Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

CWT based cross attention

MD. SAJID AL MAHMUD[1], (Member,IEEE), TAMIM HASAN BHUIYAN[1],(MEMBER,IEEE), AND SHAIKH ANOWARUL FATTAH.[1], (Senior Member, IEEE)

[1]Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh

Corresponding author: First A. Author (e-mail: author@ boulder.nist.gov).

This paragraph of the first footnote will contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

ABSTRACT Motor Imagery (MI) classification in EEG based Brain Computer Interfaces (BCIs) presents significant challenges due to the spatiotemporal, spectral, and inter-domain complexities inherent in EEG signals. This study proposes an end-to-end deep learning framework featuring a dual-path feature extraction paradigm to address these challenges. The framework leverages Continuous Wavelet Transform (CWT) to derive high-resolution time-frequency representations for multi-scale spectral analysis. Raw EEG signals and their CWT-transformed counterparts are processed in parallel through two specialized convolutional blocks: a Time-Series Convolution (TS-Conv) block for temporal and spatial feature extraction and a Time-Frequency Convolution (TF-Conv) block for hierarchical spectral feature analysis. The extracted features are fused using a Transformer Encoder module with multi-head cross-attention, aligning temporal and frequency-specific features to enhance cross-domain synergy. Additionally, temporal dependencies are further refined using a Temporal Convolutional Network (TCN). Finally, a Global Average Pooling (GAP) layer distills the features, which are classified through a dense softmax layer. In subject-specific evaluations, this framework achieved impressive average accuracies of 87.85% on the BCI IV-2a dataset and 91.55% on the BCI IV-2b dataset, with corresponding Kappa values of 0.84 and 0.83. This model sets a new standard in motor imagery classification, achieving superior accuracy with only 82k parameters, far more efficient than existing approaches. Its innovative use of cross-attention mechanism has opened the realm for lightweight, high-performance real time EEG-based BCI frameworks.

INDEX TERMS Brain-Compter Interface, Continuous Wavelet Transform, Cross Attention, Motor Imagery, spatiotemporal, transformer.

I. INTRODUCTION

BRAIN Computer Interface (BCI) establishes a direct communication pathway between the human brain and external devices by leveraging neural signals. This technology has a wide range of applications, including advancing research, enhancing cognitive and motor rehabilitation, enabling hands-free control of assistive devices, restoring lost sensory functions, and expanding human capabilities for direct mental interaction with technology [1]. Based on the electrophysiological or hemodynamic nature of the command various forms of brain signals can be utilized in Brain-Computer Interfaces (BCI), including Electroencephalography (EEG) [2], Magnetoencephalography (MEG) [3], and Functional Magnetic Resonance Imaging (fMRI) [4]. Among these, EEG signals are the most commonly employed due to their non-invasive nature, portability, low cost, minimal risk, high precision and superior temporal resolution in real time applications [5].

Motor imagery (MI) is a widely studied paradigm in BCI research, where individuals mentally simulate physical actions without executing them physically. It involves imagining movement of specific body parts, triggering sensorymotor rhythms (SMRs) in the sensorimotor cortex, which can be decoded by event-related desynchronization(ERD) [6] to identify intended movements [7]. In the context of MI-EEG (motor imagery integrated with EEG), these brain signals have shown great promise in a variety of applications, including neurological rehabilitation, stroke recovery [8], and assistive devices like robotic arms and exoskeletons [9]. Additionally, it is also developing in non-medical fields such as gaming [10], virtual reality, drone control [11] and smart home applications. Despite advancements, MI-EEG systems face challenges due to the low signal-to-noise ratio, non-stationary nature of EEG signals, and various artifacts from biological (e.g., muscle, eye movement) and non-biological sources (e.g., electronic devices, environmental noise). Ad-

ditionally, individual variability, subject dependency, nonlinearity, and high dimensionality of multichannel data complicate accurate feature extraction. These factors hinder the performance, generalization, and efficiency of decoding algorithms, limiting their applicability in real-world scenarios.

In the preliminary stage, researchers used machine learning (ML) to decodes the motor imagery through feature extraction and feature classification. Common Spatial Pattern (CSP) remains a widely used feature extraction method, optimizing spatial filters to enhance task discrimination by maximizing the variance difference across EEG signals. [12]. Later, Filter Bank Common Spatial Pattern (FBCSP) was developed to extract more detailed spatial filtering features [13]. Fast Fourier Transform [14] and Wavelet Transform [15] techniques has also been used for spatial-temporal feature extraction. In conjunction with these methods, classifiers like Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), and Random Forest (RF) are frequently used to differentiate between MI tasks. SVM is beneficial in handling high-dimensional EEG features [16], while LDA's linear discriminative capabilities are useful for distinguishing classes in lower-dimensional spaces [17]. To mitigate the singularity problem in classical LDA, Fu et al. [18] proposed the regularized LDA (RLDA) algorithm. Based on the traditional SVM, Dong et al. [19] proposed a hierarchical SVM (HSVM) algorithm to solve the EEG-based MI classification task. Luo et al. [20] proposed an MI EEG decoding method by combining the dynamic frequency feature selection (DFFS) method with an RF classifier. RF's ensemble nature helps mitigate the effect of noisy data by aggregating multiple decision trees. However, while effective, these traditional methods depend on hand-engineered features and require expert knowledge, often struggling with the non-stationary and artifact-prone nature of EEG signals, thus presenting challenges for consistency and generalization in real-world MI-BCI applications.

Deep learning (DL) revolutionizes the approach to traditional machine learning (ML) challenges by enabling the automatic extraction of high-dimensional features and capturing complex intrinsic relationships directly from raw EEG data, bypassing the need for labor-intensive manual feature extraction [21].

Convolutional Neural Networks (CNNs), inspired by the structure and function of the brain's visual cortex [22], are highly effective at capturing local spatial features. Through convolutional layers, filters slide across inputs to detect key patterns, leveraging sparse connections and parameter sharing to reduce the computational load. This architecture enables CNNs to autonomously learn intricate hierarchical features, making them particularly well-suited for complex classification and regression tasks [23], [24]. However, the fixed kernel sizes in CNNs limit their ability to capture long-term dependencies.

Addressing the shortcomings of CNNs in sequential tasks, Recurrent Neural Networks (RNNs) are designed to capture temporal dependencies in EEG time-series data. Long Short-term Memory (LSTM) [25] enhances RNNs by effectively managing long-term dependencies with gating mechanisms, making them suitable for MI-EEG classification, though they come with high computational demands and slow training times. Gated Recurrent Units(GRUs) offer a faster, more efficient alternative with reduced computational load but struggle to capture as deep long-term dependencies as LSTMs [26]. Despite the advancements of LSTMs and GRUs, RNNs face inherent limitations in parallel training and scalability, often making them less efficient for large-scale tasks, which leads to the exploration of more advanced architectures.

Temporal Convolutional Networks (TCNs) are a specialized variant of CNNs designed for time-series modeling and classification. Unlike RNNs, TCNs offer shorter training times, fewer parameters, and enhanced parallel computing capabilities. They efficiently capture both local and global features and can exponentially increase their receptive field size with only a linear increase in parameters [27]. However, while TCNs excel at discerning temporal dependencies, they sometimes struggle to extract the same level of complex features as RNNs, potentially limiting their performance on more intricate tasks.

The rise of Transformer models, driven by the effectiveness of the self-attention mechanism [28], has marked a breakthrough in MI-EEG classification by enabling the capture of global dependencies through a broad receptive field. Unlike CNNs and RNNs, Transformers process entire inputs simultaneously, excelling at identifying long-range relationships within the data. Despite their advantages, Transformers often struggle to capture local features, require substantial computational resources, and depend on large datasets for effective training, posing challenges for smaller datasets or real-time applications [29], [30].

With a growing potential, transfer learning has shown promise in motor imagery (MI) EEG classification by addressing subject variability. Hang et al. [31] developed a domain adaptation network to transfer data from one subject to another, improving cross-subject learning but requiring substantial data from the source subject. Wei et al. [32] proposed a multi-subject transfer learning network, combining data from multiple subjects to overcome data scarcity, but it may overlook subject-specific differences. Liang and Ma [33] used a Riemannian geometry alignment algorithm to select source subjects and fine-tune the model with new data, improving adaptability but requiring careful source selection. While effective, transfer learning's success depends on data availability and alignment between source and target subjects.

To leverage the strengths of individual techniques, hybrid models integrating diverse architectures have been developed. Li et al. [34] integrated wavelet transform and attention mechanisms in the Time-Frequency Compression Activation Feature Fusion Network for improved EEG classification. Altaheri et al. [35] combined multi-head self-attention and convolutional sliding windows to enhance MI classification accuracy. She et al. used Wasserstein distance in a domain

adaptation network to improve performance with labeled data from multiple subjects [36]. The primary challenges associated with these architectures include high complexity, large model sizes, high parameter count and significant processing time, all of which can impact accuracy and overall performance.

