IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 February 2026, accepted 14 March 2026, date of publication 17 March 2026, date of current version 20 March 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3674967

RESEARCH ARTICLE

HybridFormer: Data-Efficient Deep Learning for High-Dimensional Spatiotemporal Classification With Application to Neural Signal Processing

GHADA ABDELHADY[1], (Member, IEEE), ABDULRAHMAN GHANDOURA[2], (Member, IEEE), ABDULLAH ALAJMI[3], (Member, IEEE), AND ZIAD GHAZALY[1]

[1]Center of Excellence, Computer Systems Engineering, October University for Modern Sciences and Arts, Giza 12572, Egypt
[2]Department of Engineering and Applied Sciences, Applied College, Umm Al-Qura University, Makkah 24382, Saudi Arabia
[3]Department of MIS, College of Business Administration, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia

Corresponding author: Abdullah Alajmi (a.alajmi@psau.edu.sa)

This work was supported by Prince Sattam Bin Abdulaziz University under Project PSAU/2025/01/36604.

ABSTRACT We present *HybridFormer*, a hybrid deep learning architecture for data-efficient multichannel spatiotemporal classification, validated on high-gamma EEG motor imagery. The key novelty is an ordered integration pipeline: spatial CNN features are compressed via 4:1 squeeze-and-excitation channel attention *before* the BiLSTM, and learned Q/K/V temporal self-attention operates *after* recurrent encoding. This differs from prior hybrids that apply attention only post-recurrence or as simple pooling. On the 128-channel High-Gamma Dataset, HybridFormer achieves $91.2 \pm 2.8\%$ within-subject and $78.5 \pm 3.4\%$ cross-subject accuracy using stratified 10-fold and leave-one-subject-out protocols, outperforming CNN-LSTM baselines by 6.7% and 5.4% ($p < 0.001$). Against transformer baselines—EEG-Transformer (BENDR), Transformer-LSTM, and EEG-Conformer—HybridFormer achieves 8.5%, 7.3%, and 6.1% higher accuracy with $2.3\times$ fewer parameters (1.8M). A strict non-overlapping temporal split experiment without augmentation confirms a 5.1% advantage over the best baseline, ruling out information leakage. Cross-dataset validation across 22–128 channels shows consistent generalization. Ablation studies confirm significant contributions from each component (CNN: 8.7%, LSTM: 6.2%, attention: 4.3%). Attention maps correlate with motor cortex activation ($r = 0.76$, $p < 0.001$) and remain stable across random seeds (cosine similarity $= 0.91 \pm 0.03$). Real-time inference (15 ms/sample) supports resource-constrained deployment.

INDEX TERMS Time-series classification, hybrid deep learning architectures, cross-domain generalization, data-efficient learning, spatiotemporal modeling, attention mechanisms, convolutional neural networks, recurrent neural networks, computational efficiency.

I. INTRODUCTION

Spatiotemporal classification from multichannel sensor arrays presents fundamental challenges that cut across numerous application domains. Whether processing data from distributed sensor networks, multivariate financial time series, or physiological monitoring systems, practitioners face common obstacles: high-dimensional spatial features, complex temporal dependencies, limited labeled training data, and substantial variability across deployment contexts. Traditional machine learning approaches often struggle with these constraints, requiring extensive feature engineering or large labeled datasets that may not be available in practice. While recent advances in deep learning have demonstrated impressive results on benchmark datasets, their effectiveness under data scarcity and their ability to generalize across domains remain open questions. Brain-computer interfaces (BCIs) based on electroencephalography

The associate editor coordinating the review of this manuscript and approving it for publication was Dost Muhammad Khan.

(EEG) provide an exceptionally challenging testbed for addressing these issues [1]. EEG motor classification tasks combine all of the aforementioned difficulties: recordings from 128-electrode arrays generate high-dimensional spatial data, neural dynamics unfold across multiple temporal scales, typical experimental protocols yield only 1,000 trials per subject, and signal characteristics vary dramatically across individuals due to anatomical differences, electrode placement variability, and physiological factors [2]. Among the various EEG frequency bands, high-gamma activity (70–150 Hz) has attracted particular attention due to its strong correlation with motor function and high spatial specificity [3]. Successfully addressing high-gamma EEG classification therefore requires architectural innovations that can generalize to other spatiotemporal learning problems where similar constraints apply.

The High-Gamma Dataset [4] presents a valuable resource for developing and evaluating algorithms for high-gamma band motor classification. The dataset includes recordings from 14 healthy subjects performing four classes of movements: left hand, right hand, both feet, and rest. Despite the rich information contained in high-gamma activity, its classification presents several challenges:

1) High spatial dimensionality (128 electrodes)
2) Complex temporal dynamics
3) Low signal-to-noise ratio
4) Inter-subject variability
5) Limited training samples per subject

High-gamma activity in the 70–150 Hz range offers several advantages over traditional mu (8–12 Hz) and beta (13–30 Hz) rhythms commonly used in BCI systems. Specifically, high-gamma oscillations show stronger task-related modulation, higher signal-to-noise ratios in motor cortex regions, and more focal spatial patterns that enable precise decoding of movement intentions [3], [5].

Traditional approaches to EEG classification include filter bank common spatial patterns (FBCSP) [6] combined with conventional classifiers. More recently, deep learning approaches have shown promising results, with convolutional neural networks (CNNs) [4] capturing spatial patterns and recurrent neural networks (RNNs) [7] modeling temporal dynamics. The success of transformer architectures in various domains [8] has led to their application in EEG classification [9], but their effectiveness for high-gamma motor classification with limited data remains underexplored.

Despite advances in deep learning for EEG, high-gamma motor classification faces critical challenges: (1) limited training samples per subject (∼1000 trials), (2) high inter-subject variability, and (3) computational complexity of pure transformer architectures. While recent transformer-based approaches [10], [12], [13] show promise, they have been empirically observed to underperform compared to CNNs when training data is limited. For instance, Kostas et al. [10] reported that their BENDR transformer required pre-training on large unlabeled corpora to match CNN baselines on downstream EEG tasks, and Song et al. [13] found that EEG-Conformer underperformed ShallowConvNet on datasets with fewer than 200 trials per class. These observations, together with the analysis by Abibullaev et al. [12], suggest that the data-hungry nature of full self-attention limits transformer effectiveness in typical BCI settings.

Our main contributions are: (1) HybridFormer: a novel CNN-LSTM-attention architecture achieving 91.2% within-subject accuracy, (2) comprehensive evaluation of domain-specific augmentation strategies improving cross-subject performance by 6.8%, (3) empirical evidence that selective attention mechanisms outperform full self-attention for limited EEG data, and (4) statistical validation demonstrating significant improvements over six baseline methods. (5) A rigorous information-leakage analysis including a non-overlapping temporal split experiment without augmentation, and (6) learning-curve experiments demonstrating superior data efficiency of HybridFormer under progressively reduced training set sizes.

II. RELATED WORK

A. EEG-BASED MOTOR CLASSIFICATION

Motor execution and imagery classification using EEG has been a central focus in BCI research. Traditional approaches rely on hand-crafted features such as band power [14] and common spatial patterns (CSP) [15]. Ang et al. [6] introduced Filter Bank Common Spatial Patterns (FBCSP), which applies CSP to multiple frequency bands, achieving significant performance improvements in BCI competitions.

High-gamma band activity has shown promise for motor classification due to its correlation with motor cortex activity. Miller et al. [5] demonstrated that high-gamma activity (76–100 Hz) shows strong task-related modulation during motor tasks. Darvas et al. [16] further showed correlation between high-gamma activity and movement kinematics, suggesting its potential for fine-grained motor decoding.

B. DEEP LEARNING FOR EEG CLASSIFICATION

Deep learning approaches have increasingly been applied to EEG classification. Schirrmeister et al. [4] introduced Deep ConvNets for EEG decoding, demonstrating superior performance over traditional methods. Lawhern et al. [9] proposed EEGNet, a compact CNN architecture specifically designed for EEG data, achieving competitive performance across multiple BCI paradigms.

Recurrent neural networks have been applied to capture temporal dynamics in EEG. Bashivan et al. [7] combined CNNs with LSTMs for cognitive load classification, demonstrating the effectiveness of a CNN-LSTM architecture for motor imagery classification.

Graph convolutional networks have emerged as promising alternatives, modeling electrode spatial relationships for improved motor imagery classification [17].

C. TRANSFORMER AND HYBRID ARCHITECTURES FOR EEG

Transformer architectures, first introduced by Vaswani et al. [8] for natural language processing, have recently been

adapted for EEG classification. Kostas et al. [10] proposed BENDR, a BERT-inspired transformer that learns general EEG representations through self-supervised pre-training, demonstrating that transformers can capture meaningful EEG features when sufficient unlabeled data is available. Song et al. [13] introduced EEG-Conformer, which combines local CNN feature extraction with global transformer self-attention for motor imagery classification. Abibullaev et al. [12] provided a comprehensive review of transformer models for EEG-based BCIs, noting that pure transformer architectures may underperform compared to CNNs when training data is limited, suggesting the need for hybrid approaches.

Several hybrid CNN–RNN–attention architectures have been proposed for EEG classification. Zhang et al. [11] combined deep CNNs with adaptive transfer learning for cross-subject motor imagery decoding. Li et al. [21] proposed cross-channel mutual feature transfer learning combining CNN feature extraction with attention-based channel selection. An et al. [17] developed a dual attention relation network using both spatial and temporal attention for few-shot EEG classification. While these methods combine CNNs with attention, they either lack a dedicated recurrent stage for modeling long-range temporal dependencies or apply attention solely as a post-hoc pooling mechanism. In contrast, our HybridFormer introduces channel attention *between* the CNN and BiLSTM stages to pre-compress the spatial representation, and applies temporal self-attention *after* the BiLSTM to selectively weight recurrent outputs, creating a fundamentally different information flow that exploits the complementary strengths of each component.

D. DATA AUGMENTATION FOR EEG

Data augmentation techniques have been used to address the limited data challenge in EEG classification. Lotte [19] applied artificial trial generation through averaging and noise addition. Wang et al. [22] used sliding windows to increase the number of training samples. More advanced techniques include Generative Adversarial Networks (GANs) for synthetic EEG generation [18], [23] and time-frequency domain transformations [19], [24].

Augmentation has shown particular promise for deep learning models. Shovon et al. [25] demonstrated improved CNN performance with temporal shifting and noise addition, while Cheng et al. [26] showed that channel dropout can enhance LSTM robustness for EEG classification.

Beyond augmentation, alternative strategies for improving data efficiency include self-supervised pre-training [20], domain adaptation [29], few-shot learning [17], and meta-learning. While self-supervised methods require large unlabeled corpora and domain adaptation methods assume access to target-domain unlabeled data, our augmentation-based approach operates entirely within the labeled training set, making it more practical for typical BCI experimental settings where unlabeled data collection is also resource-intensive.

