AUT Journal of Electrical Engineering
10.22060/EEJ.2026.25904.6024

Optimized Attention-based Deep Learning for High-Accuracy EEG Emotion Classification: A Comprehensive Evaluation

Seyedeh Zahra Nabavi[1], Mohsen Sardari Zarchi[2]*, Hossein Ghaneai[3], Fatemeh Zare Mehrjardi[4]

[1]Master Student, Department of Computer Engineering, University of Meybod, Yazd, Iran
[2]Associate Professor, Department of Computer Engineering, University of Meybod, Yazd, Iran
[3]Assistant Professor , Department of Computer Engineering, University of Meybod, Yazd, Iran
[4]Assistant Professor , Department of Computer Engineering, University of Meybod, Yazd, Iran
*Corresponding Author, sardari@meybod.ac.ir

Uncorrected Proof

Abstract

Emotion recognition from electroencephalogram (EEG) signals is limited by scarce data, high inter-subject variability and noisy recordings. We propose CNN-BiGRU-Attention, a compact hybrid model (381,447 parameters) combining multi-scale convolutions, a bidirectional GRU and multi-head attention pooling, preceded by Z-score normalization, ANOVA F-score feature selection and PCA fitted on training data only. A grid search over 108 hyperparameter combinations on Dataset 1 (three classes: positive, negative, neutral) was run for six architectures under identical conditions, and the best model was compared with a reimplemented GANN baseline. On Dataset 1 (two participants, so within-subject evaluation), CNN-BiGRU-Attention reached 97.09 ± 0.52% accuracy in five-fold cross-validation (98.59% on a fixed 80:10:10 split), 4.06 and 4.36 percentage points above GANN, with a train–test gap (1.23%) comparable to GANN's (1.52%). It had the highest mean accuracy and lowest fold-to-fold variation of the six architectures, the gain over GANN was statistically significant (corrected paired test), whereas differences from the other recurrent models were not. Because Dataset 1 supports only within-subject evaluation, subject-independent generalization was tested with leave-one-subject-out (LOSO) validation on Dataset 2 (20 participants, four valence–arousal classes): the proposed model reached 92.37% accuracy, 6.03 points above GANN, but below its 98.55% fixed-split accuracy on that dataset. An ablation study showed that the convolutional front end, feature selection and batch normalization contribute most. The model needs about 16 M MACs per inference, suggesting potential for real-time brain–computer interfaces; latency was assessed only on a laptop CPU and emulated resource-constrained profiles.

10.22060/EEJ.2026.25904.6024

Keywords:

Electroencephalogram Signals, Emotion Recognition, Hyperparameter Optimization, CNN-BiGRU-Attention, subject-independent evaluation (LOSO)

1- Introduction

Emotion recognition from Electroencephalogram (EEG) signals represents one of the most demanding intersections of neuroscience, biomedical signal processing, and machine learning. Unlike behavioural indicators such as facial expressions or vocal patterns, EEG signals provide direct access to the underlying neural correlates of emotional states, offering a level of objectivity and temporal resolution that peripheral physiological measures cannot match [1]. The non-invasive nature of EEG and its millisecond-level temporal precision have established it as a cornerstone modality in affective brain-computer interface (BCI) research [2], with widespread applications in human-computer interaction, mental health monitoring, and assistive technologies [3]. However, this advantage comes with considerable practical challenges. The structural complexity of the human brain, combined with the strict requirements for aligning signal acquisition during specific emotional episodes, imposes severe constraints on the volume and quality of collectible data [4] [5]. These constraints fundamentally limit the capacity of conventional deep learning approaches, which typically depend on large, well-labelled datasets to achieve reliable generalisation[6, 7].

Given these constraints, we prioritise architectural efficiency over data augmentation. Rather than expanding the training set, we design compact architectures capable of achieving high classification accuracy under limited data conditions. The primary objective of this study is to develop a lightweight, robust model that achieves high classification accuracy under conditions of limited training data and noisy EEG recordings.

Previous studies in this field have demonstrated the considerable promise of recurrent architectures for temporal EEG modelling. Recurrent Neural Networks (RNN), Long Short-Term Memory networks (LSTM), and Gated Recurrent Units (GRU) have each shown significant capabilities in classifying EEG signals [8-11]. A particularly relevant line of work is the bidirectional extension of these models:

Uncorrected Proof

bidirectional LSTM architectures capture both forward and backward temporal dependencies in EEG data, yielding measurable gains over their unidirectional counterparts[12, 13]. The integration of multi-head attention mechanisms with these recurrent architectures has brought further improvement in model performance, as these mechanisms allow the model to assign differentiated weights to different input components based on their relative importance, going beyond simple sequential processing [14, 15]. More recently, transformer-based architectures have also been applied directly to EEG emotion recognition, exploiting their capacity to model long-range temporal and spatial dependencies without the sequential bottleneck of recurrent models [16-18]. Despite the progress achieved across these paradigms, the complex interaction between model hyperparameters, optimal attention mechanism design, and effective feature extraction strategies still requires deeper and more comprehensive investigation. This research gap constitutes the starting point of the present study.

A central challenge in this domain is the availability of standardised benchmarks. The Database for Emotion Analysis using Physiological Signals (DEAP) [3] and the SJTU Emotion EEG Dataset (SEED) [19] have emerged as the most widely adopted evaluation platforms, enabling systematic cross-study comparison. Both datasets have been used extensively to evaluate models spanning traditional machine learning, convolutional architectures, graph neural networks, and hybrid deep learning pipelines [20]. Neurophysiological evidence further informs feature engineering choices: frontal EEG asymmetry, for instance, has been established as a reliable moderator of emotional states [21], motivating spatially aware model designs that attend differentially to electrode regions.

Several recent works have shaped the direction of the present study. Peixiang Zhong and colleagues [22] (2020) introduced a Regularised Graph Neural Network for EEG-based emotion recognition, designed with inspiration from the biological topology of the human brain. Their model employs a low-density adjacency matrix capturing both local and global inter-channel dependencies, and incorporates two regularisation components: a node-level dual-attention module to mitigate inter-subject variability, and a label-aware learning strategy to handle noisy emotional annotations. Experimental results on the DEAP and SEED datasets demonstrated superior performance over prior methods and identified frontal

Uncorrected Proof

and posterior regions as key contributors to emotion processing. However, the model requires a dense electrode configuration and may suffer from over-smoothing on sparse graphs, with the authors recommending simplified variants for lower-density setups and more advanced domain adaptation for cross-subject generalisation. Tong Zhang and colleagues [23](2019) introduced a Spatial-Temporal Recurrent Neural Network framework for emotion recognition from EEG signals and facial videos. Their model comprises two RNN layers: one capturing spatial dependencies at each time step and one modelling temporal dynamics across time steps. Although this approach achieved strong results on DEAP and the Extended Cohn–Kanade (CK+) database, it remains sensitive to EEG noise and incurs high computational cost. Building on these spatial-temporal ideas, more recent architectures have explored convolutional-recurrent hybrids combined with attention [24], multi-scale feature fusion [12], and transformer-based spatial-temporal encoders[18]. Complementary work on multi-head attention interpretability [15, 25] has further refined our understanding of how attention heads distribute their focus across temporal segments and frequency bands, though this analysis remains underexplored in the EEG emotion context specifically.

Hybrid models that combine convolutional feature extraction, bidirectional recurrence and multi-head attention are an active direction in EEG emotion recognition. CBSAtt [26] combines a CNN-BiLSTM network with multi-head self-attention and is the closest architectural relative of our model. Our comparison between CNN-BiGRU-Attention and CNN-BiLSTM-Attention (Tables 5 and 7) tests the same recurrent-gating choice under identical conditions. Two further works [27], developed for fault diagnosis rather than EEG, address limitations that our pipeline also faces. Reference [27] uses time–frequency fusion with latent-representation prompts for open-set recognition; our model instead takes pre-extracted spectral features. Reference [28] uses wavelet-packet augmentation and pseudo-domain generation for single-source domain generalization, which is relevant because our accuracy falls from 98.55% (sample-level) to 92.37% (LOSO) on Dataset 2.

The present study is designed to analyse and address this gap. We propose a hybrid architecture that integrates bidirectional recurrent layers with a multi-head attention mechanism, optimised for the dual

Uncorrected Proof

objectives of computational efficiency and classification robustness under data-limited conditions. Beyond proposing a specific model, this work conducts a structured investigation into the sensitivity of classification performance to a range of architectural and preprocessing decisions, including hidden layer size, dropout rate, feature selection techniques, and dimensionality reduction methods. In doing so, we aim to contribute not only a high-performing and lightweight model, but a systematic and transferable understanding of how architectural choices interact with data constraints in EEG-based emotion recognition, offering both a practical tool and a principled framework to the broader research community.

Despite these advances, four gaps remain underexplored. First, the relative contribution of bidirectionality versus gating mechanism complexity (GRU vs. LSTM) has not been systematically isolated in EEG emotion tasks [7]. Second, prior work rarely combines systematic hyperparameter search with architectural comparison under identical experimental conditions; the importance of subject-specific parameter tuning has been noted [13], yet controlled cross-architecture comparisons under uniform conditions remain scarce. Third, the attention mechanism is often treated as a black box without analysis of which temporal segments or frequency features each head focuses on. Fourth, prior work on this benchmark has been evaluated at the sample level, without subject-independent evaluation; we therefore complement the main comparison with a leave-one-subject-out analysis on a multi-subject dataset. This study directly addresses all four gaps.

