IRJAEH

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5355 - 5363
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0700

An Explainable CNN–Transformer Framework for Subject-Independent EEG-Based Emotion Recognition

Venkatalakshmi M [1], Bhavani M[2], Pavithra M[3], Archana R[4], Velkumar K[5]

[1,4,5]Assistant Professor, Computer Science and Engineering, Nadar Saraswathi College of Engineering and Technology, Theni, Tamilnadu

[2]Assistant Professor, Information Technology, Nadar Saraswathi College of Engineering and Technology, Theni, Tamilnadu

[3] Assistant Professor, Artificial Intelligence and Data Science, Nadar Saraswathi College of Engineering and Technology, Theni, Tamilnadu

Emails: *venkatalakshmims@gmail.com[1], gmbhavani1990@gmail.com[2], pavithramanibtech@gmail.com[3], archanacse.nscet @gmail.com[4], velkumarskc @gmail.com[5]*

Abstract

Recognizing emotional states from electroencephalogram (EEG) signals is an active area of research in affective computing and brain–computer interaction. EEG is useful for this task because it records brain activity with high temporal resolution, but the signals are complex and can vary considerably from one person to another. These variations make it difficult to build a model that performs consistently on participants who were not seen during training. In addition, deep learning models can provide accurate predictions without clearly showing which parts of an EEG recording influenced the decision. This paper presents an explainable CNN–Transformer framework for subject-independent EEG emotion recognition. The CNN component learns local temporal and spatial patterns, while the Transformer encoder uses self-attention to model relationships across the learned representation. The two representations are fused before classification, and an attention stage is used to emphasize informative features. The DEAP dataset is considered as the primary benchmark, with EEG-only binary valence classification used as the initial task. A Leave-One-Subject-Out (LOSO) strategy is adopted so that the participant used for testing is excluded from model training. Performance is assessed using accuracy, precision, recall, F1-score, and ROC-AUC. SHAP and Integrated Gradients are considered for examining the contribution of EEG channels and temporal regions. Baseline and ablation experiments are also included to study the value of the individual model components. The numerical values currently included in the manuscript are provisional development-stage values and must be replaced by measurements from the completed DEAP experiment before publication.

Keywords: *EEG; Emotion Recognition; CNN; Transformer; Explainable Artificial Intelligence; SHAP; Integrated Gradients; DEAP; Subject-Independent Learning; Affective Computing.*

1. Introduction

Human emotions affect communication, decision-making, learning, and interaction with the surrounding environment. This has created growing interest in automatic emotion recognition within affective computing, human–computer interaction, and brain–computer interface research. Several information sources have been investigated, including facial expressions, speech, text, and physiological measurements. Among these sources, electroencephalography (EEG) is particularly useful because it records electrical brain activity directly and provides high temporal resolution [1], [2], [4]. An EEG recording contains information that changes across time and across electrode locations. Emotion-related patterns may therefore appear as short-term changes, relationships between channels, or differences in frequency-related activity. Public datasets have made it possible to study these patterns under common experimental conditions. The DEAP dataset, introduced by Koelstra et al. [7], is one of the most widely used benchmarks. It contains recordings from 32 participants who watched 40 music-video excerpts and provided ratings for valence, arousal, dominance, liking, and familiarity. Early EEG emotion-recognition systems often depended on handcrafted statistical, spectral, or time–frequency features. Such features can be useful, but their selection usually requires domain knowledge and may not transfer equally well between participants [1], [4]. Deep learning provides another route by learning representations directly from the signal. CNNs have therefore become popular because convolution can identify local temporal and spatial patterns without requiring every feature to be manually defined [9], [14]. Recurrent and hybrid models have also been investigated. LSTM-based approaches can model temporal information, while graph convolution, capsule networks, and attention mechanisms have been combined with deep networks to capture additional EEG structure [10]–[13]. However, local convolution or sequential recurrence may not fully describe relationships between distant parts of an EEG representation. Transformers provide a different mechanism through self-attention. Instead of relying only on nearby operations, self-attention allows different elements of a feature sequence to interact directly. This has encouraged the use of Transformer-based and convolutional Transformer models for EEG emotion recognition [15], [16]. TC-Net [17] is another example of Transformer-based EEG processing. A CNN–Transformer combination is attractive because the two components have complementary roles. CNNs can focus on local patterns, whereas self-attention can capture broader relationships. Recent work has already demonstrated CNN–Transformer approaches for EEG emotion recognition [16], [19]. Therefore, the combination itself should not be presented as the only source of novelty. In this study, the architecture is considered together with explainability and subject-independent evaluation. Interpretability is especially relevant for EEG applications. A prediction such as high or low valence does not explain which electrodes or temporal portions of the signal contributed to the decision. Recent work, including ERTNet, has started to address this issue by examining the interpretability of Transformer-based EEG models [18]. Attribution techniques such as SHAP and Integrated Gradients can add another layer of information by identifying features that have a strong influence on a prediction.