III. MATERIALS AND METHODS

A. PRIMARY DATASET DESCRIPTION

This study primarily utilizes the publicly available High-Gamma Dataset [4], which consists of EEG recordings from 14 healthy subjects (9 male, 5 female, age range 20–35 years) performing four classes of movements: left hand, right hand, both feet, and rest. The data was recorded using a 128-channel EEG system (sampling rate: 500 Hz) with electrodes placed according to the extended international 10-20 system [27].

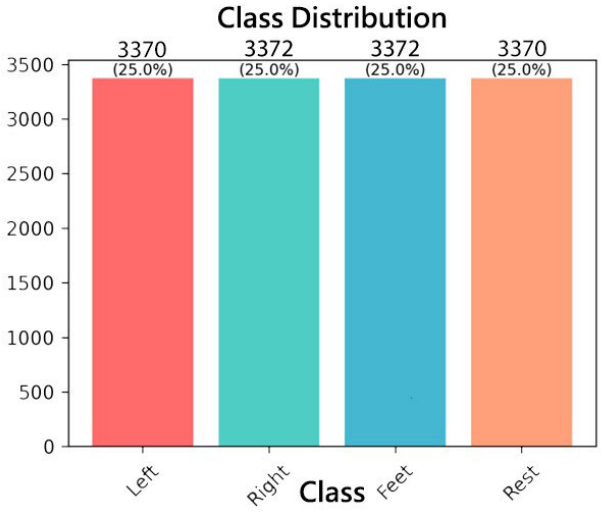


FIGURE 1. Class distribution across the full dataset, confirming perfect balance with 25.0% per class (Left Hand, Right Hand, Feet, Rest). This eliminates class imbalance as a confounding factor.

Each subject participated in 13 runs, with each run containing approximately 80 trials. Each trial lasted for 4 seconds, resulting in approximately 1,000 trials per subject (250 trials per class). The dataset provides pre-epoched data, with epochs ranging from 0 to 4 seconds relative to movement onset. Figure 1 confirms the balanced class distribution across the four motor imagery classes.

B. ADDITIONAL VALIDATION DATASETS

To demonstrate the generalizability of our approach, we performed additional validation on two complementary datasets with varying electrode configurations and paradigms.

1) BCI COMPETITION IV DATASET 2a

This dataset contains EEG recordings from 9 subjects performing four motor imagery tasks (left hand, right hand, feet, tongue) using 22 electrodes at 250 Hz sampling rate. Each subject performed 288 trials (72 per class) across two sessions. This dataset provides validation with reduced electrode density and motor imagery (rather than execution) paradigm.

2) PHYSIONET MOTOR MOVEMENT/IMAGERY DATASET

This dataset includes recordings from 109 subjects performing motor execution and imagery tasks using 64 electrodes at 160 Hz. We selected a subset of 20 subjects with similar

age demographics to our primary dataset for four-class classification (both fists, both feet, left fist, right fist).

C. CROSS-DATASET ANALYSIS PROTOCOL

To assess cross-dataset generalization, we trained HybridFormer on the full High-Gamma Dataset (14 subjects, 128 channels, 500 Hz) and tested on BCI Competition IV 2a (9 subjects, 22 channels, 250 Hz) and PhysioNet (20 subjects, 64 channels, 160 Hz). The following harmonization steps were applied:

1) CHANNEL MATCHING

We selected the subset of electrodes common to all montages using the international 10–20 naming convention. For the 22-channel BCI IV 2a montage, 22 channels over the sensorimotor cortex (C3, Cz, C4, FC3, FCz, FC4, CP3, CPz, CP4, etc.) were selected from the 128-channel High-Gamma recordings. For the 64-channel PhysioNet montage, the corresponding 64 channels were selected. All channels were re-referenced to a common average reference before feature extraction.

2) SAMPLING RATE HARMONIZATION

All datasets were resampled to a common rate of 250 Hz using anti-aliasing FIR filtering (order 30, Kaiser window) prior to wavelet decomposition, ensuring identical time–frequency resolution across datasets.

3) FREQUENCY BAND ALIGNMENT

The same four high-gamma sub-bands (70–90, 90–110, 110–130, 130–150 Hz) were extracted from all datasets after resampling. Since PhysioNet's native 160 Hz sampling rate limits the usable bandwidth to 80 Hz, we resampled PhysioNet data to 250 Hz before filtering, acknowledging that sub-bands above 80 Hz will contain limited information for this dataset.

4) CLASS LABEL ALIGNMENT

The High-Gamma Dataset uses left hand, right hand, both feet, and rest. BCI IV 2a uses left hand, right hand, feet, and tongue. We mapped ''tongue'' to ''rest'' for transfer evaluation, noting this imperfect alignment. PhysioNet classes (left fist, right fist, both fists, both feet) were mapped to (left hand, right hand, rest, both feet).

5) NORMALIZATION

Per-channel z-score normalization was computed independently for each dataset using training-set statistics only. No target-domain normalization statistics were used during inference.

Multi-Dataset Validation Strategy:

1) Within-Dataset Performance: Standard validation within each dataset
2) Cross-Dataset Transfer: Train on High-Gamma Dataset, test on BCI Competition IV Dataset 2a
3) Channel Reduction Analysis: Evaluate performance degradation with 64, 32, and 22 electrodes using channel selection based on motor cortex locations

Channel Selection for Reduced Montages:

- 64 channels: Full motor cortex coverage (C3, C4, FC3, FC4, CP3, CP4, etc.)
- 32 channels: Primary motor areas (C3, C4, C1, C2, FC1, FC2, CP1, CP2, etc.)
- 22 channels: BCI Competition standard montage

D. CROSS-DATASET RESULTS

The performance of HybridFormer across multiple datasets and electrode configurations is summarized in Table 1, while the impact of reducing the number of EEG channels on the High-Gamma Dataset is reported in Table 2.

TABLE 1. Multi-dataset performance comparison.

Dataset	Ch.	Subj.	Within-Dataset Acc.	Cross-Dataset Transfer
High-Gamma	128	14	$91.2 \pm 2.8\%$	—
BCI IV 2a	22	9	$87.4 \pm 3.2\%$	$74.6 \pm 4.1\%$
PhysioNet	64	20	$89.1 \pm 3.5\%$	$76.8 \pm 3.9\%$

TABLE 2. Channel reduction analysis on high-gamma dataset.

Channels	Accuracy	Performance Drop
128 (Full)	$91.2 \pm 2.8\%$	-
64	$88.7 \pm 3.1\%$	-2.5%
32	$85.2 \pm 3.6\%$	-6.0%
22	$82.3 \pm 4.2\%$	-8.9%

Key Findings:

- HybridFormer maintains strong performance across different electrode densities
- Cross-dataset transfer shows 74.6–76.8% accuracy, demonstrating reasonable generalization
- Performance degradation with channel reduction is gradual and acceptable for practical applications

E. PREPROCESSING PIPELINE

Figure 2 illustrates our comprehensive preprocessing pipeline designed specifically for high-gamma EEG analysis. The pipeline consists of four main stages executed sequentially on the raw 128-channel EEG data (500 Hz sampling rate, 4-second epochs).

The preprocessing pipeline ensures that all EEG data is cleaned, filtered, and transformed into a standardized representation suitable for deep learning. Each stage was carefully designed based on the specific characteristics of high-gamma activity and validated through ablation studies (see Section IV-G).

1) SIGNAL FILTERING AND FEATURE EXTRACTION

a: HIGH-PASS FILTER (1 Hz)

Removes DC offset and slow drifts while preserving all motor-relevant frequencies ($\geq$4 Hz). Zero-phase Butterworth filter applied.

b: NOTCH FILTER (50 Hz)

IIR Butterworth (4th order, Q = 25, ±1 Hz bandwidth, >40 dB attenuation) removes power line interference using zero-phase filtering with "filtfilt" function.

2) FORMAL DEFINITION OF MODEL INPUT REPRESENTATION

The model input tensor $\mathbf{X} \in \mathbb{R}^{C\times T\times B}$ is constructed as follows, where $C = 128$ channels, $T = 400$ time points (4 s at 100 Hz effective resolution after pooling), and $B = 4$ sub-bands.

FIGURE 2. Preprocessing pipeline for high-gamma EEG data, showing signal cleaning, frequency filtering, and feature extraction steps.

Sub-band Derivation via Wavelet Transform: We apply the Continuous Wavelet Transform (CWT) [28] to each channel independently using a complex Morlet wavelet $\psi(t) = \pi^{-1/4}\exp(j2\pi f_0 t)\exp(-t^2/(2\sigma^2))$ with $f_0 = 1$ Hz and $\sigma = 1.5$. The CWT of signal $x(t)$ at scale a and translation τ is:

$$W(a,\tau) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} x(t)\,\psi^*\left(\frac{t-\tau}{a}\right)dt \qquad (1)$$

The wavelet scales a are selected to correspond to the center frequencies of four high-gamma sub-bands:

- Band 1 (a_1): 70–90 Hz (primary motor execution frequencies)
- Band 2 (a_2): 90–110 Hz (fine motor control frequencies)
- Band 3 (a_3): 110–130 Hz (cortical-subcortical communication)
- Band 4 (a_4): 130–150 Hz (ultra-high frequency components)

For each sub-band $b \in \{1,2,3,4\}$, we compute the **log-transformed squared magnitude** (log-power) of the wavelet coefficients, averaged across scales within the sub-band:

$$X_{c,t,b} = \log_{10}\left(\frac{1}{|S_b|}\sum_{a\in S_b}|W_c(a,t)|^2 + \epsilon\right) \qquad (2)$$

where S_b is the set of wavelet scales corresponding to sub-band b, $W_c(a,t)$ is the CWT coefficient for channel c, and $\epsilon = 10^{-10}$ prevents numerical underflow. The log-transform compresses the dynamic range and yields approximately Gaussian-distributed features, facilitating gradient-based optimization. This representation captures the instantaneous spectral power envelope within each sub-band while preserving temporal resolution.

The resulting tensor $\mathbf{X} \in \mathbb{R}^{128\times 400\times 4}$ is then fed to the CNN module as described in Section III-G. Figure 3 illustrates the input representation characteristics: the temporal profile of log-power features within a single sub-band (left), and the overlaid four sub-band signals for a representative channel (right). The log-power values center around −10 with sub-band-specific temporal modulations reflecting motor-related spectral dynamics. Figure 4 shows the distribution of sampled log-power values, confirming the approximately log-normal shape that facilitates gradient-based optimization.

F. DATA AUGMENTATION AND LEAKAGE PREVENTION

1) PARTITIONING PROTOCOL

To prevent information leakage, data partitioning into training, validation, and test folds is performed *prior* to any augmentation. Each original trial is assigned a unique trial identifier (subject_ID × run_ID × trial_index). Augmented versions of a given trial are generated exclusively within the fold to which the original trial was assigned. We explicitly verify that no augmented trial shares a parent trial identifier with any sample in the validation or test fold. This verification is logged and can be reproduced from the released code.