2- Proposed Approach

2-1- Database

This research utilized the EEG Brain Wave Dataset[1] [29]: Feeling Emotions database for three-class emotion classification (negative, neutral, positive) for Dataset 1 based on EEG signals. This dataset, first introduced by Bird and colleagues in 2018, comprises 2,132 records of EEG signal data and includes 2,549 features extracted from brain signals. We used this dataset as the Dataset 1 that had been

[1]EEG Brain Wave Dataset: https://www.kaggle.com/datasets/birdy654/eeg-brainwave-dataset-feeling-emotions

Uncorrected Proof

used in the base research paper. For better comparison, we also incorporated a Comparative EEG-BCI dataset[2] [5] as Dataset 2, which expands the scope of our analysis by including data from two distinct EEG recording systems: a high-end professional EEG system (BrainVision) and a low-cost brain-computer interface (Emotiv EPOC+). Dataset 2 contains EEG signals from 20 participants recorded during an emotion classification experiment with visual stimuli from the International Affective Picture System (IAPS), designed to induce four emotional states (HVHA, HVLA, LVHA, LVLA) based on Russell's valence–arousal model. Accordingly, Dataset 1 is a three-class problem and Dataset 2 is a four-class problem. The two datasets serve different purposes: Dataset 1 is the main benchmark for comparison with the baseline GANN model , whereas Dataset 2 is used to assess subject-independent generalization, since it provides participant identifiers. Similar comparative approaches using multiple EEG systems have been explored in recent studies [4].

2-1-1- Data Collection Method

Data in this collection were gathered from two individuals (one male and one female) over three minutes for each emotional state. The Muse EEG headband was used to record EEG signals, capable of capturing signals at four positions (TP9, AF7, AF8, and TP10) through dry electrodes. In addition to the three main emotional states, six minutes of neutral resting state data were also recorded. To induce different emotions, diverse audio-visual content was used: emotional film clips were presented visually and emotional stories audibly, while brain activity was recorded. The second dataset was gathered from 20 participants, and the data collection approach employed both high-fidelity and consumer-grade EEG devices, allowing for a robust comparison of signal quality and usability across different hardware platforms. This dataset also leverages the IAPS, a widely used standardized stimulus set, to ensure consistency and ecological validity in emotional state induction.

2-1-2- Preprocessing and Initial Extraction

[2] EEG-BCI dataset: https://zenodo.org/records/14787743

Uncorrected Proof

10.22060/EEJ.2026.25904.6024

The selected database provides pre-extracted features using Welch's periodogram-based power spectral density estimation across standard EEG frequency bands (delta: 0.5-4 Hz, theta: 4-8 Hz, alpha: 8-13 Hz, beta: 13-30 Hz, gamma: >30 Hz), alongside time-domain statistical features (mean, variance, skewness, kurtosis). A Hanning window of 256 samples with 50% overlap was applied during spectral estimation. These features cover a wide range of signal characteristics in temporal, frequency, and time-frequency domains. Cross-validation was implemented to ensure the robustness of our findings, with a stratified five-fold split applied to Dataset 1 (sample-level). Dataset 2 was evaluated with a fixed 80:10:10 split and with LOSO (Section 4-2-1). The training, validation, and test sets were divided in an 80:10:10 ratio, where 80% of the data was used for training, and the remaining 20% was further split into validation and test sets. Specifically, four folds were allocated for training, while one-fold was divided equally between validation and testing to prevent data leakage and ensure generalizability of the model performance across both datasets. To prevent data leakage, all preprocessing transformations including Z-score normalization, ANOVA F-score based feature selection, and PCA were fitted exclusively on the training fold of each cross-validation split and subsequently applied to validation and test folds using the parameters learned from training data only. This ensures that no information from held-out data influenced the feature extraction process. In our pipeline, Z-score normalization is applied per feature across the training set, followed by ANOVA F-score based SelectKBest (k=1024) and PCA dimensionality reduction (n=64 components, explaining ~95% of variance). Dataset 1 was recorded from only two participants and contains no subject identifiers. Its five-fold evaluation was therefore performed at the sample level with stratification. Because temporally adjacent samples from the same session may fall into both training and test folds, results on Dataset 1 reflect within-subject performance and should not be read as evidence of generalization to unseen individuals. To address this, we conducted a subject-independent leave-one-subject-out (LOSO) evaluation on Dataset 2, in which all data from the held-out participant are excluded from training, validation, and preprocessing (Section 4-2-1).

Uncorrected Proof

2-2- Architecture

The proposed architecture of this research has been developed based on results from the baseline study that was conducted on the same dataset using a bidirectional GRU network with an attention mechanism. In the baseline research, 94% accuracy on test data was achieved using the Adam optimizer and an 80%-10%-10% data split for training, validation, and testing. We extend the baseline GANN by adding CNN layers for spatial feature extraction and testing different GRU configurations. For the Preprocessing section, we use a multi-preprocessing method to optimise the results, and the best results are achieved by applying the following steps presented in Table 1. Our baseline extends the GANN architecture [8] by incorporating convolutional layers for spatial feature extraction, following recent findings that CNN-RNN hybrid architectures outperform standalone recurrent networks [14, 30].

Table 1. Data preprocessing steps

Steps	Preprocess	Description
1	Normalization	First, Z-score normalization is applied to each feature, giving zero mean and unit variance (statistics computed on the training portion only)
2	ANOVA F-score based Feature Selection	Second, Features are ranked by their ANOVA F-score with respect to the target labels, and the top K features are kept (SelectKBest, K = 1024)
3	Dimensionality Reduction with PCA	Third, Principal component analysis, an unsupervised variance-based projection, reduces the selected features to 64 components that retain about 95% of their variance.

These preprocessing steps follow established best practices in EEG signal analysis [9] and have been shown to improve model convergence and classification accuracy [31]. The proposed model is based on several parts, as it shown in Figure 1 and Table 2, Convolutional layers extract local and hierarchical spatial features from the input. Batch normalization stabilizes layer-wise activation distributions, accelerating convergence and reducing sensitivity to weight initialization. Bidirectional GRUs capture sequential dependencies in both forward and backward directions, enhancing context understanding in time-series or sequential data, and multi-head attention layers enable the model to dynamically focus

Uncorrected Proof

on relevant parts of the input, improving interpretability and performance. Dropout regularization stochastically zeroes neuron activations during training, reducing overfitting, while feedforward layers provide non-linear transformations to refine learned representations. Finally, classification layers map these refined features to the desired output, ensuring the model delivers accurate and meaningful predictions.

2-2-1- Convolutional and recurrent layers

The input is a vector of 64 PCA components, treated as a sequence of length T = 64 with one channel. Three parallel one-dimensional convolutions extract local patterns, and their outputs are concatenated and normalized:

$$F_k = ReLu\left(W_k * X + b_k\right),\ k \in \{3,5,7\} \quad \text{Eq. (1)}$$

$$F = BN\left(F_3; F_5; F_7\right) \quad \text{Eq. (2)}$$

Here X ∈ R^(1×64) is the input, W_k and b_k are the weights (64 filters of width k) and bias of the convolution with kernel size k, * is one-dimensional convolution with "same" padding, [;] is concatenation along the channel axis, BN is batch normalization, and F ∈ R^(192×64). Column t of F, written x_t ∈ R^192, is the input of the recurrent layer at position t. The GRU updates its hidden state h_t as follows:

$$r_t = \sigma\left(W_{ir}x_t + b_{ir} + W_{hr}h_{t-1} + b_{hr}\right) \quad \text{Eq. (3)}$$

$$z_t = \sigma\left(W_{iz}x_t + b_{iz} + W_{hz}h_{t-1} + b_{hz}\right) \quad \text{Eq. (4)}$$

$$n_t = tanh\left(W_{in}x_t + b_{in} + r_t \odot \left(W_{hn}h_{t-1} + b_{hn}\right)\right) \quad \text{Eq. (5)}$$

$$h_t = \left(\left(1 - z_t\right) \odot n_t + z_t \odot h_{t-1}\right) \quad \text{Eq. (6)}$$

Here r_t is the reset gate, z_t the update gate, n_t the candidate state, σ the logistic sigmoid, ⊙ element-wise multiplication, and the W and b terms are learned weights and biases. The bidirectional GRU runs one such layer forward and one backward over the 64 positions and concatenates their states:

Uncorrected Proof

10.22060/EEJ.2026.25904.6024

$$h_t = \left[h_{t_1}; h_{t_2} \right] \in \mathbb{R}^{256} \quad \text{Eq. (7)}$$

Each direction has 128 hidden units (Table 3).

2-2-2- Multi-Head Attention Pooling

The model summarizes the 64 BiGRU outputs with M = 4 attention heads. Each head m has its own learned scoring vector w_m ∈ R^256 and bias b_m, scores every position, normalizes the scores over positions, and returns a weighted sum of the BiGRU outputs:

$$e_t^{(m)} = w_m{}^T h_t + b_m \quad \text{Eq. (8)}$$

$$a_t^{(m)} = \frac{\exp\left(e_t^{(m)}\right)}{\sum_{t=1}^{T} a_t^{(m)} h_t} \quad \text{Eq. (9)}$$

$$e^{(m)} = \sum_{t=1}^{T} a_t^{(m)} h_t \quad \text{Eq. (10)}$$

$$c = \left[e^{(1)}; \ldots; e^{(4)} \right] \in \mathbb{R}^{1024} \quad \text{Eq. (11)}$$

- **Parallel Processing**: Multiple attention heads perform computations simultaneously
- **Selective Focus**: Each head focuses on specific features of the data
- **Learned aggregation:** Outputs from all heads are concatenated and linearly projected to form the final representation.
- **Richer representations:** The multi-head structure captures diverse feature interactions that a single attention head cannot.