Generalization across participants is equally important. EEG characteristics can differ because of individual physiology, electrode responses, recording conditions, and differences in emotional reactions [4]–[6]. If samples from the same participant are used for both training and testing, the model may learn participant-specific characteristics that do not transfer to a new person. A subject-wise evaluation such as LOSO provides a stricter test of cross-subject generalization. Based on these considerations, this study proposes an explainable CNN–Transformer framework for subject-independent EEG emotion recognition. The framework uses CNN layers for local feature extraction, Transformer encoders for broader dependency modeling, feature fusion for combining the learned representations, and an explainability stage for examining influential EEG features.

1.1.Related Work and Literature Review

1.1.1. EEG Emotion Recognition and DEAP Dataset

Research on EEG emotion recognition has progressed from handcrafted features to deep neural networks that learn representations directly from recordings [1]–[6]. DEAP has played an important role because it contains synchronized EEG and peripheral physiological recordings together with self-reported emotion ratings [7]. The dataset includes 32 participants and 40 music-video trials, making it useful for comparing different learning strategies.

1.1.2. CNN-Based and Hybrid Local Feature Learning

CNNs are widely applied to EEG because they can learn local temporal and spatial patterns. TSception, for example, was designed to capture temporal dynamics and spatial asymmetry [9]. Multi-scale convolution and feature-fusion methods have also been explored [14]. Other studies have combined convolution with graph networks, LSTM units, capsule networks, and attention mechanisms [10]–[13]. These studies demonstrate the value of learned local representations, although convolution alone

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5355 - 5363
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0700

may not provide a direct mechanism for modeling long-range dependencies.

1.1.3. Transformer-Based Global Feature Learning

Transformers use self-attention to establish relationships between different elements of a representation. Vision Transformer approaches have been investigated for EEG emotion recognition [15], while attention-based convolutional Transformer models have combined convolutional processing with global attention [16]. TC-Net [17] explored another Transformer-based direction by integrating capsule representations. These studies indicate that self-attention can complement conventional EEG feature extraction.

1.1.4. Explainable Artificial Intelligence for EEG

As deep models become more complex, understanding how they arrive at a prediction becomes increasingly important. ERTNet [18] investigated an interpretable Transformer framework for EEG emotion recognition, while attention-based methods have also been used to inspect important EEG representations [11], [16], [20]. Such approaches provide information beyond a single classification score and can help identify influential input regions.

1.1.5. Subject-Independent Evaluation

Inter-subject variability remains one of the major challenges in EEG emotion recognition. EEG patterns that are useful for one participant may not transfer directly to another participant [4]–[6]. A subject-independent evaluation keeps the test participant completely separate from training. LOSO is therefore useful when the research question concerns generalization to an unseen individual.

1.1.6. Research Gap

The reviewed literature suggests several connected gaps. CNNs are effective for local representation learning but may not fully model distant dependencies. Transformers can model global relationships but still depend on the quality of their input representation. Recent CNN–Transformer studies show that hybrid architectures are feasible, so the combination alone is not sufficient to establish novelty [16], [19]. In addition, predictive performance and interpretability are not always examined together under a strict subject-independent protocol. The present study therefore treats hybrid representation learning, explainability, and subject-independent evaluation as connected parts of one experimental framework.

