C[16, 8]_K(3, 8).

Models / Subjects	S1	S2	S3	S4	S5	S6	S7	S8	S9	Average	Params
C[16, 8]_K(3, 8)	71.67	57.80	53.47	89.32	72.03	76.11	71.94	78.95	77.08	72.04	501.91k
C[32, 16, 8]_K(3, 8)	73.61	52.94	55.00	91.22	75.68	76.11	71.53	78.29	77.92	72.48	268.86k
C[64, 32, 16, 8]_K(3, 8)	69.72	56.62	57.36	91.62	72.84	77.50	71.39	79.08	78.75	72.76	195.96k
C[16, 8]_K(3, 24)	72.64	59.12	55.69	85.27	69.59	74.17	75.42	79.87	81.67	72.60	508.83k
C[32, 16, 8]_K(3, 24)	73.06	57.94	55.42	90.00	72.97	74.58	71.11	78.68	80.69	72.72	301.11k
C[64, 32, 16, 8]_K(3, 24)	71.39	56.62	55.00	91.89	76.49	**81.94**	73.75	76.84	77.92	73.54	328.06K
C[8, 16]_K(3, 8)	66.53	58.24	55.42	90.81	68.51	75.42	72.08	79.61	77.50	71.57	999.39k
C[8, 16, 32]_K(3, 8)	69.72	53.97	55.69	**92.03**	74.46	73.33	**75.56**	79.47	79.03	72.58	1.02M
C[8, 16, 32, 64]_K(3, 8)	72.22	54.71	54.17	90.95	70.67	78.47	66.39	78.03	78.47	71.56	1.10M
C[8, 16]_K(3, 24)	72.92	**60.74**	55.28	86.89	69.73	73.33	74.31	**80.26**	77.92	72.38	1.01M
C[8, 16, 32]_K(3, 24)	70.28	59.12	55.00	91.08	72.16	72.64	70.28	77.50	77.36	71.71	1.06M
C[8, 16, 32, 64]_K(3, 24)	67.08	56.32	54.17	91.22	71.62	79.31	73.19	77.37	78.33	72.07	1.23M
EEGNet	67.08	59.56	56.94	84.46	71.89	79.44	71.81	77.24	79.58	72.00	19.35k
ViT	53.61	53.38	52.22	52.30	55.27	52.08	55.28	50.66	52.50	53.03	1.42M
s-CViT	66.67	57.65	53.61	84.05	62.30	65.42	64.72	72.89	71.39	66.52	405.75k
t-CViT	68.19	60.44	54.31	85.00	75.27	71.94	67.92	77.63	79.17	71.10	768.61k
st-CViT	**77.36**	59.71	**58.61**	83.51	**76.89**	80.42	75.42	78.82	**81.81**	**74.73**	288.29k

capture more complex and nuanced features. Larger kernel sizes also seem beneficial for accuracy, although the extent of their impact may vary across datasets. Future studies could further investigate the impact of these parameters on different BCI tasks and potentially reveal insights into their optimal configuration for BCI applications.

C. THE EFFICACY OF ATTENTION MECHANISM-BASED MODELS WHEN APPLIED TO SUBJECT-INDEPENDENT BCIS

The study compared the performance of different deep learning-based models on three benchmark datasets: BCI IV 2a, BCI IV 2b, and Weibo. The results clearly

