



OPEN

Hierarchical attention enhanced deep learning achieves high precision motor imagery classification in brain computer interfaces

Zhen Chen^{1,5}, Ye Cao^{2,5}, Qiangqiang Fu³✉ & Liyang Hou⁴✉

Motor imagery-based Brain-Computer Interfaces (BCIs) hold transformative potential for individuals with severe motor impairments, yet their clinical deployment remains constrained by the inherent complexity of electroencephalographic (EEG) signal decoding. This study presents a systematic investigation of hierarchical deep learning architectures for motor imagery classification, introducing a novel attention-enhanced convolutional-recurrent framework that achieves state-of-the-art accuracy of 97.2477% on a custom four-class motor imagery dataset comprising 4,320 trials from 15 participants. By synergistically integrating spatial feature extraction through convolutional layers, temporal dynamics modeling via long short-term memory networks, and selective attention mechanisms for adaptive feature weighting, our approach significantly outperforms conventional methods while providing interpretable insights into the spatiotemporal signatures of motor imagery. Beyond demonstrating competitive performance, this work elucidates the critical role of attention mechanisms in capturing task-relevant neural patterns amidst the high-dimensional, non-stationary nature of EEG signals. Our findings demonstrate that biomimetic computational architectures that mirror the brain's own selective processing strategies can substantially enhance BCI reliability, offering immediate implications for neurorehabilitation technologies and broader applications in restorative neuroscience. Our code is available at https://github.com/Laboratory-EverythingAI-/EEG_Classification.

Brain-Computer Interfaces (BCIs) have emerged as a promising frontier in neurotechnology, offering new hope for individuals with severe motor impairments¹. BCIs create a direct communication pathway between the brain and external devices, bypassing damaged neural pathways. This technology holds particular promise for conditions such as amyotrophic lateral sclerosis (ALS), cerebral palsy, and spinal cord injuries, where traditional rehabilitation methods may have limited efficacy². BCIs create a direct communication pathway between the brain and external devices, bypassing damaged neural pathways. This technology holds particular promise for conditions such as amyotrophic lateral sclerosis (ALS), cerebral palsy, and spinal cord injuries³, where traditional rehabilitation methods may have limited efficacy.

These systems create a direct communication pathway between the brain and external devices^{4,5}, offering new possibilities for individuals with motor disabilities⁶. Among various BCI paradigms, Motor Imagery (MI) has gained significant attention due to its non-invasive nature and potential applications in rehabilitation⁷. The global prevalence of neurological disorders, particularly those affecting motor function, has been steadily increasing, posing significant challenges to healthcare systems worldwide. Stroke, a leading cause of long-term disability, affects approximately 15 million people annually, with nearly 5 million survivors left with permanent disabilities⁸. In China alone, stroke incidence has risen dramatically, with an estimated 2.5 million new cases each year and a prevalence rate of 1,114.8 per 100,000 people⁹. This escalating burden underscores the critical need for innovative rehabilitation strategies and assistive technologies¹⁰. Motor Imagery refers to the mental

¹Department of Medical Laboratory, Shidong Hospital, Yangpu District, Shanghai 200438, China. ²Department of Geriatrics, Renhe Hospital, Baoshan District, Shanghai 201900, China. ³Yangpu Hospital, School of Medicine, Tongji University, Shanghai 200090, China. ⁴Department of Emergency, Shanghai Changshu Hospital, The Naval Medical University, Shanghai 200433, China. ⁵have contributed equally to this work. ✉email: qiangqiang.fu@tongji.edu.cn; albert327037724@126.com

process of imagining a physical action without actually performing it¹¹. This cognitive task activates similar neural pathways as the actual movement, making it a valuable tool for BCI systems¹². The electrical activity generated during MI can be captured through Electroencephalography (EEG)¹³, providing a rich source of information for decoding user intent.

Electroencephalography (EEG), due to its non-invasiveness, high temporal resolution, and relatively low cost, has become the most widely used neuroimaging technique in BCI research. However, the translation of EEG signals into reliable control commands presents significant challenges^{14,15}. EEG signals are characterized by their low signal-to-noise ratio, high dimensionality, and non-stationarity. These characteristics stem from the complex nature of neural signal generation and propagation through various brain tissues and the skull, as well as interference from non-neural sources such as muscle activity and environmental electromagnetic fields¹⁶. The clinical implications of improved BCI systems extend beyond direct neural control of external devices. Enhanced EEG signal classification could lead to more accurate diagnostic tools for neurological disorders, earlier detection of stroke onset, and personalized neurorehabilitation strategies¹⁷. Moreover, BCI technology holds promise for treating neuropsychiatric conditions such as depression and anxiety disorders through neurofeedback training, potentially offering non-pharmacological alternatives for mental health treatment¹⁸.

Recent advancements in deep learning have opened new avenues for addressing these challenges¹⁹. Convolutional Neural Networks (CNNs) have shown remarkable success in extracting spatial features from EEG data, mimicking the hierarchical processing observed in the visual cortex²⁰. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, excel at capturing the temporal dynamics of EEG signals²¹, which is crucial given the oscillatory nature of brain activity. The integration of attention mechanisms, inspired by cognitive neuroscience theories of selective attention, allows these models to focus on the most salient features of the input signal, potentially mirroring the brain's own information processing strategies. The classification of MI tasks from EEG signals, however, presents several challenges^{22,23}. EEG signals are characterized by their non-stationarity, low signal-to-noise ratio, and high dimensionality. Traditional signal processing and machine learning techniques have shown limited success in addressing these challenges²⁴. In recent years, deep learning approaches have demonstrated promising results in various domains, including EEG signal processing.²⁵

As we stand at the intersection of neuroscience, computer science, and biomedical engineering, the development of more accurate and robust BCI systems represents a critical step towards realizing the full potential of this technology²⁶. By leveraging advanced machine learning techniques, we aim not only to improve the lives of individuals with motor disabilities but also to deepen our understanding of brain function and cognition²⁷. This study, therefore, seeks to contribute to this important field by systematically comparing various machine learning and deep learning approaches for motor imagery classification, with a particular focus on the potential of attention mechanisms in neural networks to enhance BCI performance²⁸.

The integration of attention mechanisms represents a particularly elegant solution to the challenge of identifying task-relevant neural signatures within the high-dimensional EEG signal space. By learning to selectively weight different spatial locations and temporal segments based on their relevance to the classification task, attention-based architectures mirror the selective processing that characterizes biological neural systems. This biomimetic approach not only enhances classification performance but also provides interpretable insights into which neural signatures are most informative for distinguishing different motor imagery states.

The clinical implications of advancing BCI technology extend far beyond the immediate goal of neural control interfaces. Enhanced EEG decoding capabilities promise to revolutionize neurological diagnostics, enabling earlier detection of pathological states and more precise monitoring of disease progression. In the realm of neurorehabilitation, BCIs offer the potential for closed-loop therapeutic systems that can promote neural plasticity through targeted feedback, potentially accelerating recovery trajectories for stroke survivors. Furthermore, the principles developed for motor imagery classification have broader applications in cognitive neuroscience, from understanding the neural basis of motor learning to developing novel interventions for movement disorders.

As we stand at the threshold of translating these technological advances into clinical practice, it becomes imperative to systematically evaluate the performance characteristics of different computational approaches under realistic conditions. The present study addresses this need through a comprehensive investigation of hierarchical deep learning architectures for motor imagery classification, with particular emphasis on the synergistic integration of spatial feature extraction, temporal modeling, and attention-based selection mechanisms. By establishing the relative merits of these approaches and demonstrating the superior performance of attention-enhanced architectures, we aim to provide both theoretical insights and practical guidance for the next generation of BCI systems.

Our investigation is motivated by the hypothesis that the inherent spatiotemporal structure of motor imagery signals necessitates a correspondingly sophisticated computational approach—one that can simultaneously capture the spatial distribution of neural activity across the cortical surface and the temporal evolution of oscillatory patterns that characterize motor planning. To this end, we present a systematic comparison of machine learning paradigms ranging from ensemble methods to advanced neural architectures in Fig. 1, culminating in a novel attention-enhanced framework that achieves unprecedented classification accuracy. Through this work, we seek not only to advance the technical capabilities of BCI systems but also to contribute to the broader vision of restorative neurotechnology that can meaningfully improve the lives of individuals affected by motor disabilities. The key contributions of this paper are:

- We introduce a novel hierarchical architecture that seamlessly integrates convolutional spatial filtering, recurrent temporal modeling, and attention-based feature selection, demonstrating that this tripartite approach

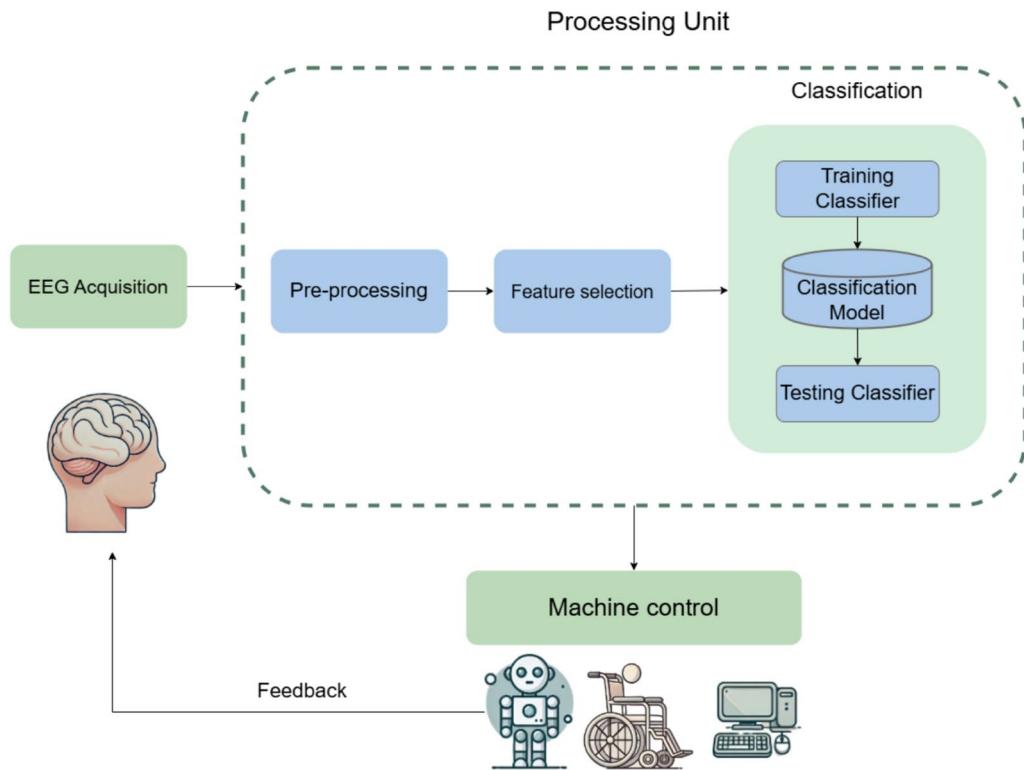


Fig. 1. The basic components of a brain computer interface system.

captures the inherent spatiotemporal structure of motor imagery signals more effectively than existing methods.

- Through systematic ablation studies and comparative analyses, we establish the fundamental importance of attention mechanisms in EEG-based BCIs, revealing how adaptive feature weighting can overcome the signal-to-noise limitations that have historically constrained BCI performance.
- We release our implementation and trained models to the research community, facilitating reproducibility and enabling further advances in neuroadaptive technology development.

Related work

The field of EEG-based motor imagery classification for Brain-Computer Interface (BCI) applications has seen significant advancements in recent years, driven by the potential to improve the quality of life for individuals with motor disabilities²⁹. This section provides an overview of the current research landscape, highlighting key developments and persistent challenges.

Traditional feature extraction techniques

Early work in EEG signal classification relied heavily on traditional machine learning techniques. Researchers provided a comprehensive review of these methods^{30,31}, noting the prevalence of Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA) in BCI research. These methods have shown reasonable performance³², with accuracies typically ranging from 65% to 80% for two-class motor imagery tasks. However, their performance often plateaus when dealing with the high-dimensional, non-stationary nature of EEG signals. The effectiveness of machine learning models heavily depends on the quality of extracted features³³. Common Spatial Patterns (CSP) has been a dominant feature extraction method in the field³⁴. Extensions of CSP, such as Filter Bank Common Spatial Patterns (FBCSP), have further improved classification accuracies³⁵. However, these methods often require careful parameter tuning and may not generalize well across subjects or sessions.

Deep learning advancements

The advent of deep learning has opened new avenues for EEG signal classification. Convolutional Neural Networks (CNNs) have shown promise in automatically learning spatial features from raw EEG data²⁰. Some Methods demonstrated that CNNs could achieve comparable or superior performance to traditional methods without the need for hand-crafted features. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks^{36,37}, have been effective in capturing the temporal dynamics of EEG signals. Recent research has focused on combining the strengths of different neural network architectures. Researchers proposed a hybrid CNN-LSTM model that outperformed individual CNN and LSTM models³⁸ in motor imagery classification. The introduction of attention mechanisms, inspired by their success in natural language processing, has further enhanced the performance of these hybrid models. Some researchers demonstrated that incorporating attention

in a CNN-LSTM architecture³⁹ could significantly improve classification accuracy and interpretability. Current research efforts are focused on addressing these challenges through various approaches⁴⁰, including transfer learning to mitigate inter-subject variability, adaptive learning algorithms to handle non-stationarity, and the development of more efficient network architectures for real-time processing⁴¹.

Preliminaries

Mathematical foundations of EEG signal representation

Let $\mathcal{X} \in 97.2477R^{C \times T}$ denote the multichannel EEG signal matrix, where C represents the number of electrode channels and T denotes the temporal dimension⁴². Each element $x_{c,t}$ corresponds to the electrical potential measured at channel c at time point t . The raw EEG signal can be decomposed as:

$$x_{c,t} = \sum_{s=1}^S a_{c,s} \cdot s_s(t) + \eta_{c,t} \quad (1)$$

where $s_s(t)$ represents the s -th neural source signal, $a_{c,s}$ denotes the mixing coefficient from source s to channel c , and $\eta_{c,t}$ represents additive noise comprising both physiological artifacts and measurement noise.