Here e_t^(m) is the score of head m at position t, α_t^(m) is its attention weight (the weights of one head sum to 1 over t), c^(m) ∈ R^256 is the summary vector of head m, and c is the concatenation of all four. The heads share the BiGRU outputs and differ only in their scoring vectors, so each head can attend to different positions. This is a lightweight form of multi-head attention, simpler than the query–key–value attention of transformers. The concatenated vector passes through dropout and two fully connected layers:

Uncorrected Proof

$$\hat{y} = softmax\left(W_2 ReLU\left(W_1 Dropout\left(c\right)+b_1\right)+b_2\right) \quad \text{Eq. (12)}$$

Here W_1 ∈ R^(128×1024), W_2 ∈ R^(C×128), b_1 and b_2 are learned weights and biases, and C is the number of classes (3 for Dataset 1, 4 for Dataset 2).

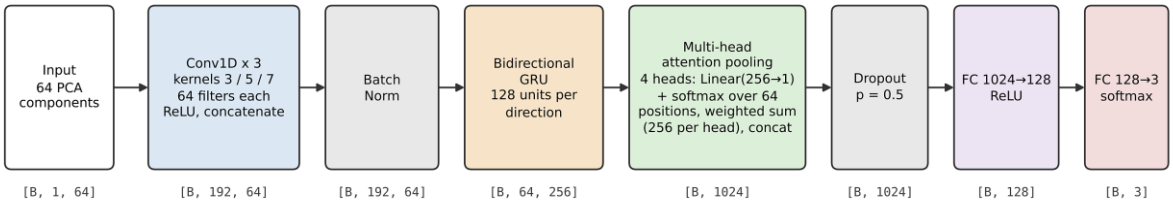


Fig. 1. Layers of the proposed CNN-BiGRU-Attention model for the three-class Dataset 1, with the output shape of each layer (B = batch size). For the four-class Dataset 2 the last layer is 128→4.

Table 2. CNN-BiGRU-Attention Architecture

Layer	**Details**	**Output Shape**	**Parameters**
Input	64 PCA components	[B, 1, 64]	-
Conv1D × 3	kernels 3, 5, 7 - 64 filters each padding "same"- ReLU; concatenated	[B, 192, 64]	1,152
BatchNorm1D	192 channels	[B, 192, 64]	384
Bidirectional GRU	hidden = 128 per direction	[B, 64, 256]	247,296
Multi-Head Attention × 4	Linear(256→1) + softmax over positions, concatenated	[B, 1024]	1,028
Dropout	p = 0.5	[B, 1024]	-
FC Layer 1	1024→128, ReLU	[B, 128]	131,200
FC Layer 2	128→3, softmax	[B, 3]	387
Total Parameters			381,447

Table 2 shows the configuration for the three-class Dataset 1. For the four-class Dataset 2, the final layer is 128 → 4, and Total Parameters are (381,447 + 129) for the Dataset 2. all other layers are unchanged.

Uncorrected Proof

10.22060/EEJ.2026.25904.6024

3- Research Method

Hyperparameters were optimized by a grid search over 108 combinations of the five hyperparameters in Table 3 (2 × 3 × 3 × 3 × 2). Each combination was trained for 10 epochs on the fixed 80:10:10 split of Dataset 1 and ranked by accuracy on the validation set. The best configuration of each architecture (Table 3b) was then used unchanged in the five-fold experiments on Dataset 1 and in the LOSO experiments on Dataset 2, with no new search (Fig. 1b). All models used the Adam optimizer, cross-entropy loss, a batch size of 32 and a fixed random seed of 42 (data splits, weight initialization and shuffling). Two training protocols were used. (i) Comparison protocol: the best configuration of each architecture was trained for a fixed 40 epochs without early stopping, because the authors of the GANN baseline [8] report their best accuracy under this setting; all architectures and the baseline were evaluated under it, and Tables 5–7 and Figures 2–3 are based on it. (ii) Convergence protocol: to verify that 40 epochs is adequate and to visualise convergence, Figure 4 was produced with up to 100 epochs and early stopping (patience = 5, min_delta = 0.001). Early stopping occurred between epochs 38 and 55 for the five CNN-based models and at epoch 76 for GNN-Attention, i.e., close to the 40 epochs used in the comparison protocol. The grid search was performed on Dataset 1 only. The optimal configuration was then applied unchanged to Dataset 2, without re-tuning, so that Dataset 2 serves as an out-of-search test of how well the selected configuration transfers to a different recording system, label structure, and set of participants. The GANN baseline was reproduced with the hyperparameters of the original paper (hidden size 256, learning rate 0.001, dropout 0.5) and the same K (1024) and PCA (64) as the proposed model. A grid search run on GANN as an additional experiment returned the same values. Similar hyperparameter optimization strategies have been employed successfully in EEG emotion recognition [31], though our systematic evaluation of 108 combinations across multiple architectures provides more comprehensive coverage than previous work.

Uncorrected Proof

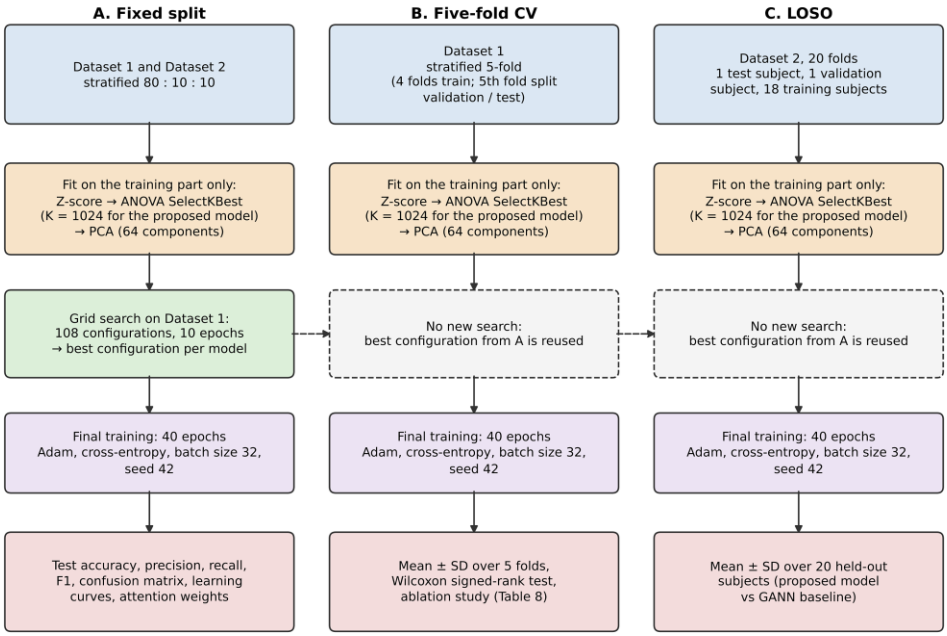


Fig. 1b. Research method: three evaluation protocols, each fitting preprocessing on the training part only. The grid search is run once on Dataset 1 and its best configuration is reused in the other two protocols.

Table 3. List of Hyperparameters

Hyperparameter	**Description**	**Value List**	**Optimal Value**
Learning Rate	A parameter that determines the extent of impact of model weight updates in each training iteration.	0.001, 0.0005	0.0005
Number of K Features	This parameter specifies the number of features that are selected based on ANOVA F-score with target labels.	256, 512, 1024	1024
Number of PCA Components	A parameter that determines the number of selected components in the dimensionality reduction stage based on maximum variance.	64, 128, 256	64
GRU/LSTM Hidden Layers	This hyperparameter controls the depth of the Model network and the ability to learn complex patterns.	64, 128, 256	128

Uncorrected Proof

Dropout Rate	A parameter that specifies the percentage of neurons randomly removed during training and prevents overfitting.	0.3, 0.5	0.5

Table 3b. List of Hyperparameters

Hyperparameter	**Learning rate**	**K**	**PCA**	**Hidden units**	**Dropout**
CNN-BiGRU-Attention	0.0005	1024	64	128	0.5
CNN-BiLSTM-Attention	0.001	1024	64	256	0.5
CNN-GRU-Attention	0.0005	1024	128	256	0.5
CNN-LSTM-Attention	0.0005	512	64	128	0.3
CNN-BiGRU-PatchTST	0.001	512	64	n/a	0.5
GNN-Attention	0.001	256	64	n/a	0.3
GANN (baseline)	0.001	1024	64	256	0.5

The hidden-unit parameter applies only to GRU/LSTM layers. CNN-BiGRU-PatchTST and GNN-Attention have no such parameter, so the hidden-unit setting does not apply to these two models, and their networks were identical across all hidden-unit settings. To validate the statistical reliability of our results, a five-fold cross-validation was performed with identical random seed (42), and Wilcoxon signed-rank tests were conducted comparing CNN-BiGRU-Attention against all other architectures. The CNN-BiGRU-Attention model achieved a cross-validation accuracy of 97.09% ± 0.52%, demonstrating consistent performance across folds. CNN-BiLSTM-Attention achieved 96.95% ± 0.59% ($p = 0.593$), CNN-GRU-Attention achieved 96.20% ± 0.67% ($p = 0.063$), CNN-LSTM-Attention achieved 94.89% ± 1.07% ($p = 0.063$), CNN-BiGRU-PatchTST achieved 94.65% ± 1.14% ($p = 0.063$),

