A Multi-Scale Inverted Spatial-Temporal Network for EEG-Based Emotion Recognition

Vinod R. Kokitkar
Department of Master of Computer Applications (M.C.A), K.L.S Gogte Institute of Technology, (Affiliated to Visvesvaraya Technological University, Belagavi), Udyambag, Belagavi, Karnataka, 590008, India
vpkokitkar@gmail.com (corresponding author)

Anand Ghuli
Department of Master of Computer Applications (M.C.A), B.L.D.E.A's V.P. Dr. P.G. Halakatti College of Engineering and Technology, (Affiliated to Visvesvaraya Technological University, Belagavi), Vijayapur, Karnataka, 586103, India
mca.anand@bldeacet.ac.in

Received: 21 January 2026 | Revised: 31 March 2026 and 22 April 2026 | Accepted: 23 April 2026

 DOI: https://doi.org/10.48084/etasr.17684

ABSTRACT

Understanding human emotional states through Electroencephalography (EEG) signals has gained significant attention due to its applications in healthcare, human-computer interaction, and affective computing. However, existing approaches often struggle to model temporal dynamics and spatial dependencies effectively, which limits recognition accuracy. The primary research gap lies in the inability of conventional and recent models to simultaneously capture multi-scale temporal patterns while preserving channel-specific information over time. To address this limitation, this study proposes a Multi-Scale Inverted Spatial-Temporal Network (MIST-E) for EEG-based emotion recognition. MIST-E constructs multi-scale representations and employs an inverted embedding strategy to maintain temporal continuity and spatial channel relationships. In addition, a newly designed CNN is used to extract discriminative features for reliable classification. Experimental results on the DEAP dataset demonstrate that MIST-E effectively captures complex spatial-temporal dependencies, achieving 90.56±1.02% accuracy for valence and 91.12±0.98% for arousal. These findings indicate that MIST-E provides improved accuracy compared to existing methods.

Keywords-EEG signals; emotion recognition; deep learning; multi-scale learning; inverted embedding

I. INTRODUCTION

Analysis and understanding of human emotional behavior have become increasingly important in psychology, neuroscience, and computational intelligence, as emotions strongly influence decision-making, cognitive processes, and social interaction. Accurate emotion recognition has wide-ranging applications in healthcare monitoring, Brain-Computer Interfaces (BCIs), adaptive learning environments, and human-computer interaction systems [1, 2]. Among the available sensing modalities, Electroencephalography (EEG) has emerged as one of the most reliable tools for emotion analysis because it provides high temporal resolution and captures neural activity directly from the brain. Unlike behavioral cues such as facial expressions or speech, EEG signals are less susceptible to intentional manipulation, making them a more objective and dependable indicator of affective states [3]. Recent advances in Artificial Intelligence (AI), particularly in Machine Learning (ML) and Deep Learning (DL), have significantly improved the ability to automatically detect emotions from EEG signals. Traditional ML techniques, such as Support Vector Machines (SVM) [4] and K-Nearest Neighbors (KNN) [5], rely on handcrafted features, including Power-Spectral Density (PSD) and statistical descriptors. Although these methods provide baseline performance, they struggle to capture complex nonlinear and temporal relationships present in EEG data. To address these shortcomings, DL models such as Convolutional Neural Networks (CNNs) [6, 7], Recurrent Neural Networks (RNNs) [8], and transformer-based architectures [9, 10] have been proposed to learn spatial and temporal patterns directly from raw signals. Despite these advances, accurately modelling multi-scale temporal dependencies and spatial relationships across EEG channels remains a challenging problem.

Several studies have attempted to enhance emotion recognition by introducing different modeling strategies. For instance, in [11], multiple Electro-Dermal Activity (EDA) decomposition techniques were combined with SVM and Random Forest (RF) classifiers, achieving moderate performance on the DEAP dataset. In [12], a reinforcement-

learning-based Frontal-Lobe Double-Dueling Q-Network (FLD3QN) incorporated prior neurological knowledge but demonstrated limited generalization across brain regions. Connectivity-based approaches such as the Graph CNN Brain-Function Connectivity (GERBN) model [13] capture inter-electrode relationships effectively but introduce computational overhead. Other works investigated physiological coupling using phase synchronization [14], Contrastive Learning for brain-region Representation (CLRA) [15], multimodal graph-based learning [16], and multi-scale temporal modelling frameworks such as miMamba [17]. Domain adaptation approaches such as Multi-Task Adversarial-Domain Adaptation (MTADA) [18] and hierarchical transformer frameworks [19] attempted to address cross-subject variability, but recognition accuracy and robustness remained inconsistent across datasets.