1.1.7. Summary and Positioning of the Proposed Work

The proposed study is positioned as an explainable subject-independent EEG framework rather than simply another CNN–Transformer classifier. Its emphasis is on combining local and global representation learning with feature attribution and a participant-wise evaluation protocol. The objective is to obtain a model whose predictive behavior can be examined alongside its classification performance.

2. Proposed Methodology

The proposed framework is designed around a simple principle: local and global information in EEG should be learned together, while the final predictions should also be examined for interpretability. The CNN learns local patterns from the multichannel signal, the Transformer processes the learned sequence and models broader relationships, and a feature-fusion stage combines the two representations before classification. An explainability stage is then used to investigate the input information associated with model decisions. Overall processing pipeline: DEAP EEG → Preprocessing → CNN Local Feature Extraction → Feature Projection/Token Formation → Transformer Encoder → CNN–Transformer Feature Fusion → Attention-Based Feature Weighting → Classification → Emotion Prediction → SHAP/Integrated Gradients.

2.1. Overall Framework Architecture

The EEG segment first passes through preprocessing, where the signal is prepared for model input. The processed segment is then supplied to the CNN branch for local feature extraction. The resulting feature map is[8] rearranged into a sequence and projected into the embedding dimension required by the Transformer. The Transformer encoder uses self-attention to learn broader dependencies. CNN and Transformer representations are then fused and passed to the classifier. Finally, SHAP or Integrated Gradients can be applied to obtain attribution

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5355 - 5363
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0700

information. As shown in Figure 1 Proposed Explainable CNN–Transformer Architecture for EEG Emotion Recognition.

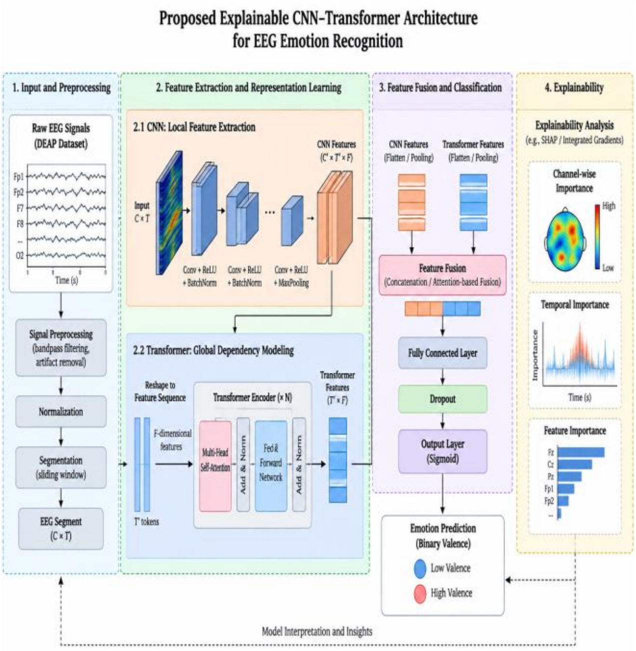

Figure 1 Proposed Explainable CNN–Transformer Architecture for EEG Emotion Recognition

2.2.EEG Input Representation and Preprocessing

Only the EEG portion of the DEAP dataset is used in the initial implementation. The 32 EEG channels are retained, while peripheral physiological signals are excluded because the first objective is to evaluate EEG-only emotion recognition. The input can be represented as $X \in R^{(C \times T)}$, where C denotes the number of EEG channels and T denotes the number of time samples in the selected segment. EEG recordings can contain noise, amplitude differences, and recording-related artifacts. The preprocessing pipeline therefore includes signal preparation, normalization, and temporal segmentation. The same procedure should be applied consistently to all participants. Parameters learned from data should be estimated only from the training portion of each LOSO split to prevent information leakage[5].

2.3.CNN-Based Local Feature Extraction

The CNN receives an EEG segment and learns local temporal and spatial patterns through convolutional filters. A typical CNN block can contain convolution, normalization, nonlinear activation, and pooling. The CNN output is used as a learned[3] representation rather than as the final prediction.

$$F = \sigma(W * X + b)$$

Here, X is the input EEG representation, W is the convolution kernel, b is the bias, * denotes convolution, and σ is the activation function.