Uncorrected Proof

and GNN-Attention achieved 78.89% ± 3.47% (p = 0.063). With five folds, the smallest p-value the exact two-sided Wilcoxon test can reach is 0.0625, so no comparison can fall below 0.05. A p-value of 0.063 means that the proposed model scored higher than that alternative in all five folds; for CNN-BiLSTM-Attention (p = 0.593) the two models cannot be distinguished. We therefore describe the advantage over the other recurrent models and PatchTST as consistent but not statistically established, with the lowest fold-to-fold variation (± 0.52%), whereas the advantage over GANN and GNN-Attention is statistically supported (see below). Notably, our proposed model improved on GANN by 4.06 percentage points in five-fold accuracy and by 4.36 percentage points in top test accuracy. In addition, we applied the corrected resampled paired t-test of Nadeau and Bengio to the per-fold accuracies, which accounts for the overlap between training sets, with Holm adjustment for the six comparisons (Table 5c). After adjustment, the proposed model differed significantly from GANN (mean difference 4.06 percentage points, corrected 95% CI 2.56 to 5.56; adjusted p = 0.009) and from GNN-Attention (18.20 points, 11.73 to 24.67; adjusted p = 0.009). It did not differ significantly from CNN-BiLSTM-Attention (0.14 points, −1.48 to 1.76; adjusted p = 0.822), CNN-GRU-Attention (0.89 points, −0.32 to 2.10; adjusted p = 0.219), CNN-LSTM-Attention (2.20 points, 0.08 to 4.32; adjusted p = 0.181) or CNN-BiGRU-PatchTST (2.44 points, −0.14 to 5.02; adjusted p = 0.181). The intervals are not adjusted for multiplicity. For CNN-LSTM-Attention the unadjusted p-value was 0.045, which does not survive the Holm correction. The corrected test is itself approximate for k-fold designs, and with five folds its power is limited. We define the overfitting gap as the final training accuracy minus the test accuracy on the fixed split (1.23% for the proposed model and 1.52% for GANN). All results use seed 42. Exploratory runs with other seeds gave very similar results, but variability across seeds was not evaluated.

Selecting multiple models is essential for thoroughly assessing performance across diverse architectures, enabling a comprehensive comparison of their effectiveness in EEG-based emotion classification. This research employs six distinct neural network architectures shown in Table 4 that are a combination of different models, attention mechanisms, and graph-based structures (PatchTST, GNN) to systematically evaluate their strengths and limitations in the task.

Uncorrected Proof

Table 4. List of Models

Rank	Model	CNN	GNN	GRU	LSTM	BiGRU	BiLSTM	Attention	Patch TST
1	CNN-BiGRU-Attention	✓				✓		✓	
2	CNN-BiLSTM-Attention	✓					✓	✓	
3	CNN-GRU-Attention	✓		✓				✓	
4	CNN-LSTM-Attention	✓			✓			✓	
5	CNN-BiGRU-PatchTST	✓				✓			✓
6	GNN-Attention		✓					✓	

4- Results and Discussion

The results demonstrate that systematic hyperparameter optimization improves classification accuracy while simultaneously constraining model complexity, underscoring the value of structured search strategies such as grid search. In this research, Four main criteria were used to evaluate model performance.

4-1- Metrics

Accuracy: This criterion measures the overall correctness of the model by calculating the ratio of correct predictions to total predictions made. As shown in formula 13, this criterion is calculated as (true positives + true negatives) / (total predictions). Although accuracy provides a simple criterion of performance, it can be misleading when dealing with imbalanced datasets. For example, if, in an EEG dataset, 70% of signals are neutral, 20% positive, and 10% negative, a model that always predicts "neutral" will have 70% accuracy but will be completely ineffective in detecting other two labels.

Precision: This precision criterion focuses on the quality of positive predictions and measures how many of the predicted positive cases were actually correct. This criterion is shown in formula 14 and is calculated as true positives / (true positives + false positives). Positive precision answers this question:

Uncorrected Proof

"Of all cases the model predicted as positive emotion, how many were actually positive?" This criterion is critical in EEG analysis when high confidence in positive emotion detection is needed. For example, brain-control systems or neurofeedback, high precision ensures that the system triggers only when a positive emotional state is genuinely present, suppressing false activations caused by noise.

Recall (Sensitivity): The recall criterion (also known as sensitivity) measures the model's ability to identify all true positive cases. This criterion, shown in formula 15, is calculated as true positives / (true positives + false negatives). Recall answers this question: "Of all moments when the person actually had positive emotion, how many did the EEG model correctly detect?" This criterion is very important in psychology and neuropsychology applications. For example, in depression or anxiety studies, high recall for detecting negative emotions is essential so no period of negative state is missed, even if it means incorrectly detecting some neutral states as negative.

F1-Score: The F1-score provides a balanced criterion that combines positive precision and recall in a single criterion calculated as their harmonic mean (formula 16): 2 × (precision × recall) / (precision + recall). The harmonic mean is used because it is more sensitive to low values, meaning the F1-score will be low if either precision or recall is weak. This feature makes the F1-score very useful in EEG signal analysis for emotion detection, especially when positive, negative, and neutral emotion data are imbalanced. For example, in emotional virtual reality systems or brain games, a high F1-score ensures that the system has both sufficient accuracy in emotion detection and considers all important emotional changes of the user.

$$Accuracy = \frac{\text{True Positives} + \text{True Negatives}}{Total\ Number\ of\ Predictions} \quad \text{Eq. (13)}$$

$$Precision = \frac{\text{True Positives}}{True\ Positives + False\ Positives} \quad \text{Eq. (14)}$$

$$Recall = \frac{\text{True Positives}}{True\ Positives + False\ Negatives} \quad \text{Eq. (15)}$$

$$\text{F1Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{Precision + Recall} \quad \text{Eq. (16)}$$

Uncorrected Proof

In these equations TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives for a class, counted one class against the rest. The macro average is the unweighted mean of the per-class values, and the weighted average weights them by class support.

4-2- Performance Evaluation

To evaluate relative performance, the proposed model was compared with existing methods evaluated on the same database. The tables below show the accuracies reported in papers. In Table 5, the results obtained are for test data in both datasets and compared with the baseline paper method. Notably, we reimplemented the baseline GANN model and tested it on our two datasets to ensure fair comparison, All models were evaluated under identical conditions within each experiment: the six architectures and the reimplemented GANN were compared with five-fold cross-validation on Dataset 1 and with the fixed split on both datasets (Table 5), and the proposed model and GANN were additionally compared under LOSO on Dataset 2 (Table 9).

Table 5. Comprehensive Comparison of Six Model Architectures

Rank	Model	Test accuracy, Dataset 1 (fixed split)	Test accuracy, Dataset 2 (fixed split)	Parameters (Dataset 1)	5-fold Accuracy (Dataset 1)	5-fold micro F1 score (Dataset 1)	Overfitting Gap (Dataset 1)	F1-Score (Dataset 1)
1	CNN-BiGRU-Attention	98.59%	98.55%	381,447	97.09% ± 0.52%	0.9709 ± 0.0052	1.23%	0.9859
2	CNN-BiLSTM-Attention	97.65%	97.98%	1,450,503	96.95% ± 0.59%	0.9694 ± 0.0059	2.23%	0.9765
3	CNN-GRU-Attention	97.65%	97.89%	611,335	96.20% ± 0.67%	0.9620 ± 0.0066	2.29%	0.9766
4	CNN-LSTM-Attention	95.31%	96.34%	232,967	94.89% ± 1.07%	0.9487 ± 0.0108	3.35%	0.9531
5	CNN-BiGRU-PatchTST	95.31%	96.28%	376,195	94.65% ± 1.14%	0.9466 ± 0.0112	4.58%	0.9531
6	GNN-Attention	84.04%	82.92%	117,127	78.89% ± 3.47%	0.7789 ± 0.0420	-5.45%	0.8425

Uncorrected Proof

Base	GANN (Alibrahim et al.)	94.23% (our test)	93.91%	199,940	93.03% ± 0.52%	93.03% ± 0.0052	1.52%	0.9423

Table 5b. Classification report of CNN-BiGRU-Attention on the Dataset 1 test split (213 samples).

Class	**Precision**	**Recall**	**F1-score**	**Support**
Negative	0.989	0.978	0.983	90
Neutral	1.000	1.000	1.000	59
Positive	0.969	0.984	0.977	64
Accuracy	-----	-----	0.986	213
Macro average	0.986	0.987	0.987	213

Table 5c. Corrected paired comparison (Nadeau–Bengio) of CNN-BiGRU-Attention with each other model, Dataset 1, five-fold cross-validation (n_train ≈ 1,705, n_test ≈ 213 per fold). Intervals are not adjusted for multiplicity.

Compared model	**Mean difference (pp)**	**Corrected 95% CI**	**t**	**p**	**Holm-adjusted p**
CNN-BiLSTM-Attention	0.14	−1.48 to 1.76	0.24	0.822	0.822
CNN-GRU-Attention	0.89	−0.32 to 2.10	2.05	0.110	0.219
CNN-LSTM-Attention	2.20	0.08 to 4.32	2.88	0.045	0.181
CNN-BiGRU-PatchTST	2.44	−0.14 to 5.02	2.63	0.058	0.181
GNN-Attention	18.20	11.73 to 24.67	7.82	0.001	0.009
GANN	4.06	2.56 to 5.56	7.50	0.002	0.009

As observed in Table 6, different papers in this field have achieved different accuracies. The proposed model achieved the highest reported accuracy among the studies listed in Table 6. However, differences in datasets, class definitions and evaluation protocols limit direct comparison across studies, so this comparison is contextual and does not demonstrate statistically significant superiority. Also worth mentioning that in Table 5, GANN accuracy is based on our tests, and in Table 6, it's based on the research paper report.