Despite these efforts, several critical challenges persist in EEG-based emotion recognition. EEG signals are inherently noisy, high-dimensional, and non-stationary, which complicates the extraction of discriminative features. Many existing methods either fail to model temporal patterns at multiple scales or merge channel information too early, causing loss of physiological specificity and misalignment between channels. These limitations lead to reduced classification performance and poor generalization across subjects and experimental conditions. To address these challenges, this work proposes a novel Multi-Scale Inverted Spatial-Temporal Network for Emotion Recognition (MIST-E). The MIST-E first performs frequency-domain analysis to identify dominant spectral components and derive adaptive temporal patches. A Multiple-Scale Perception (MSP) mechanism captures both local and global temporal dependencies, while an inverted embedding strategy reorganizes EEG sequences to preserve temporal continuity within each channel and maintain spatial separation across electrodes. Furthermore, a Region-Specific State (RSS) module dynamically enhances spatial-temporal interactions, and a CNN-based classifier learns discriminative features for accurate emotion classification. This integrated design enables MIST-E to effectively address noise, channel misalignment, and multi-scale temporal variability, leading to more robust and interpretable emotion recognition. The main contributions of the proposed MIST-E framework can be summarized as follows:

- A frequency-domain driven approach derives dominant spectral components and defines adaptive temporal patch sizes. In addition, an MSP mechanism employs convolutional kernels of varying sizes to capture intra-patch and inter-patch dependencies.
- An inverted embedding strategy aggregates temporal sequences within each channel while preserving spatial separation across electrodes. In addition, an RSS module dynamically models spatial-temporal interactions between EEG channels.
- A CNN-based classification approach jointly learns spatial and temporal correlations for improved emotion recognition performance.

II. METHODOLOGY

A. Architecture

The MIST-E architecture, shown in Figure 1, is designed to capture spectral and spatial-temporal dynamics in EEG signals in a structured manner. The process begins with EEG acquisition and preprocessing, followed by transformation of signals into the frequency domain to extract informative spectral components. The signals are segmented into non-overlapping patches and reshaped into two-dimensional representations to model both local variations and global dependencies. Multi-scale features are then extracted and refined through separate temporal and spatial processing streams. An RSS module further enhances temporal-spatial interactions. The optimized features are passed to a CNN, followed by fully connected layers and a softmax classifier to predict arousal and valence.

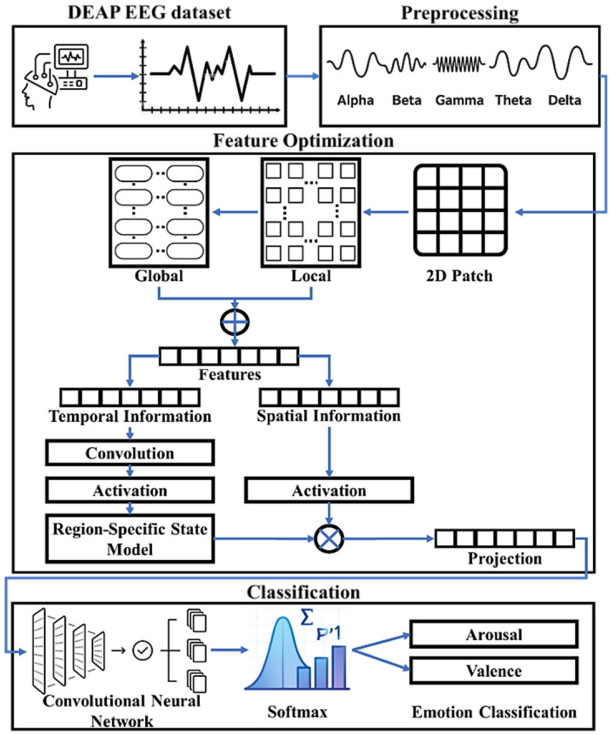

Fig. 1. Multi-scale Inverted spatial-temporal network for emotion recognition (MIST-E) architecture.

Table I provides all architectural details for the MSP and RSS modules, including kernel sizes, number of filters, activation functions, and output dimensions. The number of dominant frequencies k was adaptively selected based on an energy threshold retaining 90% of the cumulative spectral power. EEG signals are segmented into fixed-length windows of $L = 128$ samples with non-overlapping patches, and zero-padding is applied as necessary to preserve dimensional consistency during reshaping and convolution operations.

TABLE I. NETWORK ARCHITECTURE

Layer name	Kernel size	Number of filters	Activation	Output dimensions
MSP_{Conv1}	3×3	32	$ReLU$	$(128, 32)$
MSP_{Conv2}	3×3	64	$ReLU$	$(128, 64)$
MSP_{Pool}	2×2	–	–	$(64, 64)$
RSS_{Conv1}	3×3	64	$ReLU$	$(64, 64)$
RSS_{Conv2}	3×3	128	$ReLU$	$(64, 128)$
RSS_{Pool}	2×2	–	–	$(32, 128)$
Flatten	–	–	–	$(4096,)$
Dense1	–	256	$ReLU$	$(256,)$
Dense2	-	$num_{classes}$	$Softmax$	$(num_{classes},)$

B. DEAP EEG Dataset

This study utilized the DEAP EEG dataset [20] as the primary source of data for training and testing MIST-E. The DEAP dataset is publicly available and used as per its usage policy [21]. The preprocessed version of the dataset was utilized, which included EEG signals downsampled at 128 Hz frequency and reduced channels to 40 in total. Among these, 32 channels correspond to EEG electrodes positioned on the scalp following the standard international 10-20 electrode placement system, while the remaining 8 channels were PPG signals. As this study focused on analyzing emotional responses based solely on EEG activity, only 32 EEG channels were considered for further processing and analysis. Table II provides comprehensive details on the DEAP dataset.