To address these challenges, we propose an innovative end-to-end deep learning framework tailored for motor imagery (MI) classification in EEG-based brain-computer interfaces (BCIs). The framework is designed to address spatiotemporal, spectral, and inter-domain dependencies inherent in EEG signals. It leverages temporal, frequency, and hybrid feature extraction techniques to decode neural patterns associated with MI tasks.

The framework employs Continuous Wavelet Transform (CWT) to generate high-resolution time-frequency representations of raw EEG signals while preserving raw spatiotemporal information. These inputs are processed through dual-path feature extraction blocks and fused using a Transformer Encoder module that leverages positional encoding, multi-head cross-attention and feed-forward mechanism with residual connections. A Temporal Convolutional Network (TCN) is then utilized to capture long-term temporal patterns, with final features classified via a Global Average Pooling (GAP) layer and a dense softmax classifier. The key contributions of this multidomain feature integration framework are as follows:

- **Dual-Branch Feature Extraction Mechanism**: Introduces a novel architecture for complementary spatiotemporal and spectral feature extraction.
- **Advanced Temporal and Frequency Modeling**: Combines temporal, depthwise, and spatial convolutions to effectively capture temporal dynamics and spatial dependencies, alongside separable and spectral convolutions to highlight time-frequency patterns with reduced computational overhead.
- **Transformer Encoder for Cross-Domain Integration**: Utilizes a Transformer Encoder with learnable positional encodings and multi-head cross-attention mechanisms, enabling seamless integration of temporal and frequency features for superior neural pattern decoding.
- **Temporal Convolutional Network (TCN) for Long-Term Dependencies**: Incorporates a TCN block with dilated causal convolutions and Gated Linear Units (GLUs) to adaptively model task-relevant temporal patterns, ensuring robust representation of long-term temporal dependencies.
- **High-Performance Classification with Regularization Techniques**: Integrates spatial dropout, residual connections, and a Global Average Pooling (GAP) layer for effective regularization and feature condensation, culminating in a dense softmax classifier for precise motor imagery classification.

II. METHODOLOGY

By synergizing convolutional and attention mechanisms, this architecture achieves efficient and comprehensive learning of complex temporal-spatial dependencies inherent in MI-EEG data, providing an end-to-end framework Fig. 1 that obviates the need for handcrafted feature engineering. The details of Net blocks are described in the following sections.

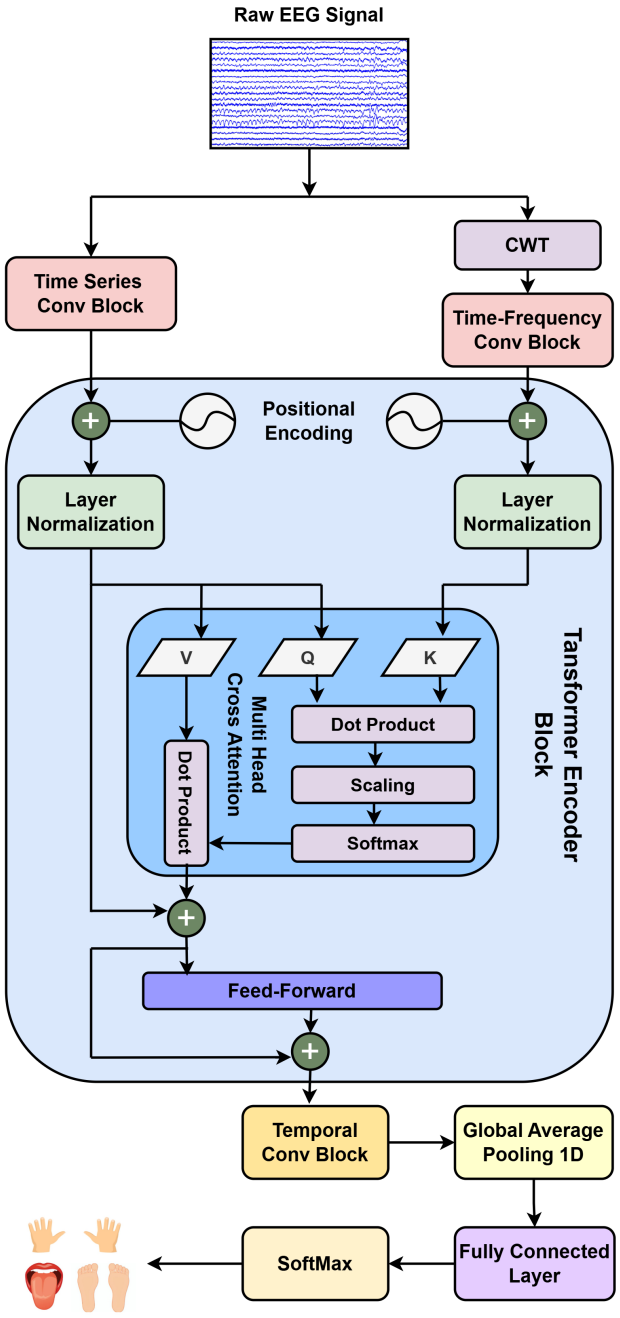

FIGURE 1. Main Model

A. PREPROCESSING

The preprocessing phase begins with loading and segmenting raw EEG signals, employing the full frequency band and all EEG channels without artifact removal. Each motor imagery (MI) trial is extracted from predefined time windows—2 to 6 seconds for the BCI2a dataset and 3 to 7 seconds for BCI2b—encapsulating task-relevant EEG signals. This approach ensures that the data captures the most significant components of the MI task. Subsequently, the extracted signals are standardized channel-wise using Z-score normalization, defined as:

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

where x_i and x_o denote the input and output signals, respectively, and μ and σ are the mean and standard deviation calculated from the training data. This step, set empirically, mitigates fluctuations and addresses the inherent nonstationarity of EEG signals. To prevent data leakage, the standardization parameters derived from the training data are consistently applied to the test data. The preprocessed data is reshaped to conform to the model input format, $(N, 1, C, T)$, where N is the number of trials, C is the number of channels, and T is the number of time points. Labels are converted into a one-hot encoded format for compatibility with the classification model.

B. DATA AUGMENTATION

To address the challenge of limited labeled EEG data, data augmentation techniques tailored for time-domain signals are applied exclusively to the training set. The following augmentation strategies are employed:

1) Gaussian Noise Addition

Random Gaussian noise is added to each signal, simulating variability while preserving its essential characteristics. The augmented signal is represented as:

$$x_{\text{aug}} = x_{\text{orig}} + \eta, \tag{2}$$

Where $\eta \sim \mathcal{N}(0, \sigma^2)$ is the Gaussian noise with a mean of 0 and a standard deviation scaled by a predefined noise factor of 0.001.

2) Scaling

Signals are scaled by a random factor to emulate amplitude variations. The augmented signal is expressed as:

$$x_{\text{aug}} = \alpha \cdot x_{\text{orig}}, \tag{3}$$

Where $\alpha \sim \text{Uniform}(\text{min\_scale}, \text{max\_scale})$ is a random scaling factor. We use $\text{min\_scale} = 0.9$ and $\text{max\_scale} = 1.1$.

3) Time Masking

Random time segments of the signal are masked (set to zero) to simulate missing or occluded signal parts. Let t_{start} and t_{end} represent the start and end points of the mask. The augmented signal can be defined as:

$$x_{\text{aug}}[t] = \begin{cases} 0, & \text{if } t_{\text{start}} \le t \le t_{\text{end}}, \\ x_{\text{orig}}[t], & \text{otherwise.} \end{cases} \tag{4}$$

4) Combined Augmentation

Multiple augmentations are combined sequentially, such as Gaussian noise addition followed by scaling and time masking, to increase diversity in the training data.The augmented dataset is constructed by concatenating the original signals with their augmented versions. For each trial, four augmented variations are generated, leading to a fivefold increase in the size of the training set. Formally, the final augmented dataset is expressed as:

$$X_{\text{aug}} = \text{Concat}(X_{\text{orig}}, X_{\text{noise}}, X_{\text{scale}}, X_{\text{noise+scale}}, X_{\text{noise+scale+mask}}) \tag{5}$$

$$y_{\text{aug}} = \text{Tile}(y_{\text{orig}}, 5) \tag{6}$$

Where X_{aug} and y_{aug} are the augmented training data and training labels, respectively. These augmentation strategies enhance the diversity of the training set, improving the model's robustness to signal variability and noise.