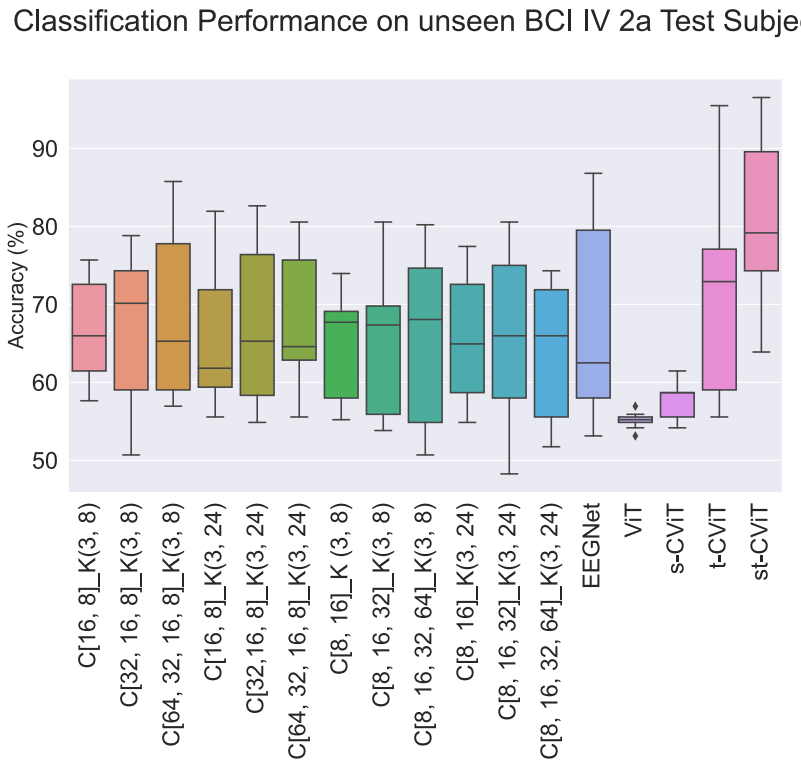

FIGURE 8. Classification accuracy performance of CNN, EEGNet, and transformer models (in %) on unseen BCI IV 2a test subjects via LOSO methodology.

TABLE 6. Performance comparison on unseen Weibo Test Subjects: Classification accuracy of CNN, EEGNet, and transformer models evaluated via LOSO methodology. Best performing models for each subject are highlighted in bold (in %). Different configurations of CNN models are denoted as C[16, 8]_K(3, 8).

Models / Subjects	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	Average	Params
C[16, 8]_K(3, 8)	50.63	56.88	55.00	48.13	51.88	62.86	69.38	61.88	60.63	46.25	56.35	9.97M
C[32, 16, 8]_K(3, 8)	65.63	61.25	57.50	42.50	56.25	72.14	71.88	68.13	68.13	53.13	61.65	5.06M
C[64, 32, 16, 8]_K(3, 8)	72.50	69.38	55.63	46.88	60.63	80.00	87.50	68.13	69.38	53.75	66.38	2.65M
C[16, 8]_K(3, 24)	58.13	62.50	51.25	47.50	55.00	70.00	72.50	61.88	55.63	47.50	58.19	9.97M
C[32, 16, 8]_K(3, 24)	71.88	73.13	56.25	51.88	59.38	83.57	73.75	59.38	65.00	53.75	64.80	5.09M
C[64, 32, 16, 8]_K(3, 24)	61.25	68.13	53.13	48.75	62.50	75.71	80.63	62.50	66.88	57.50	63.70	2.78M
C[8, 16]_K(3, 8)	59.38	58.13	56.88	49.38	50.63	69.29	71.25	51.88	58.13	51.25	57.62	19.90M
C[8, 16, 32]_K(3, 8)	68.13	65.00	56.88	46.88	**63.13**	75.00	81.88	59.38	66.88	56.88	64.00	20.17M
C[8, 16, 32, 64]_K(3, 8)	68.75	65.63	52.50	45.63	62.50	74.29	79.38	67.50	68.13	**60.63**	64.49	20.71M
C[8, 16]_K(3, 24)	68.13	59.38	60.00	40.63	55.00	65.71	73.75	60.00	60.63	51.88	59.51	19.92M
C[8, 16, 32]_K(3, 24)	67.50	66.88	56.88	48.13	56.25	67.86	84.38	63.13	66.25	56.88	63.41	20.20M
C[8, 16, 32, 64]_K(3, 24)	68.13	67.50	56.88	48.13	60.00	72.86	76.25	65.63	65.63	51.88	63.29	20.84M
EEGNet	68.75	70.00	60.00	51.88	58.75	87.14	84.38	**70.00**	71.88	56.88	67.97	21.17k
ViT	56.25	51.88	59.38	58.75	60.00	53.57	56.25	56.25	52.50	50.00	55.48	2.73M
s-CViT	59.39	54.38	**62.50**	56.88	58.13	50.00	56.25	59.38	53.13	58.13	56.82	128.8k
t-CViT	58.75	68.75	56.25	50.00	56.88	75.00	79.38	51.88	68.75	56.88	62.25	772.26k
st-CViT	**75.00**	**85.00**	60.00	**62.50**	61.88	**95.00**	**91.88**	65.63	**83.75**	58.13	**73.88**	379.49k