TABLE II. DEAP DATASET

	Feature	Data	Description
1	Total Participants	32	16 Male, 16 Female, No mental illness
2	Age	19 to 37	Age of participants
3	Video clips	40	Selected from 120 video clips
4	Sampling rate	128 Hz	Downsampling of EEG
5	Length of each EEG signal for each video	63 s	Baseline=3 s; Actual data=60 s
6	Label ratings	1-9	Ratings for arousal and valence
7	EEG data length	8064 samples/s	Base = 384, Trail = 7680
8	Data array	40×32×8064	Video/Trial×Subjects×data
9	Labels	40×2	Video/Trial×Label (Arousal, Valence)

C. Preprocessing

The DEAP dataset was further processed to extract meaningful frequency-based features, which are important for emotion recognition using EEG signals. During preprocessing, the initial 3-second baseline was removed. Also, no augmentation and channel selection were applied. Initially, the EEG data were band-limited to a frequency range of $0 - 45$ Hz, as this interval has the most significant brainwave activities important for emotion/cognitive analysis while removing irrelevant higher-frequency noise. Fast-Fourier Transform (FFT) was applied to analyze the spectral properties of signals, which converted time-domain EEG data to frequency-domain representation. Following this, the PSD was computed for quantifying the power distribution of EEG signals across different frequency components. Based on standard EEG frequency bands, the signals were divided into five different ranges: delta (δ) ranging from 0.5 to 4 Hz, associated with deep sleep and unconscious states; theta (θ) spanning 4 to 8 Hz, typically linked with relaxation and drowsiness; alpha (α) within 8 to 12 Hz, connected to calmness and reduced mental effort; beta (β) covering 12 to 30 Hz, which reflects active thinking, concentration, and alertness; and gamma (γ) from 30 to 45 Hz, commonly related to higher-order cognitive processes such as learning and memorizing. These frequency bands provided crucial insights to assess emotional behaviour.

D. Feature Optimization

Let the EEG data for each subject/participant be expressed as $3D$ matrix $S \in \mathbb{R}^{M \times T \times C}$, where M denotes the number of trials, T denotes the number of time-samples, and C denotes the number of channels. In this work, the signal was divided into N non-overlapping segments of length L, such that $T = N \times L$. These segments form dataset I as

$$I = \{(X_{ij}, Y_{ij}) |\ i = 1,2, \dots, M; j = 1,2, \dots, N\} \quad (1)$$

where each segment $X_{ij} \in \mathbb{R}^{L \times C}$ corresponds to the label $Y_{ij} \in \mathbb{R}$. For a given trial, all N segments inherit the same label. A single segment is denoted as $X^{1D} \coloneqq X_{ij}$. In this work, the task of feature optimization in EEG-based emotion recognition is to learn how the model predicts the label Y_{ij} from the input segment X^{1D}. Consider X^{1D}, which represents an EEG segment of length L with C channels. For constructing a multi-scale representation of a given EEG signal, appropriate patch sizes have to be determined. This is done by analyzing the EEG signal in the frequency domain as

$$\begin{aligned} A &= A(FFT(X^{1D})) \\ \{f_1, f_2, \dots, f_k\} &= argTop_k(A) \\ p_i &= \left[\frac{L}{f_i}\right], \qquad i \in \{1, \dots, k\} \end{aligned} \quad (2)$$

where FFT denotes the FFT process, and A denotes the amplitude corresponding to each frequency component. As higher-frequency ranges contain more noise, only the top k frequencies with the largest amplitudes were selected for patches. These dominant frequencies $\{f_1, f_2, \dots, f_k\}$ correspond to periods $\{p_1, p_2, \dots, p_k\}$ and amplitudes $\{A_{f_1}, A_{f_2}, \dots, A_{f_k}\}$. The computed periods serve as the basis for defining patch sizes. Further, the relative importance of these frequencies is quantified utilizing softmax weighting W_{f_i} as

$$W_{f_i} = \{W_{f_1}, \dots, W_{f_k}\} = Softmax(A_{f_1}, \dots, A_{f_k}) \quad (3)$$

The original signal X^{1D} is then segmented into patches of size p_i and reshaped into $2D$ structure using X_i^{2D}:

$$\begin{aligned} &X_i^{2D} = \\ &Reshape_{p_i, f_i}(Padding(X^{1D})), i \in \{1, \dots, k\} \end{aligned} \quad (4)$$