C. TEMPORAL-FREQUENCY TRANSFORMATION MODULE

The temporal-frequency transformation module processes EEG signals to extract time-frequency features using the Continuous Wavelet Transform (CWT). This transformation enables a detailed analysis of signal dynamics by decomposing time-domain signals into their frequency components. Mathematically, the CWT is defined as:

$$\text{CWT}(s, \tau) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t - \tau}{s} \right) dt. \tag{7}$$

Where $\text{CWT}(s, \tau)$ represents the wavelet coefficients at scale s and translation τ, $x(t)$ is the input signal, and $\psi^*(t)$ is the complex conjugate of the mother wavelet $\psi(t)$. The scale s adjusts the wavelet's compression, and τ shifts it in time. For this study, the Complex Morlet wavelet was employed due to its superior localization in time and frequency domains, defined as:

$$\psi(t) = \frac{1}{\sqrt{\pi B}} e^{-\frac{t^2}{B}} e^{2\pi i C t}. \tag{8}$$

Where B is the bandwidth, C is the center frequency, and i is the imaginary unit. We use '`cmor1.5-1.0`' to correspond to the center frequency and bandwidth of the wavelet, 1.5 and 1.0 respectively.

Frequency components were extracted across five distinct bands to capture the temporal-frequency characteristics of the EEG signals: 0.5–8 Hz, 8–30 Hz, 30–50 Hz, 50–85 Hz, and 85–100 Hz. A total of 32 frequencies were distributed across these bands with varying emphasis to reflect their relative significance. Specifically, 5% of the total frequencies were allocated to the 0.5–8 Hz band (dominated by delta and theta waves), 40% to the 8–30 Hz band (alpha and beta waves critical for motor imagery tasks), 30% to the 30–50 Hz band

(gamma waves), 5% to the 50–85 Hz band, and the remaining 20% to the 85–100 Hz band to balance the resolution across different frequency ranges. This non-uniform distribution ensures that lower-frequency bands, which are often more informative in EEG analysis, are represented adequately without overwhelming higher-frequency components. The choice of 32 frequencies for CWT was motivated by the necessity of balancing temporal resolution and frequency decomposition in EEG signals. A higher number of frequencies may enhance spectral detail but risk redundancy and overfitting, especially in datasets with limited samples. Conversely, lower frequencies sacrifice spectral information crucial for motor imagery tasks. This selection provided an optimal trade-off, capturing the critical spectral patterns of motor imagery-related EEG signals while maintaining computational efficiency.

The CWT coefficients, derived using these frequencies, were normalized via z-score normalization to ensure consistency across features. A batch-wide CWT was computed channel-wise, producing 2D tensors with dimensions (B, F, T, C) representing batch size, frequencies, time points, and channels respectively. These tensors provide a structured representation of the temporal-frequency features, enabling parallel processing in subsequent convolutional layers. This efficient framework facilitates the extraction of critical frequency-domain features while maintaining computational feasibility, thereby enhancing the model's ability to perform robust time-frequency analysis.

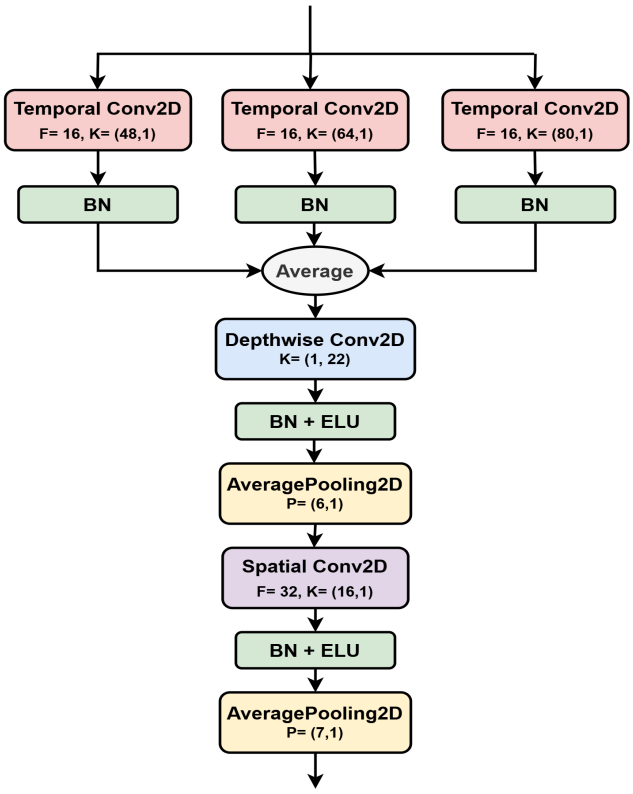

FIGURE 2. Time Series Convolution Block

D. TIME-SERIES CONVOLUTION (TS-CONV) BLOCK

The TS-Conv block depicted in Fig. 2, is designed to efficiently extract both temporal and spatial features from EEG signals, utilizing a structured sequence of convolutional operations. Initially, the first layer performs temporal convolutions across three parallel pathways using F_1 filters of sizes $(K_{T1}, 1)$, $(K_{T1} + 16, 1)$, and $(K_{T1} - 16, 1)$, where K_{T1} is set to be 64 (close to one-fourth of the sampling rate). This approach incorporates multi-kernel sizes to capture temporal patterns at varying scales. The outputs of these pathways are subsequently averaged, resulting in F_1 temporal feature maps that coalesce multi-scale temporal information.

Following this, he second layer applies depthwise convolution with $F_2 = F_1 \times D$ filters of size $(1, C)$, where C represents the number of EEG channels, and D is the number of filters linked to each temporal feature map in the previous layer. This operation extracts spatial features specific to each temporal feature map, learning spatial dependencies between EEG electrodes. The spatially enriched feature maps are then abstracted using an average pooling layer of size $(P_{T1}, 1)$, reducing the temporal resolution while preserving critical spatial information.

Subsequently, third layer refines the feature representations using F_2 filters of size $(K_{T2}, 1)$, further emphasizing task-relevant features. This is followed by another average pooling operation with a kernel size of $(P_{T2}, 1)$, which reduces the feature map resolution for computational efficiency. The second and third layers include batch normalization and ELU activations for faster convergence and nonlinearity.

The TS-Conv block outputs a sequence $\mathbf{z}_i \in \mathbb{R}^{T_c \times d}$, where T_c is the downsampled temporal length, and $d = F_2 = F_1 \times D$ represents the spatial-temporal feature dimensions. The length of the temporal sequence is determined by $T_c = T/(P_{T1} \times P_{T2})$, where T is the original number of time points. This block efficiently captures both temporal and spatial characteristics of EEG signals, making it well-suited for the next Transformer Encoder block for better classification.

E. TIME-FREQUENCY CONVOLUTION (TF-CONV) BLOCK

The TF-Conv block is specifically designed to extract time-frequency representations from EEG signals, aligning these with the demands of attention-based architectures. This is achieved through a carefully structured sequence of convolutional operations, as illustrated in Fig. 3, which focus on temporal, spatial, and spectral feature extraction to ensure robustness and efficiency.

The process begins with a separable convolution in the first layer, utilizing F_1 filters of size $(1, K_T)$, where K_T represents the kernel length along the temporal axis, set to 10 to align with the high-resolution temporal characteristics of EEG data. This approach decomposes the convolution into depthwise and pointwise operations, enabling precise

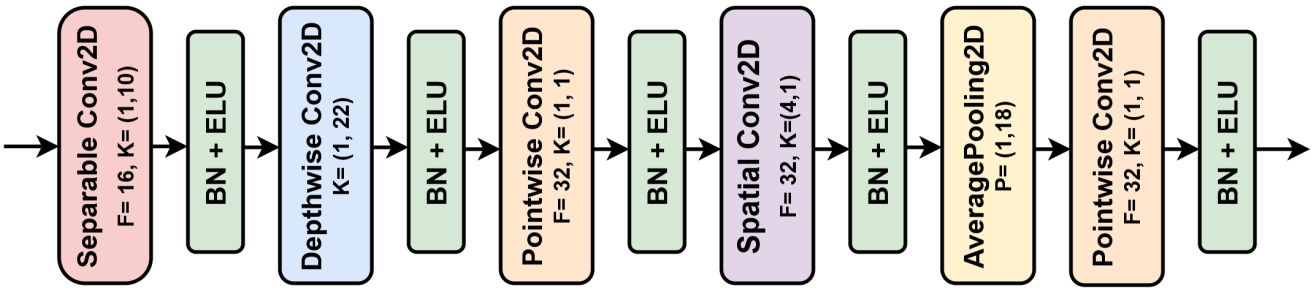

FIGURE 3. Time Frequency Convolution Block

temporal feature extraction while minimizing computational overhead.

Next, a depthwise convolution with filters of size $(1, C)$, where C is the number of EEG channels, captures spatial dependencies between electrodes. This step is crucial for learning region-specific neural activity patterns associated with motor imagery tasks.