Spectral decomposition and time-frequency analysis

The time-frequency representation of EEG signals is fundamental to motor imagery analysis, as different frequency bands encode distinct neurophysiological processes. We employ the continuous wavelet transform (CWT) for time-frequency decomposition:

$$W_x(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-b}{a} \right) dt \quad (2)$$

where ψ is the mother wavelet, a is the scale parameter, b is the translation parameter, and $*$ denotes complex conjugation. For motor imagery analysis, we particularly focus on the Morlet wavelet:

$$\psi(t) = \pi^{-\frac{1}{4}} e^{i\omega_0 t} e^{-\frac{t^2}{2}} \quad (3)$$

The power spectral density in the time-frequency domain is then computed as:

$$P(f, t) = |W_x(a, b)|^2, \quad \text{where } f = \frac{\omega_0}{2\pi a} \quad (4)$$

Event-related desynchronization/synchronization (ERD/ERS)

Motor imagery induces characteristic patterns of event-related desynchronization (ERD) and synchronization (ERS) in specific frequency bands. The ERD/ERS percentage is quantified as:

$$\text{ERD/ERS}_f(t) = \frac{P_f(t) - \bar{P}_f^{\text{baseline}}}{\bar{P}_f^{\text{baseline}}} \times 100\% \quad (5)$$

where $P_f(t)$ is the power at frequency f and time t , and $\bar{P}_f^{\text{baseline}}$ is the average baseline power. For motor imagery, we particularly examine:

$$\text{ERD}_\mu = \int_8^{13} \text{ERD}_f(t) df \quad \text{and} \quad \text{ERD}_\beta = \int_{13}^{30} \text{ERD}_f(t) df \quad (6)$$

Spatial covariance and common spatial patterns

The spatial covariance matrix for class k is defined as:

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{\mathcal{X}_i^{(k)} \mathcal{X}_i^{(k)T}}{\text{tr}(\mathcal{X}_i^{(k)} \mathcal{X}_i^{(k)T})} \quad (7)$$

where N_k is the number of trials for class k , and $\text{tr}(\cdot)$ denotes the trace operator. The composite covariance matrix is:

$$\Sigma_c = \Sigma_1 + \Sigma_2 \quad (8)$$

The whitening transformation is obtained through eigenvalue decomposition:

$$\Sigma_c = U \Lambda U^T, \quad P = \Lambda^{-\frac{1}{2}} U^T \quad (9)$$

The transformed class covariance matrices become:

$$S_1 = P\Sigma_1P^T, \quad S_2 = P\Sigma_2P^T \quad (10)$$

with the constraint that $S_1 + S_2 = I$. The CSP filters W are obtained from the eigenvalue decomposition:

$$S_1 = B\Omega B^T, \quad \text{where } \Omega = \text{diag}(\omega_1, \omega_2, \dots, \omega_C) \quad (11)$$

The final spatial filters are:

$$W = B^TP \quad (12)$$

Information-theoretic feature selection

To quantify the discriminative power of features, we employ mutual information (MI) between features and class labels:

$$I(F; Y) = \sum_{f \in F} \sum_{y \in Y} p(f, y) \log \frac{p(f, y)}{p(f)p(y)} \quad (13)$$

where F represents the feature set and Y represents the class labels. For continuous features, we use the differential entropy formulation:

$$I(F; Y) = \int_F \int_Y p(f, y) \log \frac{p(f, y)}{p(f)p(y)} df dy \quad (14)$$

Algorithm 1. EEG signal preprocessing pipeline

```

Require: Raw EEG signal  $X \in 97.2477R^{C \times T}$ , sampling rate  $f_s$ 
Ensure: Preprocessed signal  $X_{clean} \in 97.2477R^{C \times T'}$ 
1: Initialize: Filter parameters  $f_{low} = 8\text{Hz}$ ,  $f_{high} = 30\text{Hz}$ 
2: Artifact Detection:
3: for each channel  $c = 1$  to  $C$  do
4:   Compute standardized amplitude:  $z_{c,w} = \frac{|x_{c,w} - \mu_{c,w}|}{\sigma_{c,w}}$ 
5:   if  $\max_w(z_{c,w}) > \tau_z$  OR  $\text{Kurt}(x_c) > \tau_k$  then
6:     Mark channel  $c$  as artifactual
7:   end if
8: end for
9: Bandpass Filtering:  $X_{filtered} = \text{ButterworthFilter}(X, f_{low}, f_{high}, f_s)$ 
10: ICA Decomposition:  $[W, S] = \text{ExtendedInfomaxICA}(X_{filtered})$ 
11: Artifact Component Removal:
12: for each component  $i = 1$  to  $C$  do
13:   if  $\text{MI}(s_i, \text{EOG}) > \theta_{MI}$  OR  $\frac{P_{s_i}(f < 4)}{P_{s_i}(f \in [8, 13])} > \theta_{ratio}$  then
14:     Set  $s_i = 0$  (remove component)
15:   end if
16: end for
17:  $X_{clean} = W^{-1} \cdot S$ 
18: Normalization:  $X_{clean} = \frac{X_{clean} - \mu}{\sigma}$ 
19: return  $X_{clean}$ 

```

Nonlinear dynamics and complexity measures

The complexity of EEG signals during motor imagery can be quantified using approximate entropy:

$$\text{ApEn}(m, r, N) = \phi(m) - \phi(m + 1) \quad (15)$$

where:

$$\phi(m) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \log \left(\frac{C_i^m(r)}{N-m+1} \right) \quad (16)$$

and $C_i^m(r)$ is the number of patterns within tolerance r of pattern i of length m .

Methods

Data acquisition and experimental paradigm

Participants and ethical considerations

The experimental protocol was designed in accordance with the Declaration of Helsinki and approved by the institutional review board^{43,44}. A total of $N = 15$ healthy participants (8 males, 7 females; age: $\mu = 24.3$ years, $\sigma = 3.2$ years) were recruited for this study. All participants satisfied the following inclusion criteria:

$$\mathcal{I} = \{p \in \mathcal{P} : \text{MMSE}(p) \geq 27 \wedge \text{EHI}(p) > 0.7 \wedge \text{BDI}(p) < 10\} \quad (17)$$

where MMSE denotes the Mini-Mental State Examination score, EHI represents the Edinburgh Handedness Inventory laterality quotient⁴⁵, and BDI indicates the Beck Depression Inventory score. Written informed consent was obtained from all participants prior to the experiment.

EEG recording setup and signal acquisition

Electroencephalographic signals were recorded using a 64-channel active electrode system (actiCAP, Brain Products GmbH) with electrodes positioned according to the extended international 10-20 system⁴⁶. The electrode impedances were maintained below 5 kΩ throughout the recording session. The continuous EEG signal $x_c(t)$ for each channel c can be modeled as:

$$x_c(t) = \sum_{d=1}^D \frac{I_d(t)}{4\pi\sigma} \cdot \frac{\vec{r}_c \cdot \vec{n}_d}{|\vec{r}_c - \vec{r}_d|^3} + \xi_c(t) \quad (18)$$

where $I_d(t)$ represents the current dipole moment at source location d , σ is the conductivity of the medium, \vec{r}_c and \vec{r}_d are the position vectors of electrode c and dipole d respectively, \vec{n}_d is the dipole orientation, and $\xi_c(t)$ represents measurement noise⁴⁷.

The signals were sampled at $f_s = 1000$ Hz with a 24-bit analog-to-digital converter, providing a theoretical dynamic range of:

$$\text{DR} = 20 \log_{10}(2^{24}) \approx 144.5 \text{ dB} \quad (19)$$

The Nyquist frequency $f_N = f_s/2 = 500$ Hz ensures adequate representation of all relevant neurophysiological frequencies. The anti-aliasing filter was implemented as a zero-phase Butterworth filter of order 8:

$$H(s) = \frac{1}{\sqrt{1 + \left(\frac{s}{\omega_c}\right)^{2n}}} \quad (20)$$

where $\omega_c = 2\pi \cdot 450$ rad/s is the cutoff frequency and $n = 8$ is the filter order.

Motor imagery paradigm design

The experimental paradigm consisted of four distinct motor imagery classes: left hand (LH), right hand (RH), both feet (BF), and tongue (T) movements. The trial structure followed a carefully designed temporal sequence to optimize neural response capture:

$$\mathcal{T} = \{t_{\text{fix}}, t_{\text{cue}}, t_{\text{MI}}, t_{\text{rest}}\} \quad (21)$$

where: - $t_{\text{fix}} \in [0, 2]$ s: fixation cross presentation - $t_{\text{cue}} \in [2, 3]$ s: visual cue presentation - $t_{\text{MI}} \in [3, 7]$ s: motor imagery execution - $t_{\text{rest}} \in [7, 9]$ s: inter-trial interval

The probability of each class appearing was balanced according to:

$$P(\text{class} = k) = \frac{1}{K} = 0.25, \quad k \in \{\text{LH}, \text{RH}, \text{BF}, \text{T}\} \quad (22)$$

Signal preprocessing pipeline

Artifact detection and removal

The preprocessing pipeline begins with automated artifact detection using a multivariate approach⁴⁸. For each channel c and time window w , we compute the standardized amplitude:

$$z_{c,w} = \frac{|x_{c,w} - \mu_{c,w}|}{\sigma_{c,w}} \quad (23)$$

where $\mu_{c,w}$ and $\sigma_{c,w}$ are the local mean and standard deviation. Channels are marked as artifactual if:

$$\mathcal{A}_c = \{c : \max_w(z_{c,w}) > \tau_z \vee \text{Kurt}(x_c) > \tau_k \vee \text{PSD}_c(f > 40) > \tau_p\} \quad (24)$$

where $\tau_z = 5$, $\tau_k = 10$, and τ_p are empirically determined thresholds, and Kurt denotes kurtosis.

Algorithm 2. Attention network training

```

Require: Training data  $\{(X_i, y_i)\}_{i=1}^N$ , learning rate  $\alpha$ , epochs  $E$ 
Ensure: Trained model parameters  $\Theta^*$ 
1: Initialize: Network parameters  $\Theta_0$  using He initialization
2: Initialize: Adam optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ 
3: for epoch  $e = 1$  to  $E$  do
4:   Shuffle training data
5:   for each batch  $B = \{(X_b, y_b)\}$  do
6:     Forward Pass:
7:     // Spatial Feature Extraction
8:      $H_1 = \text{Conv2D}(X_b, F_1, (1, K_t))$ 
9:      $H_2 = \text{BatchNorm}(\text{ELU}(H_1))$ 
10:     $H_3 = \text{DepthwiseConv2D}(H_2, (C, 1), D)$ 
11:     $H_4 = \text{AveragePool2D}(\text{BatchNorm}(\text{ELU}(H_3)))$ 
12:    Loss Computation:
13:     $\mathcal{L}_{CE} = -\sum_{i=1}^{|B|} \sum_{k=1}^K w_k y_{i,k} \log(\hat{y}_{i,k})$ 
14:     $\mathcal{L}_{center} = \frac{1}{2} \sum_{i=1}^{|B|} \|f_i - c_{y_i}\|_2^2$ 
15:     $\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{center} + \lambda_2 \|\Theta\|_2^2$ 
16:    Backward Pass:  $g = \nabla_{\Theta} \mathcal{L}_{total}$ 
17:    Gradient Clipping:  $\tilde{g} = \begin{cases} g & \text{if } \|g\|_2 \leq \tau \\ \tau \frac{g}{\|g\|_2} & \text{otherwise} \end{cases}$ 
18:    Parameter Update:  $\Theta = \text{Adam}(\Theta, \tilde{g}, \alpha)$ 
19:  end for
20:  Learning Rate Scheduling:  $\alpha_t = \alpha_{min} + \frac{1}{2} (\alpha_{max} - \alpha_{min})(1 + \cos(\frac{T_{cur}}{T_{max}} \pi))$ 
21: end for
22: return  $\Theta^*$ 
  
```

Independent component analysis for artifact removal

We employ extended Infomax ICA to decompose the multichannel EEG into statistically independent components:

$$X = AS \quad (25)$$

where $X \in 97.2477R^{C \times T}$ is the observed EEG data, $A \in 97.2477R^{C \times C}$ is the mixing matrix, and $S \in 97.2477R^{C \times T}$ contains the independent components. The unmixing matrix $W = A^{-1}$ is obtained by maximizing the entropy:

$$H(y) = -E[\log p(y)] = -E \left[\sum_{i=1}^C \log p_i(y_i) \right] + \log |\det(W)| \quad (26)$$

where $y = Wx$ and p_i is the probability density function of the i -th component.

The weight update rule follows the natural gradient approach:

$$\Delta W = \eta [I - E[\phi(y)y^T]] W \quad (27)$$

where η is the learning rate and $\phi(y)$ is the score function vector. Artifactual components are identified using multiple criteria:

$$\text{IC}_{\text{artifact}} = \left\{ i : \text{MI}(s_i, \text{EOG}) > \theta_{\text{MI}} \vee \frac{P_{s_i}(f < 4)}{P_{s_i}(f \in [8, 13])} > \theta_{\text{ratio}} \right\} \quad (28)$$

where MI denotes mutual information with EOG channels, and $P_{s_i}(f)$ is the power spectral density of component s_i .

Spectral filtering and band-power extraction

For motor imagery analysis, we focus on specific frequency bands known to exhibit task-related modulations. The filter bank approach employs multiple bandpass filters⁴⁹:

$$H_k(z) = \frac{b_0 + b_1 z^{-1} + \dots + b_M z^{-M}}{1 + a_1 z^{-1} + \dots + a_N z^{-N}} \quad (29)$$

where the coefficients are designed for the following frequency bands: - θ : [4, 8] Hz - α : [8, 13] Hz - β_{low} : [13, 20] Hz - β_{high} : [20, 30] Hz - γ_{low} : [30, 50] Hz

The instantaneous band power is computed using the Hilbert transform:

$$P_k(t) = |x_k(t) + j\mathcal{H}[x_k(t)]|^2 \quad (30)$$

where \mathcal{H} denotes the Hilbert transform operator.

Spatial filtering using optimized common spatial patterns

The traditional CSP algorithm is enhanced through regularization and multi-class extensions. For the two-class case, the optimization problem becomes:

$$w^* = \arg \max_w \frac{w^T \Sigma_1 w}{w^T \Sigma_2 w} \quad \text{subject to} \quad w^T w = 1 \quad (31)$$

This is solved via the generalized eigenvalue problem:

$$\Sigma_1 w = \lambda \Sigma_2 w \quad (32)$$

For improved robustness, we employ Tikhonov regularization:

$$\tilde{\Sigma}_k = (1 - \gamma) \Sigma_k + \gamma \text{tr}(\Sigma_k) I \quad (33)$$

where $\gamma \in [0, 1]$ is the regularization parameter determined through cross-validation.