In this step, zero-padding is employed to ensure divisibility into integer-sized patches. The resulting 2D tensors are denoted as $X_i^{2D} \in \mathbb{R}^{p_i \times f_i \times C}$. In this work, using the 2D structure, the vertical axis captures variations within patches (local), while the horizontal axis reflects dependencies across patches (global). This process yields a collection of multi-scale 2D tensors $\{X_1^{2D}, X_2^{2D}, \dots, X_k^{2D}\}$, each encoding temporal dependencies at different scales. When the p_i value is larger, it

emphasizes long-range temporal structures, while smaller ones capture finer details. This 2D reshaping also applies convolution operations, which efficiently extract both local and global features from the EEG signal. The reshaped EEG tensors are further processed utilizing a novel MSP approach, mathematically formulated as:

$$X_i^{2D} = Reshape_{1,p_i \times f_i}\left(MSP(X_i^{2D})\right), i \in \{1, \dots, k\} \quad (5)$$

In this MSP, convolutional filters of varying kernel size are applied. This provides a network to capture both intra-patch dynamics (local variations in segment) and inter-patch relationships (patterns across segments having similar phase). After convolution, the outputs are reshaped back to the original 1D form, denoted as X_i^{1D}. To emphasize the contribution of features extracted from different frequency-specific patches, a weighted summation is performed across all scales:

$$X^{1D} = \sum_{i=1}^{k} W_{f_i} \times X_i^{1D} \quad (6)$$

where W_{f_i} denotes the weight assigned to features corresponding to frequency f_i. The fusion strategy ensures that information from multiple scale features is effectively integrated, resulting in a more discriminative representation of the EEG signal, improving emotion recognition. Once the multi-scale feature representation of the EEG signal is obtained, the next step is to model the interaction between temporal and spatial information.

Traditional embedding strategies usually merge signals from different channels recorded at the same time step into a single token. However, this approach is problematic because electrodes capture distinct physiological activities, and their signals can exhibit opposite behavior (like one channel peaking while another is at a trough). Compressing such different events into a single token not only obscures useful information but also presents misalignment in representation. Hence, to address this issue, an inverted embedding approach is utilized. Instead of grouping signals across channels at the same time point, this method aggregates multiple consecutive time steps from the same channel into one token. In doing so, the representation becomes more event-driven, i.e., it preserves temporal continuity within each channel while maintaining individuality of different channels through separate tokens. This structure strengthens MIST-E's ability to capture long-term temporal dependencies while also retaining spatial distinctions between channels, providing a more interpretable representation of EEG signals for emotion recognition.

The inverted embedding mechanism in MIST-E differs from conventional spatial-temporal embeddings by reorganizing the representation of EEG signals before feature learning. Moreover, the proposed inverted embedding aggregates consecutive temporal samples within each channel to form tokens while maintaining separation between channels. This design preserves temporal continuity and channel individuality, enabling MIST-E to capture channel-specific temporal patterns more effectively. Thus, the CNN can learn more meaningful spatial-temporal correlations. Formally, a multiscale EEG feature representation X^{1D} is reorganized as:

$$\hat{X}^{1D} = Reshape_{C,L}(X_{1D}) \quad (7)$$

where $\hat{X}^{1D}$ denotes inverted embedding, where time-steps and channels are rearranged to emphasize temporal-spatial interactions. To further capture dynamic dependencies between temporal-spatial components, this work presents an RSS model, which is applied to $\hat{X}^{1D}$. By combining inverted embedding with RSS-based fusion, MIST-E achieves a more expressive and comprehensive representation of EEG dynamics.

E. Classification

After constructing the optimized multi-scale representation and applying the inverted embedding approach, the final step is to perform classification for EEG-based emotion recognition. A novel CNN was designed to capture spatial-temporal correlations in the feature space. Moreover, due to the operation of inverted embedding, the input $\hat{X}^{1D}$ preserves temporal continuity within each channel while maintaining inter-channel distinctions. This structure enabled the CNN to simultaneously extract temporal dynamics (using convolution along the time axis) and spatial dependencies (using convolution across channels). Using the input $\hat{X}^{1D} \in \mathbb{R}^{C \times L}$, the CNN performs the convolution operation F:

$$F = \sigma\left(Conv\left(\hat{X}^{1D}; \theta_c\right)\right) \quad (8)$$

where $Conv(\cdot)$ denotes convolution using learnable parameters θ_c and $\sigma(\cdot)$ denotes a non-linear activation function (ReLU). The resulting feature maps F capture discriminative spatiotemporal patterns. The extracted features are then flattened and passed through FCL to obtain logits $z \in \mathbb{R}^n$, where n denotes the number of emotion categories (two classes), evaluated as z using:

$$z = W_f \cdot Flatten(F) + b_f \quad (9)$$

Here, W_f and b_f denote a trainable weight matrix and a bias vector, respectively. For computing class probability, a softmax function is applied to logits using:

$$\hat{Y}_i = \frac{\exp(z_i)}{\sum_{j=1}^{n} \exp(z_j)}, \; i \in \{1, \dots, n\} \quad (10)$$

where $\hat{Y}_i$ denotes the predicted probability for class i. Finally, the MIST-E training is guided by the cross-entropy loss function, which measures divergence between predicted probabilities and ground-truth label distribution:

$$L_{cls} = -\sum_{i=1}^{n} 1[y = i] \cdot \log\left(\hat{Y}_i\right) \quad (11)$$

where $1[y = i]$ is an indicator function that equals 1 when for the true-class label i and 0 otherwise. Through this classification, the CNN effectively leverages inverted embedding representation to learn joint temporal-spatial features, while the softmax classifier and cross-entropy loss ensure accurate discrimination of emotion categories.