Additionally, a third layer employs a pointwise convolution with F_2 filters of size $(1, 1)$, followed by a second convolution of size $(K_F, 1)$ to further refine the frequency feature representations. These layers emphasize time-frequency patterns crucial for distinguishing between motor imagery classes. The resulting spatial-temporal feature maps are downsampled through average pooling along the channels with a kernel size of $(1, P_T)$, where P_T is set to 18, effectively reducing the channel resolution while preserving temporal frequency context. At last, the final layer applies another pointwise convolution with F_2 filters to enhance feature representation. All the convolution layers are followed by batch normalization and an exponential linear unit (ELU) activation, ensuring numerical stability and nonlinearity.

As an output, the TF-Conv block produces a feature tensor $\mathbf{z}_i \in \mathbb{R}^{T_c \times d}$, where T_c represents the frequency resolution, and $d = F_2$ denotes the feature dimensionality along time. This tensor serves as an enriched time-frequency representation of the EEG signal, optimized for the following attention-based mechanisms.

F. TRANSFORMER ENCODER BLOCK

The Transformer Encoder block in the proposed architecture is specifically designed to capture both local and global dependencies within EEG data. This is achieved by leveraging positional encoding, multi-head cross-attention, and a feed-forward network. A key innovation of this study lies in utilizing frequency features as the key in the cross-attention mechanism, which enables the model to learn rich inter-domain correlations between temporal and frequency representations.

1) Positional Encoding

To retain sequential order information, positional embeddings are added to the input feature representations. For an input temporal feature map $\mathbf{Z} \in \mathbb{R}^{T \times d_{\text{model}}}$ from TS-Conv's output and frequency feature map $\mathbf{F} \in \mathbb{R}^{F \times d_{\text{model}}}$ from TF-Conv's output, the positional encoding is defined as:

$$\mathbf{Z}_{\text{pos}} = \mathbf{Z} + \text{PE}_{\text{time}}, \quad \mathbf{F}_{\text{pos}} = \mathbf{F} + \text{PE}_{\text{freq}} \tag{9}$$

where PE_{time} and PE_{freq} are learnable embeddings indexed by the temporal and frequency positions, respectively. This approach ensures that the model preserves structural information, while simultaneously facilitating effective attention operations. Subsequently, the embeddings are layer-normalized to stabilize the training process.

2) Multi-Head Cross-Attention

The central component of the Transformer Encoder is the multi-head cross-attention mechanism, where temporal features $\mathbf{Z}_{\text{pos}}$ are processed as queries ($\mathbf{Q}$) and values ($\mathbf{V}$), whereas frequency features $\mathbf{F}_{\text{pos}}$ are utilized as keys ($\mathbf{K}$). This configuration aligns with the inherent properties of EEG data, where frequency-domain features provide a stable reference for attention mechanisms to focus on dynamic spatio-temporal variations. The queries, keys, and values are derived through linear transformations:

$$\mathbf{Q} = \mathbf{Z}_{\text{pos}}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{F}_{\text{pos}}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{F}_{\text{pos}}\mathbf{W}_V \tag{10}$$

Where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}$ are learnable weight matrices, and $d_{\text{head}} = d_{\text{model}}/\text{num\_heads}$ is the dimensionality of each attention head.

The scaled dot-product attention for each head is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_{\text{head}}}}\right)\mathbf{V} \tag{11}$$

where $d_k = d_{\text{model}}/\text{num\_heads}$ is the scaling factor for numerical stability. The attention scores are normalized using the Softmax function to generate a weighting matrix that emphasizes significant temporal-frequency correlations.

The output from all attention heads is concatenated and projected back to the original embedding size using:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}_O \tag{12}$$

where $\mathbf{W}_O \in \mathbb{R}^{hd_{\text{head}} \times d_{\text{model}}}$ is the projection matrix.

For input dimensions of $\mathbf{Z} \in \mathbb{R}^{T \times d_{\text{model}}}$ and $\mathbf{F} \in \mathbb{R}^{F \times d_{\text{model}}}$, queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ are reshaped for h heads,

resulting in $(B, h, T, d_{\text{head}})$ for $\mathbf{Q}$ and $\mathbf{V}$, and $(B, h, F, d_{\text{head}})$ for $\mathbf{K}$.

Attention outputs are computed for each head, then reshaped back to (B, T, d_{model}) after concatenation. This cross-attention mechanism allows the temporal features to selectively refine information from frequency features, capturing task-relevant dependencies essential for EEG signal decoding.

Feed-Forward Network

The feed-forward block consists of two dense layers with an intermediate expansion factor e, embedding size, and a dropout probability p:

$$\begin{aligned} \mathbf{H}_1 &= \text{GELU}(\mathbf{Z}_{\text{out}}\mathbf{W}_1 + \mathbf{b}_1), \\ \mathbf{H}_2 &= \text{Dropout}(\mathbf{H}_1)\mathbf{W}_2 + \mathbf{b}_2 \end{aligned} \tag{13}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_{\text{model}} \times e^{d_{\text{model}}}}$, $\mathbf{W}_2 \in \mathbb{R}^{e^{d_{\text{model}}} \times d_{\text{model}}}$, and $\mathbf{b}_1, \mathbf{b}_2$ are biases.

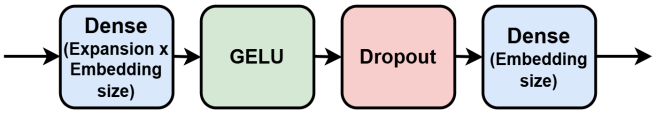

FIGURE 4. Feed Forward Block

The non-linearity and dropout regularization improve generalization while refining task-relevant features. Residual connections are applied around both the attention and feed-forward components. These connections preserve gradient flow, enabling deeper model training and enhancing convergence stability.

G. TEMPORAL CONVOLUTIONAL NETWORK BLOCK

The Temporal Convolutional Network (TCN) block processes sequential embeddings output by the transformer encoder, denoted as $H_{\text{enc}} \in \mathbb{R}^{T \times d}$, where T is the sequence length and d is the embedding dimension. This block is designed to capture long-term temporal dependencies using dilated causal convolutions while leveraging gated mechanisms and residual connections for improved feature extraction and efficient gradient propagation. The diagram of this block is given in Fig.5.

The TCN block comprises a series of residual layers, each built with dilated causal convolutional layers. Causal convolutions ensure that the output at time t depends solely on inputs from time t and earlier, preserving temporal order. The dilation factor d increases exponentially with layer depth, i.e., $d = 2^l$ for the l-th layer, exponentially expanding the receptive field without significantly increasing computational complexity. The output of a dilated causal convolution is defined as:

$$y[t] = \sum_{k=0}^{K-1} W_k \cdot H[t - d \cdot k] \tag{14}$$

Where K is the kernel size, and W_k are the convolutional filters.

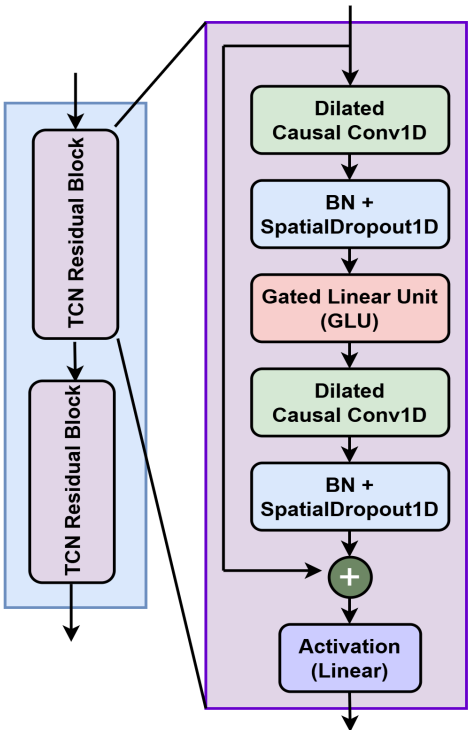

FIGURE 5. Temporal Convolution Block

To enhance temporal feature extraction, Gated Linear Units (GLUs) are integrated into each layer. The convolutional output, $Z \in \mathbb{R}^{T \times 2f}$, is split into two equal parts, Z_1 and Z_2, and processed as:

$$\text{GLU}(Z) = Z_1 \odot \sigma(Z_2) \tag{15}$$

Where σ is the sigmoid function, and $\odot$ denotes element-wise multiplication. GLUs adaptively gate features, focusing on relevant temporal patterns.

Residual connections play a crucial role in the TCN block by mitigating gradient vanishing issues, enabling deeper architectures. When the dimensions of the input and convolutional output differ, a 1×1 convolution is applied to align them before the residual addition:

$$Y_{\text{res}} = \text{Conv1D}(H_{\text{enc}}) + Y \tag{16}$$

Batch normalization stabilizes training, and SpatialDropout1D ensures effective regularization by randomly dropping entire feature maps. The receptive field size (RFS) of the TCN block, determined by the kernel size K and the number of residual blocks L, is given by:

$$\text{RFS} = 1 + 2(K-1)(2^L - 1) \tag{17}$$

In the proposed model, the input sequence length of TCN is 23 and $L = 2$. Information is not ignored only if the RFS is greater than the length of the input sequence. Therefore, we set K to 5 (RFS $= 25 > 23$) for all convolution layers.