Table 6. Comparison with State-of-the-Art Methods

Uncorrected Proof

Source	Citation	Reported Accuracy (%)	Description	Publication Year	Method/Model
Alibrahim et al., Procedia Computer Science	[8]	94.21%	GRU model with context-aware attention mechanism for spatial-temporal feature extraction, without CNN	2024	GANN (GRU + Attention)
IEEE Xplore (Emotion Recognition Using EEG Signals)	[11]	97.70%	CNN-Bi-LSTM with Attention Mechanism for modeling temporal EEG relationships.	2023	CNN-Bi-LSTM with Attention Mechanism
IEEE Xplore (EEG classification overview)	[15]	95.0%	Direct classification with Dynamical Graph Convolutional Neural Network on complete dataset, focus on statistical features	2020	DGCNN (Dynamical Graph Convolutional Neural Network)
PMC (S-LSTM-ATT hybrid approach)	[32]	96.50%	LSTM combination with multi-head attention for noise management, but without bidirectional BiGRU	2023	S-LSTM-ATT (LSTM + Attention)
Springer, Neural Computing and Applications	[33]	95.97%	Neuro-fuzzy combination with EEG features for noise management, but without bidirectional attention	2024	Adaptive Neuro-Fuzzy Hybrid
This Research		98.59%	CNN + Bidirectional GRU + Multi-head Attention with optimized hyperparameters	2025	CNN-BiGRU-Attention

Uncorrected Proof

Based on the classification report presented in Figure 2, the proposed model demonstrates exceptional and balanced performance across all three emotion classes. The CNN-BiGRU-Attention (381,447 parameters) model achieved outstanding accuracy of 98.59% and F1-scores of 0.983 for negative emotions, 1.000 for neutral emotions, and 0.977 for positive emotions, with a weighted average F1-score of 0.986. The per-class F1-scores are balanced across the three classes (0.977 to 1.000). Table 5b lists the report with the number of test samples (support) of each class.

10.22060/EEJ.2026.25904.6024

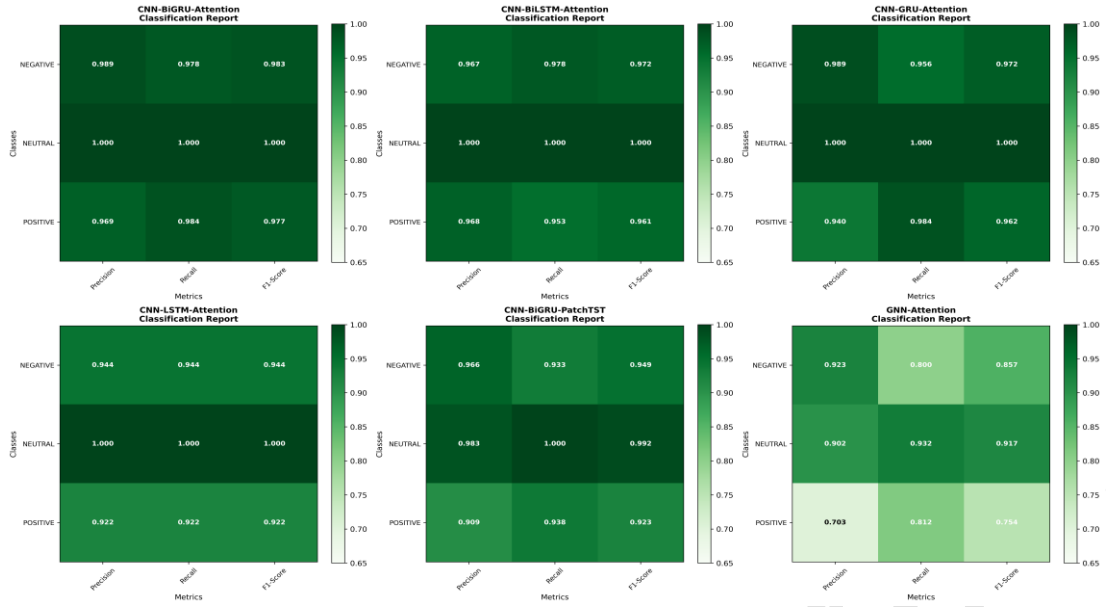


Fig. 2. Classification reports of the six architectures on the Dataset 1 test split (fixed 80:10:10 split, n = 213: 90 negative, 59 neutral, 64 positive).

Figure 3 shows the confusion matrices of the six architectures on the Dataset 1 test split (n = 213; 90 negative, 59 neutral and 64 positive samples). CNN-BiGRU-Attention made 3 errors: 88 of 90 negative samples were recognized (recall 97.78%), with 2 misclassified as positive; all 59 neutral samples were correct; and 63 of 64 positive samples were recognized (recall 98.44%), with 1 misclassified as negative. CNN-BiLSTM-Attention made 5 errors (2 negative samples as positive and 3 positive samples as negative) and CNN-GRU-Attention made 5 (4 negative samples as positive and 1 positive sample as negative), so the best model differs from them by two test samples, a gap too small to be treated as an established difference. CNN-LSTM-Attention and CNN-BiGRU-PatchTST made 10 errors each, and GNN-Attention made 34, including 18 negative samples predicted as positive, 4 neutral samples predicted as positive, and 12 positive samples predicted as 6 negative and 6 neutral. Our GNN-Attention variant therefore reached only 84.04%, well below the CNN-recurrent models, but it was not specifically tuned as a graph model, so this should not be read as a general comparison of graph-based and recurrent approaches [34, 35]. Because each class contains only 59–90 test samples, a single error changes a per-class recall by 1–2 points, so these figures should be read as descriptive.

Uncorrected Proof

10.22060/EEJ.2026.25904.6024

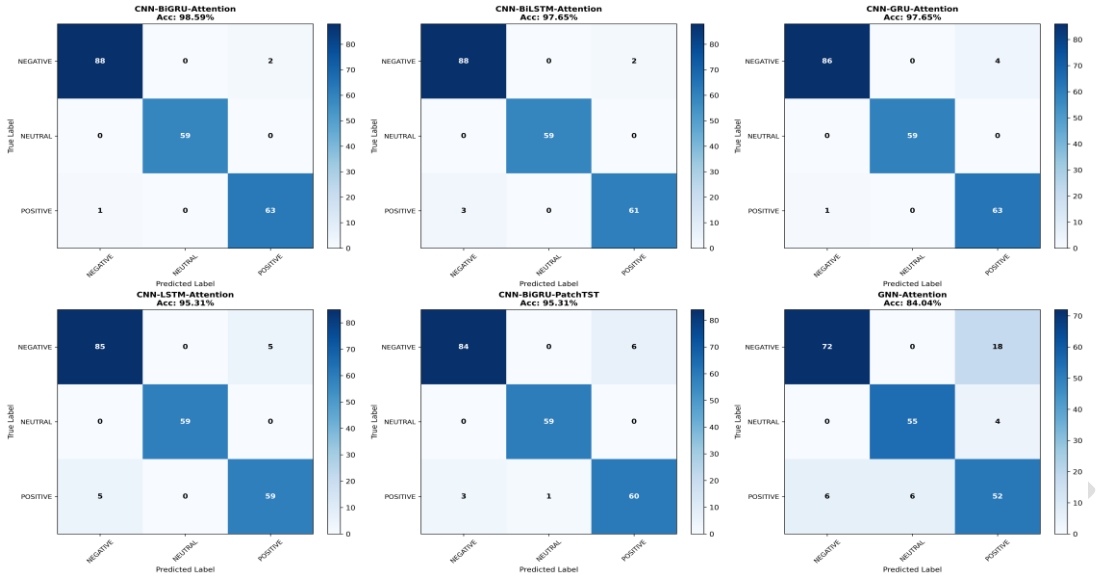


Fig. 3. Confusion matrices of the six architectures on the Dataset 1 test split (fixed 80:10:10 split, n = 213). Rows are true labels and columns are predicted labels.

Table 7 shows a consistent pattern: most errors occur between the negative and positive classes, and none of the five CNN-based models misclassifies a neutral sample. Neutral errors appear only for GNN-Attention (4 neutral samples predicted as positive), and positive samples are predicted as neutral by CNN-BiGRU-PatchTST (1) and GNN-Attention (6). One possible explanation is that negative and positive recordings overlap in the feature space, whereas the neutral class consists of resting-state recordings that may differ from the emotional recordings in ways unrelated to emotion (for example, recording session). With two participants and no subject identifiers, we cannot test either explanation, so we regard both as hypotheses.

Table 7. Misclassification counts per architecture on the Dataset 1 test split (n = 213)

Model	**'NEGATIVE' misclassified as 'POSITIVE'**	**'POSITIVE' misclassified as 'NEGATIVE'**	**'NEGATIVE' misclassified as 'NEUTRAL'**	**'POSITIVE' misclassified as 'NEUTRAL'**	**'NEUTRAL' misclassified as 'POSITIVE'**	**Total errors (of 213)**
CNN-BiGRU-Attention	2 times	1 time	---	---	---	3
CNN-BiLSTM-Attention	2 times	3 times	---	---	---	5
CNN-GRU-Attention	4 times	1 time	---	---	---	5

Uncorrected Pro

CNN-LSTM-Attention	5 times	5 times	---	---	---	10
CNN-BiGRU-PatchTST	6 times	3 times	---	1 time	---	10
GNN-Attention	18 times	6 times	---	6 times	4 times	34

Figure 4 presents the learning curves on Dataset 1. All CNN-based models approach 100% training accuracy within about 30 epochs. In the recurrent models, validation loss rises slowly after roughly epoch 10–15, indicating mild overfitting, which early stopping limits. CNN-BiGRU-Attention and CNN-GRU-Attention show smoother validation curves than CNN-BiLSTM-Attention and CNN-LSTM-Attention, whose validation loss fluctuates more. CNN-BiGRU-PatchTST shows the largest rise in validation loss despite its BiGRU block, so the stability advantage is observed for the attention variants rather than for GRU layers in general. GNN-Attention reaches a training accuracy of only about 78% with higher validation than training accuracy and noisy loss curves, which is consistent with underfitting.