III. RESULTS AND DISCUSSION

A. Experimental Setup

MIST-E was developed in Python using the TensorFlow and PyTorch libraries. All experiments were carried out on a high-performance workstation running Windows 11, configured with 32 GB RAM, an NVIDIA GeForce RTX 4060 GPU (8 GB GDDR6), and a 1 TB SSD, ensuring efficient data management and faster computation. This setup provided robust support for training, validation, and testing of the model on the DEAP dataset, enabling large-scale computations with reduced runtime overhead. To provide a more reliable assessment of MIST-E generalization, it was further evaluated using five k-fold cross-validation under a subject-independent setting. Performance metrics were averaged across folds and reported as mean±standard deviation to account for variability in EEG signals across individuals. To evaluate the performance of MIST-E, standard metrics such as accuracy, precision, recall, and F1-score were adopted, as defined in (12)–(15), where TP denotes True Positives, TN denotes True Negatives, FP refers to False Positives, and FN denotes False Negatives.

$$\text{Accuracy (A)} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (12)$$

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP+FP}} \quad (13)$$

$$\text{Recall (R)} = \frac{\text{TP}}{\text{TP+FN}} \quad (14)$$

$$\text{F1 - Score (FS)} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} \quad (15)$$

B. Performance Evaluation

The DEAP dataset was employed to evaluate MIST-E, with the preprocessing steps described above. A subject-independent evaluation was adopted to ensure an independent examination of generalizability, where MIST-E was trained and tested on data from different participants. This strategy reduced subject-specific bias and enabled MIST-E to learn invariant spatiotemporal patterns across individuals. Furthermore, multiscale feature extraction and inverted embedding contributed to robustness by preserving channel-specific dynamics while capturing global temporal trends, improving stability against inter-subject variability. This setup enabled the model to generalize across unseen individuals and mitigate subject-specific biases. The key hyperparameters of MIST-E, including learning rate, kernel sizes, and number of filters, were empirically selected and further tuned, as shown in Table III.

TABLE III. HYPERPARAMETERS CONFIGURED FOR MIST-E

Hyperparameter	Value
Learning rate	0.001
Optimizer	Adam
Batch size	64
Epochs	100
Dropout rate	0.3
Convolution kernel sizes	[3, 5, 7]
Number of Filters (CNN)	64, 128, 256
Activation Function	ReLU
Pooling Strategy	Max Pooling
Fully Connected Layer (FCL)	256 neurons
Loss Function	Cross-Entropy
Evaluation Metrics	Accuracy, Precision, Recall, F1

The experimental results (Figure 2) show superior performance in recognizing emotions from EEG signals across both valence and arousal dimensions. For valence classification, MIST-E achieved 90.56%±1.02 accuracy, and for arousal obtained 91.12%±0.98 accuracy. Table IV presents the outcomes achieved for five k-fold cross-validation, while Figures 3 and 4 show the confusion matrix and ROC-curve for valence and arousal, respectively. The improvement of MIST-E is due to multi-scale feature optimization that extracts both fine-grained and long-range temporal dependencies, while the inverted embedding mechanism ensures that spatial channel-specific patterns are preserved without introducing misalignment. Additionally, the integration of convolutional layers within the classification stage enables MIST-E to effectively model spatiotemporal correlations, enhancing its ability to distinguish between subtle emotional variations.

TABLE IV. PERFORMANCE OF MIST-E UNDER 5-FOLD CROSS-VALIDATION ON DEAP

Fold	Valence accuracy (%)	Arousal accuracy (%)
Fold 1	89.72	90.35
Fold 2	90.11	91.04
Fold 3	91.02	91.58
Fold 4	90.36	90.89
Fold 5	91.59	91.74
Mean±Std	**90.56 ± 0.67**	**91.12 ± 0.54**

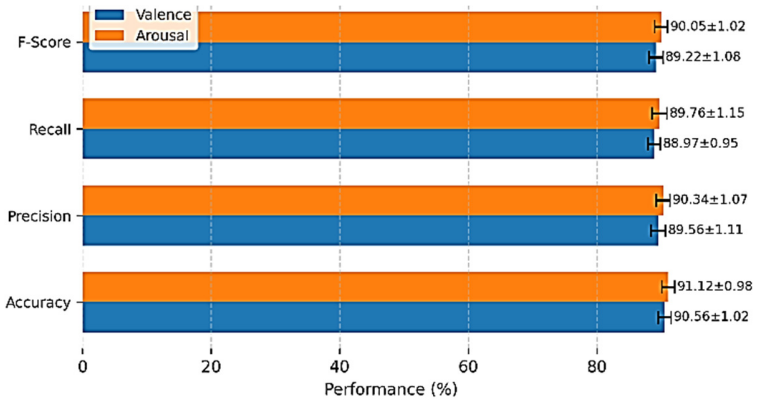

Fig. 2. Performance of MIST-E on the DEAP dataset.