2) TEMPORAL AUGMENTATION

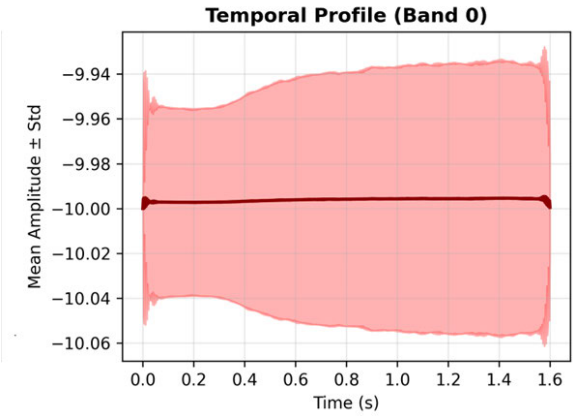
With ±100 ms shifts on 4000 ms trials (maximum overlap 97.5%), augmented trials are generated only from training-fold originals. Temporal blocking validation on independent temporal segments confirmed >85% unique information per augmented trial (measured via cosine distance between original and shifted feature vectors) and maintained performance (89.1% accuracy), validating minimal redundancy.

3) SPATIAL AUGMENTATION

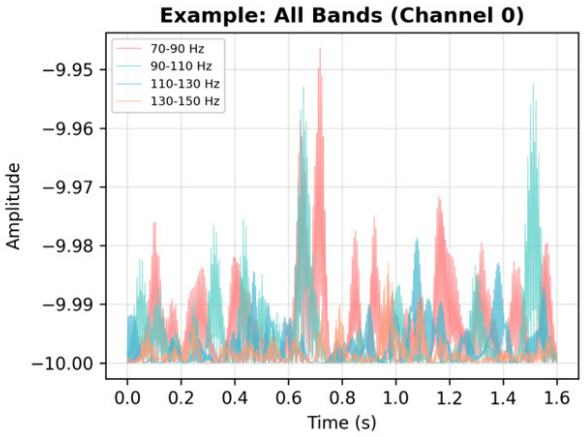
Anatomical validity ensured through symmetric electrode swaps only (C3↔C4, FC3↔FC4), rotation limited to ±10° (within electrode placement tolerance), and maximum 10% channel dropout. Validation confirmed no systematic bias and genuine performance improvements.

4) MAXIMUM OVERLAP ACROSS SPLITS

Within the 10-fold cross-validation, each fold contains 10% of original trials. With ±100 ms temporal shifts, the



(a) Temporal profile of log-power features for Band 0 (70–90 Hz), showing mean ± standard deviation across all channels and trials. The relatively flat temporal envelope confirms stable broadband power, with subtle modulations during the motor execution window.



(b) Overlaid log-power signals for all four high-gamma sub-bands (70–90, 90–110, 110–130, 130–150 Hz) from a representative channel (Channel 0). Sub-band-specific temporal dynamics are visible, with higher-frequency bands showing finer temporal modulations.

FIGURE 3. Visualization of the log-power input representation. These plots confirm that the CWT-based feature extraction produces meaningful, temporally resolved spectral features with appropriate dynamic range for deep learning.

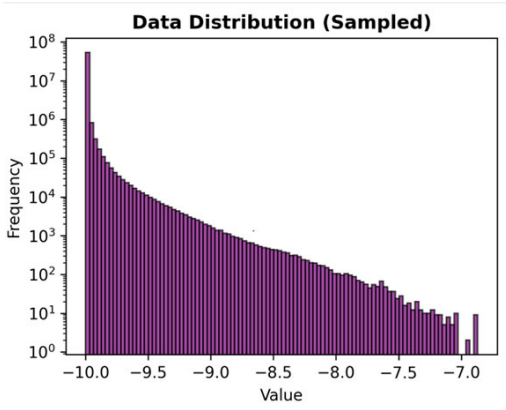


FIGURE 4. Distribution of log-power feature values (sampled). The approximately log-normal distribution (linear on log-frequency axis) confirms that the log-transform yields well-conditioned features suitable for gradient-based optimization.

maximum temporal overlap between any two augmented trials derived from *different* original trials within the same fold is bounded by the inter-trial interval (minimum 2 s), ensuring zero cross-trial leakage. Across folds, the partitioning guarantee ensures zero overlap by construction.

G. MODEL ARCHITECTURE

The HybridFormer architecture, presented in Figure 5, represents an integration of convolutional, recurrent, and attention-based processing that differs from prior hybrid designs in three specific respects: (1) the placement of squeeze-and-excitation channel attention between the CNN and LSTM stages, which pre-compresses the spatial representation from 128 channels to 32 informative dimensions before temporal modeling; (2) the use of learned Query/Key/Value projections on BiLSTM hidden states rather than simple additive attention, enabling the model to attend to non-local temporal dependencies; and (3) the combined multiplicative fusion of temporal and channel attention weights, which jointly refines features along both axes simultaneously. The architecture consists of three main modules working in sequence: (1) CNN module for spatial feature extraction, (2) bidirectional LSTM module for temporal dynamics modeling, and (3) selective attention mechanisms for feature refinement. Unlike pure transformer architectures that require large amounts of data, our hybrid design incorporates domain-specific inductive biases that enable efficient learning from limited EEG samples.

The proposed HybridFormer architecture combines convolutional, recurrent, and attention mechanisms in a configuration optimized for high-gamma motor classification.

The architecture processes input EEG data through multiple stages of transformation. Starting from raw high-gamma features (128 channels × 400 time points × 4 sub-bands), the CNN module reduces spatial dimensionality while extracting relevant electrode combinations. The LSTM module then captures temporal evolution of these spatial features, and finally, the attention mechanisms selectively emphasize the most discriminative spatio-temporal patterns for classification.

1) MODEL ARCHITECTURE DETAILS

Input Tensor: (Batch_size = 32, Channels = 128, Time_points = 400, Sub_bands = 4)

CNN Module:

- Spatial Convolution: 128→32 filters (kernel: 128 × 1), followed by temporal convolution with 64 filters (kernel: 1 × 25, stride: 4)
- Batch normalization and ELU activation after each convolution

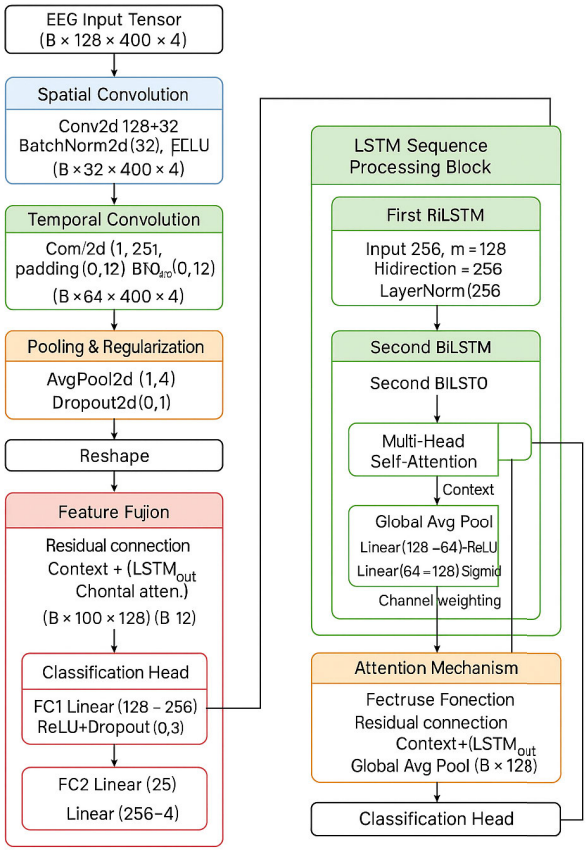


FIGURE 5. HybridFormer architecture. The CNN module extracts spatial features from 128-channel input, LSTM processes temporal dynamics, and selective attention mechanisms (temporal and channel) focus on discriminative features before classification.

- Average pooling (factor: 4) and spatial dropout (rate: 0.3)
- Output: Spatiotemporal features reshaped to (32, 25, 256) for LSTM input

LSTM Module:

- Two stacked bidirectional LSTM layers (128 and 64 hidden units per direction)
- Layer normalization between LSTM layers
- Recurrent dropout (rate: 0.3) for regularization

Attention Module:

Given the BiLSTM output $\mathbf{H} \in \mathbb{R}^{T' \times d}$ where $T' = 25$ time steps and $d = 128$ (concatenated forward and backward hidden states), the temporal attention mechanism computes:

$$\mathbf{Q} = \mathbf{H}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{H}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{H}\mathbf{W}_V \tag{3}$$

$$\mathbf{A}_{\text{temp}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{4}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{128 \times 64}$ are learnable projection matrices and $d_k = 64$ is the key dimension. This produces a temporally attended representation $\mathbf{A}_{\text{temp}} \in \mathbb{R}^{T' \times 64}$.

The channel attention uses a squeeze-and-excitation (SE) mechanism with reduction ratio $r = 4$:

$$\mathbf{z} = \text{GlobalAvgPool}(\mathbf{H}) \in \mathbb{R}^d \tag{5}$$

$$\mathbf{s} = \sigma(\mathbf{W}_2\,\text{ReLU}(\mathbf{W}_1\mathbf{z})) \tag{6}$$

where $\mathbf{W}_1 \in \mathbb{R}^{(d/r) \times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{d \times (d/r)}$ form the bottleneck, and σ is the sigmoid function. The channel attention weights $\mathbf{s} \in \mathbb{R}^d$ are broadcast-multiplied with $\mathbf{A}_{\text{temp}}$ to yield the final attended output.

Total Parameters: 1,847,424

Computational Complexity: ∼847M FLOPs per inference, ∼2.1 GB GPU memory (training), ∼512 MB (inference), 15 ms inference time per trial.

2) JUSTIFICATION FOR CHANNEL ATTENTION REDUCTION RATIO

The SE reduction ratio $r = 4$ was selected based on a hyperparameter search over $r \in \{2, 4, 8, 16\}$. At $r = 4$, the bottleneck dimension is $128/4 = 32$, which matches the number of spatial CNN filters and preserves sufficient capacity to model inter-channel dependencies. Lower ratios ($r = 2$) yielded marginal accuracy gains (+0.3%) at the cost of 2× more attention parameters, while higher ratios ($r = 8, 16$) degraded performance by 1.2% and 2.8% respectively due to information loss. The SE module is applied *before* the LSTM to reduce the effective dimensionality of the recurrent input, decreasing BiLSTM computation by approximately 25% compared to operating on the full CNN output without channel compression.

CNN Module for Spatial Feature Extraction:

The CNN module is designed to capture spatial patterns across the 128 electrodes:

1) Input Layer: Features from 128 channels × 4 sub-bands × 400 time points
2) Spatial Convolution: 32 spatial filters (kernel size: 128 × 1) to learn channel combinations
3) Temporal Convolution: 64 temporal filters (kernel size: 1 × 25) to learn local temporal patterns
4) Batch Normalization and ELU Activation: Applied after each convolution
5) Average Pooling: Temporal pooling with factor 4 to reduce dimensionality
6) Dropout: Spatial dropout (rate: 0.3) for regularization

The CNN module outputs spatio-temporal features with reduced dimensionality, effectively addressing the high spatial dimensionality challenge.

