Attention-Augmented CNN-LSTM Framework for EEG-Based Spatiotemporal Emotion Recognition

Original Scientific Paper

Ayushi Kotwal
Dept. of Computer Science & IT, University of Jammu,
Jammu, India
ayushi.kotwal@jammuuniversity.ac.in

Kush Sharma
Hands-on Solutions Expert in AI/ML for Healthcare,
Texas, United States
Kush.sharma@live.com

Jatinder Manhas*
Computer Science & IT, University of Jammu,
Jammu, India
jatindermanhas@jammuuniversity.ac.in

*Corresponding author

Abstract *– EEG emotion recognition offers a direct path to the analysis of affective brain states. This work presents an attention-enhanced CNN-LSTM model for emotion recognition from EEG in the SEED benchmark dataset (three classes: positive, neutral, negative). The suggested method combines Power Spectral Density (PSD), Differential Entropy (DE), and hemispheric asymmetry descriptors (ASM, DASM, RASM) into a single input description, which is trained through a hybrid architecture in which a CNN learns the patterns of spatial channels, stacked LSTM layers learn the patterns of temporal dependencies, and a temporal attention block highlights the most informative time segments. The model has an accuracy of 98.63% in 5-fold cross-validation across subjects, with averaged precision, recall, and F1-score exceeding 98.6. Compared with solutions from representative published SEEDs listed in Table 1, these findings indicate competitive performance and underscore the usefulness of complementary spectral and hemispheric signal fusion before attention-directed spatiotemporal modeling.*

Keywords*: SEED, Emotion Recognition, EEG, Attention, CNN, LSTM*

Received: December 24, 2025; Received in revised form: April 22, 2026; Accepted: April 23, 2026

1. INTRODUCTION

Emotion is one of the most important elements of the human experience, affecting cognition, behavior, social interaction, and decision-making. As a result, automatic emotion recognition has been a key focus in affective computing, aiming to create systems that can detect, decode, and comprehend human affective states. Physiological signals provide a less controllable yet usually more accurate representation of emotional state than speech and facial expressions, and are thus of particular importance to affective analysis [1]. Among these signals, electroencephalography (EEG) provides a direct measure of brain activity and has become a potent modality for emotion recognition. Nevertheless, EEG signals are highly nonlinear and nonstationary, and exhibit complex spatiotemporal dependencies. Earlier studies relied mainly on handcrafted descriptors combined with conventional classifiers such as Support Vector Machines (SVMs) and k-Nearest Neighbors (k-NN). Widely used features include Power Spectral Density (PSD) and Differential Entropy (DE), as well as the Fast Fourier Transform (FFT), which capture informative spectral variations in EEG signals [2], [3], [4]. Building on the neurophysiological concept of hemispheric asymmetry, researchers have also introduced Asymmetry (ASM), Differential Asymmetry (DASM), Rational Asymmetry (RASM), and Differential Causal Asymmetry (DCAU) to characterize emotion-related differences between the two hemispheres [2], [3]. Positive affect is often associated with relatively greater left-frontal activity, whereas negative affect is linked to relatively greater right-frontal activity. EEG recordings are commonly acquired using the internationally standardized 10-20 electrode placement system, which ensures broad coverage of cortical regions. Fig. 1 illustrates this electrode arrangement [5].

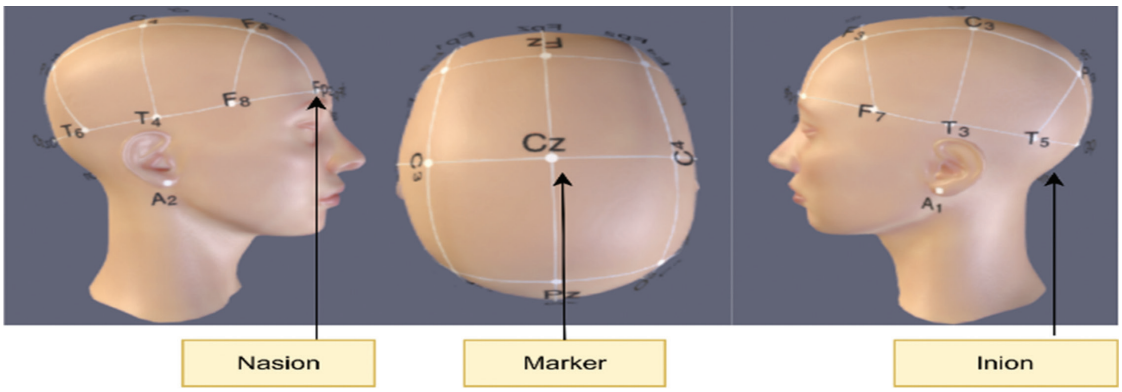


Fig. 1. Standard 10–20 international EEG electrode placement system showing spatial distribution of scalp recording channels used for brain activity acquisition