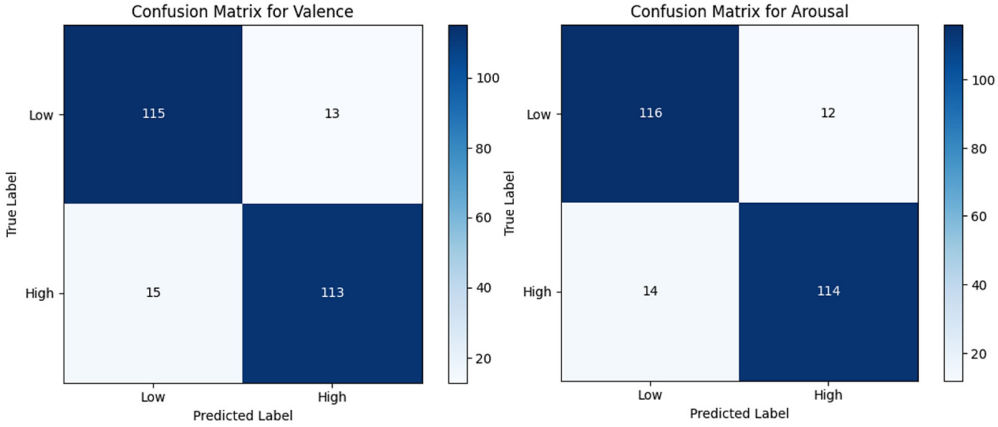

Fig. 3. Confusion matrix for valence and arousal.

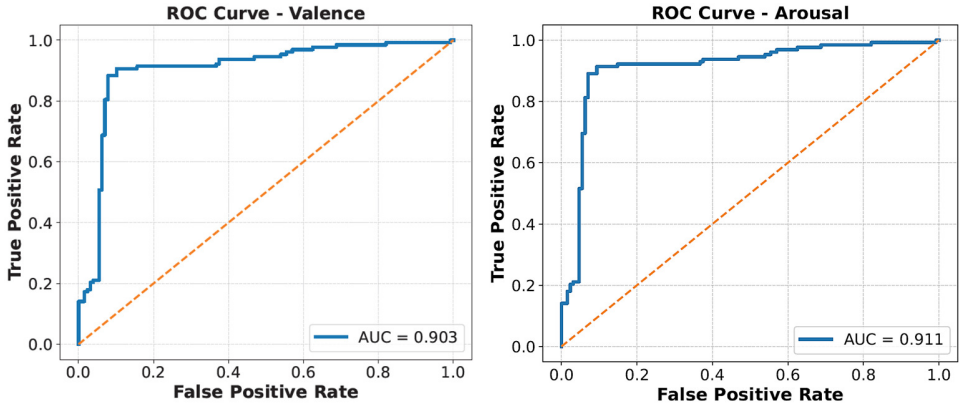

Fig. 4. ROC curve for valence and arousal.

C. Ablation Study

To further evaluate the contribution of individual components within MIST-E, an ablation study was conducted on the DEAP dataset, as presented in Table V. In this analysis, key modules were systematically removed while keeping the remaining architecture unchanged. The evaluated components include the multi-scale patching mechanism, the inverted embedding strategy, the MSP module, and the RSS module.

TABLE V. ABLATION STUDY

Model variant	Valence accuracy (%)	Arousal accuracy (%)
Baseline CNN (without the proposed modules)	82.14	83.02
Without multi-scale patching	85.21	86.08
Without inverted embedding	86.47	87.12
Without the MSP module	87.36	88.04
Without the RSS module	88.41	89.02
Full MIST-E model	**90.56**	**91.12**

The results indicate that removing any component leads to a noticeable decrease in performance. In particular, the absence of multi-scale patching reduces MIST-E's ability to capture temporal dependencies across different frequency bands, while removing inverted embedding leads to loss of channel-specific temporal continuity. Similarly, the MSP module contributes to learning intra-patch and inter-patch dependencies, and the RSS module enhances spatial-temporal alignment across EEG channels. These results confirm that the integration of these components collectively improves the robustness and accuracy of the MIST-E framework for EEG-based emotion recognition.