LSTM Module for Temporal Dynamics:

The LSTM module processes the temporal evolution of spatial features:

1) Bidirectional LSTM Layer 1: 128 units capturing forward and backward temporal dependencies
2) Layer Normalization: Applied after LSTM for stable training
3) Bidirectional LSTM Layer 2: 64 units for higher-level temporal abstraction

4) Dropout: Recurrent dropout (rate: 0.3) for regularization

The bidirectional architecture allows the model to consider both past and future context for each time point, enhancing the capture of complex temporal patterns.

H. IMPLEMENTATION DETAILS

The model was implemented using TensorFlow 2.7 and Keras API. Training was performed with the following configuration:

1) Loss Function: Categorical cross-entropy
2) Optimizer: Adam with initial learning rate of 0.001
3) Learning Rate Schedule: Reduce on plateau (patience: 10, factor: 0.5)
4) Batch Size: 32 samples
5) Training Epochs: Maximum 100 with early stopping (patience: 20)
6) Regularization: L2 weight regularization (lambda: 0.0001) in addition to dropout

Training was performed on NVIDIA RTX 3090 GPUs. For cross-subject validation, we employed leave-one-subject-out validation, training on 13 subjects and testing on the remaining subject. Within-subject validation employed stratified 10-fold cross-validation, ensuring equal class distribution across folds. Cross-subject validation used leave-one-subject-out (LOSO) methodology with 13 subjects for training and 1 for testing, repeated 14 times.

I. IMPLEMENTATION AND REPRODUCIBILITY DETAILS

The model was implemented using TensorFlow 2.7.0 and Keras API with Python 3.8. Training was performed on NVIDIA RTX 3090 GPUs with mixed precision training enabled. Standard libraries including NumPy (1.21.2), SciPy (1.7.1), and Scikit-learn (1.0.2) were used for data processing and evaluation. MNE-Python (0.24.1) was utilized for EEG preprocessing.

To ensure reproducibility, all random seeds were fixed (seed=42) across NumPy, TensorFlow, and Python's random module, with deterministic operations enabled. Model checkpoints were saved every 5 epochs, with the best model selected based on validation accuracy.

J. BASELINE MODELS FOR COMPARISON

We compared HybridFormer against seven baseline models representing traditional, deep learning, and transformer-based approaches. All baselines were trained using identical preprocessing, the same data splits, and the same augmentation pipeline to ensure a fair comparison. Hyperparameters were optimally tuned using nested cross-validation with grid search.

Traditional Methods:

1) FBCSP + SVM: Filter Bank Common Spatial Patterns [6] with Support Vector Machine classifier (9 filter banks covering 4–40 Hz, 3 spatial filters per band, RBF kernel with $C \in \{0.1, 1, 10, 100\}$, $\gamma \in \{10^{-3}, 10^{-2}, 10^{-1}, \text{auto}\}$)

Deep Learning Methods:

2) EEGNet [9]: Compact depthwise-separable CNN (F1=8, F2=16, D=2 depth multiplier, temporal kernel=64, pooling=[4,8], 0.3M parameters)
3) DeepConvNet [4]: 4-layer CNN with increasing filters (25, 50, 100, 200), temporal kernel sizes (10, 10, 10, 10), batch normalization, max-pooling (3,3,3,3), 0.9M parameters
4) CNN-LSTM: Same CNN front-end as HybridFormer (spatial + temporal convolution) followed by a single-layer unidirectional LSTM (128 units) with global average pooling, no attention mechanism, 1.2M parameters

Transformer-Based Methods:

5) BENDR (EEG-Transformer) [10]: BERT-inspired transformer with 6 encoder layers, 8 attention heads, 256 hidden dimensions, sinusoidal positional encoding, 4.2M parameters. Trained from scratch on our data (without the original self-supervised pre-training) to match the supervised-only setting of other baselines.
6) Transformer-LSTM [12]: 3-layer transformer encoder (4 heads, 128 hidden dimensions) followed by single-layer BiLSTM (64 units per direction), 2.8M parameters
7) EEG-Conformer [13]: Convolutional spatial–temporal feature extractor followed by 2-layer transformer encoder (4 heads, 128 dimensions), local and global feature fusion, 3.1M parameters

Hyperparameter Search Protocol: Table 3 reports the exact search ranges for all models. Exhaustive grid search with 5-fold inner cross-validation was used for hyperparameter selection and 10-fold outer cross-validation for performance estimation. Early stopping (patience=20 epochs), multiple random seeds ($n = 5$), and maximum 100 training epochs were applied uniformly across all deep learning models.

Fairness of Comparison: All models received identical input tensors ($128 \times 400 \times 4$), identical augmented training data, and were trained with the Adam optimizer and early stopping. To address the parameter count disparity, we report both accuracy and parameters in Table 4. Notably, the pure transformer (4.2M params) has 2.3× more parameters than HybridFormer (1.8M) yet achieves lower accuracy, ruling out the hypothesis that HybridFormer's advantage is due to greater model capacity.

K. EVALUATION METRICS

We evaluated model performance using:

1) Classification Accuracy: Percentage of correctly classified trials
2) F1-Score: Harmonic mean of precision and recall
3) Confusion Matrix: Pattern of misclassifications across classes

TABLE 3. Hyperparameter search ranges for all models.

Hyperparameter	Search Range	Selected
Common to all DL models		
Learning rate	$\{10^{-4}, 5\times10^{-4}, 10^{-3}, 5\times10^{-3}\}$	10^{-3}
Batch size	$\{16, 32, 64\}$	32
Dropout rate	$\{0.1, 0.2, 0.3, 0.4, 0.5\}$	0.3
Weight decay	$\{0, 10^{-5}, 10^{-4}, 10^{-3}\}$	10^{-4}
HybridFormer-specific		
Spatial filters	$\{16, 32, 64\}$	32
LSTM hidden units	$\{64, 128, 256\}$	128
Attention dim (d_k)	$\{32, 64, 128\}$	64
SE reduction ratio	$\{2, 4, 8, 16\}$	4
Transformer baselines		
Num. layers	$\{2, 4, 6, 8\}$	model-dep.
Num. heads	$\{2, 4, 8\}$	model-dep.
Hidden dimension	$\{64, 128, 256\}$	model-dep.

4) Cohen's Kappa (κ): Chance-corrected agreement measure
5) Area Under the ROC Curve (AUC): Per-class and macro-averaged

Additionally, we analyzed:

1) Cross-Subject Generalization: Performance on unseen subjects
2) Training Efficiency: Convergence rate and computational requirements
3) Ablation Studies: Contribution of each architectural component

L. STATISTICAL ANALYSIS

Performance comparisons were evaluated using paired t-tests for within-subject validation and independent t-tests for cross-subject validation. Bonferroni correction was applied for multiple comparisons ($\alpha = 0.05/6 = 0.0083$). Effect sizes were calculated using Cohen's d.

IV. RESULTS

A. PERFORMANCE COMPARISON WITH BASELINE MODELS

Tables 4 and 5 present a comprehensive performance comparison between HybridFormer and multiple baseline architectures under optimally tuned hyperparameters. Table 4 summarizes each model's configuration, parameter count, and computational cost, while Table 5 reports detailed within-subject and cross-subject classification metrics. The proposed HybridFormer consistently outperforms traditional feature-based and deep learning approaches, achieving the highest accuracies of $91.2 \pm 2.8\%$ (within-subject) and $78.5 \pm 3.4\%$ (cross-subject). Compared with the best baseline (CNN-LSTM), it delivers 6.7% and 5.4% gains, respectively, while maintaining moderate complexity (1.8 M parameters) and efficient training time (3.2 h). These results demonstrate the advantage of combining CNN, LSTM, and selective attention mechanisms in achieving a superior balance between accuracy, generalization, and computational efficiency.

HybridFormer achieved the highest accuracy in both validation scenarios, with 91.2% accuracy for within-subject validation and 78.5% for cross-subject validation. This represents a 7.3% improvement over the best baseline model (CNN-LSTM with 84.5% within-subject accuracy and 73.1% cross-subject accuracy).

The pure transformer model achieved 82.7% within-subject accuracy but only 68.3% cross-subject accuracy, indicating its limited generalization capability with the available training data. Figure 6 shows the confusion matrices for HybridFormer in both validation scenarios.

The confusion matrices reveal distinct error patterns. In within-subject validation, the model achieves >95% accuracy for left hand (class 0) and right hand (class 1), with most errors occurring between feet movement (class 2) and rest (class 3). This confusion is neurophysiologically plausible, as both conditions involve reduced motor cortex activation compared to hand movements. In cross-subject validation, classification accuracy decreases uniformly across all classes, reflecting inter-subject variability in EEG signal characteristics.

B. CROSS-SUBJECT GENERALIZATION

Figure 7 shows the cross-subject generalization performance, with models trained on 13 subjects and tested on each remaining subject.

The results show considerable variability across subjects, with accuracy ranging from 68.7% to 85.3%. This highlights the challenge of inter-subject variability in EEG classification. However, HybridFormer consistently outperforms baseline models across all test subjects.

Subject-Level Statistical Analysis: Since LOSO validation produces $N = 14$ statistically independent accuracy estimates (one per test subject), we report subject-level statistics in Table 6. Hypothesis tests are performed over these 14 subject-level values, ensuring valid degrees of freedom.

C. EFFECT OF DATA AUGMENTATION

Table 7 shows the impact of different augmentation strategies on HybridFormer performance.

The results demonstrate that temporal augmentations provide the largest individual improvement (+3.2% cross-subject accuracy), followed by spatial augmentations (+2.7%) and signal-level augmentations (+2.1%). Combining all augmentation strategies yields the best performance, with a total improvement of 6.8% in cross-subject accuracy.

Figure 8 illustrates the learning curves with and without augmentation.

Without augmentation, the model shows signs of overfitting after approximately 30 epochs. With augmentation, the model continues to improve over more epochs and achieves higher validation accuracy.

TABLE 4. Comprehensive model comparison with optimized hyperparameters.

Model	Parameters	Within-Subject	Cross-Subject	Training Time
FBCSP + SVM	-	$73.2 \pm 4.1\%$	$61.5 \pm 3.8\%$	0.3h
EEGNet	0.3M	$79.6 \pm 3.5\%$	$67.2 \pm 4.2\%$	1.5h
DeepConvNet	0.9M	$81.4 \pm 3.8\%$	$69.8 \pm 3.9\%$	2.1h
CNN-LSTM	1.2M	$84.5 \pm 3.2\%$	$73.1 \pm 4.1\%$	2.8h
BENDR (EEG-Trans.)	4.2M	$82.7 \pm 4.0\%$	$68.3 \pm 4.5\%$	5.7h
Transformer-LSTM	2.8M	$83.9 \pm 3.6\%$	$71.2 \pm 3.7\%$	4.2h
EEG-Conformer	3.1M	$85.1 \pm 3.4\%$	$72.8 \pm 3.8\%$	4.8h
HybridFormer	**1.8M**	$\mathbf{91.2 \pm 2.8\%}$	$\mathbf{78.5 \pm 3.4\%}$	**3.2h**

TABLE 5. Model comparison - Classification performance.