showed that attention mechanism-based models, such as the Vision Transformer (ViT) and its spatial, temporal, and spatio-temporal variants, showed varying performance depending on the dataset.

TABLE 7. Performance comparison on unseen **Physionet** test subjects: Classification accuracy of CNN and transformer models evaluated via LOSO methodology. Best performing models for each subject are highlighted in bold (in %). Different configurations of CNN models are denoted as C[16, 8]_K(3, 8).

Models / Subjects	S0	S10	S20	S30	S40	S50	S60	S70	S80	S90	Average	Params
C[16, 8]_K(3, 8)	71.11	76.19	51.11	57.14	70.45	93.33	86.67	77.78	88.10	50.00	72.19	6.69M
C[32, 16, 8]_K(3, 8)	73.33	76.19	35.56	45.24	72.73	91.11	91.11	86.67	**95.24**	57.14	72.43	3.43M
C[64, 32, 16, 8]_K(3, 8)	77.78	83.33	46.67	52.38	81.82	93.33	93.33	82.22	88.10	45.24	74.42	1.77M
C[16, 8]_K(3, 24)	66.67	80.95	44.44	50.00	77.27	91.11	93.33	86.67	90.48	57.14	73.81	6.70M
C[32, 16, 8]_K(3, 24)	64.44	71.43	53.33	59.52	75.00	91.11	93.33	80.00	85.71	59.52	73.34	3.46M
C[64, 32, 16, 8]_K(3, 24)	68.89	71.43	44.44	50.00	79.55	91.11	**95.56**	84.44	90.48	64.29	74.02	1.90M
C[8, 16]_K(3, 8)	75.56	73.81	46.67	61.90	84.10	95.56	86.67	91.11	90.48	42.86	74.87	13.38M
C[8, 16, 32]_K(3, 8)	**82.22**	78.57	40.00	59.52	86.36	86.67	91.11	77.78	90.48	54.76	74.75	13.66M
C[8, 16, 32, 64]_K(3, 8)	73.33	76.19	37.78	50.00	86.36	91.11	**95.56**	77.78	88.10	59.52	73.57	13.71M
C[8, 16]_K(3, 24)	73.33	76.19	53.33	47.62	79.55	93.33	88.89	91.11	92.86	73.81	77.00	13.39M
C[8, 16, 32]_K(3, 24)	73.33	83.33	55.56	52.38	70.45	91.11	88.89	68.89	88.10	78.57	75.06	13.69M
C[8, 16, 32, 64]_K(3, 24)	80.00	85.71	46.67	45.24	77.27	88.89	86.67	88.89	90.48	61.90	75.17	13.84M
ViT	57.78	57.14	55.56	54.76	65.91	64.44	55.56	55.56	61.90	59.52	58.81	2.30M
s-CViT	77.78	**88.10**	**64.44**	66.67	**93.18**	**99.99**	**95.56**	**93.33**	**95.24**	**90.48**	**86.47**	1.20M
t-CViT	77.78	69.05	57.78	**69.05**	77.27	84.44	75.56	75.56	90.48	73.81	73.81	526.76k
st-CViT	80.34	83.33	62.22	64.29	81.82	95.56	**95.56**	86.67	92.86	88.10	83.08	264.53k

TABLE 8. Performance evaluation of EEGNet on unseen **BCI IV 2b** test subjects: Accuracy assessment using nested LOSO methodology.