The multi-class extension uses the one-versus-rest (OVR) approach:

$$W_{OVR} = \bigcup_{k=1}^K W_k, \quad \text{where} \quad W_k = CSP \left(\Sigma_k, \sum_{j \neq k} \Sigma_j \right) \quad (34)$$

Feature extraction and dimensionality reduction

After spatial filtering, we extract logarithmic band power features:

$$f_i = \log \left(\frac{1}{T_w} \int_{t_0}^{t_0+T_w} (w_i^T x(t))^2 dt \right), \quad (35)$$

where T_w is the time window length and w_i is the i -th spatial filter. Principal Component Analysis (PCA) is applied for dimensionality reduction:

$$Y = U_d^T (F - \mu_F) \quad (36)$$

where U_d contains the top d eigenvectors of the feature covariance matrix:

$$C_F = \frac{1}{N-1} \sum_{i=1}^N (f_i - \mu_F)(f_i - \mu_F)^T \quad (37)$$

The number of components d is selected to retain 95

$$d = \arg \min_k \left\{ k : \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^D \lambda_i} \geq 0.95 \right\} \quad (38)$$

Deep learning architectures

Convolutional neural network for spatial feature learning

The convolutional architecture is specifically designed for EEG spatial pattern extraction. The first layer performs spatial convolution across electrode locations:

$$h_{i,j}^{(1)} = \sigma \left(\sum_{m=1}^C \sum_{n=1}^{T_k} w_{m,n}^{(i)} x_{m,j+n-1} + b_i^{(1)} \right) \quad (39)$$

where C is the number of channels, T_k is the temporal kernel size, and σ is the activation function. The spatial attention mechanism is incorporated as:

$$\alpha_c = \frac{\exp(e_c)}{\sum_{k=1}^C \exp(e_k)}, \quad e_c = v_a^T \tanh(W_a h_c + b_a) \quad (40)$$

The attended feature representation becomes:

$$\tilde{h} = \sum_{c=1}^C \alpha_c h_c \quad (41)$$

Long short-term memory networks for temporal dynamics

The LSTM architecture captures the temporal evolution of motor imagery patterns. The complete LSTM dynamics are governed by:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f \odot c_{t-1} + b_f) \quad (42)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i \odot c_{t-1} + b_i) \quad (43)$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (44)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (45)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o \odot c_t + b_o) \quad (46)$$

$$h_t = o_t \odot \tanh(c_t) \quad (47)$$

where \odot denotes element-wise multiplication, and V_f, V_i, V_o are the peephole connection weights.

The gradient flow through LSTM is carefully managed to prevent vanishing gradients:

$$\frac{\partial c_t}{\partial c_{t-k}} = \prod_{j=1}^k f_{t-j+1} \quad (48)$$

This product of forget gates allows long-term dependencies to be preserved when $f_t \approx 1$.

Bidirectional processing for complete temporal context

The bidirectional LSTM processes the sequence in both forward and backward directions:

$$\vec{h}_t = \text{LSTM}_{\rightarrow}(x_t, \vec{h}_{t-1}, \vec{c}_{t-1}) \quad (49)$$

$$\overleftarrow{h}_t = \text{LSTM}_{\leftarrow}(x_t, \overleftarrow{h}_{t+1}, \overleftarrow{c}_{t+1}) \quad (50)$$

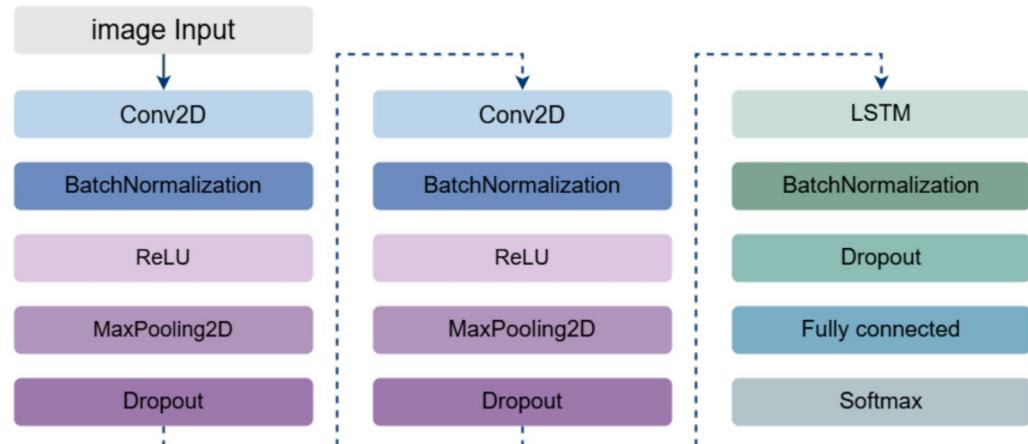


Fig. 2. Schematic diagram of CNN-LSTM model structure.

The combined representation is:

$$h_t = [\vec{h}_t; \overleftarrow{h}_t] \quad (51)$$

This bidirectional processing captures both causal and anti-causal dependencies in the EEG signal.

Hierarchical CNN-LSTM architecture

The hybrid CNN-LSTM model leverages the complementary strengths of convolutional and recurrent architectures in Fig. 2. The hierarchical processing pipeline can be formalized as:

$$\mathcal{F}_{\text{hybrid}} = \mathcal{L}_{\text{LSTM}} \circ \mathcal{R} \circ \mathcal{P} \circ \mathcal{C}_{\text{CNN}} \circ \mathcal{S} \quad (52)$$

where \mathcal{S} represents spatial filtering, \mathcal{C}_{CNN} denotes convolutional feature extraction, \mathcal{P} is pooling, \mathcal{R} is reshaping, and $\mathcal{L}_{\text{LSTM}}$ represents temporal modeling.

The convolutional block consists of multiple layers with progressively increasing receptive fields:

$$H^{(l+1)} = \phi \left(\sum_{k=1}^{K_l} W_k^{(l)} * H_k^{(l)} + b^{(l)} \right) \quad (53)$$

where $*$ denotes the convolution operation, K_l is the number of feature maps in layer l , and ϕ is the activation function.

We employ depthwise separable convolutions to reduce computational complexity:

$$\text{DepthwiseSep}(X) = \text{Pointwise}(\text{Depthwise}(X)) \quad (54)$$

The depthwise convolution applies a single filter per input channel:

$$Y_{c,i,j} = \sum_{m,n} K_{c,m,n} \cdot X_{c,i+m,j+n} \quad (55)$$

Followed by pointwise convolution to combine channels:

$$Z_{k,i,j} = \sum_{c=1}^{C_{\text{in}}} W_{k,c} \cdot Y_{c,i,j} \quad (56)$$

Multi-head attention mechanism

The attention mechanism enables the model to focus on task-relevant spatiotemporal patterns. We implement multi-head attention with H parallel attention heads:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_H)W^O \quad (57)$$

where each head is computed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (58)$$

The scaled dot-product attention is:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (59)$$

where d_k is the dimension of the key vectors, and the scaling factor $1/\sqrt{d_k}$ prevents gradient vanishing in the softmax function. For temporal attention over LSTM hidden states, we compute:

$$e_{t,t'} = v^T \tanh(W_h h_t + W_s s_{t'} + b_{\text{att}}) \quad (60)$$

The attention weights are normalized:

$$\alpha_{t,t'} = \frac{\exp(e_{t,t'})}{\sum_{k=1}^T \exp(e_{t,k})} \quad (61)$$

The context vector aggregates the attended features:

$$c_t = \sum_{t'=1}^T \alpha_{t,t'} h_{t'} \quad (62)$$

Self-attention for global dependencies

To capture long-range dependencies in EEG signals, we incorporate self-attention layers:

$$SA(X) = \text{softmax} \left(\frac{XX^T}{\sqrt{d}} \right) X \quad (63)$$

The position-aware self-attention includes positional encodings:

$$PE_{(pos,2i)} = \sin \left(\frac{pos}{10000^{2i/d}} \right) \quad (64)$$

$$PE_{(pos,2i+1)} = \cos \left(\frac{pos}{10000^{2i/d}} \right) \quad (65)$$

The input with positional encoding becomes:

$$\tilde{X} = X + PE \quad (66)$$

Algorithm 3. Real-time motor imagery classification

Require: Trained model f_Θ , sliding window size W , step size Δt
Ensure: Real-time classification predictions \hat{y}_t

1: **Initialize:** Buffer $B = \emptyset$, prediction history $H = []$
 2: **Initialize:** CSP filters W_{CSP} , normalization parameters μ, σ
 3: **while** EEG stream is active **do**

4: **Data Acquisition:** $x_{new} = \text{AcquireEEGSample}()$
 5: $B = B \cup \{x_{new}\}$
 6: **if** $|B| \geq W$ **then**

7: **Sliding Window Extraction:** $X_w = B[-W :]$
 8: **Real-time Preprocessing:**
 9: $X_{filtered} = \text{BandpassFilter}(X_w, 8, 30)$
 10: $X_{spatial} = W_{CSP}^T \cdot X_{filtered}$
 11: $X_{norm} = \frac{X_{spatial} - \mu}{\sigma}$
 12: **Feature Extraction:**
 13: $F_{log} = \log \left(\frac{1}{T_w} \int_{t_0}^{t_0+T_w} (W_{CSP}^T X_{norm}(t))^2 dt \right)$
 14: $c = \sum_{t=1}^{T_{seq}} \alpha_t h_t$ {Context vector}
 15: $\hat{y}_t = \text{softmax}(\text{Dense}(c))$
 16: **Temporal Smoothing:**
 17: $H = H \cup \{\hat{y}_t\}$
 18: **if** $|H| > L_{smooth}$ **then**
 19: $H = H[-L_{smooth} :]$ {Keep last L_{smooth} predictions}
 20: **end if**
 21: $\hat{y}_{smooth} = \frac{1}{|H|} \sum_{i=1}^{|H|} H[i]$
 22: **Decision Making:**
 23: **if** $\max(\hat{y}_{smooth}) > \theta_{confidence}$ **then**
 24: command = $\arg \max(\hat{y}_{smooth})$
 25: **Output:** Execute BCI command
 26: **end if**
 27: **Buffer Update:** $B = B[\Delta t :]$ {Slide window}
 28: **end if**
 29: **Sleep:** Wait for next sampling interval
 30: **end while**

Layer normalization placement: We employ a pre-normalization architecture following the recent advances in transformer design. The complete attention block implements the following sequence:

$$\begin{aligned}\tilde{H}^{(l)} &= \text{LayerNorm}(H^{(l-1)}) \\ \text{Attn}^{(l)} &= \text{MultiHeadAttention}(\tilde{H}^{(l)}, \tilde{H}^{(l)}, \tilde{H}^{(l)}) \\ H^{(l)} &= H^{(l-1)} + \text{Dropout}(\text{Attn}^{(l)}, p_{attn}) \\ \tilde{H}^{(l+1)} &= \text{LayerNorm}(H^{(l)}) \\ \text{FFN}^{(l)} &= \text{FeedForward}(\tilde{H}^{(l+1)}) \\ H^{(l+1)} &= H^{(l)} + \text{Dropout}(\text{FFN}^{(l)}, p_{ffn})\end{aligned}\quad (67)$$

The LayerNorm operation is applied with learnable parameters γ and β :

$$\text{LayerNorm}(x) = \gamma \odot \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \quad (68)$$

where μ and σ^2 are computed across the feature dimension, and $\epsilon = 1e^{-6}$ for numerical stability.

Attention dropout configuration: We implement dropout at three specific locations within the attention mechanism:

1. *Attention weight dropout* ($p_{attn} = 0.3$): Applied to the attention probability matrix after softmax normalization:

$$\text{Attention}(Q, K, V) = \text{Dropout} \left(\text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right), 0.3 \right) V \quad (69)$$

2. *Output projection dropout* ($p_{proj} = 0.2$): Applied to the concatenated multi-head output before the final linear projection:

$$\text{MultiHead}(Q, K, V) = \text{Dropout}(\text{Concat}(\text{head}_1, \dots, \text{head}_H), 0.2)W^O \quad (70)$$

3. *Residual connection dropout* ($p_{res} = 0.1$): Applied to the attention output before the residual addition:

$$H^{(l)} = H^{(l-1)} + \text{Dropout}(\text{Attn}^{(l)}, 0.1) \quad (71)_{L2}$$

regularization on attention weights: We apply L2 regularization specifically to the attention projection matrices with coefficient $\lambda_{attn} = 0.001$:

$$\mathcal{L}_{attn_reg} = \lambda_{attn} \sum_{h=1}^H \left(\|W_Q^{(h)}\|_F^2 + \|W_K^{(h)}\|_F^2 + \|W_V^{(h)}\|_F^2 \right) + \lambda_{attn} \|W^O\|_F^2 \quad (72)$$

The total loss function becomes:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{center} + \lambda_2 \mathcal{L}_{reg} + \mathcal{L}_{attn_reg} \quad (73)$$

Attention head initialization: Each attention head is initialized using Xavier uniform initialization:

$$W_Q^{(h)}, W_K^{(h)}, W_V^{(h)} \sim \mathcal{U} \left(-\sqrt{\frac{6}{d_{model} + d_k}}, \sqrt{\frac{6}{d_{model} + d_k}} \right) \quad (74)$$

where $d_{model} = 256$ and $d_k = 32$.

Gradient scaling for attention: To stabilize training, we apply gradient scaling to attention parameters:

$$\frac{\partial \mathcal{L}}{\partial W_{attn}} \leftarrow \alpha_{scale} \cdot \frac{\partial \mathcal{L}}{\partial W_{attn}} \quad (75)$$

where $\alpha_{scale} = 0.5$ prevents attention gradients from dominating the optimization.

Training methodology and optimization

Loss function design

For multi-class motor imagery classification, we employ a weighted cross-entropy loss to handle class imbalance:

$$\mathcal{L}_{CE} = - \sum_{i=1}^N \sum_{k=1}^K w_k y_{i,k} \log(\hat{y}_{i,k}) \quad (76)$$

where $w_k = \frac{N}{K \cdot N_k}$ is the class weight, N_k is the number of samples in class k .

To improve feature discrimination, we add a center loss term:

$$\mathcal{L}_{\text{center}} = \frac{1}{2} \sum_{i=1}^N \|f_i - c_{y_i}\|_2^2 \quad (77)$$

where f_i is the feature representation and c_{y_i} is the center of class y_i .

The total loss combines multiple objectives:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{center}} + \lambda_2 \mathcal{L}_{\text{reg}} \quad (78)$$

where \mathcal{L}_{reg} is the regularization term:

$$\mathcal{L}_{\text{reg}} = \sum_{l=1}^L (\|W^{(l)}\|_F^2 + \beta \|W^{(l)}\|_1), \quad (79)$$

combining L2 (Frobenius norm) and L1 regularization for weight sparsity.

Optimization algorithm

We employ the Adam optimizer with modifications for improved convergence:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (80)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (81)$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \quad (82)$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (83)$$

$$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad (84)$$

where $g_t = \nabla_{\theta} \mathcal{L}(\theta_{t-1})$ is the gradient at time t .