The output of the TCN block, $Y \in \mathbb{R}^{T \times f}$, where f is the number of filters, is passed to subsequent layers for classification. This design ensures robust temporal feature extraction

while maintaining computational efficiency, making it well-suited for motor imagery classification.

H. CLASSIFIER MODULE

The classifier module serves as the final stage of the framework, translating the high-level temporal features extracted by the preceding network into actionable classification outputs. It comprises two primary components: a Global Average Pooling (GAP) layer and a fully connected (dense) layer, followed by a Softmax activation.

The GAP layer aggregates temporal information from the output of the Temporal Convolutional Network (TCN) block, denoted as $H_{\text{TCN}} \in \mathbb{R}^{T \times f}$, where T is the sequence length and f is the number of filters. By computing the mean across the temporal dimension, the GAP operation condenses the feature map into a vector $z \in \mathbb{R}^f$, ensuring the network maintains global temporal context without introducing additional parameters. This operation is defined as:

$$z[j] = \frac{1}{T} \sum_{t=1}^{T} H_{\text{TCN}}[t, j], \quad \text{for } j = 1, \ldots, f \tag{18}$$

The condensed representation z is then passed through a dense layer with n_{classes} units, each corresponding to a target class in the motor imagery (MI) task. The dense layer, parameterized with $W \in \mathbb{R}^{f \times n_{\text{classes}}}$ and a bias term $b \in \mathbb{R}^{n_{\text{classes}}}$, performs the linear transformation:

$$\hat{y} = \text{softmax}(W^\top z + b) \tag{19}$$

where the Softmax function ensures that the output $\hat{y}$ represents a valid probability distribution over the classes.

The network is trained using the categorical cross-entropy loss, defined as:

$$\mathcal{L} = -\frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{c=1}^{n_{\text{classes}}} y_{ic} \log(\hat{y}_{ic}) \tag{20}$$

where N_b is the number of trials in a batch, y_{ic} is the true label, and $\hat{y}_{ic}$ is the predicted probability for the c-th class of the i-th trial.

III. RESULT AND ANALYSIS

A. DATASET DESCRIPTION

- **Dataset I** : The Graz BCI Competition IV dataset IIA consists of data from nine participants who performed four distinct motor imagery tasks: imagining movements of the left hand, right hand, both feet, and the tongue. Each trial was preceded by visual and auditory cues, followed by a six-second motor imagery task. Additionally, recordings for eye-related artifacts were captured under various conditions, including eyes open, eyes closed, and eye movements. To ensure robustness, each participant completed two sessions on different days, with each session comprising six runs of 48 trials (12 trials per task), totaling 288 trials per session.Data acquisition involved 22 EEG electrodes and 3 EOG channels, with the EEG electrodes arranged according to the international 10-20 system. The signals were sampled at 250 Hz and underwent preprocessing, including bandpass filtering between 0.5 Hz and 100 Hz, as well as a 50 Hz notch filter to eliminate power line noise.
- **Dataset II** : The Graz BCI Competition IV dataset IIB contains EEG recordings from nine right-handed participants performing motor imagery (MI) tasks for left- and right-hand movements. Participants completed five sessions: two without feedback (screening) and three with feedback using a smiley indicator. Each session included six runs of 20 trials for screening, and four runs of 20 trials for feedback.Each trial involved a fixation cross and auditory cues, followed by a 4-second MI task with randomized breaks. EEG was recorded from three bipolar electrodes (C3, Cz, C4) and three monopolar EOG electrodes, sampled at 250 Hz and filtered (0.5–100 Hz bandpass, 50 Hz notch).

B. PERFORMANCE MERTRIC

The key metrics used to evaluate model performance include Accuracy, Cohen's Kappa, Precision, Recall, and F1-Score.Accuracy measures the overall correctness of the model but can be misleading, especially in imbalanced datasets. Cohen's Kappa adjusts for chance agreement and provides a more reliable measure of agreement between predicted and actual labels. Precision focuses on the proportion of correctly predicted positive instances among all predicted positives, while Recall (or Sensitivity) measures the proportion of actual positives correctly identified by the model. Finally, the F1-Score is the harmonic mean of Precision and Recall, offering a balance between the two when both false positives and false negatives matter. The equations are,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{21}$$

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{22}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{23}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{24}$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{25}$$

Here **TP** (True Positive) refers to the correctly predicted positive samples, **TN** (True Negative) indicates the correctly predicted negative samples, **FP** (False Positive) occurs when the model incorrectly predicts a sample as positive **FN** (False Negative) refers to when the model fails to identify a positive sample

$\mathbf{P}_0$ means the proportion of times the model's predictions match the actual labels and $\mathbf{P}_e$ means the proportion of agreement expected by chance calculated based on the model's predictions. The best training result from both the data set is given below in Table I,II:

TABLE I. Performance Metric on Dataset I

Subject	Accuracy	Kappa	Precision	Recall	F1
1	91.32	0.884	0.915	0.913	0.913
2	78.47	0.713	0.81	0.784	0.775
3	97.67	0.969	0.979	0.978	0.978
4	89.24	0.856	0.898	0.892	0.893
5	82.99	0.773	0.857	0.831	0.827
6	77.43	0.699	0.79	0.77	0.78
7	94.44	0.926	0.946	0.945	0.945
8	88.19	0.843	0.884	0.882	0.882
9	90.97	0.885	0.916	0.914	0.914
Average	**87.8578**	**0.8387**	**0.8883**	**0.8788**	**0.8786**

TABLE II. Performance Metric on Dataset II

Subject	Accuracy	Kappa	Precision	Recall	F1
1	84.06	0.681	0.858	0.841	0.839
2	78.93	0.579	0.796	0.789	0.788
3	89.69	0.794	0.897	0.897	0.897
4	99.06	0.981	0.991	0.991	0.991
5	99.38	0.988	0.994	0.994	0.994
6	91.56	0.831	0.918	0.916	0.916
7	95.31	0.906	0.955	0.953	0.953
8	95.94	0.919	0.959	0.959	0.959
9	90.00	0.800	0.900	0.900	0.900
Average	**91.5478**	**0.831**	**0.91867**	**0.91556**	**0.91522**

C. TRAINING

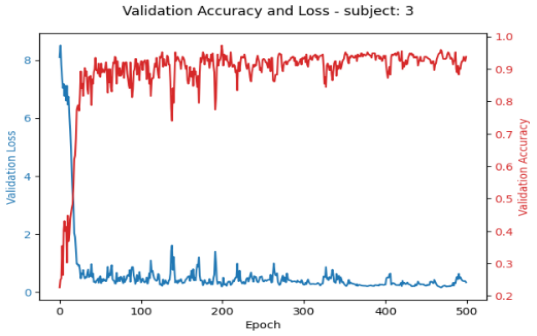

FIGURE 6. Learning Curve

The proposed model was trained on an NVIDIA Tesla P100-PCIE GPU with 16 GB of RAM, using the Python's TensorFlow framework. The training data comprised all EEG channels, excluding extra EOG channels, with no explicit artifact removal. Training was conducted for a maximum of 1000 epochs with a batch size of 64, using the Adam optimizer and a categorical cross-entropy loss function. To prevent overfitting, an early stopping mechanism with a patience of 300 epochs was employed, along with a Reduce-on-Plateau learning rate scheduler. The initial learning rate was set to 0.0009, ensuring gradual convergence. A random seed of 42 was used to ensure reproducibility. To enhance model reliability, training was repeated for 10 independent runs. As shown in the learning curves for subject-3 in Fig. 6, the model demonstrated rapid convergence within the first 100 epochs, with validation accuracy stabilizing at its maximum value and validation loss plateauing at low levels. Minor fluctuations in later epochs indicated potential overfitting, thereby confirming the effectiveness of the early stopping strategy. These results highlight the model's ability to learn robust features from EEG data while maintaining generalization to unseen samples, offering an efficient training framework for Motor Imagery classification.

D. COMPARISON

From the TABLE III and IV, we can say that our model outperformed all others, achieving a classification accuracy of 87.85% for Dataset I and 91.55% for Dataset II. It achieved the highest subject-wise accuracy for subjects 1, 3, 4, 6, and 7 in Dataset I, and 2, 3, 4, 5, 7, and 9 in Dataset II, with kappa scores of 0.84 and 0.82, respectively. The model's superiority is further validated by the lowest standard deviation. Also our proposed model has less parameter and computational cost with less complexity which is depicted in table 4.