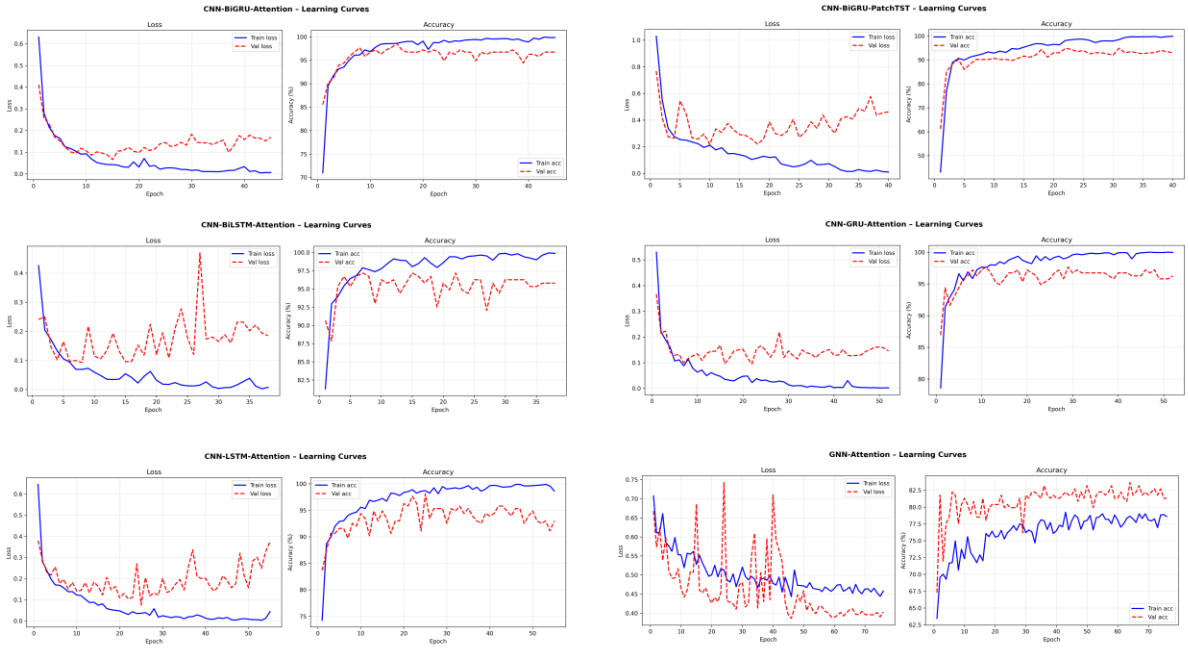


Fig. 4. Learning curves of the six architectures on Dataset 1 (fixed 80:10:10 split) under the convergence protocol: up to 100 epochs with early stopping (patience 5, min_delta 0.001). The comparison results in Tables 5–7 use a fixed 40 epochs.

Proof

10.22060/EEJ.2026.25904.6024

Figure 5 shows the attention weights of the four heads of CNN-BiGRU-Attention for each class, averaged over the Dataset 1 test samples (n = 90, 59 and 64). The x-axis is the index of the 64 PCA components that form the input sequence; the components are ordered by explained variance, so index 0 is the highest-variance component. For negative and positive samples, all four heads place most of their weight on the first few components (Head 2 is broader for negative samples). For neutral samples, Heads 1–3 also spread weight over components around 30–45. The attention patterns for negative and positive samples are therefore more similar to each other than to those of neutral samples, which is consistent with the confusion pattern in Table 7. We present this as an observation and not as proof of what each head represents.

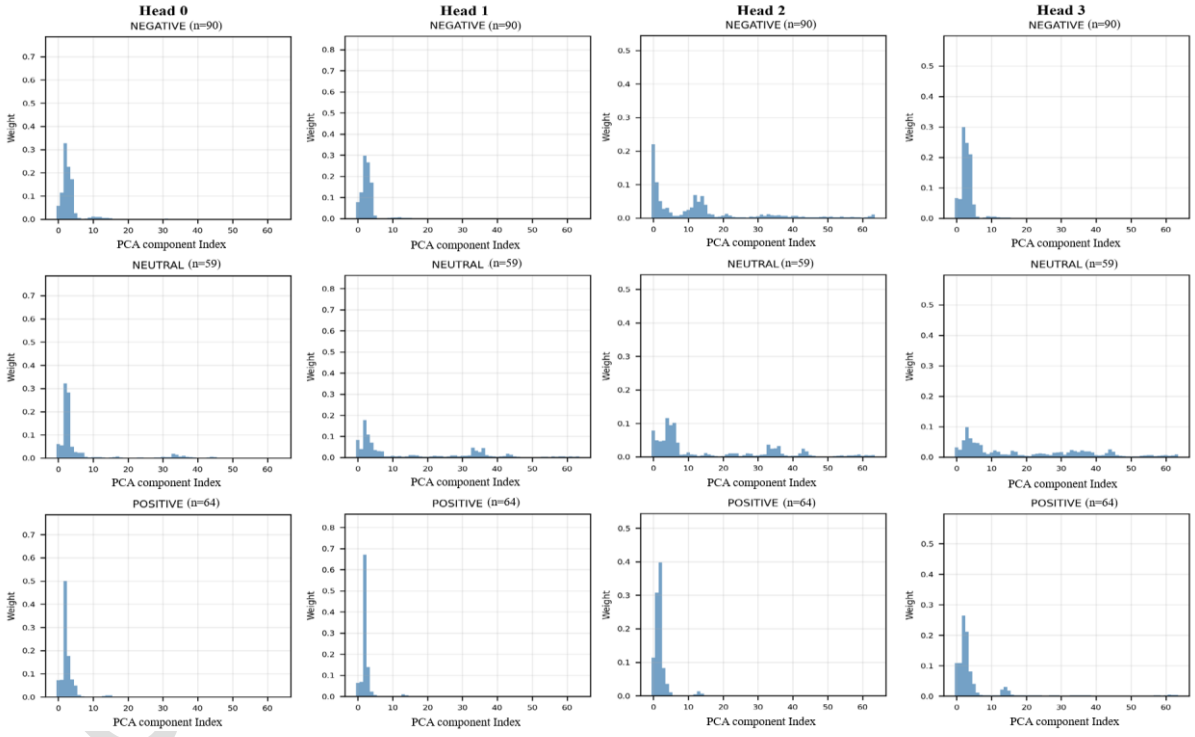


Fig. 5. Class-averaged attention weights of the four heads of CNN-BiGRU-Attention over the 64 PCA components, on the Dataset 1 test split (n = 90, 59 and 64 for negative, neutral and positive).

Table 8. Variant of the best model on Ablation of the proposed model on Dataset 1, five-fold cross-validation

Model Variant	Change	Accuracy (mean±SD)	Macro-F1	Δ vs full
Full model	none	97.09% ± 0.52	0.970	-
Full model	2-Head Attention	96.42% ± 0.52	0.942	-0.67
Full model	8-Head Attention	96.60% ± 0.52	0.959	-0.49
A	No attention (mean pooling)	96.21% ± 0.70	0.962	-0.88
B	No CNN (BiGRU on raw sequence)	95.41% ± 1.54	0.954	-1.68
C	Single kernel (5) instead of 3/5/7	96.63% ± 2.02	0.966	-0.49
D	Unidirectional GRU	96.65% ± 1.11	0.966	-0.44
E	No BatchNorm	95.78% ± 1.43	0.957	-1.31
F	No PCA	96.37% ± 1.08	0.963	-0.72
G	No feature selection	95.53% ± 0.59	0.955	-1.56

To verify that the advantage of CNN-BiGRU-Attention is robust, we removed or modified one component at a time and evaluated each variant with the same five folds, seed and training settings as the full model (Table 8).

4-2-1- Subject-Independent Evaluation (LOSO) on Dataset 2

To assess generalization to unseen individuals, we performed leave-one-subject-out (LOSO) cross-validation on Dataset 2 (20 participants, four classes). This evaluation was applied to the proposed CNN-BiGRU-Attention model and the GANN baseline, to test whether the advantage observed on Dataset 1 persists when the test participant is entirely unseen. In each of the 20 folds, one participant was held out for testing, one participant was used for validation and early stopping, and the remaining 18 participants were used for training. The validation participant was the next participant in a fixed order. Z-score normalization, feature selection, and PCA were fitted exclusively on the training participants of each fold. Hyperparameters were fixed to the optimal values obtained on Dataset 1 (Table

Uncorrected Proof

3) and were not re-tuned per fold. Table 9 and Fig 6 are reports the mean accuracy across the 20 folds. The proposed model achieved 92.37% [± 0.79] accuracy, compared with 86.34% [± 0.91] for GANN, an improvement of 6.03 percentage points. This is lower than the sample-level result on Dataset 2 (98.55%), which is expected since inter-subject variability makes cross-subject generalization harder.