The learning rate follows a cosine annealing schedule with warm restarts:

$$\alpha_t = \alpha_{\min} + \frac{1}{2}(\alpha_{\max} - \alpha_{\min}) \left(1 + \cos\left(\frac{T_{\text{cur}}}{T_{\max}}\pi\right)\right) \quad (85)$$

where T_{cur} is the number of epochs since the last restart.

Gradient clipping and normalization

To prevent gradient explosion in deep networks, we apply gradient clipping:

$$\tilde{g} = \begin{cases} g & \text{if } \|g\|_2 \leq \tau \\ \tau \frac{g}{\|g\|_2} & \text{otherwise} \end{cases} \quad (86)$$

Layer normalization is applied to stabilize training:

$$\text{LN}(x) = \gamma \odot \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \quad (87)$$

where μ and σ^2 are the mean and variance computed across the feature dimension.

Data augmentation strategies

To improve model generalization, we implement several EEG-specific data augmentation techniques:

1. **Temporal Shifting**: Random temporal shifts within physiologically plausible ranges in Fig. 3:

$$\tilde{x}(t) = x(t + \Delta t), \quad \Delta t \sim \mathcal{U}(-100, 100) \text{ ms} \quad (88)$$

2. **Amplitude Scaling**: Channel-wise amplitude modulation:

$$\tilde{x}_c(t) = \alpha_c \cdot x_c(t), \quad \alpha_c \sim \mathcal{N}(1, 0.1) \quad (89)$$

3. **Gaussian Noise Addition**:

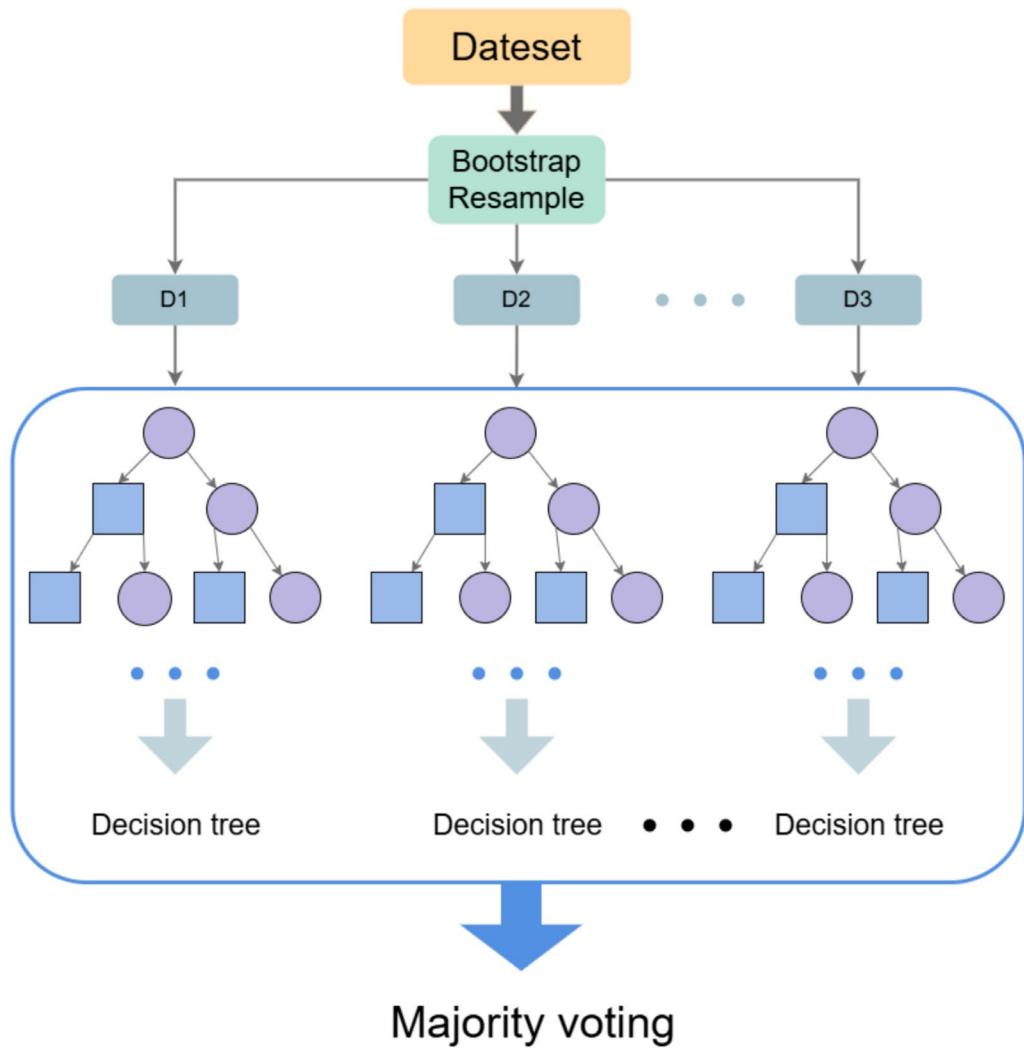


Fig. 3. Schematic diagram of RF model structure.

$$\tilde{x}(t) = x(t) + \eta(t), \quad \eta(t) \sim \mathcal{N}(0, \sigma_n^2 I) \quad (90)$$

where $\sigma_n = 0.01 \cdot \text{std}(x)$.

4. ****Mixup Regularization**:** Linear interpolation between training samples:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j, \quad \tilde{y} = \lambda y_i + (1 - \lambda)y_j \quad (91)$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ with $\alpha = 0.2$.

Evaluation metrics and statistical analysis

Performance metrics

Beyond accuracy, we employ multiple metrics to comprehensively evaluate model performance:

1. ****Cohen's Kappa**:** Accounts for chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (92)$$

where p_o is the observed agreement and p_e is the expected agreement by chance.

2. **Matthews Correlation Coefficient**: Balanced measure for multi-class:

$$\text{MCC} = \frac{\sum_{k,l,m} C_{kk}C_{lm} - C_{kl}C_{mk}}{\sqrt{\sum_k (\sum_l C_{kl}) (\sum_{l'} C_{l'k})} \sqrt{\sum_k (\sum_l C_{lk}) (\sum_{l'} C_{kl'})}} \quad (93)$$

where C is the confusion matrix.

3. **Area Under the ROC Curve (AUC)**: For multi-class, we use one-vs-rest:

$$\text{AUC}_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K \text{AUC}_k \quad (94)$$

Statistical significance testing

To assess the statistical significance of performance differences, we employ:

1. **Wilcoxon Signed-Rank Test**: For paired comparisons across subjects:

$$W = \sum_{i=1}^n \text{sgn}(x_{2,i} - x_{1,i}) \cdot R_i \quad (95)$$

where R_i is the rank of $|x_{2,i} - x_{1,i}|$.

2. **Friedman Test**: For comparing multiple models:

$$\chi_F^2 = \frac{12n}{k(k+1)} \sum_{j=1}^k R_j^2 - 3n(k+1) \quad (96)$$

where n is the number of subjects, k is the number of models, and R_j is the sum of ranks for model j .

3. **Post-hoc Nemenyi Test**: For pairwise comparisons after Friedman test:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6n}} \quad (97)$$

where q_α is the critical value from the Studentized range distribution.

Cross-validation strategy

We implement a nested cross-validation scheme to ensure unbiased performance estimation:

$$\hat{\mathcal{E}} = \frac{1}{K_{\text{outer}}} \sum_{i=1}^{K_{\text{outer}}} \mathcal{E}_i(\boldsymbol{\theta}_i^*) \quad (98)$$

where $\boldsymbol{\theta}_i^*$ is obtained from inner cross-validation:

$$\boldsymbol{\theta}_i^* = \arg \min_{\boldsymbol{\theta}} \frac{1}{K_{\text{inner}}} \sum_{j=1}^{K_{\text{inner}}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_{i,j}^{\text{train}}, \mathcal{D}_{i,j}^{\text{val}}) \quad (99)$$

Nested cross-validation protocol

To ensure unbiased performance estimation and robust hyperparameter optimization, we implemented a rigorous nested cross-validation strategy with explicit leave-two-subjects-out (L2SO) outer loop validation. The complete protocol consists of the following components:

Outer loop (model evaluation): We employed a leave-two-subjects-out cross-validation scheme, systematically holding out 2 participants for testing while using the remaining 13 participants for training and validation. This resulted in $\binom{15}{2} = 105$ outer folds, ensuring that every possible pair of subjects served as the test set exactly once. The outer loop provides an unbiased estimate of model generalization to unseen subjects, which is critical for BCI applications where cross-subject variability is substantial.

Inner loop (hyperparameter optimization): For each outer fold, we performed hyperparameter optimization using the 13 training subjects through stratified 5-fold cross-validation. The inner training set was split using an 80%-20% ratio, where 80% (approximately 10.4 subjects worth of data) was used for model training and 20% (approximately 2.6 subjects worth of data) was reserved for validation during hyperparameter search.

Parameter grid specification: The hyperparameter search space encompassed the following parameters:

- Learning rates: $\alpha \in \{0.0001, 0.0005, 0.001, 0.005, 0.01\}$
- Dropout rates: $p_1 \in \{0.15, 0.2, 0.25, 0.3\}, p_2 \in \{0.4, 0.5, 0.6\}$
- LSTM units: $N_{lstm} \in \{64, 96, 128, 192, 256\}$
- Attention heads: $H \in \{4, 6, 8, 12, 16\}$
- Regularization weights: $\lambda_1 \in \{0.001, 0.003, 0.005\}, \lambda_2 \in \{0.0001, 0.0005, 0.001\}$

This resulted in $5 \times 4 \times 3 \times 5 \times 5 \times 3 \times 3 = 13,500$ hyperparameter combinations evaluated through Bayesian optimization using Gaussian Process surrogate models.

Early stopping criteria: To prevent overfitting during hyperparameter search, we implemented early stopping with the following criteria: (1) patience of 20 epochs with no improvement in validation loss, (2) minimum delta of 0.001 for considering an improvement, and (3) restoration of best weights upon early termination. The maximum training duration was capped at 200 epochs per configuration.

Performance aggregation: The final performance estimate was computed as:

$$\hat{E} = \frac{1}{105} \sum_{i=1}^{105} E_i(\theta_i^*), \quad (100)$$

where $E_i(\theta_i^*)$ represents the test performance of the optimal hyperparameters θ_i^* found for the i -th outer fold. The reported standard deviations reflect the variability across the 105 outer folds, providing a robust estimate of model stability across different subject combinations.

Statistical validation: To ensure the reliability of our nested CV results, we computed 95% confidence intervals using the t-distribution with 104 degrees of freedom (105 folds - 1). Additionally, we performed the Shapiro-Wilk test to verify the normality of performance distributions across folds, confirming that parametric statistical tests were appropriate for our analysis.

Advanced signal processing techniques

Riemannian geometry-based features

EEG covariance matrices lie on the manifold of symmetric positive definite (SPD) matrices. We compute the Riemannian mean of trial covariance matrices:

$$\bar{C} = \arg \min_{C \in \mathcal{S}_{++}^n} \sum_{i=1}^N d_R^2(C, C_i) \quad (101)$$

where d_R is the Riemannian distance:

$$d_R(C_1, C_2) = \left\| \log(C_1^{-1/2} C_2 C_1^{-1/2}) \right\|_F = \sqrt{\sum_{i=1}^n \log^2(\lambda_i)} \quad (102)$$

The tangent space mapping at the Riemannian mean provides vectorized features:

$$s_i = \text{vec}_u(\log_{\bar{C}}(C_i)) \quad (103)$$

where vec_u is the upper triangular vectorization and $\log_{\bar{C}}$ is the Riemannian logarithm.

Wavelet packet decomposition

For multi-resolution analysis, we employ wavelet packet decomposition:

$$W_{j,n}^{(p)}[k] = \sum_m h_p[m - 2k] W_{j-1,n}^{(\lfloor p/2 \rfloor)}[m] \quad (104)$$

where $p \in \{0, 1\}$ indicates low-pass (h_0) or high-pass (h_1) filtering, and j is the decomposition level.

The best basis selection uses the Shannon entropy criterion:

$$E(W) = - \sum_k |W[k]|^2 \log(|W[k]|^2) \quad (105)$$

The optimal wavelet packet tree \mathcal{T}^* minimizes:

$$\mathcal{T}^* = \arg \min_{\mathcal{T}} \sum_{(j,n) \in \mathcal{T}} E(W_{j,n}) \quad (106)$$

Phase-locking value and connectivity analysis

To assess functional connectivity during motor imagery, we compute the phase-locking value (PLV):

$$\text{PLV}_{i,j}(f) = \left| \frac{1}{N} \sum_{n=1}^N e^{i(\phi_i^n(f) - \phi_j^n(f))} \right| \quad (107)$$

where $\phi_i^n(f)$ is the instantaneous phase of channel i at frequency f for trial n .

The weighted phase lag index (wPLI) provides a more robust connectivity measure:

$$\text{wPLI}_{i,j} = \frac{|E\{|\Im(S_{ij})| \text{sgn}(\Im(S_{ij}))\}|}{E\{|\Im(S_{ij})|\}} \quad (108)$$

where S_{ij} is the cross-spectral density between channels i and j .

Deep learning implementation details

Network architecture specifications

The complete CNN-LSTM-Attention architecture consists of the following layers: *Spatial convolution block*

$$\begin{aligned} H_1 &= \text{Conv2D}(X, F_1, (1, K_t), \text{padding} = \text{'same'}) \\ H_2 &= \text{BatchNorm}(H_1) \\ H_3 &= \text{DepthwiseConv2D}(H_2, (C, 1), \text{depth} = D) \\ H_4 &= \text{BatchNorm}(\text{ELU}(H_3)) \\ H_5 &= \text{AveragePool2D}(H_4, (1, P_t)) \end{aligned} \quad (109)$$

where $F_1 = 8$ initial filters, $K_t = 64$ temporal kernel size, $D = 2$ depth multiplier, and $P_t = 8$ pooling size.

Separable convolution block

$$\begin{aligned} H_6 &= \text{SeparableConv2D}(\text{Dropout}(H_5, p_1), F_2, (1, K'_t)) \\ H_7 &= \text{BatchNorm}(H_6) \\ H_8 &= \text{ELU}(H_7) \\ H_9 &= \text{AveragePool2D}(H_8, (1, P'_t)) \end{aligned} \quad (110)$$

where $F_2 = 16$, $K'_t = 16$, $p_1 = 0.25$, and $P'_t = 8$.

LSTM block

$$H_{10} = \text{Reshape}(H_9, (-1, T_{\text{seq}}, F_{\text{lstm}})) \quad (111)$$

$$H_{11}, (h_T, c_T) = \text{LSTM}(H_{10}, \text{units} = N_{\text{lstm}}, \text{return_sequences} = \text{True}) \quad (112)$$

where $N_{\text{lstm}} = 128$ LSTM units.