D. Comparative Study

Table VI shows a comparative analysis of existing EEG-based emotion recognition methods, highlighting recent advances in the field while also indicating persistent challenges. All reported results for the reference models are taken directly from their respective published works, where evaluations were also conducted under a subject-independent (inter-subject) setting on the DEAP dataset, ensuring a fair comparison basis. The miMamba framework [17] employs multi-scale temporal feature extraction along with temporal-spatial fusion mechanisms, achieving 86.04% accuracy for valence and 85.94% for arousal. These results demonstrate the effectiveness of modeling EEG signals at multiple temporal resolutions. The Hierarchical Transformer approach [19] integrates respiration-based multimodal information, reporting comparatively lower performance of 72.42% for valence and 73.91% for arousal. This suggests that incorporating non-EEG modalities may not always improve performance in fine-grained emotion recognition under subject-independent conditions.

In contrast, the proposed MIST-E model consistently outperforms the referenced methods, achieving 90.56% ± 1.02 for valence and 91.12% ± 0.98 for arousal. This improvement indicates that the proposed architecture is more effective in learning discriminative EEG representations across unseen subjects. In addition to performance comparison, computational efficiency was also evaluated. The MIST-E model contains approximately 1.85 million trainable parameters and requires 0.42 GFLOPs per inference, which is lower than most transformer-based architectures. The average training time per fold on the DEAP dataset is approximately 38 minutes using an NVIDIA RTX 3080 GPU, reflecting a balanced trade-off between computational cost and classification performance.

These findings demonstrate that MIST-E provides a more robust and discriminative representation of EEG signals, thereby establishing a new benchmark for EEG-based emotion recognition.

TABLE VI. COMPARATIVE STUDY FOR SUBJECT-INDEPENDENT/INTER-SUBJECT EVALUATION ON DEAP

Ref	Model	Average classification accuracy	
		Valence	Arousal
[17]	miMamba	86.04	85.94
[19]	Hierarchical Transformer	72.42	73.91
Proposed	**MIST-E**	**90.56±1.02**	**91.12±0.98**

IV. CONCLUSION

This study presented a novel approach, MIST-E, to advance EEG-based emotion recognition, identifying key challenges such as noise, inter-subject variability, and the limited ability of conventional models to capture spatiotemporal dependencies. Existing approaches, including traditional ML, graph-based methods, and transformer models, have been shown to achieve moderate performance but often struggle with scalability, robustness, and interpretability. To address these gaps, the problem was defined as developing a model capable of effectively capturing both local and global EEG patterns, while maintaining temporal continuity and spatial specificity. MIST-E integrated multi-scale feature optimization, inverted embedding to preserve channel-specific dynamics, and CNN-based classification. Experimental evaluation on the DEAP dataset demonstrated that MIST-E achieved 90.56±1.02% accuracy for valence and 91.12±0.98% for arousal, significantly outperforming state-of-the-art methods. The results confirm that the proposed model provides a richer, more discriminative representation of EEG signals, thereby improving emotion recognition reliability. In the future, this work can be extended by considering sensory responses and presenting hybrid approaches, extending the MIST-E approach.

DECLARATION OF COMPETING INTERESTS

Not applicable to this work.

ACKNOWLEDGMENT

The authors would like to thank the developers and contributors of the DEAP dataset for making the dataset publicly available for research purposes.

DATA AVAILABILITY

The DEAP dataset used in this study is publicly available and can be accessed from the official website of the DEAP [21]. The data used in this work were obtained and utilized in accordance with the dataset usage policy.

AI USE AND DECLARATION OF GENERATIVE AI USE

During the preparation of this work, the authors used Grammarly AI for language refinement, grammar correction, and improving the readability of the manuscript. After using this tool, the authors carefully reviewed and edited the content as needed and take full responsibility for the content of the published article.

REFERENCES

[1] M. V. Cruz, S. Jamal, and S. C. Sethuraman, "A Comprehensive Survey of Brain–Computer Interface Technology in Health care: Research Perspectives," *Journal of Medical Signals & Sensors*, vol. 15, no. 6, June 2025, https://doi.org/10.4103/jmss.jmss_49_24.

[2] N. Babu, U. Satija, J. Mathew, and A. P. Vinod, "Emotion recognition in virtual and non-virtual environments using EEG signals: Dataset and evaluation," *Biomedical Signal Processing and Control*, vol. 106, Aug. 2025, Art. no. 107674, https://doi.org/10.1016/j.bspc.2025.107674.

[3] A. Raza and M. Z. Yusoff, "Development of a CNN-LSTM Deep Learning Model for Motor Imagery EEG Classification for BCI Applications," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 22705–22711, June 2025, https://doi.org/10.48084/etasr.9945.

[4] M. A. Mahmood, K. Alsalem, M. K. Elbashir, S. A. El-Ghany, and A. A. A. El-Aziz, "Segmentation-enhanced approach for emotion detection from EEG signals using the fuzzy C-mean and SVM," *Scientific Reports*, vol. 15, no. 1, Aug. 2025, Art. no. 31956, https://doi.org/10.1038/s41598-025-17220-w.