TABLE III. Comparison of Methods by Parameters and FLOPs

Method	Parameters	FLOPs
EEGNET [37]	3.58k	15.6M
EEGTCNET [38]	15.26k	7.2M
FBMSNET [39]	16k	147M
Conformer [40]	800k	201M
EEG-Inception [41]	9.1M	270M
MFRCNET [42]	13k	113M
FBCNET [43]	9.5k	210M
ATCNET [35]	115k	118M
Ours	**82k**	**190M**

From the confusion matrices (Fig. 7), left-hand imagery detection had the highest precision (89%), while tongue imagery was the most challenging, with 84% accuracy in Dataset I. For Dataset II, both tasks achieved similar accuracy levels.

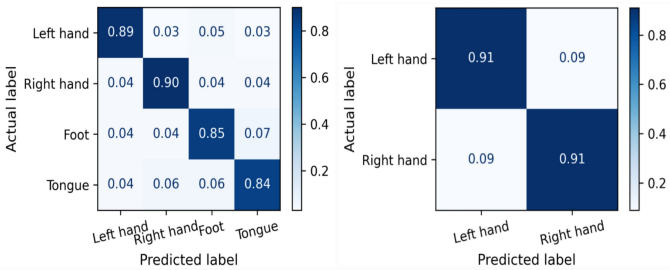

FIGURE 7. Confusion Matrix for dataset I ,II

E. ABLATION STUDY

The architecture's modular design was rigorously evaluated through an ablation study, where each block was systematically removed to assess its individual contribution. The study highlights the critical role of each block in achieving high classification performance, as evidenced by the results

TABLE IV. Comparison Results of Different Methods on Dataset I [Avg Acc: The Average Accuracy (%)]

Year	Methods	S1	S2	S3	S4	S5	S6	S7	S8	S9	Avg Acc	Std	Kappa
2017	FB-CSP [44]	76.00	56.50	81.20	61.20	60.80	55.00	45.20	82.80	81.30	66.67	13.80	0.56
2018	EEGNet [37]	83.33	63.80	85.76	61.85	58.87	59.34	84.81	82.12	76.85	72.97	11.74	0.67
2020	EEGTCNet [38]	85.77	65.02	94.51	64.91	75.36	61.40	87.36	83.76	78.03	77.35	11.58	0.71
2021	TCNetFusion [45]	90.74	70.67	95.23	76.75	82.24	68.83	94.22	88.92	85.98	83.73	9.79	0.78
2022	TSF-STAN [46]	88.30	**81.70**	92.20	77.60	63.30	67.50	90.00	**95.00**	**91.70**	83.00	11.40	0.77
2022	ATCNet [35]	88.50	70.50	97.60	81.00	83.00	73.60	93.10	90.30	91.00	85.40	9.10	0.81
2023	IFNet [47]	83.68	51.74	90.83	76.25	67.85	57.50	88.75	82.15	84.17	75.88	13.89	0.68
2023	Conformer [40]	88.19	61.46	93.40	78.13	52.08	65.28	92.36	88.19	88.89	78.66	15.30	0.72
2024	DMSA-MSNet [48]	88.28	72.57	96.88	83.33	**85.07**	77.26	94.24	88.16	90.82	86.29	7.77	0.82
2024	**Proposed**	**91.32**	78.47	**97.67**	**89.24**	82.99	**77.43**	**94.44**	88.19	90.97	**87.85**	**6.91**	**0.84**

TABLE V. Comparison Results of Different Methods on Dataset II [Avg Acc: The Average Accuracy (%)]

Year	Methods	S1	S2	S3	S4	S5	S6	S7	S8	S9	Avg Acc	Std	Kappa
2017	FB-CSP [44]	70.00	60.40	60.90	97.50	93.10	80.60	78.10	92.50	87.90	80.11	13.92	0.61
2018	EEGNet [37]	70.31	70.36	78.44	95.33	93.44	82.18	91.88	87.19	71.65	82.31	10.15	0.65
2020	EEGTCNet [38]	69.69	66.07	77.22	96.14	91.13	74.57	84.96	91.25	74.38	80.60	10.62	0.64
2021	TCNetFusion [45]	82.31	77.68	89.32	98.06	95.69	83.17	92.34	**96.58**	76.34	87.94	8.71	0.75
2022	TSF-STAN [46]	**86.10**	77.90	67.60	98.50	91.70	**95.80**	93.70	90.80	89.70	88.00	9.61	0.76
2022	ATCNet [35]	78.44	74.92	86.88	98.74	94.69	87.81	93.13	93.44	86.56	88.29	7.75	0.77
2023	IFNet [47]	73.25	54.71	59.06	97.50	86.19	84.62	83.19	92.81	90.19	80.17	14.86	0.60
2023	Conformer [40]	82.50	65.71	63.75	98.44	86.56	90.31	87.81	94.38	92.19	84.63	11.50	0.69
2024	DMSA-MSNet [48]	81.87	74.64	88.75	98.75	98.69	91.00	94.69	95.00	91.25	90.52	8.06	0.81
2024	**Proposed**	84.06	**78.93**	**89.69**	**99.06**	**99.38**	91.56	**95.31**	95.94	**90.00**	**91.55**	**6.82**	**0.83**

summarized in Table V. The removal of blocks causes a noticeable drop in accuracy due to the loss of specific functionalities offered by these components. Below, we detail the impact of removing each block, emphasizing their significance in addressing the unique challenges posed by EEG-based MI classification.

The data augmentation pipeline serves to expand the diversity of training samples by creating synthetic variations while preserving the essential characteristics of MI signals. Its removal leads to a significant accuracy drop of 3.36% (Dataset I) and 2.11% (Dataset II), underscoring its role in reducing overfitting and enabling the model to generalize effectively. The absence of augmentation restricts the model's ability to learn robust features, particularly in low-resource scenarios, making it vulnerable to the inherent variability of EEG signals.

The TCN block leverages dilated causal convolutions to model long-range temporal dependencies while maintaining temporal causality. This is critical for MI tasks, where temporal dynamics hold vital discriminative information. The use of Gated Linear Units (GLUs) further enhances feature selectivity by suppressing irrelevant patterns. When removed, the model suffers a notable drop in accuracy of 2.78% (Dataset I) and 0.29% (Dataset II).

The feed-forward network, embedded within the Transformer Encoder, is responsible for refining the attention-enhanced features by mapping them into a higher-dimensional space. Its role in enhancing the discriminative power of features is evident from the 2.12% and 0.32% accuracy drop for Dataset I and Dataset II, respectively, when removed.

Positional encoding enables the Transformer to retain temporal and spatial sequence information essential for analyzing spatiotemporal dependencies in EEG signals. Without this component, the model's ability to discern the sequential structure of input data is compromised, leading to accuracy losses of 1.97% (Dataset I) and 0.38% (Dataset II).

TABLE VI. Effects of Different Block

Removed Block	Accuracy Decreased by (%)	
	Dataset I	Dataset II
Augmentation	3.36	2.11
TCN	2.78	0.29
FF	2.12	0.32
PE	1.97	0.38
Residual	1.97	0.12
Transformer	3.82	2.66

Residual connections are pivotal in facilitating efficient gradient flow and preventing vanishing gradient issues, especially in deep architectures. They also allow for the preserva-

tion of unaltered features, which can be critical for classification tasks. When residual connections are removed, accuracy decreases by 1.97% (Dataset I) and 0.12% (Dataset II).

The Transformer Encoder serves as the backbone for cross-domain feature integration, capturing both local and global dependencies through multi-head cross-attention. Removal of this module completely loses the time-frequency representations resulting in the absence of inter-domain relationships crucial for MI-EEG classification. Consequently, it results in the highest accuracy degradation after data augmentation, with decreases of 3.82% (Dataset I) and 2.66% (Dataset II) symbolizing the Transformer Encoder's unparalleled role in ensuring robust and context-aware feature representation.

The t-SNE plots in Fig. 8 provide a visual understanding of the effect of block removal on feature clustering and class separability. With all blocks intact (subplot 1), the features form compact, well-separated clusters, indicating robust class discrimination. The removal of blocks progressively disrupts this separability. After augmentation removal(subplot 2) class clusters become less distinct and interspersed, reflecting the reduced feature diversity and generalization capability. When Transformer Encoder is removed, the most significant degradation in separability occurs, with clusters overlapping extensively (subplot 6), indicating the loss of inter-domain synergy. The separability degrades incrementally, with feature clusters appearing less compact, reflecting the partial loss of functionality offered by the respective blocks (subplots 3-5).