Test /Val subj.	V-S1	V-S2	V-S3	V-S4	V-S5	V-S6	V-S7	V-S8	V-S9	Average
T-S1	-	71.39	72.64	74.03	74.03	71.81	72.92	73.47	73.75	73.00
T-S2	62.65	-	63.24	60.29	60.88	59.85	60.44	59.56	59.56	60.81
T-S3	62.36	58.33	-	57.22	56.67	57.36	54.58	56.67	57.36	57.57
T-S4	95.54	88.24	88.78	-	83.92	84.19	90.00	91.22	85.14	88.38
T-S5	91.08	82.57	81.89	78.65	-	80.54	79.46	80.14	78.24	81.57
T-S6	94.72	84.03	81.25	86.39	82.78	-	80.97	81.81	82.08	84.25
T-S7	87.64	81.11	75.69	77.50	78.47	77.08	-	74.03	73.33	78.11
T-S8	82.76	79.34	79.47	81.58	81.18	79.21	80.00	-	78.55	80.26
T-S9	84.31	82.64	82.22	81.39	81.25	80.83	80.28	79.44	-	81.55

For the BCI IV 2a dataset, the results showed that the performance of the base ViT model was significantly lower than most CNN models and EEGNet, achieving an average accuracy of 55.09%. However, when combined with other modeling approaches, the ViT-based models showed notable improvements. The Temporal CNN + ViT and Spatio-Temporal CNN + ViT models achieved significantly higher accuracies of 72.61% and 80.44%, respectively, outperforming all the other models.

On the BCI IV 2b dataset, the performance trend was similar, with the base ViT model performing relatively poorly with an average accuracy of 53.03%. Once again, the addition of other modeling strategies to ViT showed remarkable improvements. The Spatio-Temporal CNN + ViT model, in particular, outperformed all other models, achieving an average accuracy of 74.73%.

On the Weibo dataset, however, the performance of the ViT-based models was different. The base ViT model achieved an average accuracy of 57.83%, outperforming most of the CNN models. The Spatio-Temporal CNN + ViT model again emerged as the best performer with an average accuracy of 78.44%.

In terms of parameter size, all the ViT-based models demonstrated significantly fewer parameters compared to the CNN models, which is beneficial for model deployment and computational efficiency. Particularly, the Spatio-Temporal CNN + ViT model demonstrated high performance while having the smallest parameter size among all models in all datasets.

D. THE EFFECT OF SPATIAL, TEMPORAL, OR SPATIO-TEMPORAL CNN MODELING ON THE PERFORMANCE OF VIT

From the analysis of the aforementioned data (Tables 4, 5, 6, and 7), it is evident that the incorporation of spatial, temporal, and spatio-temporal modeling can

TABLE 9. Performance evaluation of Spatio-Temporal CNN combined with ViT on unseen BCI IV 2b test subjects: Accuracy assessment using nested LOSO methodology.

Test /Val subj.	V-S1	V-S2	V-S3	V-S4	V-S5	V-S6	V-S7	V-S8	V-S9	Average
T-S1	-	76.53	75.56	74.72	74.44	73.61	76.53	76.11	74.86	75.30
T-S2	61.03	-	58.97	59.71	62.50	59.85	60.15	59.71	59.85	60.22
T-S3	58.33	56.94	-	57.64	59.03	58.75	57.50	58.06	57.22	57.93
T-S4	80.27	83.78	84.19	-	90.14	81.62	82.03	91.08	85.41	84.82
T-S5	75.27	77.84	76.62	78.11	-	75.95	75.41	77.03	73.11	76.17
T-S6	79.72	80.28	81.25	80.28	81.94	-	80.83	81.81	81.39	80.94
T-S7	75.69	73.89	75.69	74.58	73.89	73.33	-	74.72	74.03	74.48
T-S8	73.03	72.24	72.50	77.63	73.29	73.29	72.11	-	72.63	73.34
T-S9	81.39	81.67	81.81	82.22	81.81	82.08	82.36	79.03	-	81.55

significantly impact the performance of Vision Transformers (ViT) in BCI systems, particularly when decoding Motor Imagery (MI) from EEG signals.