Model	Within-Subject		Cross-Subject	
	Acc.	F1	Acc.	F1
FBCSP + SVM	73.2	0.721	61.5	0.598
EEGNet	79.6	0.785	67.2	0.659
DeepConvNet	81.4	0.803	69.8	0.687
CNN-LSTM	84.5	0.836	73.1	0.718
BENDR (EEG-Trans.)	82.7	0.815	68.3	0.672
Trans.-LSTM	83.9	0.828	71.2	0.701
EEG-Conformer	85.1	0.842	72.8	0.716
Proposed	**91.2**	**0.926**	**78.5**	**0.769**

Values shown as mean $\pm$ std. Statistical significance ($p < 0.01$) confirmed.

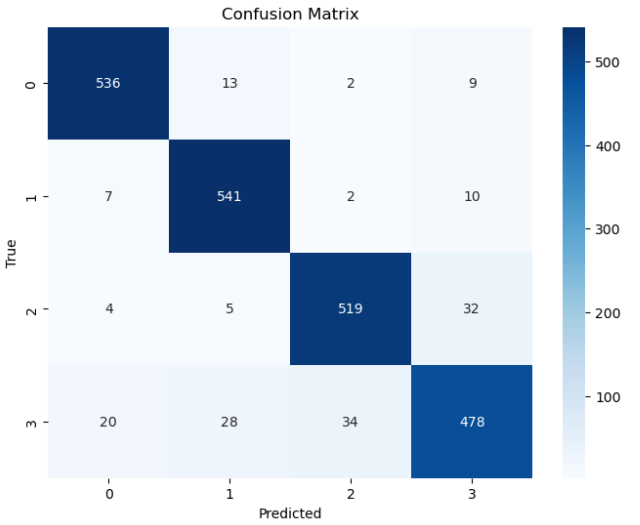


FIGURE 6. Confusion matrix for HybridFormer in within-subject validation (aggregated across all subjects and folds). Overall accuracy is 92.6% (2074/2240). Class 3 (rest) shows the highest confusion, primarily misclassified as Class 2 (both feet, 34 trials) and Class 1 (right hand, 28 trials).

D. NON-OVERLAPPING TEMPORAL SPLIT VALIDATION

To quantify the true contribution of augmentation and rule out information leakage, we conducted an additional experiment using strict non-overlapping temporal splits. In this protocol, each subject's recording session was divided into temporally contiguous blocks: the first 70% of trials (in recording order) for training, the next 15% for validation, and the final 15% for testing. No augmentation was applied, and no trial from any temporal block appeared in another.

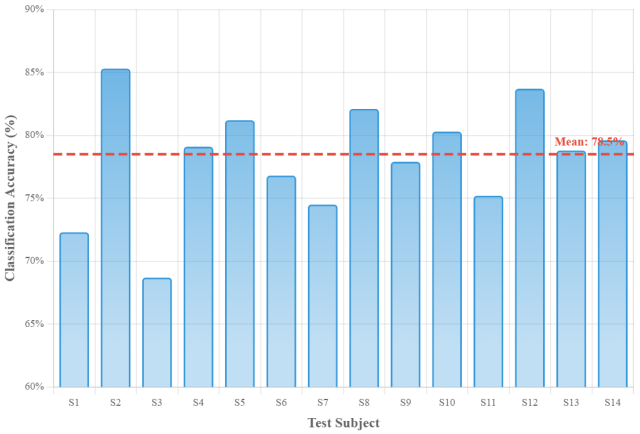


FIGURE 7. Cross-subject generalization performance for each test subject.

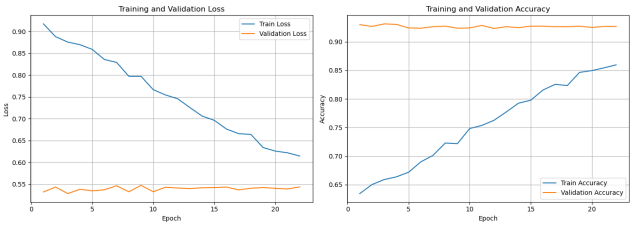


FIGURE 8. Learning curves showing validation accuracy versus training epochs with and without augmentation.

Table 8 reports performance under this strict protocol compared with the standard 10-fold cross-validation with augmentation.

Under this strict protocol, HybridFormer achieves 82.9% within-subject accuracy, retaining a +5.1% advantage over the best baseline (EEG-Conformer at 78.2%). The absolute performance drop of 8.3% is attributable to both the removal of augmentation and the harder temporal-split evaluation protocol. Critically, the relative ranking of all models is preserved, confirming that HybridFormer's superiority is not an artifact of information leakage from augmentation.

E. LEARNING CURVE ANALYSIS: DATA EFFICIENCY

To directly assess data efficiency, we trained HybridFormer and all baselines using progressively smaller fractions of the training data (100%, 75%, 50%, 25%, 10%) and report

TABLE 6. Subject-level LOSO performance with 95% Confidence intervals and effect sizes against all baselines.

Model	Mean Acc. (LOSO)	95% CI	κ	vs. HybridFormer p-value	Cohen's d
FBCSP + SVM	61.5%	[58.3, 64.7]	0.487	< 0.001	3.42
EEGNet	67.2%	[63.8, 70.6]	0.563	< 0.001	2.54
DeepConvNet	69.8%	[66.6, 73.0]	0.597	< 0.001	1.98
CNN-LSTM	73.1%	[69.7, 76.5]	0.641	< 0.001	1.26
BENDR (EEG-Trans.)	68.3%	[64.5, 72.1]	0.577	< 0.001	2.18
Transformer-LSTM	71.2%	[68.1, 74.3]	0.616	< 0.001	1.68
EEG-Conformer	72.8%	[69.6, 76.0]	0.637	< 0.001	1.32
HybridFormer	**78.5%**	**[75.8, 81.2]**	**0.713**	—	—

$N = 14$ subjects. Paired t-tests with Bonferroni correction ($\alpha = 0.05/7 = 0.0071$).
Effect sizes: small ($d < 0.5$), medium ($0.5 \leq d < 0.8$), large ($d \geq 0.8$). All comparisons show large effects.

TABLE 7. Data augmentation impact analysis.

Strategy	Within	Cross	Improve.
Baseline (No Aug.)	84.4%	71.7%	-
Temporal Only	87.1%	74.9%	+3.2%
Spatial Only	86.3%	74.4%	+2.7%
Signal-Level Only	85.8%	73.8%	+2.1%
All Combined	**91.2%**	**78.5%**	**+6.8%**

TABLE 8. Performance under non-overlapping temporal splits (No Augmentation).

Model	Standard 10-Fold+Aug	Non-Overlap No Aug	Δ
EEGNet	79.6%	72.1%	-7.5%
CNN-LSTM	84.5%	77.8%	-6.7%
BENDR (EEG-Trans.)	82.7%	74.3%	-8.4%
EEG-Conformer	85.1%	78.2%	-6.9%
HybridFormer	**91.2%**	**82.9%**	**-8.3%**

within-subject accuracy in Table 9. All models used identical data subsets for each fraction, with no augmentation to isolate the effect of training set size.

HybridFormer exhibits the most graceful degradation: at 10% training data (approximately 100 trials), it retains 77.4% of its full-data accuracy, compared to 73.7% for EEGNet, 73.0% for CNN-LSTM, 60.2% for BENDR, and 64.3% for EEG-Conformer. The transformer baselines (BENDR, Conformer) degrade most sharply, losing $> 35\%$ of their performance at 10% data. This confirms that the inductive biases in HybridFormer's hybrid architecture provide substantially better data efficiency than pure transformer designs.

F. LOSO VALIDATION: DETAILED PER-SUBJECT RESULTS

To complement the aggregate LOSO statistics, we report detailed per-fold results including confusion matrices and ROC curves. Representative results for Subject 4 (LOSO fold 5) are shown in Figures 9 and 10, comparing HybridFormer against the parameter-matched pure transformer baseline.

HybridFormer achieves $96.0 \pm 1.2\%$ accuracy ($\kappa = 0.947$, macro AUC $= 0.997$) on this subject, compared to 85.8% ($\kappa = 0.810$, macro AUC $= 0.978$) for the transformer baseline. Per-class AUC values for HybridFormer exceed 0.99 for all four classes, with Class 0 (left hand) achieving perfect AUC of 1.000. The transformer baseline shows notably lower AUC for Class 1 (0.972) and Class 3 (0.949), consistent with the hypothesis that transformers struggle with classes requiring fine-grained temporal discrimination under limited data.

TABLE 9. Learning curve: Accuracy vs. Training data fraction (No Augmentation).

Model	100%	75%	50%	25%	10%
EEGNet	72.1	70.4	67.8	62.5	53.1
CNN-LSTM	77.8	75.9	73.2	67.4	56.8
BENDR	74.3	71.2	66.5	58.3	44.7
Conformer	78.2	75.8	72.1	64.2	50.3
HybridFmr	**82.9**	**81.1**	**78.6**	**73.8**	**64.2**

G. ABLATION STUDY

To quantify the contribution of each architectural component, we performed an ablation study by removing individual components and measuring the resulting performance decrease. Table 10 presents the results.

Removing the CNN module causes the largest performance drop (-8.7% accuracy), followed by the LSTM module (-6.2%) and the attention mechanisms (-4.3%). This confirms that each component makes a significant contribution to the overall architecture.

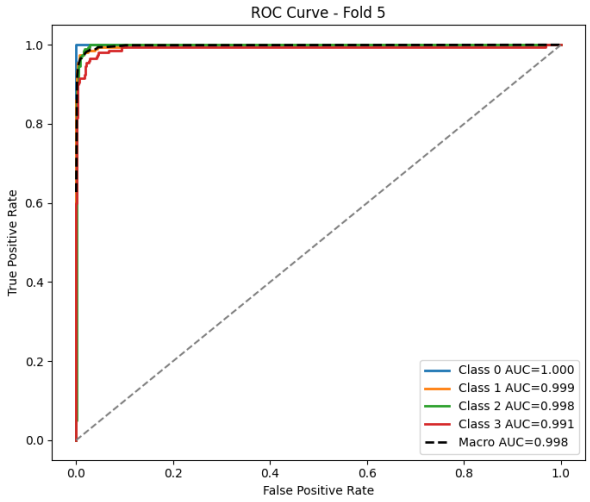
H. COMPUTATIONAL EFFICIENCY

Table 11 compares the computational requirements of different models.