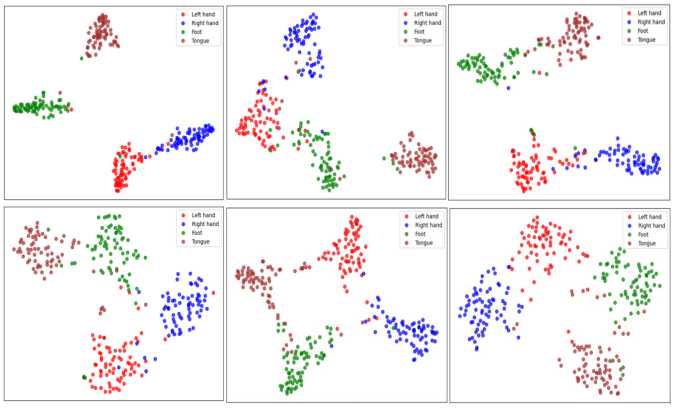

FIGURE 8. t-SNE Plots after Removal of (1) No Block Removed, (2) Augmentation (3) TCN (4) FF (5) PE (6) Transformer (Numbered Left to Right)

Overall, the t-SNE plots validate the synergistic contribution of all blocks, demonstrating how their interplay ensures the extraction of robust and distinct features, critical for MI classification.

F. EFFECT OF HYPER PARAMETER

In this section, we evaluate the selection and impact of key hyperparameters in our proposed architecture. The selections are grounded in both empirical evidence and theoretical insights, ensuring an optimal balance between model complexity and generalization. Fig. 9 illustrates the performance trends under varying hyperparameter configurations.

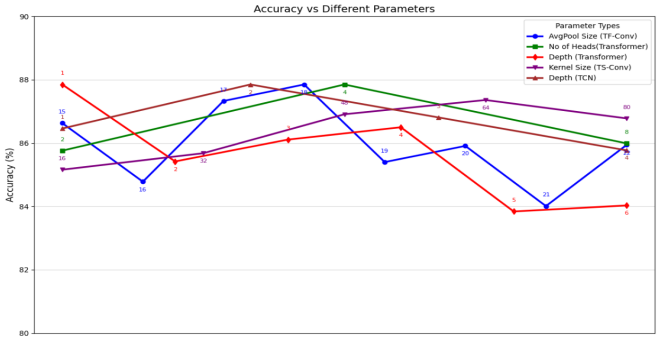

FIGURE 9. Parameters effect on Accuracy

The use of parallel multi-kernel convolutions with sizes 48, 64, and 80 reflects a deliberate design choice to capture temporal features across multiple resolutions. Kernel size 48 focuses on fine-grained temporal details, while 80 emphasizes broader temporal dependencies, with 64 balancing these scales yielding the highest accuracy (Fig. 9). This parallel structure enables the extraction of complementary temporal dynamics, enriching the feature space and enhancing discriminative capability. The combination leverages the diversity of receptive fields to process intricate temporal patterns without over-relying on a single scale. The depthwise convolution with a kernel size of (1,22) aligns with the dataset's 22-channel EEG configuration, effectively modeling spatial relationships across channels. Following this, we applied progressive average pooling with sizes 6 and 7. The pooling size of 6 downsampled temporal dimensions, preserving relevant information while eliminating noise. The intermediate convolution block with kernel size (16,1) further refined the features, emphasizing temporal specificity before applying pooling with size 7. This sequential strategy ensures effective hierarchical feature abstraction while aligning feature dimensions for seamless integration with subsequent modules.

The separable convolution with kernel size (1,10) was selected after observing that larger kernels (e.g., 16, 32, 64) led to substantial accuracy drops due to over-smoothing, while smaller kernels (e.g., 6, 8) retained excessive noise. The kernel size of 10 struck a balance, effectively capturing local temporal dependencies. Then the depthwise convolution with kernel size (1, 22) isolates inter-channel interactions, effectively processing multi-channel EEG data. The frequency-wise convolution kernel size of (4,1) emerged as optimal for capturing inter-frequency interactions. Smaller kernels (e.g., 2) failed to generalize adequately, while larger ones (e.g., 8) introduced redundancies. The average pooling kernel size of (1,18) was found to maximize performance by smoothing noisy features while retaining salient temporal patterns while reducing channel dimension to 1. As demonstrated in Fig. 6, smaller and larger pooling sizes degraded accuracy, un-

derscoring the need for a carefully tuned pooling operation to align with the inherent time-frequency resolution of EEG data.

The selection of 4 attention heads and an embedding size of 32 was grounded in the trade-off between model complexity and feature discrimination. Increasing the number of heads beyond four led to diminishing returns, while fewer heads hindered the ability to capture diverse aspects of the feature space. The transformer depth of 1 was optimal for this task, as deeper layers introduced feature redundancy and diluted critical temporal patterns extracted from earlier layers. This aligns with findings in prior EEG research, where shallow transformer architectures excel in decoding relevant and highly informative signals without overfitting.

To combat overfitting, we applied a dropout rate of 0.3 in TS-Conv and TCN, 0.5 before average pooling, and 0.6 after average pooling in TF-Conv. These values were chosen to balance generalization and feature retention, mitigating overfitting risks while preserving critical information flow. And higher dropout after pooling effectively reduces redundancy in the compact feature space. Additionally, a weight decay of 0.009 and max-norm constraints of 0.6 were used to stabilize training and prevent gradient explosions. The weight decay penalizes large weights, encouraging smoother model updates, while max-norm constraints enforce upper bounds on weight magnitudes, further enhancing training stability.The rigorous selection of hyperparameters and architectural choices reflects the interplay between theoretical considerations and empirical evidence. By tailoring kernel sizes, pooling strategies, and regularization parameters to the unique characteristics of EEG data, the proposed model achieves robust and efficient motor imagery classification. The modular design ensures that each component contributes optimally to the overall performance.

IV. LIMITATIONS AND FUTURE WORK

While the proposed architecture demonstrates robust performance and computational efficiency for motor imagery (MI) classification, several limitations warrant attention. Firstly, the model's reliance on CWT for feature extraction, while effective, introduces a computational overhead and longer processing time requirement that may limit its applicability in real-time, resource-constrained BCI systems. Additionally, the fixed selection of 32 frequencies in CWT may not fully capture task-specific spectral characteristics, potentially reducing generalization across diverse datasets. Furthermore, the model relies heavily on data augmentation for improved generalization, indicating potential limitations in learning from small training datasets without augmentation. Finally, the absence of subject-independent validation in the current work raises concerns regarding the robustness and adaptability of the architecture to unseen subjects.

Future work will explore alternative signal processing techniques, including Discrete Wavelet Transform (DWT), Short-Time Fourier Transform (STFT), and Fast Fourier Transform (FFT), alongside CWT, to further enhance frequency-domain feature extraction. DWT offers localized time-frequency decomposition, STFT effectively analyzes non-stationary signals, and FFT improves computational efficiency, making these methods promising for augmenting performance. Additionally, optimizing the time-frequency features jointly with the deep learning pipeline could better integrate its contribution to classification accuracy. Further advancements could involve combining time, frequency, and non-linear features, such as those derived from Empirical Mode Decomposition (EMD), to construct a richer and more diverse feature set. Also, incorporating domain adaptation techniques or transfer learning strategies could help address inter-subject variability and improve practical applicability. This comprehensive approach aims to develop more accurate, adaptive, and computationally efficient BCI systems, driving future progress in motor imagery classification.

V. CONCLUSION

This study presents a novel dual-path EEG classification framework that synergizes time-domain and frequency-domain feature extraction to achieve state-of-the-art performance in motor imagery classification.The integration of multi-kernel convolutions, separable convolutions, and cross-domain attention mechanisms further enhances the architecture's ability to extract multi-scale, domain-specific features efficiently. Comprehensive evaluations revealed that the selective use of kernel sizes, pooling strategies, and optimal Transformer depth, receptive field balances feature extraction and generalization. This work not only advances motor imagery classification but also provides a scalable foundation for adaptive and efficient BCI systems, paving the way for future innovations in neural interface technology.

.

Appendixes, if needed, appear before the acknowledgment.

REFERENCES

[1] M. L. Martini, E. K. Oermann, N. L. Opie, F. Panov, T. Oxley, and K. Yaeger, "Sensor modalities for brain-computer interface technology: a comprehensive literature review," Neurosurgery, vol. 86, no. 2, pp. E108–E117, 2020.
[2] 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," Journal of neural engineering, vol. 15, no. 3, p. 031005, 2018.
[3] 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, 2007.
[4] N. Weiskopf, K. Mathiak, S. W. Bock, F. Scharnowski, R. Veit, W. Grodd, R. Goebel, and N. Birbaumer, "Principles of a brain-computer interface (bci) based on real-time functional magnetic resonance imaging (fmri)," IEEE transactions on biomedical engineering, vol. 51, no. 6, pp. 966–970, 2004.
[5] J. Shin and W. Chung, "Multi-band cnn with band-dependent kernels and amalgamated cross entropy loss for motor imagery classification," IEEE journal of biomedical and health informatics, 2023.
[6] G. Pfurtscheller and F. L. Da Silva, "Event-related eeg/meg synchronization and desynchronization: basic principles," Clinical neurophysiology, vol. 110, no. 11, pp. 1842–1857, 1999.