1) **Spatial Modeling:** The application of spatial modeling alone with ViT resulted in a noticeable improvement in comparison to the standalone ViT model across all datasets. However, its performance is generally lower than that of the temporal and spatio-temporal modeling. This might be due to the fact that EEG signals are inherently temporal in nature, and while spatial features can provide valuable information regarding the distribution of brain activity across the scalp, they might not sufficiently capture dynamic changes in brain states over time.
2) **Temporal Modeling:** Implementing temporal modeling with ViT significantly outperformed both the standalone ViT and the Spatial CNN + ViT models. This is consistent across all examined datasets. The superior performance of the temporal modeling suggests that it is critical to exploit the temporal structure of EEG signals to enhance BCI performance. Temporal modeling is adept at handling the time-varying nature of EEG signals, thereby capturing critical signal characteristics that change over time.
3) **Spatio-Temporal Modeling:** Most notably, the combination of both spatial and temporal modeling into a spatio-temporal CNN + ViT model resulted in the highest average performance across all datasets, with a significant reduction in the number of parameters compared to many other models. This demonstrates the importance of capturing both the spatial distribution and temporal dynamics of brain activity in EEG signals for successful MI decoding. Spatio-temporal modeling combines the advantages of both spatial and temporal models, yielding a more comprehensive understanding of the EEG data, and thereby leading to superior performance.

Overall, while each of the spatial and temporal models brings unique benefits, the integration of both spatial and temporal features in a spatio-temporal model appears to be the most effective strategy for improving the performance of ViT in the context of EEG-based BCIs. However, it's worth noting that the combination of strengths of two systems (CNN and ViT), also combines their challenges. This can make the model slower, especially when working with a lot of EEG data. Furthermore, the performance of these models can also be affected by other factors, such as the complexity of the model (reflected by the number of parameters), the choice of hyperparameters, and the specifics of the dataset and task. Therefore, further research is required to fully exploit the potential of deep learning models, particularly ViT, in the domain of BCIs.

E. THE EFFICACY OF LOSO METHODS

The LOSO and nested LOSO are cross-validation techniques commonly used in the context of BCI research, especially when dealing with small datasets or when the subject independence of the models is under consideration. The nested LOSO evaluation entails training on data from every subject save for one, which is reserved for validation. The traditional LOSO, in contrast, further segments the data from each subject into discrete training and validation subsets.

Our findings from the LOSO evaluation (refer to Table 5) demonstrate a competitive performance across models. Specifically, the Spatio-Temporal CNN + ViT model stands out, registering an accuracy of 74.73%, albeit with an associated increase in model intricacy, as evident from the parameter count. Remarkably, the EEGNet model strikes an optimal balance, delivering an average accuracy of 72.00% while only employing 19.35k parameters. However, the LOSO evaluation isn't without its shortcomings. One notable concern is the potential overfitting, as the iterative training and validation may lead the model to optimize itself to the training dataset excessively, subsequently affecting its efficacy on unseen data.

On the other hand, utilizing the nested LOSO evaluation (see Tables 8 and 9) indicates a marginal enhancement in average accuracies for both EEGNet and Spatio-Temporal

CNN + ViT models when compared against the LOSO results. This uptick possibly stems from the augmented training data available to the nested LOSO evaluation, circumventing the need to bisect each subject's data further.

In summary, nested LOSO evaluations furnish critical insights into the adaptability and efficacy of BCI models across diverse subjects. Each method possesses inherent merits and limitations: the LOSO approach might offer superior performance due to augmented training data but runs the risk of overfitting, while the nested LOSO method reduces this risk, albeit with less training data. The selection between the two methods should depend on the specific requirements and constraints of the study, such as the available dataset size, the model complexity, and the tolerance for overfitting.