2.4.Feature Projection and Token Formation

The CNN produces a feature map, whereas the Transformer expects a sequence of feature vectors. The CNN output is therefore rearranged into a sequence of tokens and projected into the Transformer embedding dimension. Positional information can be added so that the temporal order of the feature sequence is retained.

2.5.Transformer-Based Global Dependency Modeling

The projected feature sequence is passed through a Transformer encoder. Self-attention allows the model to compare different parts of the sequence and learn relationships that may not be captured by local convolution alone.

$$\text{Attention}(Q,K,V) = \text{softmax}(QK^T / \sqrt{d_k})V$$

Q, K, and V denote the query, key, and value matrices, while d_k is the key-vector dimension. Multi-head self-attention uses several attention heads so that different relationships can be learned in parallel.

2.6.CNN–Transformer Feature Fusion

The CNN and Transformer provide complementary information. CNN features retain local characteristics, whereas Transformer features represent broader relationships. The two representations are combined before classification.

$$F\_fusion = [F\_CNN \,||\, F\_Trans]$$

The symbol || denotes feature concatenation. In an attention-based implementation, the fused representation can subsequently be reweighted so that informative components receive greater emphasis.

2.7.Attention-Based Feature Weighting

An attention layer can assign different weights to the fused representations. The purpose is not to manually select features, but to allow the network to learn which representations are most useful for separating

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5355 - 5363
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0700

the target emotion classes.

$$\alpha_i = \exp(e_i) / \Sigma_j \exp(e_j)$$

$$F\_att = \Sigma_i \alpha_i F_i$$

Binary Emotion Classification

For binary valence classification, the final representation is passed to a fully connected layer followed by a sigmoid function. The output represents the probability of the high-valence class. Binary cross-entropy is used as the training objective.

$\hat{y} = \sigma(W\_cF\_att + b\_c)$

$L = -(1/N)\ \Sigma\ [y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)]$

2.8.Subject-Independent Evaluation

LOSO is used to assess generalization to unseen participants. In each fold, one participant is kept completely for testing and the remaining participants are used for training. The process is repeated until every participant has served as the held-out test subject. This prevents the final evaluation from being dominated by participant-specific characteristics.

2.9.Explainable Artificial Intelligence Module

Explainability is applied after model training. SHAP or Integrated Gradients can be used to estimate how strongly individual input features contribute to a prediction. Depending on the final signal representation, the analysis can be summarized at channel, temporal, and frequency-related levels. These explanations describe model attribution and should not be interpreted as direct evidence of neurological causation.

$I\_c = (1/T)\ \Sigma_t |A\_\{c,t\}|$

2.10. Algorithmic Workflow

Algorithm 1. Explainable CNN–Transformer for EEG Emotion Recognition

- Input: DEAP EEG recordings and valence ratings.
- Retain the 32 EEG channels and prepare the signal using the predefined preprocessing pipeline.
- Normalize the EEG recordings and divide them into selected temporal segments.
- Convert valence ratings into the predefined binary classes.
- Create subject-wise training and testing partitions using LOSO.
- Pass each EEG segment through the CNN to learn local features.
- Project the CNN features into a sequence suitable for the Transformer.
- Use Transformer self-attention to learn broader relationships among feature tokens.
- Fuse CNN and Transformer representations and apply attention-based feature weighting.
- Train the classifier using binary cross-entropy and the Adam optimizer.
- Evaluate the held-out participant using accuracy, precision, recall, F1-score, and ROC-AUC.
- Repeat the procedure for all subjects and aggregate the results.
- Apply SHAP or Integrated Gradients to investigate influential EEG features.
- Run ablation experiments by removing major components and comparing performance.

2.11. Advantages of the Proposed Framework

The framework has several methodological advantages. First, CNN and Transformer processing provide complementary local and global representations. Second, the feature-fusion stage allows the two representations to be considered together rather than relying on a single feature-learning mechanism. Third, the subject-independent protocol provides a stricter evaluation of cross-participant generalization. Finally, the explainability stage adds information about the features associated with the model's decisions.