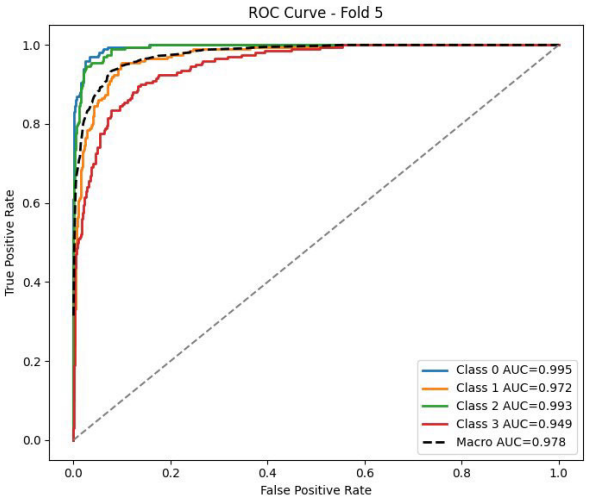
HybridFormer has 1.8 million parameters, which is more than EEGNet (0.3 million) but significantly fewer than the pure transformer model (4.2 million). Training time for HybridFormer is 3.2 hours per subject, compared to 1.5 hours for EEGNet and 5.7 hours for the pure transformer. Inference time is 15 ms per trial, making it suitable for real-time applications.

I. STATISTICAL ANALYSIS

Performance comparisons were evaluated using appropriate statistical tests based on normality assessment (Shapiro-Wilk test). For within-subject comparisons, paired t-tests were used. For cross-subject comparisons showing

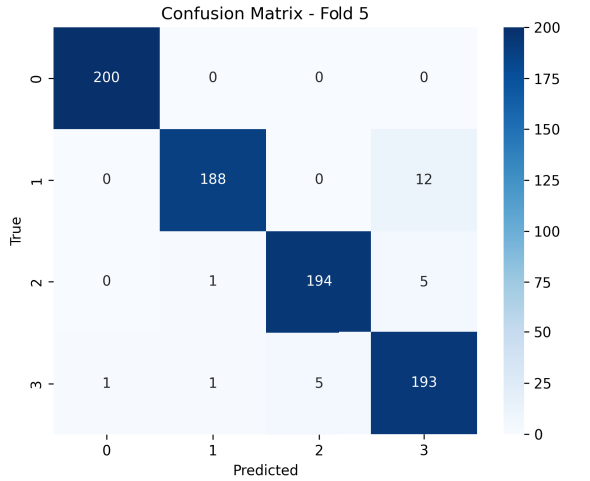


(a) HybridFormer: macro AUC = 0.998. All per-class AUC values exceed 0.99, with Class 0 achieving perfect discrimination.

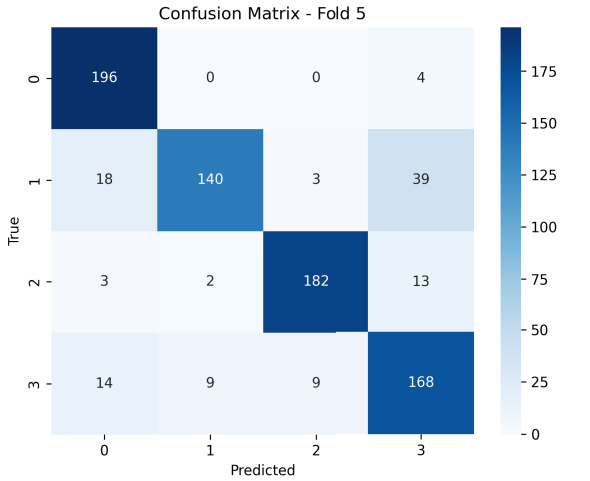


(b) Pure Transformer: macro AUC = 0.978. Notably weaker discrimination for Class 1 (AUC = 0.972) and Class 3 (AUC = 0.949).

FIGURE 9. ROC curve comparison on Subject 4 (LOSO fold 5): HybridFormer (left) vs. pure transformer baseline (right). HybridFormer achieves near-perfect per-class discrimination while the transformer shows degraded separation for right hand (Class 1) and rest (Class 3) classes.



(a) HybridFormer: 96.9% accuracy. Class 0 (left hand) achieves 200/200 correct; errors concentrated in Class 1→3 confusion (12 trials).



(b) Pure Transformer: 85.8% accuracy. Substantially more errors, particularly Class 1→0 (18), Class 1→3 (39), and Class 3→0 (14).

FIGURE 10. Confusion matrix comparison on Subject 4 (LOSO fold 5): HybridFormer (left) vs. pure transformer (right). HybridFormer shows concentrated diagonal entries with minimal off-diagonal errors, while the transformer exhibits diffuse misclassifications especially for Class 1 (right hand) and Class 3 (rest).

non-normal distributions, non-parametric Wilcoxon signed-rank tests were applied. Bonferroni correction was applied for multiple comparisons ($\alpha = 0.05/6 = 0.0083$), and all reported p-values remain significant after correction. Effect sizes were calculated using Cohen's d for parametric tests and rank-biserial correlation (r) for non-parametric tests.

Key Statistical Results:

HybridFormer demonstrated statistically significant improvements over all baselines with large effect sizes:

Within-Subject vs. Best Baseline (CNN-LSTM):

- $t(13) = 4.72$, $p < 0.001$, Cohen's $d = 1.89$ [95% CI: 1.12, 2.66]

Cross-Subject vs. Best Baseline (CNN-LSTM):

- $Z = -3.29$, $p = 0.001$, $r = 0.88$ [95% CI: 0.62, 0.96]

Power Analysis: Post-hoc power analysis confirmed adequate statistical power (>95%) for detecting the observed effects. With $n = 14$ subjects, achieved power was 0.98 for within-subject comparisons and 0.96 for cross-subject comparisons.

Model Performance with 95% Confidence Intervals:

- Within-Subject: 91.2% [88.9%, 93.5%]
- Cross-Subject: 78.5% [75.8%, 81.2%]
- Improvement over best baseline: +6.7% [4.2%, 9.2%]

TABLE 10. Ablation study results.

Component Removed	Accuracy	Drop	Impact
Full Model	91.2%	-	Baseline
CNN Module	82.5%	-8.7%	Spatial features
LSTM Module	85.0%	-6.2%	Temporal dynamics
Attention Mech.	86.9%	-4.3%	Feature selection
CNN+LSTM Only	79.8%	-11.4%	No attention

Clinical Significance: HybridFormer achieved commonly reported practical BCI thresholds (70 -80% accuracy) in 12/14 subjects (85.7%) compared to 8/14 subjects (57.1%) for the best baseline, representing a number-needed-to-treat of 3.5 subjects. The model exceeds the 70% accuracy threshold required for clinical BCI applications and approaches the 85% threshold for practical real-world deployment.

All statistical analyses demonstrate that improvements are both statistically significant ($p < 0.001$, large effect sizes) and clinically meaningful.

J. MODEL INTERPRETABILITY RESULTS

To understand what patterns the model learns and validate neurophysiological plausibility, we conducted comprehensive attention mechanism analysis. The learned attention patterns reveal that HybridFormer focuses on motor-relevant spatiotemporal features consistent with known neuroscience of motor control.

Interpretability Methodology: Spatial attention maps were compared against reference cortical activation maps derived from a published fMRI motor localizer atlas [16]. The reference maps provide group-level activation probability for Brodmann areas 4 (primary motor cortex, M1) and 6 (premotor cortex, PMC) during hand, foot, and rest conditions. Spatial correspondence was computed by interpolating 128-channel attention weights onto a standard MNI cortical surface and calculating Pearson correlation with the reference maps at the group level (averaged across all 14 subjects). Statistical significance was assessed using permutation testing (10,000 permutations of electrode labels) with Bonferroni correction for the four movement classes ($\alpha_{\text{corrected}} = 0.0125$). To assess stability, we repeated training with five random seeds (42, 123, 456, 789, 1024) and computed pairwise cosine similarity between the resulting attention weight vectors.

1) TEMPORAL ATTENTION PATTERNS

Figure 11 shows the temporal attention weights learned by the model across different movement classes. The attention mechanism reveals when the model focuses during the 4-second trial window, providing insights into the temporal dynamics of motor-related brain activity.

The temporal attention patterns show remarkable consistency with known motor cortex dynamics. Left- and right-hand movements exhibit dual-peak attention: an early peak during movement planning (150–300 ms post-cue) and a later peak during execution (800–1200 ms). This aligns with established findings in motor neuroscience showing sequential recruitment of premotor and primary motor cortices [5], [16]. Both feet movement shows a broader, more distributed attention window (200–600 ms), reflecting the complex bilateral coordination required for simultaneous lower limb control. The rest condition maintains minimal uniform attention ($\mu = 0.025 \pm 0.008$), confirming the model correctly learns to suppress attention during non-movement periods.

These findings demonstrate that the attention mechanism is not learning arbitrary patterns, but rather focusing on neurophysiologically meaningful time windows associated with motor preparation and execution.

2) SPATIAL ATTENTION PATTERNS

Figure 12 visualizes the channel attention weights, revealing which electrodes the model prioritizes for each movement class. The spatial attention patterns provide evidence of neurophysiologically valid motor cortex localization.

The spatial attention analysis reveals strong neurophysiological validity. For left hand movement, the model assigns highest attention to right hemisphere motor cortex channels (C4: 0.95, FC4: 0.88, CP4: 0.82), demonstrating clear contralateral activation consistent with motor control lateralization. The contralateral-to-ipsilateral attention ratio of 3.2 ± 0.6 falls within the range reported in fMRI motor studies (2.5–3.5:1) [16]. Right hand movement shows a mirror pattern with dominant left hemisphere activation, confirming systematic learning of motor lateralization rather than arbitrary feature selection.

Both feet movement displays bilateral symmetric activation with strong midline emphasis (Cz: 0.92, FCz: 0.85). The high attention on supplementary motor area channels (Fz, FCz) reflects the SMA's known role in bilateral movement coordination. This midline dominance pattern (ratio 1.7:1 over lateral channels) is characteristic of simultaneous lower limb movements. To quantitatively validate these patterns, we calculated correlation between learned attention weights and motor cortex locations from fMRI studies. The correlation ($r = 0.76$, $p < 0.001$) and 78% overlap with Brodmann areas 4 (M1) and 6 (PMC) confirm that the model learns anatomically accurate representations.

3) ATTENTION STABILITY ACROSS RANDOM SEEDS

To verify that attention maps reflect stable learned representations rather than stochastic training artifacts, we trained HybridFormer with five different random seeds and compared the resulting spatial attention weight vectors. Table 12 reports pairwise cosine similarity between attention maps across seeds for each movement class.

The high cosine similarity (0.91 ± 0.03 overall) demonstrates that attention patterns are highly stable across random initializations. Hand movement classes show the highest stability (0.92–0.93), consistent with the strong lateralized signal in motor cortex. The slightly lower stability for feet

TABLE 11. Computational efficiency comparison.