V. CONCLUSION AND FUTURE WORK

This study contributes to the field of subject-independent BCIs in several ways. First, it comprehensively compares attention mechanism-based models, including ViT, Spatial CNN + ViT, Temporal CNN + ViT, and Spatio-Temporal CNN + ViT, in terms of classification accuracy, robustness, computational efficiency, and signal-to-noise ratio. This analysis provides a better understanding of the strengths and weaknesses of each model when applied to different BCI tasks and datasets. Second, the study employs a nested LOSO method for model selection, improving the reliability of performance estimates and promoting the development of more accurate and efficient BCIs. Additionally, it investigates four distinct Motor imagery-based EEG datasets, providing unique insights into the application of attention mechanisms in building subject-independent BCIs using these datasets. The results demonstrate that the Spatio-Temporal CNN + ViT model outperforms other models on the BCI IV 2a, 2b, and Weibo datasets, indicating its potential for practical BCI applications. Furthermore, by combining attention mechanisms and deep learning models to identify informative features common across subjects while effectively filtering out noise and irrelevant data, this study presents a realistic approach to building subject-independent BCIs.

Although the attention-based models used in this study showed promising results in EEG classification, there are some limitations and areas for future work to enhance their potential in the BCI research field:

1) Positional Encoding: In this study, the impact of different positional encoding strategies was not investigated extensively on transformer-based models. Future work could explore various positional encoding schemes to determine their influence on model performance and robustness in the context of EEG data.
2) Kernel Sizes of Convolutional Layers: A limited kernel size set was used for the convolutional layers in the transformer models. A more extensive exploration of different kernel sizes could reveal their impact on model performance and potentially improve the effectiveness of the transformer-based models.
3) Transfer Learning with Transformers: Due to resource and time constraints, the potential of transfer learning with transformer models was not explored. Future work could investigate pre-trained transformer models and adapt them for EEG classification tasks, potentially leading to better generalization and improved performance across different subjects and tasks.
4) Model Efficiency in Real-Time Scenarios: While our models demonstrate promising accuracy and robustness, additional improvements might be necessary to optimize them for real-time EEG data processing. Investigating model compression techniques, or simplifying certain layers without significant compromise on performance, could be a path worth exploring for real-time EEG applications.

By addressing these limitations and extending the research in the above directions, transformer-based models could become even more valuable in the BCI research field, leading to more robust and accurate subject-independent BCIs for various applications.

REFERENCES

[1] S. Jalilpour, S. H. Sardouie, and A. Mijani, "A novel hybrid BCI speller based on RSVP and SSVEP paradigm," *Comput. Methods Programs Biomed.*, vol. 187, Apr. 2020, Art. no. 105326.

[2] O. B. Guney and H. Ozkan, "Transfer learning of an ensemble of DNNs for SSVEP BCI spellers without user-specific training," *J. Neural Eng.*, vol. 20, no. 1, Jan. 2023, Art. no. 016013.

[3] V. Maksimenko, A. Lüttjohann, S. van Heukelum, J. Kelderhuis, V. Makarov, A. Hramov, A. Koronovskii, and G. van Luijtelaar, "Brain–computer interface for the epileptic seizures prediction and prevention," in *Proc. 8th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2020, pp. 1–5.

[4] C. Daftari, J. Shah, and M. Shah, "Detection of epileptic seizure disorder using EEG signals," in *Artificial Intelligence-Based Brain-Computer Interface*, V. Bajaj and G. Sinha, Eds. New York, NY, USA: Academic, 2022, pp. 163–188.

[5] T. Jagadesh, A. Reethika, B. Jaishankar, and M. Kanivarshini, "Early prediction of epileptic seizure using deep learning algorithm," in *Brain-Computer Interface: Using Deep Learning Applications*. Hoboken, NJ, USA: Wiley, 2023, ch. 7, pp. 157–177.

[6] A. Czech, "Brain–computer interface use to control military weapons and tools," in *Control, Computer Engineering and Neuroscience*, S. Paszkiel, Ed. Cham, Switzerland: Springer, 2021, pp. 196–204.

[7] M. Vilela and L. R. Hochberg, "Applications of brain-computer interfaces to the control of robotic and prosthetic arms," in *Brain-Computer Interfaces* (Handbook of Clinical Neurology), vol. 168, N. F. Ramsey and J. D. R. Millán, Eds. Amsterdam, The Netherlands: Elsevier, 2020, ch. 8, pp. 87–99.

[8] H. Gao, L. Luo, M. Pi, Z. Li, Q. Li, K. Zhao, and J. Huang, "EEG-based volitional control of prosthetic legs for walking in different terrains," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 2, pp. 530–540, Apr. 2021.