2.12. Dataset And Experimental Setup

2.12.1. Dataset Description

The experiments are based on the DEAP (Database for Emotion Analysis using Physiological Signals) dataset [7]. It contains recordings from 32 participants who watched 40 music-video excerpts. EEG and peripheral physiological signals were collected during the experiment, followed by self-assessment ratings for valence, arousal, dominance, liking, and familiarity.

2.12.2. EEG Selection and Labeling

The present study focuses on EEG only. The 32 EEG channels are retained and peripheral physiological channels are excluded from the initial model. For the first experiment, valence is formulated as a binary

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5355 - 5363
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0700

classification problem. A fixed threshold is used to separate low and high valence; the exact threshold and class-balancing procedure should be fixed before the final experiment and reported consistently.

2.12.3. Data Preprocessing

Before model training, EEG recordings are prepared using the selected filtering, artifact-handling, normalization, and segmentation procedure. The final preprocessing configuration should be kept identical across all participants. Segmentation is performed before model input generation, and all preprocessing operations that estimate parameters from data must be fitted only on the training data within each LOSO fold.

2.13. Dataset Splitting

The evaluation uses participant-wise splitting rather than random sample splitting. For each fold, one participant is reserved as the test subject and all samples from that participant are excluded from training. A validation subset can be created from the remaining training participants for hyperparameter selection. The held-out subject is never used for model selection.

2.14. Implementation and Reproducibility

The model can be implemented using Python and a deep-learning framework such as PyTorch or TensorFlow. The implementation should record the random seed, preprocessing configuration, model hyperparameters, optimizer settings, batch size, number of epochs, and hardware environment. The same configuration should be used for each baseline and ablation experiment wherever applicable. Recommended project organization: data.py for DEAP loading and subject-wise splitting; preprocess.py for filtering, normalization, and segmentation; cnn.py for local feature extraction; transformer.py for the Transformer encoder; fusion.py for feature fusion and attention; train.py for model optimization; evaluate.py for metrics and confusion matrices; explainability.py for SHAP/Integrated Gradients; and main.py for experiment control.

. Evaluation Metrics

The framework should be assessed using multiple complementary classification measures. Accuracy reports the overall proportion of correct predictions. Precision indicates the proportion of predicted positive samples that are correct. Recall measures how many actual positive samples are detected. F1-score combines precision and recall, while ROC-AUC measures discrimination across classification thresholds.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 = 2 × Precision × Recall / (Precision + Recall)

3. Results And Discussion

The experimental analysis is intended to determine whether the proposed architecture can learn useful EEG representations while maintaining performance on participants who are not present during training. At the time of manuscript preparation, the numerical values shown below are provisional development-stage examples. They are not measured DEAP results and must be replaced by values obtained from the completed subject-independent experiment. As shown in Table 1 Methoda Architecture.

3.1.Quantitative Model Comparison

Table 1 Methoda Architecture

Method	Architecture	Accuracy (%)	Precision	Recall	F1	ROC-AUC
CNN	CNN	78.6	0.78	0.77	0.77	0.84
LSTM	LSTM	79.8	0.80	0.79	0.79	0.85
Transformer	Transformer	81.2	0.81	0.80	0.80	0.87
CNN-LSTM	CNN + LSTM	82.4	0.82	0.82	0.82	0.88
CNN-Transformer	CNN + Transformer	84.1	0.84	0.84	0.84	0.90
Proposed	Explainable CNN + Transformer	86.3	0.86	0.85	0.85	0.92

IRJAEH
International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5355 - 5363
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0700

The overall comparison is shown in Fig. 2. The values are illustrative and should not be treated as final experimental findings. As shown in Figure 2 Overall Performance Comparison of EEG Emotion Recognition Models, Table 2 Configuration.