[7] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. L. Da Silva, "Mu rhythm (de) synchronization and eeg single-trial classification of different motor imagery tasks," NeuroImage, vol. 31, no. 1, pp. 153–159, 2006.

[8] M. A. Cervera, S. R. Soekadar, J. Ushiba, J. d. R. Millán, M. Liu, N. Birbaumer, and G. Garipelli, "Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis," Annals of clinical and translational neurology, vol. 5, no. 5, pp. 651–663, 2018.

[9] S. Moghimi, A. Kushki, A. Marie Guerguerian, and T. Chau, "A review of eeg-based brain-computer interfaces as access pathways for individuals with severe disabilities," Assistive Technology, vol. 25, no. 2, pp. 99–110, 2013.

[10] N. Padfield, J. Zabalza, H. Zhao, V. Masero, and J. Ren, "Eeg-based brain-computer interfaces using motor-imagery: Techniques and challenges," Sensors, vol. 19, no. 6, p. 1423, 2019.

[11] A. Nourmohammadi, M. Jafari, and T. O. Zander, "A survey on unmanned aerial vehicle remote control using brain–computer interface," IEEE Transactions on Human-Machine Systems, vol. 48, no. 4, pp. 337–348, 2018.

[12] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b," Frontiers in neuroscience, vol. 6, p. 39, 2012.

[13] ——, "Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b," Frontiers in neuroscience, vol. 6, p. 39, 2012.

[14] A. S. Al-Fahoum and A. A. Al-Fraihat, "Methods of eeg signal features extraction using linear analysis in frequency and time-frequency domains," International Scholarly Research Notices, vol. 2014, no. 1, p. 730218, 2014.

[15] Y. Wang, K. C. Veluvolu, and M. Lee, "Time-frequency analysis of band-limited eeg with bmflc and kalman filter for bci applications," Journal of neuroengineering and rehabilitation, vol. 10, pp. 1–16, 2013.

[16] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," IEEE transactions on Neural Networks, vol. 13, no. 2, pp. 415–425, 2002.

[17] P. Gaur, H. Gupta, A. Chowdhury, K. McCreadie, R. B. Pachori, and H. Wang, "A sliding window common spatial pattern for enhancing motor imagery classification in eeg-bci," IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1–9, 2021.

[18] R. Fu, Y. Tian, T. Bao, Z. Meng, and P. Shi, "Improvement motor imagery eeg classification based on regularized linear discriminant analysis," Journal of medical systems, vol. 43, pp. 1–13, 2019.

[19] E. Dong, C. Li, L. Li, S. Du, A. N. Belkacem, and C. Chen, "Classification of multi-class motor imagery with a novel hierarchical svm algorithm for brain–computer interfaces," Medical & biological engineering & computing, vol. 55, pp. 1809–1818, 2017.

[20] J. Luo, Z. Feng, J. Zhang, and N. Lu, "Dynamic frequency feature selection based approach for classification of motor imageries," Computers in biology and medicine, vol. 75, pp. 45–53, 2016.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, pp. 436–444, 2015.

[22] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," Journal of big Data, vol. 8, pp. 1–74, 2021.

[23] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: a review," Journal of neural engineering, vol. 16, no. 3, p. 031001, 2019.

[24] B. Zhang, W. Wang, Y. Xiao, S. Xiao, S. Chen, S. Chen, G. Xu, and W. Che, "Cross-subject seizure detection in eegs using deep transfer learning," Computational and Mathematical Methods in Medicine, vol. 2020, no. 1, p. 7902072, 2020.

[25] S. Kumar, A. Sharma, and T. Tsunoda, "Brain wave classification using long short-term memory network based optical predictor. sci rep 9: 9153," 2019.

[26] S. Lian and Z. Li, "An end-to-end multi-task motor imagery eeg classification neural network based on dynamic fusion of spectral-temporal features," Computers in Biology and Medicine, p. 108727, 2024.

[27] X. Jia, Y. Song, L. Yang, and L. Xie, "Joint spatial and temporal features extraction for multi-classification of motor imagery eeg," Biomedical Signal Processing and Control, vol. 71, p. 103247, 2022.

[28] A. Vaswani, "Attention is all you need," Advances in Neural Information Processing Systems, 2017.

[29] B. Abibullaev, A. Keutayeva, and A. Zollanvari, "Deep learning in eeg-based bcis: a comprehensive review of transformer models, advantages, challenges, and applications," IEEE Access, 2023.

[30] A. Keutayeva and B. Abibullaev, "Data constraints and performance optimization for transformer-based models in eeg-based brain-computer interfaces: A survey," IEEE Access, 2024.

[31] W. Hang, W. Feng, R. Du, S. Liang, Y. Chen, Q. Wang, and X. Liu, "Cross-subject eeg signal recognition using deep domain adaptation network," IEEE Access, vol. 7, pp. 128 273–128 282, 2019.

[32] X. Wei, P. Ortega, and A. A. Faisal, "Inter-subject deep transfer learning for motor imagery eeg decoding," in 2021 10th international IEEE/EMBS conference on neural engineering (NER). IEEE, 2021, pp. 21–24.

[33] Y. Liang and Y. Ma, "Calibrating eeg features in motor imagery classification tasks with a small amount of current data using multisource fusion transfer learning," Biomedical Signal Processing and Control, vol. 62, p. 102101, 2020.

[34] Y. Li, L. Guo, Y. Liu, J. Liu, and F. Meng, "A temporal-spectral-based squeeze-and-excitation feature fusion network for motor imagery eeg decoding," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 1534–1545, 2021.

[35] H. Altaheri, G. Muhammad, and M. Alsulaiman, "Physics-informed attention temporal convolutional network for eeg-based motor imagery classification," IEEE transactions on industrial informatics, vol. 19, no. 2, pp. 2249–2258, 2022.

[36] Q. She, T. Chen, F. Fang, J. Zhang, Y. Gao, and Y. Zhang, "Improved domain adaptation network based on wasserstein distance for motor imagery eeg classification," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 1137–1148, 2023.

[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," Journal of neural engineering, vol. 15, no. 5, p. 056013, 2018.

[38] T. M. Ingolfsson, M. Hersche, X. Wang, N. Kobayashi, L. Cavigelli, and L. Benini, "Eeg-tcnet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces," in 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2020, pp. 2958–2965.

[39] K. Liu, M. Yang, Z. Yu, G. Wang, and W. Wu, "Fbmsnet: A filter-bank multi-scale convolutional neural network for eeg-based motor imagery decoding," IEEE Transactions on Biomedical Engineering, vol. 70, no. 2, pp. 436–445, 2022.

[40] Y. Song, Q. Zheng, B. Liu, and X. Gao, "Eeg conformer: Convolutional transformer for eeg decoding and visualization," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 710–719, 2022.

[41] C. Zhang, Y.-K. Kim, and A. Eskandarian, "Eeg-inception: an accurate and robust end-to-end neural network for eeg-based motor imagery classification," Journal of Neural Engineering, vol. 18, no. 4, p. 046014, 2021.

[42] X. Li, Z. Yang, X. Tu, J. Wang, and J. Huang, "Mfrc-net: Multi-scale feature residual convolutional neural network for motor imagery decoding," IEEE Journal of Biomedical and Health Informatics, 2024.

[43] R. Mane, E. Chew, K. Chua, K. K. Ang, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan, "Fbcnet: A multi-view convolutional neural network for brain-computer interface," arXiv preprint arXiv:2104.01233, 2021.

[44] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b," Frontiers in neuroscience, vol. 6, p. 39, 2012.

[45] Y. K. Musallam, N. I. AlFassam, G. Muhammad, S. U. Amin, M. Alsulaiman, W. Abdul, H. Altaheri, M. A. Bencherif, and M. Algabri, "Electroencephalography-based motor imagery classification using temporal convolutional network fusion," Biomedical Signal Processing and Control, vol. 69, p. 102826, 2021.

[46] X. Jia, Y. Song, L. Yang, and L. Xie, "Joint spatial and temporal features extraction for multi-classification of motor imagery eeg," Biomedical Signal Processing and Control, vol. 71, p. 103247, 2022.

[47] J. Wang, L. Yao, and Y. Wang, "Ifnet: An interactive frequency convolutional neural network for enhancing motor imagery decoding from eeg," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 1900–1911, 2023.

[48] K. Yang, J. Wang, L. Yang, L. Bian, Z. J. Luo, and C. Yang, "A diagonal masking self-attention-based multi-scale network for motor imagery classification," Journal of Neural Engineering, 2024.