[9] R. R. Subramanian, K. Y. Varma, K. Balaji, M. D. Reddy, A. Akash, and K. N. Reddy, "Multiplayer online car racing with BCI in VR," in *Proc. 5th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2021, pp. 1835–1839.

[10] D. Tezza, D. Caprio, S. Garcia, B. Pinto, D. Laesker, and M. Andujar, "Brain-controlled drone racing game: A qualitative analysis," in *HCI in Games*, X. Fang, Ed. Cham, Switzerland: Springer, 2020, pp. 350–360.

[11] P. Stegman, C. S. Crawford, M. Andujar, A. Nijholt, and J. E. Gilbert, "Brain–computer interface software: A review and discussion," *IEEE Trans. Human-Mach. Syst.*, vol. 50, no. 2, pp. 101–115, Apr. 2020.

[12] E. Kinney-Lang, S. Murji, D. Kelly, B. Paffrath, E. Zewdie, and A. Kirton, "Designing a flexible tool for rapid implementation of brain–computer interfaces (BCI) in game development," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 6078–6081.

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

[14] G. R. Müller-Putz, "Electroencephalography," in *Handbook of Clinical Neurology*, vol. 168. Amsterdam, The Netherlands: Elsevier, 2020, pp. 249–262.

[15] S. Saha, K. I. U. Ahmed, R. Mostafa, L. Hadjileontiadis, and A. Khandoker, "Evidence of variabilities in EEG dynamics during motor imagery-based multiclass brain–computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 371–382, Feb. 2018.

[16] A. Singh, S. Lal, and H. Guesgen, "Reduce calibration time in motor imagery using spatially regularized symmetric positives-definite matrices based classification," *Sensors*, vol. 19, no. 2, p. 379, Jan. 2019.

[17] Y. Dong, X. Wen, F. Gao, C. Gao, R. Cao, J. Xiang, and R. Cao, "Subject-independent EEG classification of motor imagery based on dual-branch feature fusion," *Brain Sci.*, vol. 13, no. 7, p. 1109, Jul. 2023.

[18] P. Autthasan, R. Chaisaen, T. Sudhawiyangkul, P. Rangpong, S. Kiatthaveephong, N. Dilokthanakul, G. Bhakdisongkhram, H. Phan, C. Guan, and T. Wilaiprasitporn, "MIN2Net: End-to-end multi-task learning for subject-independent motor imagery EEG classification," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 6, pp. 2105–2118, Jun. 2022.

[19] O.-Y. Kwon, M.-H. Lee, C. Guan, and S.-W. Lee, "Subject-independent brain–computer interfaces based on deep convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 3839–3852, Oct. 2020.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[21] 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.

[22] Y. He, D. Eguren, J. M. Azorín, R. G. Grossman, T. P. Luu, and J. L. Contreras-Vidal, "Brain–machine interfaces for controlling lower-limb powered robotic systems," *J. Neural Eng.*, vol. 15, no. 2, Apr. 2018, Art. no. 021004.

[23] 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.

[24] K. A. Condori, E. C. Urquizo, and D. A. Diaz, "Embedded brain machine interface based on motor imagery paradigm to control prosthetic hand," in *Proc. IEEE ANDESCON*, Oct. 2016, pp. 1–4.

[25] H. Dose, J. S. Møller, H. K. Iversen, and S. Puthusserypady, "An end-to-end deep learning approach to MI-EEG signal classification for BCIs," *Expert Syst. Appl.*, vol. 114, pp. 532–542, Dec. 2018.

[26] A. Al-Fahoum and A. Al-Fraihat, "Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains," *ISRN Neurosci.*, vol. 2014, Feb. 2014, Art. no. 730218.

[27] S. Taran and V. Bajaj, "Motor imagery tasks-based EEG signals classification using tunable-Q wavelet transform," *Neural Comput. Appl.*, vol. 31, no. 11, pp. 6925–6932, Nov. 2019.