Asymmetry-based features were introduced to better capture the neural mechanisms underlying emotion. According to frontal alpha asymmetry theory, approach and avoidance tendencies are associated with differences in alpha-band power between the left and right prefrontal regions [5]. Features such as DASM and RASM [6], which have been successfully used to distinguish between positive and negative emotions, are practical formulations of this concept. More recent descriptors, such as ASM and DCAU, extend this line of work by modeling richer inter-hemispheric interactions [7]. The transition to deep learning began when CNNs were used to automatically extract representations from raw EEG signals or time-frequency representations [8]. CNNs are effective at learning spatial patterns across electrode channels, but they are less suited to capturing the temporal evolution of emotional responses. To address this limitation, EEG signals have also been modeled using recurrent architectures such as LSTMs and GRUs, which are well-suited to sequential data [9]. Hybrid CNN-LSTM models combine these complementary strengths by first learning spatial representations and then modeling their temporal evolution [10]. Previous analyses on cognitive workload and stress data have likewise highlighted the value of convolutional structures for capturing salient EEG patterns [11].

Using the SEED dataset, this study proposes a unified feature-fusion and spatiotemporal deep-learning framework for three-class EEG emotion recognition (positive, neutral, negative). Spectral descriptors (PSD and DE) and hemispheric asymmetry cues (ASM, DASM, RASM) are integrated into a single (T \C \F) representation, ensuring complementary frequency-domain and inter-hemispheric information is fused before modeling. An attention-augmented CNN processes the fused tensor–LSTM pipeline, in which CNN layers learn discriminative spatial channel patterns, LSTM layers capture temporal dependencies across segments, and temporal attention serves as a relevance-weighting mechanism to emphasize emotion-informative intervals during sequence aggregation. The work further strengthens rigor by explicitly specifying the SEED evaluation protocol and presenting component-level analyses that clarify the contribution of fusion, spatiotemporal modeling, and attention to robust discrimination. The main contributions of this work are:

1. A unified feature fusion framework combining PSD, DE, ASM, DASM, and RASM.
2. A structured spatiotemporal EEG representation ($T \times C \times F$).
3. An attention-augmented CNN–LSTM model for improved temporal relevance learning.
4. A comprehensive evaluation with component-level analysis on the SEED dataset.

2. RELATED WORK

Neural signal emotion recognition is a valuable area of brain-computer interface research, as it enables objective evaluation of affective behaviors across applications in cognitive neuroscience, human-computer interaction, and mental health monitoring. Xu *et al*. [12] addressed a significant problem in EEG emotion detection by achieving consistent results across subjects. In contrast, prior models often failed due to substantial inter-subject variation in brain activity. They mainly contributed a hybrid GRU-Conv architecture, unlike regular GRUs, which disregard intermediate states. The first stage of their contribution was to input the EEG into multiple GRU units, which track temporal variations across time series without discarding the outputs of all units. A better multidimensional representation is subsequently enabled by the CNN layers' extraction of spatial patterns from this temporal information. By more effectively combining spatiotemporal cues, it beat baselines such as standalone GRUs or CNNs when tested in subject-independent settings on SEED (87.04% average accuracy across three emotions) and DEAP (70.07% arousal, 67.36% valence). Zuo and Wang [13] already extended a general EEG emotion classifier, but moved to stock trading environments where rapid emotional changes affect financial behavior, unlike general applications of affective computing. They proposed EEGNet, a small convolutional network designed to work with the small-channel layout of EEG, operating directly on the raw signals of a trading simulation dataset of their own, with

valence-arousal responses from participant feedback. It achieved approximately 85-90% accuracy in detecting state differences evoked by trading, such as stress or excitement, making it effective for real-time detection. Still, its domain-specific data collection prevented general transferability, and sequential patterns in EEG are absent (e.g., no RNN), rendering it time-irrelevant.