[5] K. Sutijirapan, S. Auephanwiriyakul, and N. Theera-Umpon, "Emotion Detection from Electroencephalography Signals Using String Grammar K-Nearest Neighbors," in *2024 21st International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, May 2024, pp. 1–4, https://doi.org/10.1109/ECTI-CON60892.2024.10594829.

[6] F. Yan, Z. Guo, A. M. Iliyasu, and K. Hirota, "Multi-branch convolutional neural network with cross-attention mechanism for emotion recognition," *Scientific Reports*, vol. 15, no. 1, Feb. 2025, Art. no. 3976, https://doi.org/10.1038/s41598-025-88248-1.

[7] K. Devarajan, S. Ponnan, and S. Perumal, "Hybrid CNN-transformer architecture for enhanced EEG-based emotion recognition: capturing local and global dependencies with self-attention mechanisms," *Discover Computing*, vol. 28, no. 1, May 2025, Art. no. 87, https://doi.org/10.1007/s10791-025-09596-0.

[8] M. Li, P. Yu, and Y. Shen, "A spatial and temporal transformer-based EEG emotion recognition in VR environment," *Frontiers in Human Neuroscience*, vol. 19, Feb. 2025, Art. no. 1517273, https://doi.org/10.3389/fnhum.2025.1517273.

[9] M. M. Alam, M. A. Dini, D.-S. Kim, and T. Jun, "TMNet: Transformer-fused multimodal framework for emotion recognition via EEG and speech," *ICT Express*, vol. 11, no. 4, pp. 657–665, Aug. 2025, https://doi.org/10.1016/j.icte.2025.04.007.

[10] L. Liu, Q. Zhao, L. Liu, Y. Qiao, and J. Gao, "A Lightweight Network Based on Multi-Scale Convolutional Neural Network and Gated Transformer for EEG Emotion Classification," *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 38, no. 4, July 2025, Art. no. e70087, https://doi.org/10.1002/jnm.70087.

[11] Y. R. Veeranki, N. Ganapathy, R. Swaminathan, and H. F. Posada-Quintero, "Comparison of Electrodermal Activity Signal Decomposition Techniques for Emotion Recognition," *IEEE Access*, vol. 12, pp. 19952–19966, 2024, https://doi.org/10.1109/ACCESS.2024.3361832.

[12] D. Li, L. Xie, Z. Wang, and H. Yang, "Brain Emotion Perception Inspired EEG Emotion Recognition With Deep Reinforcement Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 9, pp. 12979–12992, Sept. 2024, https://doi.org/10.1109/TNNLS.2023.3265730.

[13] P. Gao, X. Zheng, T. Wang, and Y. Zhang, "Graph Convolutional Neural Network Based Emotion Recognition with Brain Functional Connectivity Network," *International Journal of Crowd Science*, vol. 8, no. 4, pp. 195–204, Sept. 2024, https://doi.org/10.26599/IJCS.2024.9100022.

[14] S. Banik, H. Kumar, N. Ganapathy, and R. Swaminathan, "Assessment of Emotion Elicitation Using Multimodal Physiological Sensors and Phase Synchronization," *IEEE Sensors Letters*, vol. 8, no. 8, pp. 1–4, Aug. 2024, https://doi.org/10.1109/LSENS.2024.3426562.

[15] S. Dai *et al.*, "Contrastive Learning of EEG Representation of Brain Area for Emotion Recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–13, 2025, https://doi.org/10.1109/TIM.2025.3533618.

[16] Z. Cai, H. Gao, M. Wu, J. Li, and C. Liu, "A Unified Physiological Signal Interaction Network for Cross-Dataset Emotion Recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 17, no. 6, pp. 1447–1460, Dec. 2025, https://doi.org/10.1109/TCDS.2025.3566229.

[17] X. Zhou, D. Huang, X. Peng, and L. Yin, "miMamba: EEG-Based Emotion Recognition With Multi-Scale Inverted Mamba Models," *IEEE Transactions on Affective Computing*, vol. 16, no. 4, pp. 3266–3278, Oct. 2025, https://doi.org/10.1109/TAFFC.2025.3587443.

[18] L. Qiu, Z. Ying, X. Song, W. Feng, C. Zhou, and J. Pan, "MTADA: A Multi-Task Adversarial Domain Adaptation Network for EEG-Based Cross-Subject Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 16, no. 4, pp. 3354–3368, Oct. 2025, https://doi.org/10.1109/TAFFC.2025.3595137.

[19] Y. Wang *et al.*, "Hierarchical Transformer With Auxiliary Learning for Subject-Independent Respiration Emotion Recognition," *IEEE Sensors Journal*, vol. 25, no. 16, pp. 31290–31301, Aug. 2025, https://doi.org/10.1109/JSEN.2025.3587271.

[20] S. Koelstra *et al.*, "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, Jan. 2012, https://doi.org/10.1109/T-AFFC.2011.15.

[21] S. Koelstra, "DEAP: A Dataset for Emotion Analysis using Physiological and Audiovisual Signals," 2011. https://www.eecs.qmul.ac.uk/mmv/datasets/deap/