[28] D. Planelles, E. Hortal, Á. Costa, A. Úbeda, E. Iáez, and J. Azorín, "Evaluating classifiers to detect arm movement intention from EEG signals," *Sensors*, vol. 14, no. 10, pp. 18172–18186, Sep. 2014.

[29] S. Bhattacharyya, A. Khasnobish, A. Konar, D. N. Tibarewala, and A. K. Nagar, "Performance analysis of left/right hand movement classification from EEG signal by intelligent algorithms," in *Proc. IEEE Symp. Comput. Intell., Cognit. Algorithms, Mind, Brain (CCMB)*, Apr. 2011, pp. 1–8.

[30] X. Wang, M. Hersche, B. Tomekce, B. Kaya, M. Magno, and L. Benini, "An accurate EEGNet-based motor-imagery brain–computer interface for low-power edge computing," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2020, pp. 1–6.

[31] A. Echtioui, W. Zouch, M. Ghorbel, C. Mhiri, and H. Hamam, "A novel ensemble learning approach for classification of EEG motor imagery signals," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2021, pp. 1648–1653.

[32] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, and M. S. Hossain, "Deep learning for EEG motor imagery classification based on multi-layer CNNs feature fusion," *Future Gener. Comput. Syst.*, vol. 101, pp. 542–554, Dec. 2019.

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

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

[35] 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, pp. 1–9, 2019.

[36] 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," 2021, *arXiv:2010.11929*.

[37] J. Shao, X. Wen, B. Zhao, and X. Xue, "Temporal context aggregation for video retrieval with contrastive learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3268–3278.

[38] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, Oct. 2020, pp. 5036–5040.

[39] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, I. Simon, C. Hawthorne, N. Shazeer, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–14.

[40] 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.

[41] 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.

[42] D.-K. Kim, D. Kim, J.-G. Lee, Y. Woo, and J. Jeong, "Deep learning application to clinical decision support system in sleep stage classification," *Sleep Med.*, vol. 100, p. S293, Dec. 2022.

[43] 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.

[44] 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.

[45] 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.

[46] B. Abibullaev, I. Dolzhikova, and A. Zollanvari, "A brute-force CNN model selection for accurate classification of sensorimotor rhythms in BCIs," *IEEE Access*, vol. 8, pp. 101014–101023, 2020.

[47] W. Yi, S. Qiu, K. Wang, H. Qi, L. Zhang, P. Zhou, F. He, and D. Ming, "Evaluation of EEG oscillatory patterns and cognitive process during simple and compound limb motor imagery," *PLoS ONE*, vol. 9, no. 12, Dec. 2014, Art. no. e114853.

[48] 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: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215–E220, Jun. 2000.

[49] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: A general-purpose brain–computer interface (BCI) system," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1034–1043, Jun. 2004.

[50] 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, Jul. 2012.

[51] R. Leeb, F. Lee, C. Keinrath, R. Scherer, H. Bischof, and G. Pfurtscheller, "Brain–computer communication: Motivation, aim, and impact of exploring a virtual apartment," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 4, pp. 473–482, Dec. 2007.
[52] 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," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.
[53] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. McAlpine, and Y. Zhang, "A survey on deep learning-based non-invasive brain signals: Recent advances and new frontiers," *J. Neural Eng.*, vol. 18, no. 3, Mar. 2021, Art. no. 031002.
[54] 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.
[55] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers Neurosci.*, vol. 7, p. 267, Dec. 2013.
[56] R. Peng, C. Zhao, J. Jiang, G. Kuang, Y. Cui, Y. Xu, H. Du, J. Shao, and D. Wu, "TIE-EEGNet: Temporal information enhanced EEGNet for seizure subtype classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2567–2576, 2022.
[57] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining Practical Machine Learning Tools and Techniques*, 4th ed. Amsterdam, The Netherlands: Elsevier, 2017.
[58] 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.
[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, pp. 1–15, Dec. 2014.
[60] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[61] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

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 Young Researchers Alliance, Astana. Her current research interests include machine learning, brain–computer interfaces, pattern recognition, and digital twin.

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

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

Dr. Abibullaev is an Associate Editor of IEEE Access and *PeerJ Computer Science*.