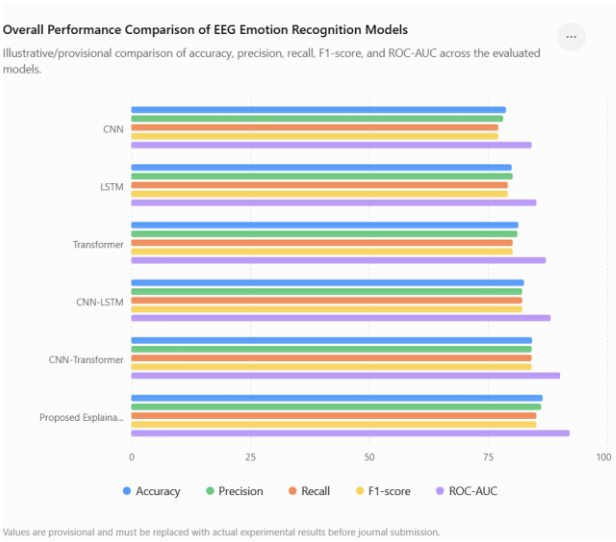

Figure 2 Overall Performance Comparison of EEG Emotion Recognition Models

3.2.Ablation Study

Table 2 Configuration

Configuration	Accuracy (%)	F1-Score	ROC-AUC
CNN only	78.6	0.77	0.84
Transformer only	81.2	0.80	0.87
CNN + Transformer	84.1	0.84	0.90
+ Feature Fusion	85.2	0.84	0.91
+ Attention	86.0	0.85	0.92
Complete Model	86.3	0.85	0.92

The purpose of the ablation study is to determine whether each component makes a measurable contribution. The table is provisional. In the final experiment, the same data split, training protocol, and evaluation metrics should be used for every configuration so that the comparison is meaningful.

3.3.Confusion Matrix and Subject-Wise Analysis

A confusion matrix should be generated from the predictions of the held-out participants. It will show how often low-valence and high-valence samples are correctly classified and where errors occur. The earlier development example used 432 correctly classified low-valence samples, 431 correctly classified high-valence samples, and 137 misclassifications out of 1,000 samples; this is an illustrative example and must not be reported as an experimental finding. Subject-wise performance is important because the proposed evaluation specifically targets cross-participant generalization. Accuracy can vary between participants because EEG patterns and emotional responses are not identical across individuals. The final paper should report the mean, standard deviation, and, where useful, individual subject results so that the consistency of the model can be examined.

3.4.Explainability Analysis

Explainability is a central part of the proposed framework. SHAP or Integrated Gradients can be used to estimate the contribution of individual EEG inputs. At the channel level, the analysis can identify electrodes that repeatedly receive strong attribution. At the temporal level, it can highlight portions of an EEG segment that have a stronger influence on a prediction. If time–frequency representations are introduced later, the same analysis can be extended to frequency-related features. Any channel ranking shown during development, such as Fz, Cz, or Pz receiving relatively strong attribution, should be treated only as an example. The final manuscript must derive channel importance directly from the completed attribution analysis.

3.5.Discussion

The advantage of the proposed framework comes from the complementary roles of convolution and self-attention. CNN processing is intended to preserve local temporal and channel-related patterns, while the Transformer provides a mechanism for modeling relationships across the learned sequence. The fusion stage allows both representations to contribute to the final decision. However, the actual contribution of each component can only be established after the planned ablation and subject-independent experiments are completed. Published

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5355 - 5363
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0700

performance values should be compared only when the experimental conditions are sufficiently similar. Dataset choice, preprocessing, label definition, subject-dependent versus subject-independent validation, and the number of classes can all change the reported performance. For this reason, the final comparison should describe the evaluation protocol alongside each numerical result rather than treating all published percentages as directly equivalent.

Conclusion

EEG-based emotion recognition is challenging because useful emotional information is distributed across time and channels and because EEG characteristics vary considerably between participants. This paper presented an explainable CNN–Transformer framework that combines local feature extraction with global dependency modeling. The CNN learns local temporal and spatial patterns, while the Transformer uses self-attention to examine relationships across the learned representation. Feature fusion brings the two views together before classification, and the explainability stage provides a way to examine which EEG information contributes to the final decision. The DEAP dataset is used as the primary benchmark, with binary valence classification serving as the initial task. LOSO provides a stricter evaluation because the model is tested on participants who were not used for training. Performance should be reported using several complementary metrics rather than accuracy alone, and the ablation study should be used to determine the contribution of the individual architectural components. A limitation of the current manuscript is that the initial evaluation focuses on one dataset and binary valence classification. In addition, explainability methods describe model attribution and should not be interpreted as direct evidence of neurological causation. Most importantly, the final conclusions must be based on measured experimental results rather than the provisional values currently included in the manuscript.