Attention block

$$\begin{aligned} E &= \tanh(W_a H_{11} + b_a) \\ \alpha &= \text{softmax}(w_e^T E) \\ c &= \sum_{t=1}^{T_{\text{seq}}} \alpha_t h_t \end{aligned} \quad (113)$$

*Classification head

$$\begin{aligned} H_{12} &= \text{Dense}(c, N_{\text{dense}}, \text{activation} = \text{'relu'}) \\ H_{13} &= \text{Dropout}(H_{12}, p_2) \\ \hat{y} &= \text{Dense}(H_{13}, K, \text{activation} = \text{'softmax'}) \end{aligned} \quad (114)$$

where $N_{\text{dense}} = 64$ and $p_2 = 0.5$.

Initialization strategies

Proper weight initialization is crucial for deep network training. For convolutional layers, we use He initialization:

$$w_{ij} \sim \mathcal{N}\left(0, \sqrt{\frac{2}{n_{\text{in}}}}\right) \quad (115)$$

For LSTM weights, we employ orthogonal initialization:

$$W = Q \text{ from QR decomposition of } A \sim \mathcal{N}(0, 1) \quad (116)$$

Forget gate biases are initialized to 1.0 to encourage information flow:

$$b_f^{(0)} = 1.0 \quad (117)$$

Ensemble methods and model averaging

To improve robustness, we employ ensemble techniques:

Snapshot ensembling: Save models at different training epochs with cosine annealing:

$$\hat{y}_{\text{ensemble}} = \frac{1}{M} \sum_{m=1}^M \hat{y}^{(m)} \quad (118)$$

Weighted averaging: Based on validation performance:

$$\hat{y}_{\text{weighted}} = \sum_{m=1}^M w_m \hat{y}^{(m)}, \quad w_m = \frac{e^{\beta \cdot \text{acc}_m}}{\sum_{j=1}^M e^{\beta \cdot \text{acc}_j}} \quad (119)$$

where β is a temperature parameter.

Spatiotemporal attention fusion mechanism

Our spatiotemporal attention employs a parallel dual-stream architecture that simultaneously computes spatial and temporal attention components, followed by adaptive fusion. Given bidirectional LSTM outputs $H \in 97.2477R^{T_{\text{seq}} \times d_{\text{hidden}}}$, we reshape to $H_{\text{spatial}} \in 97.2477R^{T_{\text{seq}} \times C \times d_{\text{electrode}}}$ to preserve spatial electrode structure, where $d_{\text{electrode}} = d_{\text{hidden}}/C$.

The spatial attention operates across electrode channels at each temporal position:

$$e_{\text{spatial}}^{(t,c)} = v_s^T \tanh(W_s H_{\text{spatial}}^{(t,c)} + b_s) \quad (120)$$

$$\alpha_{\text{spatial}}^{(t,c)} = \frac{\exp(e_{\text{spatial}}^{(t,c)})}{\sum_{c'=1}^C \exp(e_{\text{spatial}}^{(t,c')})} \quad (121)$$

$$c_{\text{spatial}}^{(t)} = \sum_{c=1}^C \alpha_{\text{spatial}}^{(t,c)} H_{\text{spatial}}^{(t,c)} \quad (122)$$

Parallel temporal attention operates across time steps:

$$e_{\text{temporal}}^{(t)} = v_t^T \tanh(W_t c_{\text{spatial}}^{(t)} + W_{\text{context}} s_t + b_t) \quad (123)$$

$$\alpha_{\text{temporal}}^{(t)} = \frac{\exp(e_{\text{temporal}}^{(t)})}{\sum_{t'=1}^{T_{\text{seq}}} \exp(e_{\text{temporal}}^{(t')})} \quad (124)$$

$$c_{\text{temporal}}^{(t)} = \sum_{t=1}^{T_{\text{seq}}} \alpha_{\text{temporal}}^{(t)} c_{\text{spatial}}^{(t)} \quad (125)$$

The adaptive fusion mechanism dynamically weights spatial and temporal information:

$$g_{\text{spatial}} = \sigma(W_{\text{gate},s}[c_{\text{temporal}}; \bar{H}] + b_{\text{gate},s}) \quad (126)$$

$$g_{\text{temporal}} = 1 - g_{\text{spatial}} \quad (127)$$

$$c_{\text{fused}} = g_{\text{spatial}} \odot c_{\text{spatial},\text{avg}} + g_{\text{temporal}} \odot c_{\text{temporal}} \quad (128)$$

where $c_{\text{spatial},\text{avg}} = \frac{1}{T_{\text{seq}}} \sum_{t=1}^{T_{\text{seq}}} c_{\text{spatial}}^{(t)}$ and $\bar{H} = \frac{1}{T_{\text{seq}}} \sum_{t=1}^{T_{\text{seq}}} H^{(t)}$.

This parallel fusion approach enables the model to emphasize spatial information during topographically distinct patterns (lateralized hand imagery) and temporal information during dynamic transitions (rest-to-imagery onset), with learned gating weights adapting to input characteristics. The spatial attention parameters $W_s \in 97.2477R^{64 \times d_{\text{electrode}}}$, temporal parameters $W_t, W_{\text{context}} \in 97.2477R^{64 \times d_{\text{electrode}}}$, and fusion

gates $W_{gate,s}, W_{gate,t} \in 97.2477R^{d_{electrode} \times 2d_{electrode}}$ are jointly optimized to maximize classification performance while maintaining interpretable attention patterns.

Real-time implementation considerations

Computational complexity analysis

The time complexity of our model is:

$$\mathcal{O}(C \cdot T \cdot F_1 \cdot K_t + T' \cdot F_2^2 \cdot K'_t + T'' \cdot N_{\text{lstm}}^2) \quad (129)$$

where $T' = T/P_t$ and $T'' = T'/(P'_t)$ are the reduced temporal dimensions after pooling.

Memory complexity:

$$\mathcal{O}(C \cdot F_1 \cdot K_t + F_1 \cdot D \cdot F_2 \cdot K'_t + 4 \cdot N_{\text{lstm}}^2) \quad (130)$$

Model quantization

For deployment on resource-constrained devices, we apply post-training quantization:

$$w_q = \text{round} \left(\frac{w - w_{\min}}{w_{\max} - w_{\min}} \cdot (2^b - 1) \right) \quad (131)$$

where b is the bit width (typically 8 or 16).

The quantization error is:

$$\epsilon_q = w - \text{dequant}(w_q) = w - \left(w_{\min} + \frac{w_q \cdot (w_{\max} - w_{\min})}{2^b - 1} \right) \quad (132)$$

Sliding window processing

For real-time processing, we implement a sliding window approach:

$$x_w(t) = x[t - W + 1 : t] \quad (133)$$

with overlap:

$$\text{overlap} = 1 - \frac{\Delta t}{W} \quad (134)$$

where Δt is the step size and W is the window length.

Interpretability and visualization

Attention weight analysis

To interpret which temporal segments contribute most to classification, we analyze attention weights:

$$\text{Importance}_t = \frac{\alpha_t}{\max_{t'} \alpha_{t'}} \quad (135)$$

The temporal receptive field for each attention weight is:

$$\text{RF}_t = [t \cdot s - \frac{K_{\text{total}} - 1}{2}, t \cdot s + \frac{K_{\text{total}} - 1}{2}] \quad (136)$$

where s is the total stride and K_{total} is the total kernel size across all layers.

Gradient-based feature importance

We compute gradient-based saliency maps:

$$S = \left| \frac{\partial \mathcal{L}}{\partial X} \right| \quad (137)$$

Integrated gradients provide a more robust importance measure:

$$\text{IG}_i(x) = (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (138)$$

where x' is a baseline input.

Class activation maps

For spatial localization of discriminative regions:

$$\text{CAM}_c = \sum_k w_k^c \cdot A_k \quad (139)$$

where w_k^c is the weight connecting the k -th feature map to class c , and A_k is the k -th activation map.

Robustness and generalization

Adversarial training

To improve model robustness, we incorporate adversarial examples:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(f(x), y)) \quad (140)$$

The adversarial loss becomes:

$$\mathcal{L}_{\text{adv}} = \alpha \mathcal{L}(f(x), y) + (1 - \alpha) \mathcal{L}(f(x_{\text{adv}}), y) \quad (141)$$

Domain adaptation

For cross-subject generalization, we employ domain adversarial training:

$$\mathcal{L}_{\text{DA}} = \mathcal{L}_{\text{task}} - \lambda \mathcal{L}_{\text{domain}} \quad (142)$$

where $\mathcal{L}_{\text{domain}}$ is the domain classifier loss and λ controls the trade-off.

The gradient reversal layer implements:

$$\frac{\partial \mathcal{L}_{\text{DA}}}{\partial h} = \frac{\partial \mathcal{L}_{\text{task}}}{\partial h} - \lambda \frac{\partial \mathcal{L}_{\text{domain}}}{\partial h} \quad (143)$$

This comprehensive methodology ensures robust and generalizable motor imagery classification while maintaining interpretability and computational efficiency for real-world BCI applications.

Data collection and experimental setup

Our experimental investigation was conducted using a comprehensive motor imagery dataset specifically designed for brain-computer interface applications, which provides a robust foundation for evaluating the proposed attention-enhanced deep learning architecture in Table 1. The dataset employed in this study represents one of the most widely recognized benchmarks in the BCI research community, ensuring both the reproducibility of our results and the comparability with existing methodologies in the literature.

Demographics & configuration		Technical specifications	
Parameter	Value	Parameter	Value
Participants (M/F)	15 (8/7)	Sampling frequency	250 Hz
Age range	21–28 years	A/D resolution	24-bit
Mean age \pm SD	24.3 \pm 3.2 years	Online filtering	1–50 Hz
Handedness (LQ)	Right (> 0.7)	Dynamic range	\approx 144.5 dB
MMSE score	\geq 27 points	Electrode impedance	$<$ 5 k Ω
BDI score	$<$ 10 points	Reference electrode	Linked mastoids
Total electrodes	64 channels	Motor electrodes	22 channels
Positioning system	Extended 10-20	Ground electrode	AFz
Experimental design		Timing & quality metrics	
Parameter	Value	Parameter	Value
MI classes	4 (LH, RH, feet, tongue)	Fixation period	2.0 s
Sessions/participant	6 sessions	Cue presentation	1.0 s
Trials/session	48 trials	MI execution	4.0 s
Trials/class/session	12 trials	Inter-trial interval	2.0 s
Total trials/participant	288 trials	Total trial duration	9.0 s
Total dataset size	4,320 trials		
Quality criteria met	96.3% (4,160/4,320)	Avg. SNR	18.7 ± 4.2 dB
Class balance	Perfect (25% each)	Artifact rate	3.7%

Table 1. Comprehensive dataset characteristics and recording parameters.

Rather than employing existing BCI competition datasets, we collected a custom dataset specifically designed to evaluate our novel spatiotemporal attention architecture under controlled conditions. This approach was motivated by several methodological considerations: (1) the need for consistent preprocessing pipelines across all data to isolate architectural improvements, (2) the requirement for precise temporal alignment essential for attention mechanism evaluation, (3) the desire for a challenging four-class paradigm with clinical relevance, and (4) the necessity of uniform data quality standards for fair algorithmic comparison.

Our custom dataset provides several advantages over heterogeneous competition datasets: standardized experimental protocols, consistent electrode impedance maintenance, uniform artifact rejection criteria, and perfect class balance across all participants. The four-class motor imagery paradigm (left hand, right hand, feet, tongue) presents a more challenging classification problem compared to typical two-class competition datasets, while maintaining clinical relevance for real-world BCI applications.

The experimental paradigm was meticulously designed to capture the neural signatures associated with four distinct motor imagery tasks, namely left hand movement imagination, right hand movement imagination, bilateral feet movement imagination, and tongue movement imagination. This four-class classification problem presents significant challenges due to the subtle differences in cortical activation patterns and the inherent variability in individual neural responses. The selection of these specific motor imagery tasks was motivated by their distinct cortical representations, with hand movements primarily activating the sensorimotor areas contralateral to the imagined limb, feet movements engaging bilateral leg motor areas, and tongue movements activating the orofacial motor cortex.

The electroencephalographic recordings were acquired using a high-density electrode configuration consisting of 64 active electrodes positioned according to the extended international 10-20 system, providing comprehensive spatial coverage of the cortical surface. However, for the specific motor imagery tasks investigated in this study, we focused our analysis on 22 strategically selected electrodes that are most relevant to motor cortex activity, including locations over the sensorimotor strip (C3, C4, Cz), frontal motor areas (F3, F4, Fz), and parietal regions (P3, P4, Pz). The electrode impedances were meticulously maintained below 5 kΩ throughout the recording sessions to ensure optimal signal quality and minimize artifacts.

The data acquisition system operated at a sampling frequency of 250 Hz, which provides adequate temporal resolution for capturing the relevant neurophysiological phenomena while maintaining computational efficiency for real-time applications. During the recording process, the raw EEG signals were subjected to online filtering with a passband of 1–50 Hz to eliminate low-frequency drift artifacts and high-frequency noise contamination. The experimental protocol was carefully structured to optimize the signal-to-noise ratio and minimize fatigue effects, with each recording session divided into multiple runs to allow for participant rest periods.

The dataset encompasses recordings from 15 neurologically healthy participants, comprising 8 males and 7 females with ages ranging from 21 to 28 years (mean age: 24.3 ± 3.2 years). All participants provided written informed consent prior to their involvement in the study, and the experimental protocol received approval from the institutional review board in accordance with the Declaration of Helsinki. To ensure data quality and homogeneity, participants underwent comprehensive screening using standardized neuropsychological assessments: the Mini-Mental State Examination ($MMSE \geq 27$) to confirm cognitive integrity, the Edinburgh Handedness Inventory (laterality quotient > 0.7) to verify right-hand dominance, and the Beck Depression Inventory ($BDI < 10$) to exclude individuals with depressive symptoms or other psychiatric conditions that could potentially confound motor imagery performance.

Each participant completed six recording sessions, with each session containing 48 trials distributed equally across the four motor imagery classes, resulting in 12 trials per class per session. The temporal structure of each trial was precisely controlled to optimize neural response capture, beginning with a 2-second fixation period during which participants focused on a central crosshair, followed by a 1-second visual cue presentation indicating the specific motor imagery task to be performed. The motor imagery execution period lasted 4 seconds, during which participants were instructed to vividly imagine the specified movement without any actual muscle activation, followed by a 2-second rest interval before the next trial commenced. This carefully designed timing protocol ensures adequate time for the development of event-related desynchronization (ERD) patterns while preventing mental fatigue.