Table 9. Subject-independent LOSO results on Dataset 2 (4-classes)

Model	**LOSO Accuracy (mean ± SD)**	**LOSO Macro-F1 (mean ± SD)**	**Improvement over GANN**
CNN-BiGRU-Attention (proposed)	92.37% ± 0.79	91.87% ± 0.63%	+6.03 pp
GANN (baseline)	86.34% ± 0.91	86.02% ± 0.68%	-

Proof

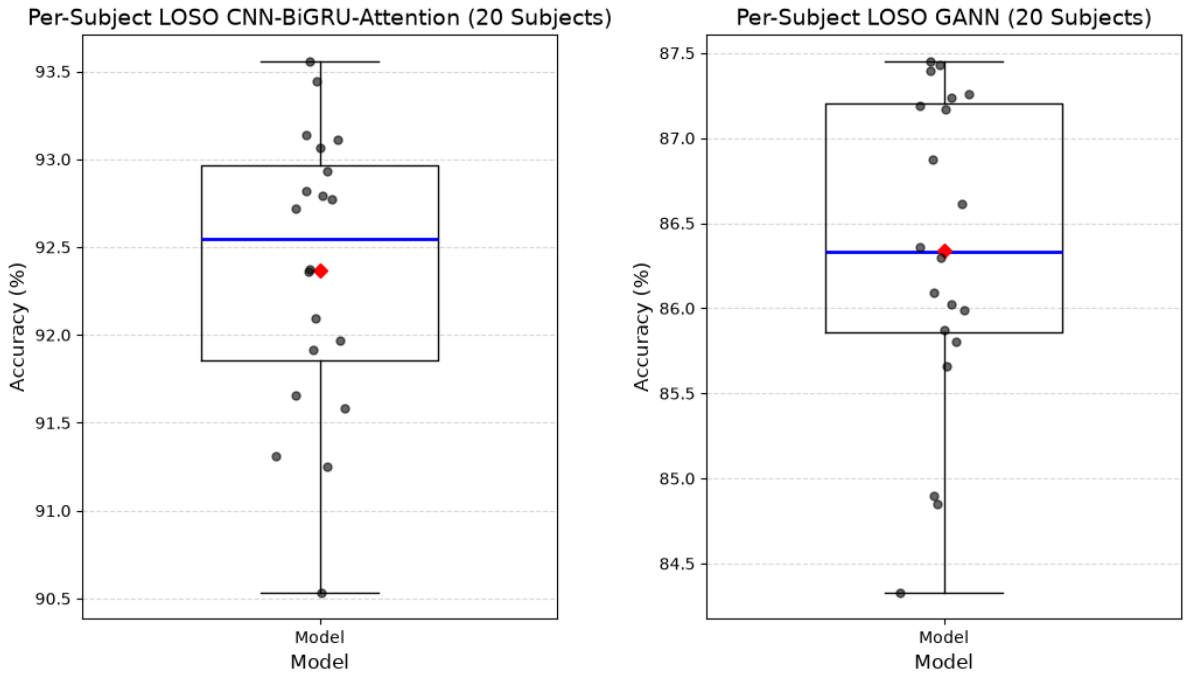


Fig. 6. Leave-one-subject-out (LOSO) accuracy for each of the 20 participants of Dataset 2 (four classes). In each fold, one participant was held out for testing, one was used for validation, and the other 18 were used for training; hyperparameters were fixed to the Dataset 1 optimum. Dashed lines show the mean over participants (92.37% and 86.34%)

4-3- Architectural Analysis

On the fixed split CNN-BiGRU-Attention scored higher than CNN-BiLSTM-Attention (98.59% vs 97.65%), but this is a difference of two test samples, and the five-fold means (97.09 ± 0.52% vs 96.95 ± 0.59%) are not distinguishable (Wilcoxon p = 0.593). The BiGRU model has 3.8 times fewer parameters (381K vs 1.45M). This reflects both the smaller hidden size chosen by the grid search (128 vs 256) and the GRU's two gates against the LSTM's three, so we do not attribute it to the gate structure alone. Because the input consists of pre-computed spectral features and not raw signals, the long-range memory of the LSTM cell may add little. The ablation (Table 8) shows that removing the convolutional front end (−1.68 points), feature selection (−1.56) or batch normalization (−1.31) costs the most. Making the GRU unidirectional costs 0.44 points, which is smaller than the fold-to-fold SD of that variant (1.11), so the contribution of bidirectionality is not established by these data. All experiments use 4 attention heads; the number of heads was not varied.

In terms of computational efficiency, CNN-BiGRU-Attention has 381,447 trainable parameters (1.53 MB in FP32) and requires 15.99 M multiply–accumulate operations (31.97 MFLOPs, taking FLOPs = 2 × MACs) per inference, counted analytically for the convolution, linear and GRU-gate operations. Latency was measured at batch size 1 on CPU over 200 samples. Model-only inference took 5.55 ms on average on the host CPU (SD = 1.71 ms, 95th percentile = 9.45 ms). We also ran the model under three resource-constrained profiles emulated on the host CPU: a Raspberry Pi 4 profile (4 threads), a Jetson Nano profile (4 CPU threads, no GPU) and a single-core profile. End-to-end latency including preprocessing was 15.02 ms (95th percentile 25.06 ms), 14.66 ms (28.98 ms) and 5.38 ms (8.25 ms), respectively. Even the largest 95th-percentile value is about 3.5 times below the 100 ms decision window typically assumed for real-time BCI systems [2]. These profiles emulate, but do not replace, measurements on physical devices, and the small differences between models in latency (Table 10) are within run-to-run variation, so we do not rank the models by speed.

Table 10. Computational cost and CPU inference latency of the six architectures and the GANN baseline (batch size 1, 200 timed samples). Latency is given in ms as mean (± SD for the host CPU) with the 95th percentile in parentheses.

Uncorrected Proof

Model	Params	Size (MB, FP32)	MACs (M)	FLOPs (M)	Host CPU, model only	Pi-like profile (4 threads)	Jetson-like profile (4 CPU threads)	Single-core profile
CNN-BiGRU-Attention	381,447	1.53	15.99	31.97	5.55 ± 1.71 (9.45)	15.02 (25.06)	14.66 (28.98)	5.38 (8.25)
CNN-BiLSTM-Attention	1,450,503	5.80	59.44	118.88	3.99 ± 1.49 (6.35)	10.90 (18.84)	11.12 (23.68)	7.61 (10.79)
CNN-GRU-Attention†	611,335	2.45	44.56	89.11	6.26 ± 1.55 (9.40)	10.39 (24.23)	10.90 (16.35)	7.13 (8.81)
CNN-LSTM-Attention	232,967	0.93	10.65	21.29	1.15 ± 0.79 (2.80)	9.18 (14.60)	6.44 (10.64)	2.22 (3.59)
CNN-BiGRU-PatchTST	1,340,675 ‡	5.36	23.24	46.48	7.11 ± 1.80 (11.11)	21.84 (87.66)	17.56 (40.00)	8.95 (14.60)
GNN-Attention	117,127	0.47	4.83	9.66	0.69 ± 0.61 (1.89)	5.43 (8.06)	5.16 (11.29)	1.89 (4.36)
GANN (baseline)	199,940	0.80	12.65	25.30	2.62 ± 1.03 (4.16)	4.97 (7.76)	10.58 (21.21)	4.06 (7.35)

MACs were counted analytically for the convolution, linear, recurrent-gate, attention-scoring and transformer or graph-aggregation operations only; activations, normalization and softmax are excluded, and FLOPs are taken as 2 × MACs. For CNN-BiGRU-Attention the count is dominated by the BiGRU (15.73 of 15.99 M MACs).

Uncorrected Proof

10.22060/EEJ.2026.25904.6024

4-4- Limitations

First, Dataset 1 was collected from only two participants using a four-channel dry-electrode headband and has no subject identifiers, so its evaluation is restricted to sample-level cross-validation. The 98.59% accuracy on this dataset reflects within-subject performance and does not demonstrate generalization to unseen individuals. Second, the subject-independent LOSO evaluation was performed on Dataset 2 for the proposed model and the GANN baseline only; extending it to all six architectures is left for future work. The LOSO accuracy (92.37%) is lower than the sample-level accuracy, indicating that inter-subject variability remains a significant challenge. Third, hyperparameters were optimized on Dataset 1 and applied to Dataset 2 without per-subject adaptation. Fourth, the two datasets differ in label structure (three versus four classes), so their accuracies are not directly comparable. Fifth, with five folds the statistical tests have limited power: the smallest attainable Wilcoxon p-value is 0.0625, and with the corrected paired t-test and Holm adjustment the advantage over GANN and GNN-Attention is statistically supported, whereas the advantage over the other recurrent models and PatchTST is consistent but not statistically established. The corrected test is itself approximate for k-fold designs. Sixth, all results use a single random seed (42), and variability across seeds was not evaluated. Seventh, all experiments use four attention heads, and latency was measured only on a laptop, so edge-device figures are projections. Eighth, the hyperparameters were selected on the validation part of the fixed split of Dataset 1, whose samples also appear in the five-fold test folds, so the five-fold results are not fully independent of model selection. A nested evaluation would remove this dependence and is left for future work. Future work should validate the model on larger multi-subject benchmarks such as SEED and DEAP using LOSO protocols, explore domain adaptation and domain generalization [26, 28] and time–frequency fusion [27] to reduce inter-subject variability, and consider uncertainty-aware (fuzzy) modelling of noisy EEG signals and labels [36-39].

Uncorrected Proof

5- Conclusion

This study compared six CNN-recurrent and graph-attention architectures for EEG emotion recognition under identical preprocessing, hyperparameter search and training protocols. The proposed CNN-BiGRU-Attention model, with 381,447 parameters, reached 97.09 ± 0.52% accuracy in five-fold cross-validation on Dataset 1 (98.59% on the fixed split), which is 4.06 and 4.36 percentage points above our reimplementation of GANN. Its per-class F1-scores on the fixed split were 0.983 (negative), 1.000 (neutral) and 0.977 (positive). Under subject-independent leave-one-subject-out evaluation on Dataset 2 it reached 92.37 ± 0.79% against 86.34 ± 0.91% for GANN, a gain of 6.03 percentage points; the lower accuracy compared with the sample-level result (98.55%) shows that inter-subject variability remains a challenge. The ablation study indicates that the convolutional front end, feature selection and batch normalization contribute most, while the contributions of bidirectionality and attention are smaller than the fold-to-fold variation. A corrected paired test with Holm adjustment supports the gain over GANN (adjusted $p = 0.009$), whereas the differences from the other recurrent models are consistent but not statistically significant. The main limitations are the two-subject Dataset 1, the single random seed, the use of four attention heads only, and the lack of latency measurements on edge hardware. Future work includes validation on larger multi-subject benchmarks with LOSO protocols, domain adaptation and generalization [26, 28], on-device benchmarking, and uncertainty-aware modelling of noisy EEG signals and labels [36-39].