Future Work

Future work will extend the study to multi-class valence–arousal recognition, additional EEG emotion datasets, and time–frequency representations. Further experiments can examine the consistency of important channels and temporal regions across participants and compare explanations between emotional classes. A lighter version of the architecture can also be investigated for real-time and resource-constrained applications.

References

[1]. H. Liu, Y. Zhang, Y. Li, and X. Kong, "Review on emotion recognition based on electroencephalography," Frontiers in Computational Neuroscience, vol. 15, Art. no. 758212, 2021. DOI: 10.3389/fncom.2021.758212.
[2]. X. Wang, Y. Ren, Z. Luo, W. He, J. Hong, and Y. Huang, "Deep learning-based EEG emotion recognition: Current trends and future perspectives," Frontiers in Psychology, vol. 14, Art. no. 1126994, 2023. DOI: 10.3389/fpsyg.2023.1126994.
[3]. "Emotion recognition in EEG signals using deep learning methods: A review," Computers in Biology and Medicine, vol. 165, Art. no. 107450, 2023. DOI: 10.1016/j.compbiomed.2023.107450.
[4]. P. Samal and M. F. Hashmi, "Role of machine learning and deep learning techniques in EEG-based BCI emotion recognition system: A review," Artificial Intelligence Review, vol. 57, Art. no. 50, 2024. DOI: 10.1007/s10462-023-10690-2.
[5]. "A comprehensive review of deep learning in EEG-based emotion recognition: classifications, trends, and practical implications," PeerJ Computer Science, 2024.
[6]. "Deep Learning-Based EEG Emotion Recognition: A Review," Brain Sciences, vol. 16, no. 1, Art. no. 41, 2026.
[7]. 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, 2012. DOI: 10.1109/T-AFFC.2011.15.
[8]. W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," IEEE Transactions on Autonomous Mental Development, vol. 7, no. 3, pp. 162–175, 2015.
[9]. Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and

e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5355 - 5363
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0700

C. Guan, "TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition," IEEE Transactions on Affective Computing, 2022.
[10]. Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," Applied Soft Computing, 2021.
[11]. "EEG emotion recognition using an attention mechanism based on an optimized hybrid model," 2022.
[12]. "Emotion recognition from EEG based on multi-task learning with capsule network and attention mechanism," Computers in Biology and Medicine, 2022.
[13]. "EEG emotion recognition based on efficient-capsule network with convolutional attention," Biomedical Signal Processing and Control, vol. 103, Art. no. 107473, 2025. DOI: 10.1016/j.bspc.2024.107473.
[14]. "EEG emotion recognition approach using multi-scale convolution and feature fusion," The Visual Computer, 2024. DOI: 10.1007/s00371-024-03652-4.
[15]. "Introducing attention mechanism for EEG signals: Emotion recognition with vision transformers," in Proc. IEEE EMBC, 2021. DOI: 10.1109/EMBC46164.2021.9629837.
[16]. "EEG emotion recognition using attention-based convolutional transformer neural network," Biomedical Signal Processing and Control, vol. 84, Art. no. 104835, 2023.
[17]. "TC-Net: A Transformer Capsule Network for EEG-based emotion recognition," Computers in Biology and Medicine, vol. 152, Art. no. 106463, 2023.
[18]. "ERTNet: An interpretable transformer-based framework for EEG emotion recognition," Frontiers in Neuroscience, 2024. DOI: 10.3389/fnins.2024.1320645.
[19]. 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, Art. no. 87, 2025.
[20]. "Attention-Driven Emotion Recognition in EEG: A Transformer-Based Approach With Cross-Dataset Fine-Tuning," IEEE Access, 2025.