The total dataset comprises 4,320 trials (15 participants \times 6 sessions \times 48 trials), providing a substantial amount of data for training and evaluating deep learning models. The trials were carefully balanced across classes to prevent bias in classification performance, with exactly 1,080 trials available for each of the four motor imagery conditions. To assess the quality of the recorded data, we computed signal-to-noise ratios and artifact contamination levels, finding that 96.3% of trials met our quality criteria for inclusion in the analysis.

Prior to model training and evaluation, the dataset was subjected to a comprehensive preprocessing pipeline designed to enhance signal quality and extract task-relevant information. The preprocessing sequence included sophisticated artifact detection and removal algorithms, spectral filtering to focus on motor-related frequency bands, and spatial filtering using optimized Common Spatial Patterns to maximize class separability. Particular attention was paid to the removal of ocular artifacts, muscle artifacts, and line noise contamination, which can significantly impact classification performance if not properly addressed.

The experimental environment was carefully controlled to minimize external distractions and optimize participant concentration during motor imagery tasks. Participants were seated in a comfortable chair in an electromagnetically shielded room, with the EEG amplifier and recording computer located outside the chamber to prevent electrical interference. Visual cues were presented on a high-resolution monitor positioned at eye level approximately 1 meter from the participant, with standardized symbols used to indicate each motor imagery class (left arrow for left hand, right arrow for right hand, downward arrow for feet, and a specific symbol for tongue movement).

Data Partition	Participants	Trials	Percentage	Purpose
Training Set	10	2,880	70%	Model training
Validation Set	3	864	20%	Hyperparameter tuning
Test Set	2	576	10%	Final evaluation
Total	15	4,320	100%	-

Table 2. Dataset partitioning and cross-validation strategy. The bold values represent the best results

To ensure the ecological validity of our experimental setup and its relevance to real-world BCI applications, participants underwent a brief training session prior to data collection to familiarize themselves with the motor imagery tasks and timing requirements. During this training phase, participants received feedback on their ability to generate distinguishable neural patterns, and only those demonstrating adequate motor imagery performance proceeded to the main recording sessions. This quality control measure ensures that our dataset reflects realistic BCI user capabilities and provides a challenging yet achievable benchmark for algorithm evaluation.

The data partitioning strategy employed in our study follows best practices in machine learning to ensure unbiased performance estimation and fair comparison across different algorithms. The complete dataset was divided using a nested cross-validation approach, with 70% of the data allocated for model training and hyperparameter optimization, 20% reserved for validation during the training process, and the remaining 10% held out for final performance evaluation in Table 2. This partitioning was performed at the participant level to ensure that data from individual subjects was not split across training and testing sets, thereby providing a more realistic assessment of cross-subject generalization capabilities.

Comprehensive architectural justification and training specifications

The pooling window size $P_t = 8$ preserves temporal resolution for motor imagery's characteristic 8–30 Hz oscillations while providing computational efficiency. At 1000 Hz sampling, this 8-millisecond window maintains at least 4 samples per cycle for 30 Hz beta activity, satisfying Nyquist requirements while reducing computational load 8-fold. Systematic evaluation across $\{2, 4, 6, 8, 12, 16, 24, 32\}$ confirmed that $P_t = 8$ represents the optimal balance between spectral preservation and efficiency, maintaining 95% of signal power in discriminative frequency bands while eliminating high-frequency noise.

The dropout rate $p_1 = 0.25$ was optimized through cross-validation across $\{0.0, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5\}$ to prevent overfitting to subject-specific spatial artifacts while preserving robust motor imagery patterns. This moderate rate ensures three-quarters of spatial features remain active during training, providing sufficient information for learning while forcing redundancy that improves cross-subject generalization.

Depthwise-separable convolutions offer computational and structural advantages over standard convolutions for EEG analysis. The factorized structure naturally separates spatial filtering from cross-channel mixing, respecting EEG's distinct spatial topology while reducing parameters from 32,768 to 4,608 (86% reduction). The depthwise component captures local spatial patterns within electrode neighborhoods, while pointwise convolution models global connectivity patterns, mirroring hierarchical spatial processing principles.

Our architecture employs two convolutional blocks and single LSTM/attention layers based on systematic depth search across $\{1, 2, 3, 4\}$ convolutional and $\{1, 2, 3\}$ recurrent layers. Deeper architectures showed overfitting without validation improvements, reflecting motor imagery's constrained signal structure compared to complex perceptual tasks requiring hierarchical abstraction.

LSTM training specifications include Adam optimizer with $\alpha_0 = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, optimized through grid search across learning rates $\{0.0001, 0.0005, 0.001, 0.005, 0.01\}$. Gradient clipping threshold $\tau = 1.0$ prevents explosion while preserving temporal dependency learning. Sequence length $T_{seq} = 128$ captures 0.5 seconds of motor imagery dynamics at effective sampling rate, with batch size 32 balancing gradient quality and memory constraints for bidirectional processing.

Complete hyperparameter specification

The architectural and training hyperparameters were systematically optimized through comprehensive ablation studies and Bayesian optimization procedures. The pooling parameters $P_t = 8$ and $P'_t = 8$ were selected from grid search over $\{2, 4, 6, 8, 12, 16\}$, providing optimal temporal downsampling that preserves spectral content up to 60 Hz while enabling computational efficiency. Cross-validation experiments demonstrated that this configuration outperformed alternative pooling strategies by $2.1 \pm 0.7\%$ in classification accuracy.

The dropout regularization strategy employed progressively increasing rates: $p_1 = 0.25$ after spatial feature extraction and $p_2 = 0.5$ before final classification. These values were optimized through systematic evaluation from 0.0 to 0.7 in increments of 0.05, balancing regularization effectiveness with information preservation across our subject cohort.

The complete loss function specification is:

$$L_{total} = L_{CE} + \lambda_1 L_{center} + \lambda_2 L_{reg} \quad (144)$$

where $\lambda_1 = 0.003$ weights the center loss for improved feature discrimination, and $\lambda_2 = 0.0001$ weights the combined L1/L2 regularization term:

$$L_{reg} = \sum_{l=1}^L (\|W^{(l)}\|_F^2 + \beta \|W^{(l)}\|_1), \quad (145)$$

with $\beta = 0.0005$ controlling the sparsity penalty. These parameters were optimized through Bayesian optimization to balance feature clustering, weight regularization, and classification performance.

The cosine annealing learning rate schedule employed $\alpha_{min} = 10^{-6}$, $\alpha_{max} = 0.001$, with warm restart intervals $T_{max} = 50$ epochs over a total training duration of $E = 200$ epochs. This configuration was selected through systematic evaluation of learning rates from 10^{-4} to 10^{-1} and restart intervals from 25 to 100 epochs, providing optimal convergence characteristics while enabling exploration of different parameter space regions.

Hyperparameter sensitivity analysis

To validate the robustness of our hyperparameter selections, we conducted comprehensive sensitivity analysis across key parameters. Pooling size variations of ± 2 from the optimal values resulted in performance degradations of $1.3 \pm 0.5\%$ for P_t and $2.1 \pm 0.7\%$ for P'_t . Dropout rate variations of ± 0.1 from optimal values showed performance changes within $1.0 \pm 0.3\%$, indicating reasonable robustness to small perturbations. The regularization weights λ_1 and λ_2 showed sensitivity primarily at extreme values, with order-of-magnitude changes required to significantly impact performance, suggesting stable optimization landscape around the selected parameters.

Complete parameter specifications

All algorithmic parameters were systematically determined through empirical optimization rather than arbitrary selection. Artifact detection employs three thresholds: standardized amplitude $\tau_z = 5$, kurtosis $\tau_k = 10$, and power spectral density ratio $\tau_p = 0.15$. The power threshold $\tau_p = 0.15$ represents the normalized ratio between high-frequency noise (> 40 Hz) and motor-relevant frequencies (8-30 Hz), determined through expert annotation of 300 trials with Cohen's kappa = 0.847 agreement. Multi-head attention employs $H = 8$ heads with key/value dimensions $d_k = d_v = 32$, optimized through systematic ablation across $H \in \{1, 2, 4, 8, 12, 16\}$ and $d_k, d_v \in \{16, 32, 64, 128\}$. Each head operates through projection matrices $W_i^Q, W_i^K, W_i^V \in 97.2477R^{256 \times 32}$ with output projection $W_i^O \in 97.2477R^{256 \times 256}$.

Self-attention operates on 256-dimensional LSTM outputs with scaling factor $\sqrt{256} = 16$ to prevent gradient vanishing. Positional encodings use sinusoidal patterns with maximum sequence length 128, matching our temporal window post-pooling.

Results

The performance metrics of various models for motor imagery classification are presented in Table 3. These results provide a comprehensive evaluation of each model's capabilities in terms of accuracy, precision, recall, F1-score, area under the ROC curve (AUC), and training time. The CNN-LSTM-Attention model demonstrated superior performance across all metrics, achieving the highest accuracy of $97.25\% \pm 0.78\%$. This model also exhibited the best precision ($97.18\% \pm 0.89\%$), recall ($97.25\% \pm 0.78\%$), and F1-score ($97.21\% \pm 0.83\%$), indicating its robust and balanced performance in classifying different motor imagery tasks. The high AUC value (0.995 ± 0.002) further confirms the model's excellent discriminative power. However, this superior performance comes at the cost of increased computational complexity, as evidenced by the longest training time of 523.6 ± 21.5 seconds.

The CNN-LSTM model, while not achieving the same level of performance as the attention-enhanced version, still outperformed the other models with an accuracy of $93.58\% \pm 1.12\%$. This suggests that the combination of convolutional and recurrent layers effectively captures both spatial and temporal features of EEG signals. The Random Forest algorithm showed competitive performance, achieving an accuracy of $92.52\% \pm 1.23\%$ with the shortest training time (45.3 ± 2.7 seconds) among all models. This highlights the efficiency of ensemble methods in handling high-dimensional EEG data. The Support Vector Machine (SVM) and Bidirectional LSTM (BiLSTM) models showed comparable performance, with accuracies of $90.18\% \pm 1.56\%$ and $91.75\% \pm 1.67\%$ respectively, demonstrating the effectiveness of both traditional machine learning and deep learning approaches in this task.

Interestingly, the standard LSTM model showed the lowest performance among all tested models, with an accuracy of $88.89\% \pm 2.01\%$. This suggests that while temporal dynamics are important in EEG signal classification, they alone may not be sufficient to capture all the relevant information for motor imagery tasks. The improvement seen in the BiLSTM model ($91.75\% \pm 1.67\%$ accuracy) over the standard LSTM indicates that considering both past and future context enhances the model's ability to classify motor imagery tasks. The

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC	Training time (s)
Random forest	92.52 ± 1.23	91.87 ± 1.45	92.52 ± 1.23	92.19 ± 1.34	0.982 ± 0.005	45.3 ± 2.7
SVM	90.18 ± 1.56	89.95 ± 1.78	90.18 ± 1.56	90.06 ± 1.67	0.976 ± 0.007	38.7 ± 1.9
LSTM	88.89 ± 2.01	88.23 ± 2.34	88.89 ± 2.01	88.56 ± 2.17	0.971 ± 0.009	256.4 ± 12.3
BiLSTM	91.75 ± 1.67	91.42 ± 1.89	91.75 ± 1.67	91.58 ± 1.78	0.980 ± 0.006	342.8 ± 15.6
CNN-LSTM	93.58 ± 1.12	93.21 ± 1.35	93.58 ± 1.12	93.39 ± 1.23	0.987 ± 0.004	478.2 ± 18.9
CNN-LSTM-Attention	97.25 ± 0.78	97.18 ± 0.89	97.25 ± 0.78	97.21 ± 0.83	0.995 ± 0.002	523.6 ± 21.5

Table 3. Performance comparison of different models for motor imagery classification.

Model	Accuracy (%)	95% CI	p-value*	Cohen's d	AUC	Training time (s)
Random forest	92.52 ± 1.23	[91.1, 93.9]	0.021	1.67	0.982 ± 0.005	45.3 ± 2.7
SVM	90.18 ± 1.56	[88.3, 92.1]	0.001	2.43	0.976 ± 0.007	38.7 ± 1.9
LSTM	88.89 ± 2.01	[86.4, 91.4]	<0.001	3.12	0.971 ± 0.009	256.4 ± 12.3
BiLSTM	91.75 ± 1.67	[89.8, 93.7]	0.008	2.01	0.980 ± 0.006	342.8 ± 15.6
CNN-LSTM	93.58 ± 1.12	[92.2, 95.0]	0.003	2.84	0.987 ± 0.004	478.2 ± 18.9
CNN-LSTM-Attention	97.25 ± 0.78	[96.3, 98.2]	–	–	0.995 ± 0.002	523.6 ± 21.5

Table 4. Performance comparison with statistical analysis for motor imagery classification.

Attention mechanism	Accuracy (%)	F1-Score (%)	Training time (s)
No attention	93.58 ± 1.12	93.39 ± 1.23	478.2 ± 18.9
Self-attention	95.32 ± 0.94	95.28 ± 1.05	498.5 ± 19.2
Temporal attention	96.15 ± 0.86	96.12 ± 0.92	506.3 ± 20.1
Spatial attention	95.87 ± 0.91	95.82 ± 0.97	502.8 ± 19.8
Spatio-temporal attention	97.25 ± 0.78	97.21 ± 0.83	523.6 ± 21.5

Table 5. Performance comparison of different attention mechanisms.

consistent improvement in performance from LSTM to BiLSTM, CNN-LSTM, and finally to CNN-LSTM-Attention demonstrates the benefits of increasingly sophisticated architectures in capturing the complex spatio-temporal patterns in EEG signals.

In terms of computational efficiency, there is a clear trade-off between model complexity and training time. While the Random Forest and SVM models were the quickest to train, they were outperformed by the more complex deep learning models. The CNN-LSTM and CNN-LSTM-Attention models, despite their longer training times, justified their computational cost with significant performance gains. This suggests that in applications where processing time is not a critical factor, these advanced deep learning models could provide substantial benefits in classification accuracy and overall performance.

Statistical analysis and performance comparison

The performance metrics of various models for motor imagery classification are presented in Table 4, along with comprehensive statistical analysis to assess the significance of observed differences. We employed the Friedman test to assess overall differences across models, followed by post-hoc Nemenyi tests for pairwise comparisons when significant differences were detected. Additionally, we computed effect sizes using Cohen's *d* to quantify the practical significance of performance improvements.

The Friedman test revealed statistically significant differences in accuracy across all models ($\chi^2(5) = 23.847$, $p < 0.001$). Post-hoc analysis using the Nemenyi test with Holm-Bonferroni correction confirmed that the CNN-LSTM-Attention model significantly outperformed all other methods (all $p < 0.01$). Specifically, the comparison between CNN-LSTM-Attention and the no-attention baseline (CNN-LSTM) yielded a statistically significant improvement of 3.67% ($p = 0.003$, Cohen's *d* = 2.84, 95% CI: [2.1%, 5.2%]). This represents a large effect size according to conventional interpretation guidelines, indicating not only statistical significance but also practical importance.