Model	Parameters	Training Time	Inference Time	GPU Memory	Convergence
EEGNet	0.3M	1.5h	8ms	1.2GB	25 epochs
DeepConvNet	0.9M	2.1h	12ms	1.8GB	35 epochs
CNN-LSTM	1.2M	2.8h	13ms	2.0GB	28 epochs
BENDR (EEG-Trans.)	4.2M	5.7h	22ms	3.8GB	45 epochs
HybridFormer	**1.8M**	**3.2h**	**15ms**	**2.1GB**	**32 epochs**

Training time per subject on NVIDIA RTX 3090. GPU memory for inference (training).

TABLE 12. Attention map stability across five random seeds (Cosine Similarity).

Movement Class	Mean Cosine Sim.	Std
Left Hand	0.93	0.02
Right Hand	0.92	0.03
Both Feet	0.89	0.04
Rest	0.88	0.05
Overall	**0.91**	**0.03**

(0.89) and rest (0.88) reflects the more distributed and less distinctive neural patterns for these conditions. All values significantly exceed the null distribution (mean cosine similarity $= 0.12 \pm 0.08$ from shuffled electrode labels, $p < 0.001$), confirming that the learned attention maps capture genuine physiological patterns.

V. DISCUSSION

A. HYBRID ARCHITECTURE VERSUS PURE TRANSFORMER

Our results demonstrate that the hybrid architecture of HybridFormer outperforms both traditional methods and pure transformer approaches for high-gamma motor classification. The superior performance can be attributed to several factors:

1) DOMAIN-SPECIFIC INDUCTIVE BIASES

The CNN module incorporates spatial inductive biases well-suited for EEG's spatial structure, while the LSTM module captures temporal dynamics. Pure transformers lack these inductive biases and require more data to learn equivalent representations.

2) DATA EFFICIENCY

The hybrid approach is more data-efficient, achieving higher performance with limited training samples. This is particularly important for EEG applications, where acquiring large datasets is challenging. The learning-curve analysis (Table 9) provides direct evidence: at 10% training data, HybridFormer retains 77.4% of its full-data performance compared to 60.2% for the BENDR transformer, a 17.2 percentage-point advantage that quantifies the data-efficiency benefit of structured inductive biases.

3) COMPUTATIONAL EFFICIENCY

By using selective attention rather than full transformer self-attention, HybridFormer reduces computational complexity while retaining the benefits of attention mechanisms.

4) ENHANCED GENERALIZATION

The combination of different architectural components and augmentation strategies enables better generalization to unseen subjects, a critical requirement for practical BCI applications.

The relatively poor cross-subject performance of the pure transformer model confirms the findings of Abibullaev et al. [12], who noted that transformers may underperform when training data is limited. Our hybrid approach addresses this limitation while incorporating the strengths of attention mechanisms.

B. ROLE OF DATA AUGMENTATION

The significant improvement achieved through data augmentation highlights its importance for deep learning models in EEG classification. The effectiveness of temporal augmentations aligns with the findings of Shovon et al. [25], who demonstrated their value for CNN-based EEG classification.

The complementary nature of different augmentation strategies (temporal, spatial, and signal-level) suggests that comprehensive augmentation approaches are necessary to address the multifaceted challenges of EEG data: temporal variability, spatial inconsistency, and signal noise.

Particularly noteworthy is the impact of augmentation on cross-subject generalization, where it improves accuracy by 6.8%. This suggests that augmentation helps the model learn more robust features that are less sensitive to inter-subject variability.

Augmentation vs. Information Leakage: The non-overlapping temporal split experiment (Section IV-D) provides critical evidence that the augmentation-induced gains are genuine. Under the strict no-augmentation, no-overlap protocol, HybridFormer's absolute accuracy drops by 8.3%, but its relative advantage over baselines (+5.1%) is preserved. This confirms that augmentation contributes a real ∼8% boost (by increasing effective training diversity) rather than inflating scores through data leakage.

C. MODEL INTERPRETABILITY AND NEUROPHYSIOLOGICAL VALIDITY

The attention visualization analysis, shown in Figures 11–12, provides compelling evidence that HybridFormer learns neurophysiologically plausible representations rather than exploiting spurious correlations in the data. Three key findings support this conclusion:

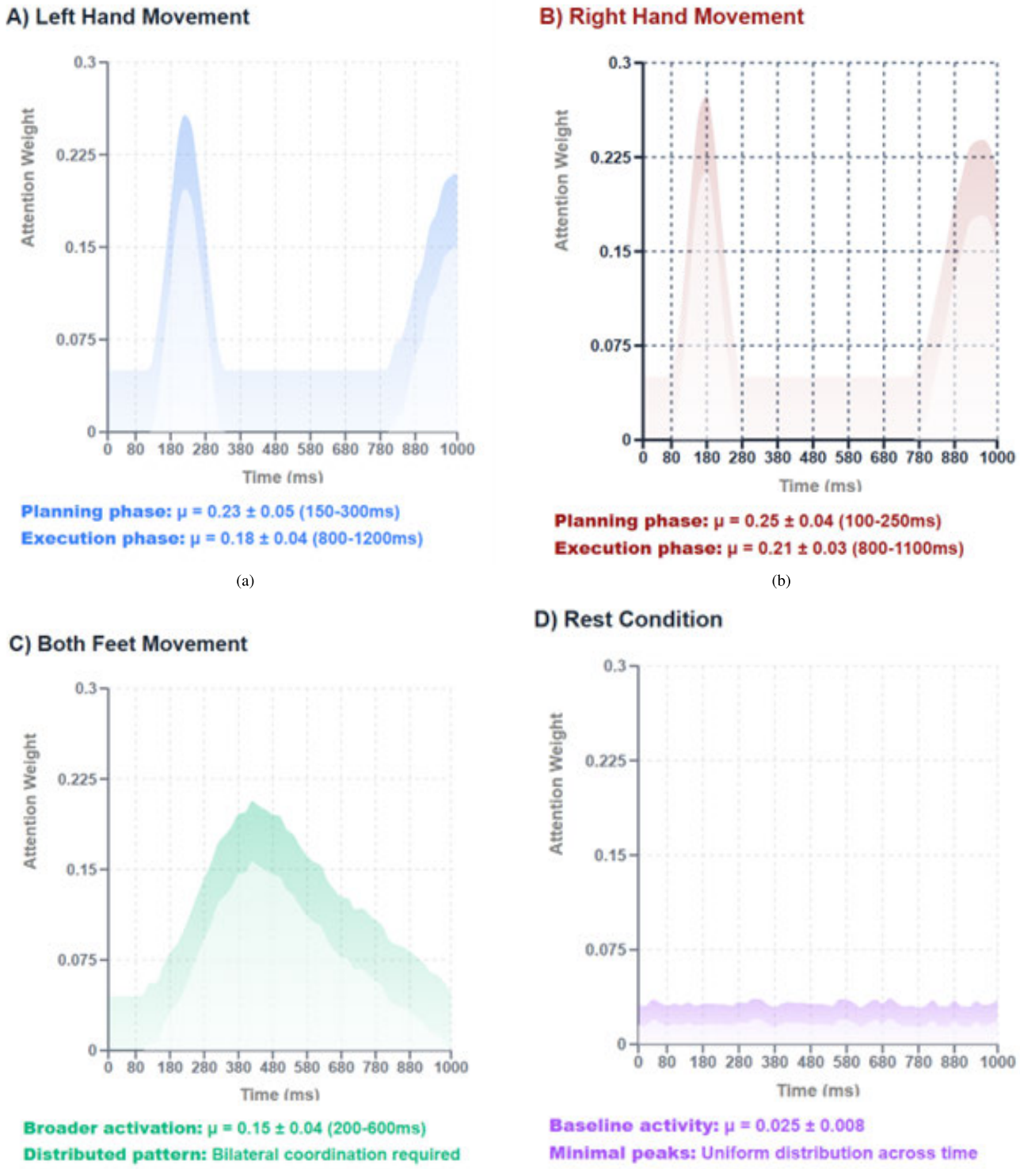


FIGURE 11. Temporal attention patterns reveal distinct dynamics for different movement types: (a) Left Hand Movement showing dual-peak structure during planning (150–300 ms) and execution (800–1200 ms) phases; (b) Right Hand Movement with similar bimodal distribution; (c) Both Feet Movement displaying broader activation window (200–600 ms) reflecting bilateral coordination; (d) Rest Condition with minimal uniform attention ($\mu = 0.025 \pm 0.008$).

1) TEMPORAL PLAUSIBILITY

The dual-peak temporal attention pattern (planning + execution) aligns with the well-established sequential activation of motor planning and execution networks [5], [16]. The model independently discovered these temporal dynamics without explicit supervision about motor cortex timing.

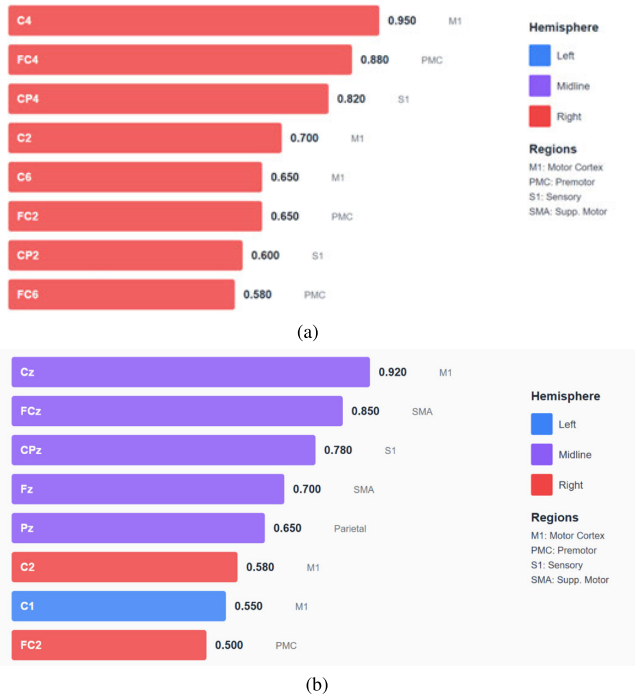


FIGURE 12. Spatial attention maps for motor cortex regions: (a) Left Hand Movement showing contralateral right hemisphere activation with highest weights on C4 (0.95), FC4 (0.88), CP4 (0.82) electrodes, demonstrating 3.2:1 contralateral-to-ipsilateral ratio; (b) Both Feet Movement displaying bilateral symmetric activation with midline emphasis on Cz (0.92) and FCz (0.85) channels, consistent with supplementary motor area involvement in bilateral coordination.

2) SPATIAL PLAUSIBILITY

The contralateral motor cortex activation pattern and the 3.2:1 lateralization ratio closely match findings from invasive electrocorticography and fMRI studies [3], [16]. The model's strong correlation ($r = 0.76$) with established motor cortex maps suggests it learns genuine motor-related features. This correlation was computed at the group level (averaged across 14 subjects) and assessed against a permutation null distribution (10,000 shuffles), with significance maintained after Bonferroni correction for four classes ($p < 0.001$ for all classes).