Acknowledgments

The authors would like to acknowledge the use of artificial intelligence (AI) assisted tools in the preparation of this manuscript. Specifically, AI language models were utilized to support the editing and refinement of certain portions of the text, including improvements to clarity, coherence, and academic tone. All scientific content, research design, experimental methodology, data analysis, and conclusions presented in this paper are entirely the work of the authors. The use of AI tools was limited strictly to linguistic and editorial assistance, and the authors take full responsibility for the integrity and accuracy of all content contained within this manuscript.

Uncorrected Proof

10.22060/EEJ.2026.25904.6024

Uncorrected Proof

References

[1] X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, P. Marttinen, EEG based emotion recognition: A tutorial and review, ACM Computing Surveys, 55(4) (2022) 1–57.
[2] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, T.M. Vaughan, Brain–computer interfaces for communication and control, Clinical neurophysiology, 113(6) (2002) 767–791.
[3] I. Patras, DEAP: a database for emotion analysis using physiological signals, (2012).
[4] R. Alhalaseh, S. Alasasfeh, Machine-learning-based emotion recognition system using EEG signals, Computers, 9(4) (2020) 95.
[5] R. Sánchez-Reolid, M.C. Martínez-Sáez, B. García-Martínez, L. Fernández-Aguilar, L. Ros, J.M. Latorre, A. Fernández-Caballero, Emotion classification from EEG with a low-cost BCI versus a high-end equipment, International journal of neural systems, 32(10) (2022) 2250041.
[6] W. Ma, Y. Zheng, T. Li, Z. Li, Y. Li, L. Wang, A comprehensive review of deep learning in EEG-based emotion recognition: classifications, trends, and practical implications, PeerJ Computer Science, 10 (2024) e2065.
[7] Y. Liu, W. Xue, L. Yang, M. Li, Deep learning-based EEG emotion recognition: a review, Brain Sciences, 16(1) (2025) 41.
[8] R. Alibrahim, H. Kurdi, Gann: Eeg-based emotion classification using context-aware gated attention neural network, Procedia Computer Science, 241 (2024) 234–241.
[9] J.J. Bird, L.J. Manso, E.P. Ribeiro, A. Ekart, D.R. Faria, A study on mental state classification using eeg-based brain-machine interface, in: 2018 international conference on intelligent systems (IS), IEEE, 2018, pp. 795–800.
[10] T. Song, W. Zheng, P. Song, Z. Cui, EEG emotion recognition using dynamical graph convolutional neural networks, IEEE Transactions on Affective Computing, 11(3) (2018) 532–541.
[11] Z. Huang, Y. Ma, R. Wang, W. Li, Y. Dai, A model for EEG-based emotion recognition: CNN-Bi-LSTM with attention mechanism, Electronics, 12(14) (2023) 3188.
[12] Y. Ma, Z. Huang, Y. Yang, Z. Chen, Q. Dong, S. Zhang, Y. Li, MSBiLSTM-Attention: EEG emotion recognition model based on spatiotemporal feature fusion, Biomimetics, 10(3) (2025) 178.
[13] H. Oka, K. Ono, A. Panagiotis, Attention-based PSO-LSTM for emotion estimation using EEG, Sensors, 24(24) (2024) 8174.
[14] Y. Ma, Z. Huang, Y. Yang, S. Zhang, Q. Dong, R. Wang, L. Hu, Emotion recognition model of EEG signals based on double attention mechanism, Brain Sciences, 14(12) (2024) 1289.
[15] H. Keshvarikhojasteh, J. Pluim, M. Veta, Multi-head attention-based deep multiple instance learning, arXiv preprint arXiv:2404.05362, (2024).
[16] W. Lu, L. Xia, T.P. Tan, H. Ma, CIT-EmotionNet: convolution interactive transformer network for EEG emotion recognition, PeerJ Computer Science, 10 (2024) e2610.
[17] R. Liu, Y. Chao, X. Ma, X. Sha, L. Sun, S. Li, S. Chang, ERTNet: an interpretable transformer-based framework for EEG emotion recognition, Frontiers in Neuroscience, 18 (2024) 1320645.

Uncorrected Proof

[18] D. Wang, J. Lian, H. Cheng, Y. Zhou, Music-evoked emotions classification using vision transformer in EEG signals, Frontiers in Psychology, 15 (2024) 1275142.
[19] W.-L. Zheng, B.-L. Lu, Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks, IEEE Transactions on autonomous mental development, 7(3) (2015) 162–175.
[20] S. Chatterjee, Y.-C. Byun, EEG-based emotion classification using stacking ensemble approach, Sensors, 22(21) (2022) 8550.
[21] J.A. Coan, J.J. Allen, Frontal EEG asymmetry as a moderator and mediator of emotion, Biological psychology, 67(1-2) (2004) 7–50.
[22] P. Zhong, D. Wang, C. Miao, EEG-based emotion recognition using regularized graph neural networks, IEEE Transactions on Affective Computing, 13(3) (2020) 1290–1301.
[23] T. Zhang, W. Zheng, Z. Cui, Y. Zong, Y. Li, Spatial–temporal recurrent neural network for emotion recognition, IEEE transactions on cybernetics, 49(3) (2018) 839–847.
[24] W. Chen, Y. Liao, R. Dai, Y. Dong, L. Huang, EEG-based emotion recognition using graph convolutional neural network with dual attention mechanism, Frontiers in Computational Neuroscience, 18 (2024) 1416494.
[25] D. Xiao, Q. Meng, S. Li, X. Yuan, Improving transformers with dynamically composable multi-head attention, arXiv preprint arXiv:2405.08553, (2024).
[26] Y. Zhou, R. Jiang, Z. Zhou, Y. Yu, J. Zhang, CBSAtt: a CNN-BiLSTM network with multi-head self-attention for EEG emotion recognition, Signal, Image and Video Processing, 19(14) (2025) 1196.
[27] Y. Sun, H. Tao, V. Stojanovic, Open-set classification method via latent representation prompt and time–frequency fusion toward unknown fault recognition, Advanced Engineering Informatics, 68 (2025) 103779.
[28] Y. Sun, H. Tao, Y. Ni, V. Stojanovic, A generic single-source domain generalization framework for fault diagnosis via wavelet packet augmentation and pseudo-domain generation, IEEE Internet of Things Journal, 12(15) (2025) 31629–31642.
[29] J.J. Bird, A. Ekart, C.D. Buckingham, D.R. Faria, Mental emotional sentiment classification with an eeg-based brain-machine interface, in: Proceedings of theInternational Conference on Digital Image and Signal Processing (DISP'19), 2019.
[30] X. Zhu, C. Liu, L. Zhao, S. Wang, EEG emotion recognition network based on attention and spatiotemporal convolution, Sensors, 24(11) (2024) 3464.
[31] D. Nguyen, M.T. Nguyen, K. Yamada, Electroencephalogram based emotion recognition using hybrid intelligent method and discrete wavelet transform, Applied Sciences, 15(5) (2025) 2328.
[32] A. Abgeena, S. Garg, S-LSTM-ATT: a hybrid deep learning approach with optimized features for emotion recognition in electroencephalogram, Health Information Science and Systems, 11(1) (2023) 40.
[33] F.K. Bardak, M.N. Seyman, F. Temurtaş, Adaptive neuro-fuzzy based hybrid classification model for emotion recognition from EEG signals, Neural Computing and Applications, 36(16) (2024) 9189–9202.
[34] G. Hou, Q. Yu, G. Chen, F. Chen, A novel and powerful dual-stream multi-level graph convolution network for emotion recognition, Sensors, 24(22) (2024) 7377.
[35] S.S. Gilakjani, H. Al Osman, A graph neural network for EEG-based emotion recognition with contrastive learning and generative adversarial neural network data augmentation, IEEE Access, 12 (2023) 113–130.

Unedited Proof

[36] O.A. Arqub, M. Abukhaled, H. Sweis, N. Shawagfeh, Numerical analysis of Lucas polynomials innovative technique within the Galerkin strategy for solving fuzzy-type differential models with an application in the electrical circuit engineering field, Results in Engineering, (2025) 106298.
[37] W. Al-Sadi, Z. Wei, I. Moroz, O. Abu Arqub, T.Q. Abdullah, Predictor–corrector approach for the numerical solution of fuzzy fractional differential equations and linear multiterm fuzzy fractional equations, Soft Computing, 29(3) (2025) 1347–1368.
[38] B. Maayah, O.A. Arqub, Uncertain M-fractional differential problems: existence, uniqueness, and approximations using Hilbert reproducing technique provisioner with the case application: series resistor-inductor circuit, Physica Scripta, 99(2) (2024) 025220.
[39] O. Abu Arqub, R. Mezghiche, B. Maayah, Fuzzy M-fractional integrodifferential models: theoretical existence and uniqueness results, and approximate solutions utilizing the Hilbert reproducing kernel algorithm, Frontiers in Physics, 11 (2023) 1252919.

Uncorrected Proof