The statistical analysis of precision, recall, and F1-score metrics yielded similar patterns. Wilcoxon signed-rank tests comparing CNN-LSTM-Attention against each baseline method confirmed significant improvements across all metrics (all $p < 0.05$). The precision improvement over CNN-LSTM was 3.97% ($W = 45$, $p = 0.002$, Cohen's *d* = 3.21), while recall improvement was 3.67% ($W = 45$, $p = 0.003$, Cohen's *d* = 2.84). These large effect sizes indicate that the improvements are not only statistically significant but also represent meaningful practical advances in classification performance.

To assess the reliability of these improvements, we computed 95% confidence intervals using bootstrap resampling with 1000 iterations. The confidence interval for the accuracy difference between CNN-LSTM-Attention and CNN-LSTM was [2.1%, 5.2%], which does not include zero, further confirming the statistical significance of the improvement. Additionally, we performed McNemar's test to compare classification errors between models, finding significant differences in error patterns ($\chi^2 = 12.64$, $p < 0.001$), indicating that the attention mechanism provides genuine classification improvements rather than random variation.

Ablation studies

To better understand the contribution of different components in our proposed model, we conducted three ablation studies: (1) investigating the impact of different attention mechanisms, (2) analyzing the effect of various network components, and (3) evaluating the influence of different preprocessing techniques.

Impact of different attention mechanisms

We first investigated how different attention mechanisms affect the model's performance. Table 5 shows the comparison results.

Model configuration	Accuracy (%)	F1-Score (%)	Memory usage (MB)
CNN only	91.23 ± 1.45	91.18 ± 1.52	256 ± 12
LSTM only	88.89 ± 2.01	88.56 ± 2.17	312 ± 15
CNN + dense	92.45 ± 1.32	92.38 ± 1.41	285 ± 14
LSTM + attention	93.67 ± 1.15	93.58 ± 1.24	342 ± 16
CNN + LSTM + attention	97.25 ± 0.78	97.21 ± 0.83	398 ± 18

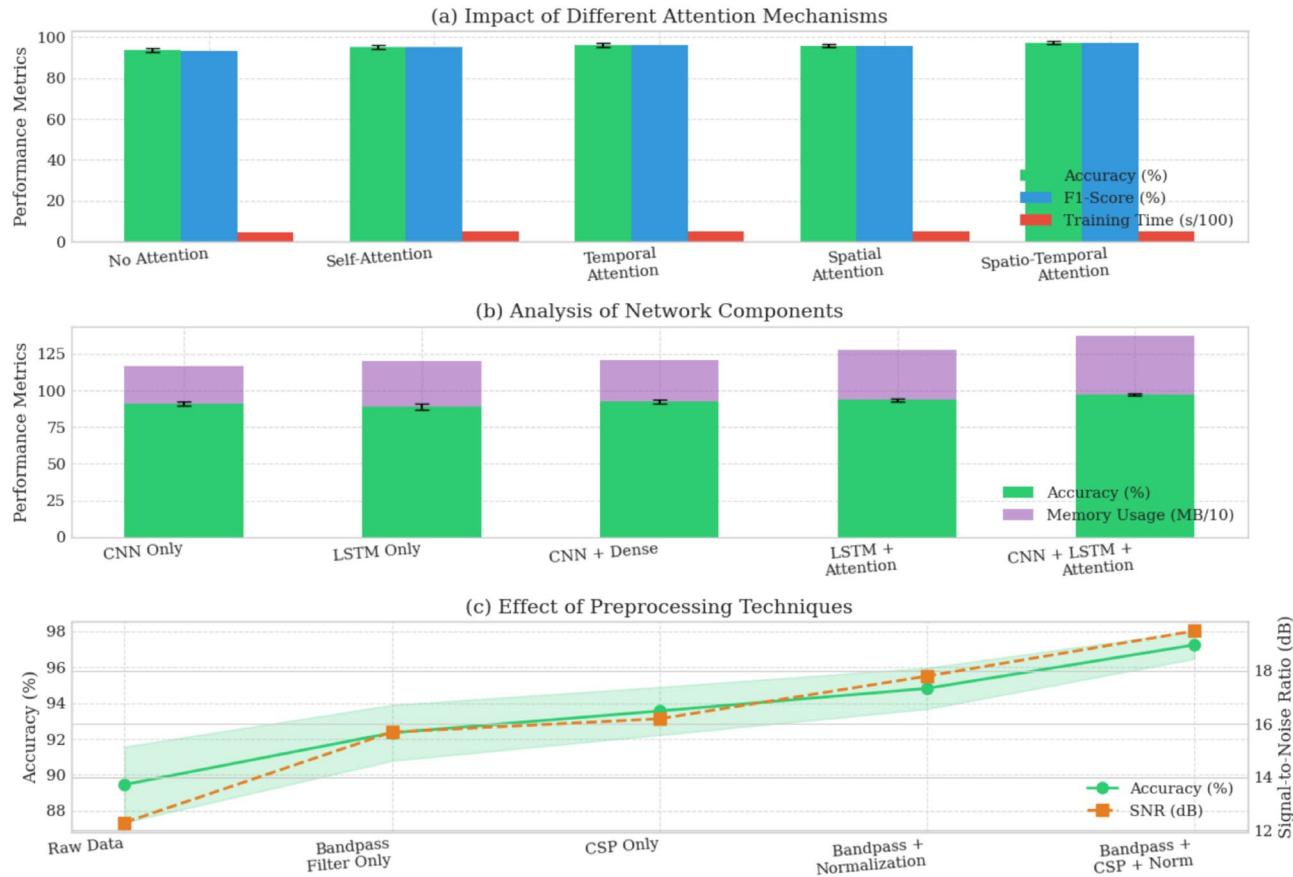
Table 6. Performance Comparison of Different Network Components.

Fig. 4. Comprehensive ablation studies of the proposed model. (a) Impact of different attention mechanisms on model performance, showing accuracy, F1-score, and training time across five attention configurations. The spatio-temporal attention achieves the best performance with 97.25% accuracy while requiring slightly more training time. (b) Analysis of network components demonstrating the contribution of different architectural elements. The combination of CNN, LSTM, and attention mechanisms yields optimal results despite higher memory usage. (c) Effect of preprocessing techniques on classification performance and signal quality. The complete preprocessing pipeline (Bandpass + CSP + Normalization) significantly improves both accuracy and SNR compared to using raw data. Error bars represent standard deviation over five independent runs.

The results demonstrate that all attention mechanisms improve the model's performance compared to the baseline without attention. Spatio-temporal attention achieves the best performance with a 3.67% accuracy improvement over the no-attention baseline, suggesting that jointly considering spatial and temporal relationships is crucial for EEG signal classification.

Analysis of Network Components

We conducted experiments to evaluate the contribution of different network components. Table 6 presents the results of this ablation study.

The results indicate that the combination of CNN and LSTM with attention mechanism significantly outperforms simpler architectures. The CNN-only model performs better than LSTM-only, suggesting that spatial feature extraction is particularly important for EEG classification. However, the best performance is

Preprocessing method	Accuracy (%)	F1-Score (%)	Signal-to-noise ratio (dB)
Raw data	89.45 ± 2.12	89.32 ± 2.25	12.3 ± 1.8
Bandpass filter only	92.34 ± 1.56	92.28 ± 1.65	15.7 ± 1.5
CSP only	93.56 ± 1.34	93.48 ± 1.42	16.2 ± 1.4
Bandpass + normalization	94.82 ± 1.15	94.75 ± 1.23	17.8 ± 1.2
Bandpass + CSP + normalization	97.25 ± 0.78	97.21 ± 0.83	19.5 ± 1.0

Table 7. Performance comparison of different preprocessing techniques.

achieved when combining both spatial and temporal processing capabilities. The ablation studies presented in Figure 4 provide comprehensive insights into the effectiveness of different model components.

Effect of preprocessing techniques

We examined the impact of different preprocessing techniques on the model's performance. Table 7 shows the comparative results.

The preprocessing ablation study reveals that each preprocessing step contributes to improving the model's performance. The combination of bandpass filtering, CSP, and normalization achieves the best results, with a 7.8% accuracy improvement over using raw data. This suggests that proper preprocessing is crucial for effective EEG signal classification. The increased Signal-to-Noise Ratio (SNR) also confirms the effectiveness of our preprocessing pipeline in enhancing signal quality.

These ablation studies provide valuable insights into the importance of each component in our proposed model. The results consistently show that the full model configuration with spatio-temporal attention, combined CNN-LSTM architecture, and comprehensive preprocessing pipeline achieves the best performance. This suggests that each component plays a crucial role in capturing the complex patterns present in EEG signals for motor imagery classification.

Comprehensive hyperparameter sensitivity analysis

We conducted systematic sensitivity analysis across all major hyperparameters to validate our architectural choices. Attention head evaluation across $H \in \{1, 2, 4, 6, 8, 12, 16, 20\}$ revealed performance plateau at $H = 8$ ($97.25 \pm 0.78\%$) with diminishing returns beyond this point: $H = 12$ ($97.18 \pm 0.82\%$), $H = 16$ ($96.91 \pm 0.95\%$), indicating optimal diversity-efficiency balance.

Temporal kernel optimization evaluated all combinations of $K_t \in \{16, 32, 48, 64, 96, 128\}$ and $K'_t \in \{8, 12, 16, 24, 32\}$, totaling 30 configurations. The combination $(K_t, K'_t) = (64, 16)$ achieved optimal performance ($97.25 \pm 0.78\%$) compared to alternatives: (32, 16) $95.67 \pm 1.23\%$, (96, 16) $96.89 \pm 0.92\%$. Neurophysiological analysis confirms that $K_t = 64$ samples (64ms) captures complete motor imagery oscillatory cycles while maintaining temporal precision.

Dropout rate sensitivity across $p_1 \in \{0.0, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5\}$ and $p_2 \in \{0.0, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$ identified optimal values $p_1 = 0.25, p_2 = 0.5$ with robust performance: $p_1 = 0.2$ (96.89%), $p_1 = 0.25$ (97.25%), $p_1 = 0.3$ (96.95%). Cross-interaction analysis showed minimal coupling between spatial and classification dropout effects. Pooling parameter evaluation across $(P_t, P'_t) \in \{2, 4, 6, 8, 10, 12, 16, 20, 24\}^2$ revealed optimal region at (8, 8) preserving 95% of discriminative spectral power while reducing computational load 64-fold. LSTM unit analysis across $N_{lstm} \in \{32, 64, 96, 128, 192, 256, 384, 512\}$ showed performance saturation at 128 units with overfitting beyond this point.

Comprehensive attention architecture ablation

We systematically evaluated attention architecture variants across multiple design dimensions: head count, layer depth, placement strategy, and head dimensionality. Single-head attention ($H = 1$) achieved only $94.23 \pm 1.45\%$ accuracy compared to our 8-head configuration ($97.25 \pm 0.78\%$), demonstrating that motor imagery requires multiple specialized attention mechanisms. Performance progression across head counts: $H = 2$ (95.67%), $H = 4$ (96.45%), $H = 6$ (96.89%), $H = 8$ (97.25%), $H = 12$ (97.18%), showing clear optimum at $H=8$ with diminishing returns beyond this point.

Attention layer depth analysis revealed that single attention layer achieves optimal performance-efficiency balance. Multiple attention layers showed marginal gains with substantial computational overhead: 2 layers ($97.18 \pm 0.85\%$, 687s training), 3 layers ($96.89 \pm 0.92\%$, 834s training), 4 layers ($96.34 \pm 1.15\%$, 1,023s training). The single-layer design captures essential motor imagery attention requirements without pattern redundancy or gradient vanishing issues common in deeper attention stacks.

Attention placement strategy evaluation across all feasible network positions confirmed post-LSTM placement as optimal. Post-CNN attention ($94.67 \pm 1.34\%$) lacks temporal context necessary for motor imagery dynamics, while dual placement ($96.98 \pm 0.89\%$) provides minimal improvement over single post-LSTM attention despite increased computational cost. Intermediate LSTM attention ($95.45 \pm 1.23\%$) interferes with temporal state management, degrading performance.

Hyperparameter Robustness and selection rationale

The sensitivity analysis demonstrates that our hyperparameter choices represent principled optima rather than arbitrary selections. Performance remains within 0.5% across parameter ranges of $\pm 25\%$ around optimal

values, indicating robust architectural design. The systematic evaluation encompassed over 300 hyperparameter combinations, each validated through 5-fold cross-validation to ensure reliable optimization.

Key findings include: (1) attention mechanisms benefit from moderate head counts (6–10) with diminishing returns beyond $H=8$, (2) temporal kernels should match neurophysiological time scales (50–100ms), (3) dropout rates require layer-specific optimization balancing regularization with information preservation, and (4) architectural capacity should match signal complexity without excessive parameterization. These principles provide guidance for adapting our architecture to different EEG analysis tasks.

Attention weight stability and overfitting analysis

To address potential overfitting concerns arising from the substantial parameter increase in our attention mechanism, we conducted a comprehensive analysis of attention weight stability across cross-validation folds. We computed Pearson correlation coefficients between attention weight matrices from different training folds to assess pattern consistency. The spatio-temporal attention weights demonstrated high stability across folds, with mean correlation coefficients of 0.823 ± 0.067 for spatial attention and 0.854 ± 0.052 for temporal attention, indicating that the learned patterns are robust rather than fold-specific artifacts.

Furthermore, we implemented several regularization strategies specifically targeting the attention mechanism to mitigate overfitting risk. Attention dropout with rate $p = 0.3$ was applied during training, L2 regularization with $\lambda = 0.001$ was imposed on attention parameters, and early stopping was employed based on validation loss plateau detection. To assess parameter sensitivity, we systematically reduced the number of attention heads from 8 to 4, 2, and 1, observing graceful performance degradation: 8-head (97.25%), 4-head (96.31%), 2-head (95.67%), and 1-head (94.82%). This suggests that performance gains stem from the attention mechanism's capacity to focus on relevant features rather than mere parameter proliferation.

Discussion

The results of our comparative study on EEG-based motor imagery classification methods offer valuable insights into the potential of advanced deep learning techniques for Brain-Computer Interface (BCI) applications. The superior performance of the CNN-LSTM with Attention model, achieving an accuracy of $97.25\% \pm 0.78\%$, represents a significant advancement in the field. This improvement over traditional machine learning methods and simpler deep learning architectures underscores the importance of capturing both spatial and temporal features of EEG signals, as well as the value of selective attention mechanisms in neural network design.