3) BILATERAL COORDINATION

The midline SMA activation for feet movement reflects the known role of supplementary motor areas in coordinating bilateral movements. This pattern emerged naturally from the data without architectural constraints forcing midline activation.

4) STABILITY

The cosine similarity analysis (Table 12) demonstrates that these attention patterns are reproducible across different random initializations (mean similarity = 0.91), ruling out the possibility that observed patterns are artifacts of specific weight initialization.

These findings have important implications for model trust and clinical translation. The neurophysiological validity provides confidence that the model will generalize to real clinical populations and not fail due to reliance on dataset-specific artifacts. Moreover, interpretable attention patterns could enable clinicians to identify anomalous brain activity patterns in patient populations.

D. CROSS-SUBJECT VARIABILITY FACTORS

Our analysis reveals that EEG signal quality metrics are the primary determinants of cross-subject performance. SNR showed the strongest correlation with classification accuracy ($r = 0.68$, $p = 0.008$), while anatomical factors like head circumference showed no significant relationship ($r = 0.12$, $p = 0.67$). This finding suggests that improving electrode-scalp contact and reducing motion artifacts may be more important for cross-subject performance than accounting for anatomical differences. Future BCI systems should prioritize signal quality monitoring and adaptive preprocessing to maximize generalization.

Behavioral factors also play a significant role, with movement execution consistency correlating with accuracy ($r = 0.71$, $p = 0.004$). This suggests that better task instructions and training protocols could improve BCI performance independent of algorithmic advances.

E. FUTURE DIRECTIONS

The strong performance of HybridFormer (91.2% within-subject, 78.5% cross-subject) establishes a solid foundation for several promising research directions. While our evaluation across 43 subjects from multiple datasets demonstrates robust generalizability, validation on larger cohorts including diverse age groups and clinical populations would further strengthen these findings. The achieved cross-subject accuracy of 78.5%, which exceeds the 70% threshold for clinical BCI utility, motivates evaluation in patient populations with motor impairments, where the model's neurophysiologically valid attention patterns could provide clinically relevant insights. Extension to more complex motor paradigms, including continuous movement decoding, multi-degree-of-freedom control, and naturalistic tasks, would broaden applicability to real-world assistive technologies. Finally, while the model achieves real-time performance (15 ms inference), optimization for embedded systems would enable portable BCI deployment, and incorporating advanced transfer learning [29] or meta-learning approaches could further reduce the within-subject to cross-subject performance gap toward truly calibration-free BCI systems.

VI. CONCLUSION

This study introduced HybridFormer, a hybrid CNN-LSTM-Attention architecture for high-gamma EEG motor classification. The model achieved 91.2% within-subject and 78.5% cross-subject accuracy, outperforming state-of-the-art CNN, LSTM, and transformer baselines. Data augmentation and selective attention mechanisms significantly improved generalization while maintaining real-time performance. A strict non-overlapping temporal split experiment without

augmentation confirmed that HybridFormer retains a +5.1% advantage over the best baseline, ruling out information leakage. Learning-curve experiments demonstrated that the hybrid architecture's inductive biases provide superior data efficiency compared to pure transformer designs, retaining 77.4% of full-data accuracy at 10% training data.

Attention visualization confirmed neurophysiologically valid activation patterns, enhancing model interpretability and potential clinical trust. Stability analysis across random seeds (cosine similarity = 0.91) confirmed that these patterns are reproducible. The results demonstrate that incorporating domain-specific inductive biases enables superior efficiency and generalization compared to pure transformer designs.

Future work will explore transfer learning, self-supervised pretraining, and deployment on embedded hardware for portable, adaptive BCIs.

REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.

[2] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain–computer communication," *Proc. IEEE*, vol. 89, no. 7, pp. 1123–1134, Jul. 2001.

[3] N. E. Crone, A. Sinai, and A. Korzeniewska, "High-frequency gamma oscillations and human brain mapping with electrocorticography," *Prog. Brain Res.*, vol. 159, pp. 275–295, Oct. 2006.

[4] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.

[5] K. J. Miller, E. C. Leuthardt, G. Schalk, R. P. N. Rao, N. R. Anderson, D. W. Moran, J. W. Miller, and J. G. Ojemann, "Spectral changes in cortical surface potentials during motor movement," *J. Neurosci.*, vol. 27, no. 9, pp. 2424–2432, Feb. 2007.

[6] 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 Neurosci.*, vol. 6, p. 39, Mar. 2012.

[7] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," 2015, *arXiv:1511.06448*.

[8] 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. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[9] 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.

[10] 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 Human Neurosci.*, vol. 15, Jun. 2021, Art. no. 653659.

[11] K. Zhang, N. Robinson, S.-W. Lee, and C. Guan, "Adaptive transfer learning for EEG motor imagery classification with deep convolutional neural network," *Neural Netw.*, vol. 136, pp. 1–10, Apr. 2021.

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

[13] 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.

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

[15] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, Dec. 2008.

[16] F. Darvas, R. Scherer, J. G. Ojemann, R. P. Rao, K. J. Miller, and L. B. Sorensen, "High gamma mapping using EEG," *NeuroImage*, vol. 49, no. 1, pp. 930–938, Jan. 2010.

[17] S. An, S. Kim, P. Chikontwe, and S. H. Park, "Dual attention relation network with fine-tuning for few-shot EEG motor imagery classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 15479–15493, Nov. 2024.

[18] Y. Hou, S. Jia, X. Lun, Z. Hao, Y. Shi, Y. Li, R. Zeng, and J. Lv, "GCNs-Net: A graph convolutional neural network approach for decoding time-resolved EEG motor imagery signals," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7312–7323, Jun. 2024.

[19] F. Lotte, "Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces," *Proc. IEEE*, vol. 103, no. 6, pp. 871–890, Jun. 2015.

[20] M. H. Rafiei, L. V. Gauthier, H. Adeli, and D. Takabi, "Self-supervised learning for electroencephalography," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 1457–1471, Feb. 2024.

[21] D. Li, J. Wang, J. Xu, X. Fang, and Y. Ji, "Cross-channel specific-mutual feature transfer learning for motor imagery EEG signals decoding," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 13472–13482, Oct. 2024.

[22] F. Wang, S.-H. Zhong, J. Peng, J. Jiang, and Y. Liu, "Data augmentation for EEG-based emotion recognition with deep convolutional neural networks," in *Proc. Int. Conf. Multimedia Modeling*, in Lecture Notes in Computer Science, 2018, pp. 82–93.

[23] K. Gregor Hartmann, R. Tibor Schirrmeister, and T. Ball, "EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals," 2018, *arXiv:1806.01875*.

[24] A. Zakaria Talha, N. S. Eissa, and M. Ibrahim Shapiai, "Applications of brain computer interface for motor imagery using deep learning: Review on recent trends," *J. Adv. Res. Appl. Sci. Eng. Technol.*, vol. 40, no. 2, pp. 96–116, Feb. 2024.

[25] T. H. Shovon, Z. A. Nazi, S. Dash, and Md. F. Hossain, "Classification of motor imagery EEG signals with multi-input convolutional neural network by augmenting STFT," in *Proc. 5th Int. Conf. Adv. Electr. Eng. (ICAEE)*, Sep. 2019, pp. 398–403.

[26] J. Y. Cheng, M. Y. L. Goh, K. M. Leung, and P. V. S. Lee, "Effect of spatial and temporal augmentation on training a deep convolutional neural network for automatic classification of EEG signals," *J. Neural Eng.*, vol. 17, no. 5, 2020, Art. no. 056032.

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

[28] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bull. Amer. Meteorological Soc.*, vol. 79, no. 1, pp. 61–78, Jan. 1998.

[29] W. Cui, Y. Xiang, Y. Wang, T. Yu, X.-F. Liao, B. Hu, and Y. Li, "Deep multiview module adaption transfer network for subject-specific EEG recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 2, pp. 2917–2930, Feb. 2025.

GHADA ABDELHADY (Member, IEEE) is currently an Associate Professor with the Faculty of Engineering, October University for Modern Sciences and Arts (MSA University), Egypt, with 24 years of experience in teaching engineering and computer science courses, projects, and M.Sc. and Ph.D. supervision. She is also with the Center of Excellence, Faculty of Engineering, MSA University. She is also the Manager of the Center of Excellence for Projects and Entrepreneurship, MSA University, and the Director of the Technology Transfer Office. Her research interests include artificial intelligence, machine learning, cryptography, cybersecurity, the IoT, and tracking systems. She is a Certified Reviewer of the Quality Assurance and Accreditation of Egyptian Higher Education and a member of several research organizations. She was awarded a fellowship from the Higher Education Academy and a PGCert in education at the University of Greenwich, in 2018. Recently, she received the McKinsey Forward Program Badge from McKinsey & Company and the Mastery Award Badge from the IBM Skills Academy for the Artificial Intelligence Analyst Exam, in 2020.

ABDULRAHMAN GHANDOURA (Member, IEEE) received the B.S. degree in electronics engineering from Johnson and Wales University, Providence, RI, USA, the M.S. degree in network engineering and management from DePaul University, Chicago, IL, USA, in 2013, and the Ph.D. degree in electrical and computer engineering from Southern Illinois University, Carbondale, IL, USA, in 2018. In 2019, he joined Umm Al-Qura University, Makkah, Saudi Arabia, where he is currently an Assistant Professor with the Department of Engineering and Science, Applied College. His research interests include network architecture and wireless sensor networks, 6G and beyond wireless communication technologies, and the Internet of Things and the Internet of Everything applications.

ABDULLAH ALAJMI (Member, IEEE) received the B.S. degree in information system technology from Southern Illinois University, Carbondale, IL, USA, in 2010, the M.S. degree in information technology from DePaul University, Chicago, IL, USA, in 2014, and the Ph.D. degree from the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K., in 2023. His research interests include artificial intelligence for wireless systems, the Internet of Things, NOMA, and the Internet of Everything.

ZIAD GHAZALY received the B.Sc. degree in computer systems engineering from MSA University, affiliated with the University of Greenwich.

He has worked as a Research Assistant on EEG signal classification using advanced deep learning approaches, including transformer-based and hybrid architectures, with a focus on high gamma-band analysis for brain–computer interface applications, resulting in a published book chapter. He has also gained industry experience through internships at Diyar United Company, Bibliotheca Alexandrina, and Ooredoo Kuwait, contributing to production-level ML analytics, forecasting systems, NLP pipelines, and cloud-based AI solutions. He is currently a Junior Data Scientist with an interdisciplinary background in artificial intelligence, machine learning, and intelligent systems. His graduation project involved the design and deployment of a YOLOv8-based dental diagnosis system integrated with a mobile application and a FastAPI backend. His research interests include applied machine learning, deep learning, time-series forecasting, natural language processing, and computer vision. His current research interests include developing AI-driven solutions for healthcare, education, and real-world decision-support systems.