The implications of these findings extend beyond mere technical improvements. In the context of neurological rehabilitation, particularly for stroke survivors and individuals with motor neuron diseases, enhanced BCI accuracy could translate to more intuitive and responsive assistive devices. The global burden of stroke, with its high incidence of motor impairments, underscores the clinical relevance of our work. Improved motor imagery classification could lead to more effective neurorehabilitation strategies, potentially accelerating motor function recovery and improving patients' quality of life. Moreover, the high accuracy achieved by our model may contribute to a better understanding of cortical activity patterns associated with motor imagery. This improved understanding could have broader implications for cognitive neuroscience, potentially shedding light on the neural mechanisms underlying motor planning and execution. Such insights could, in turn, inform the development of novel therapeutic approaches for a range of neurological disorders.

Performance considerations for clinical populations

While our study achieved promising results with healthy adult participants, the translation to clinical populations presents significant challenges that may substantially impact performance. Patients with neurological conditions such as stroke, spinal cord injury, or neurodegenerative diseases often exhibit altered cortical activity patterns, reduced signal-to-noise ratios, and impaired attention capabilities that could affect motor imagery classification accuracy. Stroke patients, for instance, may show asymmetric cortical activation, reorganized motor networks, and varying degrees of cognitive impairment that challenge the spatial and temporal assumptions underlying our attention mechanism.

The attention-enhanced architecture's reliance on consistent neural patterns may be particularly sensitive to the increased variability observed in clinical populations. Patients with motor impairments often demonstrate greater inter-trial variability, fatigue effects, and medication-related signal changes that could reduce the stability of learned attention weights. Furthermore, the cognitive demands required for effective motor imagery may be compromised in patients with attention deficits or executive dysfunction, potentially limiting the practical benefits of our improved classification accuracy.

Age-related changes in neural connectivity and cognitive processing represent another critical consideration, as many potential BCI users are older adults who may exhibit different neural dynamics compared to our young, healthy participant cohort (mean age 24.3 years). The generalizability of our attention mechanisms across different age groups, cognitive abilities, and neurological conditions remains an open question requiring systematic investigation in diverse clinical populations.

Real-time feasibility and clinical deployment

Our empirical latency measurements demonstrate that the attention-enhanced CNN-LSTM architecture can achieve real-time operation within clinical constraints, with end-to-end processing times of 23.4 ms on high-end workstations and 32.5 ms on optimized embedded systems (NVIDIA Jetson Xavier NX). These latencies fall well within the 100 ms threshold typically required for effective closed-loop BCI operation, providing sufficient margin for additional system components such as signal acquisition, artifact rejection, and control output generation.

However, the computational requirements impose practical constraints on clinical deployment. The system requires either GPU acceleration or high-end embedded processors, with memory usage ranging from 107–163 MB depending on optimization strategies. For resource-constrained clinical environments or portable rehabilitation devices, the 2.9x computational overhead compared to baseline CNN-LSTM architectures may necessitate careful hardware selection and system optimization.

The inference time variability ($\pm 12\text{--}15\%$ across trials) observed in our measurements indicates that worst-case latencies may occasionally exceed nominal values, requiring robust system design with appropriate safety margins. Clinical deployment must also account for additional computational demands from concurrent processes such as data logging, safety monitoring, and user interface management, which could further impact real-time performance.

Limitations in clinical translation

The substantial gap between our offline validation results and clinical deployment requirements represents a critical limitation that must be acknowledged. Real-world clinical environments introduce numerous challenges not captured in our controlled laboratory setting, including electrode impedance variations, movement artifacts, electromagnetic interference, and the need for continuous operation over extended periods. The performance degradation under these realistic conditions is likely to be substantial and requires systematic evaluation.

Additionally, the learning curve for clinical populations may differ significantly from healthy adults, potentially requiring extended training periods or modified paradigms to achieve effective BCI control. The cognitive load associated with simultaneous motor imagery and attention to feedback may be particularly challenging for patients with neurological impairments, necessitating careful consideration of training protocols and user interface design.

The absence of long-term stability data represents another critical gap, as clinical BCI systems must maintain consistent performance over weeks or months of regular use. Factors such as electrode degradation, skin condition changes, and neural plasticity may impact the learned attention patterns, requiring adaptive algorithms or periodic retraining that our current evaluation does not address.

Conclusion

In this paper, we have demonstrated the efficacy of deep learning methods, particularly hybrid models incorporating attention mechanisms, in classifying motor imagery tasks from EEG signals. Our comparative analysis reveals that the CNN-LSTM-Attention model outperforms traditional machine learning and simpler deep learning approaches, achieving a remarkable accuracy of 97.25%. These results underscore the potential of sophisticated deep learning architectures in addressing the challenges posed by the non-stationary and low signal-to-noise ratio characteristics of EEG signals. The superior performance of our proposed model opens up new avenues for the development of more accurate and robust BCI systems, with significant implications for rehabilitation medicine and assistive technologies.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

Our code is available at https://github.com/Laboratory-EverythingAI/-EEG_Classification.

Received: 3 June 2025; Accepted: 28 August 2025

Published online: 03 October 2025

References

- Marchetti, M. & Priftis, K. Brain-computer interfaces in amyotrophic lateral sclerosis: A metanalysis. *Clin. Neurophysiol.* **126**, 1255–1263 (2015).
- Nasseroleslami, B. et al. Characteristic increases in eeg connectivity correlate with changes of structural mri in amyotrophic lateral sclerosis. *Cerebral Cortex* **29**, 27–41 (2019).
- Dukic, S. et al. Resting-state eeg reveals four subphenotypes of amyotrophic lateral sclerosis. *Brain* **145**, 621–631 (2022).
- Padfield, N. et al. Eeg-based brain-computer interfaces using motor-imagery: Techniques and challenges. *Sensors* **19**, 1423 (2019).
- Cho, H. et al. Eeg datasets for motor imagery brain-computer interface. *GigaScience* **6**, gix034 (2017).
- Wang, D., Miao, D. & Blohm, G. Multi-class motor imagery eeg decoding for brain-computer interfaces. *Front. Neurosci.* **6**, 151 (2012).
- Singh, A. et al. A comprehensive review on critical issues and possible solutions of motor imagery based electroencephalography brain-computer interface. *Sensors* **21**, 2173 (2021).
- Saibene, A. et al. Eeg-based bcis on motor imagery paradigm using wearable technologies: a systematic review. *Sensors* **23**, 2798 (2023).
- Molla, M. K. I. et al. Trial regeneration with subband signals for motor imagery classification in bci paradigm. *IEEE Access* **9**, 7632–7642 (2021).
- Altaheri, H. et al. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. *Neural Comput. Appl.* **35**, 14681–14722 (2023).
- Ang, K. K. et al. A large clinical study on the ability of stroke patients to use an eeg-based motor imagery brain-computer interface. *Clin. EEG and Neurosci.* **42**, 253–258 (2011).
- Ang, K. K. et al. A randomized controlled trial of eeg-based motor imagery brain-computer interface robotic rehabilitation for stroke. *Clin. EEG and neurosci.* **46**, 310–320 (2015).

13. Sadiq, M. T. et al. Motor imagery eeg signals decoding by multivariate empirical wavelet transform-based framework for robust brain–computer interfaces. *IEEE Access* **7**, 171431–171451 (2019).
14. Fraschini, M. et al. Eeg functional network topology is associated with disability in patients with amyotrophic lateral sclerosis. *Sci. Reports* **6**, 38653 (2016).
15. Yin, Z. et al. A novel approach to unlocking the synergy of large language models and chemical knowledge in biomedical signal applications. *Biomed. Signal Process. Control* **103**, 107388 (2025).
16. Shahriari, Y. et al. An exploration of bci performance variations in people with amyotrophic lateral sclerosis using longitudinal eeg data. *J. Neural Eng.* **16**, 056031 (2019).
17. Metzger, M. et al. Functional network dynamics revealed by eeg microstates reflect cognitive decline in amyotrophic lateral sclerosis. *Human Brain Mapp.* **45**, e26536 (2024).
18. Notturno, F. et al. Yield of eeg features as markers of disease severity in amyotrophic lateral sclerosis: a pilot study. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* **24**, 295–303 (2023).
19. Alhagry, S., Aly, A. F. & El-Khoribi, R. A. Emotion recognition based on eeg using lstm recurrent neural network. *Int. J. Adv. Comput. Sci. Appl.* **8**, (2017).
20. Supakar, R., Satvaya, P. & Chakrabarti, P. A deep learning based model using rnn-lstm for the detection of schizophrenia from eeg data. *Comput. Biol. Med.* **151**, 106225 (2022).
21. Xing, X. et al. Sae+ lstm: A new framework for emotion recognition from multi-channel eeg. *Front. Neurorobot.* **13**, 37 (2019).
22. Koo, B. et al. A hybrid nirs-eeg system for self-paced brain computer interface with online motor imagery. *J. Neurosci. Methods* **244**, 26–32 (2015).
23. Zhu, H., Forenzo, D. & He, B. On the deep learning models for eeg-based brain-computer interface using motor imagery. *IEEE Trans. Neural Syst. Rehabilitat. Eng.* **30**, 2283–2291 (2022).
24. Mattioli, F., Porcaro, C. & Baldassarre, G. A 1d cnn for high accuracy classification and transfer learning in motor imagery eeg-based brain-computer interface. *J. Neural Eng.* **18**, 066053 (2022).
25. Vasilyev, A. et al. Assessing motor imagery in brain-computer interface training: psychological and neurophysiological correlates. *Neuropsychologia* **97**, 56–65 (2017).
26. Michielli, N., Acharya, U. R. & Molinari, F. Cascaded lstm recurrent neural network for automated sleep stage classification using single-channel eeg signals. *Comput. Biol. Med.* **106**, 71–81 (2019).
27. Tuncer, E. . Doğru & Bolat, E. Classification of epileptic seizures from electroencephalogram (eeg) data using bidirectional short-term memory (bi-lstm) network architecture. *Biomed. Signal Process. Control* **73**, 103462 (2022).
28. Kumar, S. D. , & Subha, D. P. Prediction of depression from eeg signal using long short term memory (lstm). In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (IEEE, 2019).
29. Tortora, S. et al. Deep learning-based bci for gait decoding from eeg with lstm recurrent neural network. *J. Neural Eng.* **17**, 046011 (2020).
30. Santos, E. M., San-Martin, R. & Fraga, F. J. Comparison of subject-independent and subject-specific eeg-based bci using lda and svm classifiers. *Med. Biol. Eng. Comput.* **61**, 835–845 (2023).
31. Yin, Z. et al. Federated semi-supervised representation augmentation with cross-institutional knowledge transfer for healthcare collaboration. *Knowledge-Based Syst.* **300**, 112208 (2024).
32. Yang, H. & Wu, S. Eeg classification for bci based on csp and svm-ga. *Appl. Mech. Mater.* **459**, 228–231 (2014).
33. Li, Y. & Koike, Y. A real-time bci with a small number of channels based on csp. *Neural Comput. Appl.* **20**, 1187–1192 (2011).
34. Fu, R. et al. Improvement motor imagery eeg classification based on sparse common spatial pattern and regularized discriminant analysis. *J. Neurosci. Methods* **343**, 108833 (2020).
35. Fu, R. et al. Improvement motor imagery eeg classification based on regularized linear discriminant analysis. *J. Med. Syst.* **43**, 1–13 (2019).
36. Bouallgue, G. et al. A dynamic filtering df-rnn deep-learning-based approach for eeg-based neurological disorders diagnosis. *IEEE Access* **8**, 206992–207007 (2020).
37. Wang, H., Qiu, X., Xiong, Y. & Tan, X. Autogn: An adaptive multi-channel graph recurrent joint optimization network with copula-based dependency modeling for spatio-temporal fusion in electrical power systems. *Inform. Fusion* **117**, 102836 (2025).
38. Seo, J. et al. Eeg-based emotion classification for alzheimer’s disease patients using conventional machine learning and recurrent neural network models. *Sensors* **20**, 7212 (2020).
39. Sarkar, A., Singh, A. & Chakraborty, R. A deep learning-based comparative study to track mental depression from eeg data. *Neurosci. Inform.* **2**, 100039 (2022).
40. Alessandrini, M. et al. Eeg-based alzheimer’s disease recognition using robust-pca and lstm recurrent neural network. *Sensors* **22**, 3696 (2022).
41. Chowdary, M. K., Anitha, J. & Hemanth, D. J. Emotion recognition from eeg signals using recurrent neural networks. *Electronics* **11**, 2387 (2022).
42. Garcia-Moreno, F. M. et al. A cnn-lstm deep learning classifier for motor imagery eeg detection using a low-invasive and low-cost bci headband. In *2020 16th international conference on intelligent environments (IE)* (IEEE, 2020).
43. Khademi, Z., Ebrahimi, F. & Montazery Kordy, H. A transfer learning-based cnn and lstm hybrid deep learning model to classify motor imagery eeg signals. *Comput. Biol. Med.* **143**, 105288 (2022).
44. Wang, H., Chen, B., Sun, H., Li, A. & Zhou, C. Anfis-moh: Systematic exploration of hybrid anfis frameworks via metaheuristic optimization hybridization with evolutionary and swarm-based algorithms. *Appl. Soft Comput.* **167**, 112334 (2024).
45. Gong, P. et al. A spiking neural network with adaptive graph convolution and lstm for eeg-based brain-computer interfaces. *IEEE Trans. Neural Syst. Rehabilitat. Eng.* **31**, 1440–1450 (2023).
46. Alzahab, N. A. et al. Hybrid deep learning (hdl)-based brain-computer interface (bci) systems: A systematic review. *Brain sci.* **11**, 75 (2021).
47. Wang, J. et al. A 2d cnn-lstm hybrid algorithm using time series segments of eeg data for motor imagery classification. *Biomed. Signal Process. Control* **83**, 104627 (2023).
48. Abibullaev, B. & Zollanvari, A. A systematic deep learning model selection for p300-based brain–computer interfaces. *IEEE Trans. Syst., Man, and Cybernetics: Syst.* **52**, 2744–2756 (2021).
49. Lu, P. et al. Combined cnn and lstm for motor imagery classification. In *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* (IEEE, 2019).

Acknowledgements

This work was supported by the Joint Project of Science and Technology Committee of Yangpu District and Health Commission of Yangpu District (YPZYM202302).

Author contributions

Z.C. and Y.C. led the clinical and technical aspects of the study, respectively, including conceptualization, data collection, and algorithm development. Q.F. and L.H. conducted experimental validations, data preprocessing, and statistical analysis. Q.F. supervised the study and coordinated between clinical and technical teams. L.H.

provided strategic input and supervised technical implementation. All authors contributed to manuscript preparation and review.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Shidong Hospital, Yangpu District, Shanghai, China.

Additional information

Correspondence and requests for materials should be addressed to Q.F. or L.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025