To overcome the single-model feature silos of standalone CNN or LSTM, Iyer *et al*. [14] proposed an ensemble that combines CNN-extracted frequency-spatial features with LSTM-sequenced dynamics, surpassing single-model feature silos in handling the dual spatial-temporal nature of EEG. Evaluated on publicly available EEG data such as DEAP with four emotions (happy, sad, fear, neutral) it achieved 95.56 per cent accuracy, outperforming single branch CNN (92) or LSTM (90) by weighted voting; its major strengths are that it is robust to noise by preprocessing (bandpass filtering, artifact removal) and can be scaled to multi-class tasks; its major

Li *et al*. [15] addressed the shortcomings of early deep models that did not account for the channel-specific relevance of EEG by adding an attention layer to a CNN-RNN hybrid, which weights critical bands after feature extraction, thereby establishing the first focus on homogeneous processing principles. Subject-dependently tested on SEED-IV (four emotions), it was 92-95% accurate, improved 510% by adaptive specialization on theta/alpha rhythms; its advantages include interpretability (visualized attention maps), efficiency with no undue data requirements, and limited attention to multi-scale temporal hierarchies; limitations include outdated architecture before transformers, poorer performance by cross-subjects. Yousefi *et al*. [16] aimed to address the shortcomings of classic feedforward networks in modeling sequential patterns in EEG by implementing a deep LSTM network that focused solely on the recurrent processing of temporal dependencies in a multichannel signal. They trained preprocessed EEG epochs directly in stacked LSTM layers. They tested subject-dependently on DEAP, achieving 92.5% valence accuracy and 90.1% arousal, surpassing CNN baselines by focusing on memory cells for nonstationary drifts. Its benefits are simplicity and low spatial bias, which are best suited to temporal-heavy tasks; however, it performs poorly on spatial topology without CNN fusion and tends to experience vanishing gradients when used in deeper models, indicating that hybrid attention is necessary in cross-subject settings. Yu *et al*. [17] asked the question of the delay of offline models by making a priority on real-time requirements, extending the sequential capabilities of LSTM with additive attention to dynamically weigh the recent EEG windows with instant valence-arousal information updates. Their LSTM-attention pipeline trained on SEED/SEED-IV achieved 85.4% (SEED) and 74.3% (SEED-IV) accuracy, with 0.5s refresh rates, which is 60x more detections per minute than 10s baselines. Its advantages are the ability to be deployed practically and to withstand noise over short contexts; its disadvantages are trade-offs in accuracy due to limited data per window and dataset-dependent tuning, which prevent generalization across scales without multi-scale inputs. Du *et al*. [18] targeted music-evoked emotions' subtlety, extending 1D-CNN's local pattern detection with BiLSTM for bidirectional context in valence-arousal labeling, addressing unidirectional LSTM's oversight of future cues in EEG music responses. On self-collected ancient Chinese music EEG (14 channels) and public benchmarks, the 1D-CNN-BiLSTM achieved 91% valence/arousal, boosted by repeated convolutional blocks for finer time-series abstraction. Pros: excels in domain-specific (e.g., cultural music) classification with end-to-end raw signal use; cons: heavy reliance on fixed kernel sizes risks missing variable rhythms, plus college-student bias reduces demographic robustness.Liu *et al*. [19] addressed the critiques of black-box DL by developing ERTNet, a CNN-transformer hybrid in which temporal/spatial convolutions preprocess EEG, followed by transformer self-attention that integrates global context to allow interpretable band effects (beta/gamma dominant). Subject-independent tests on DEAP (74.2%) and SEED-V (67.2) were better than pure CNN/LSTM, with Gaussian kernels auto-filtering noise; strengths: visualization of electrode contributions is encouraging you to trust BCI; limitations: Transformer quadratic scaling limits long EEG streams, and intermediate scores across subjects are evidence of a domain gap that you can bridge with your hybrid. Cheng *et al*. [20] went beyond stationary CNNs by proposing a multi-scale dynamic CNN that filters bands using a gated Transformer to sequentially gate emotional motifs, thereby addressing the limitations of uniform scaling in accounting for EEG frequency variance. Assessed on free datasets, it achieved accuracies of up to 96% on within-subject valence, using gates to inhibit irrelevant sequences. Advantages: adaptive to dynamic stimuli such as music/speech; disadvantages: adding gates increases parameters and can overfit without transfer learning, and they are not as interpretable as pure attention competitors. Kotwal *et al*. [21] added to previous EEG studies based on the analysis of raw signals or power spectral densities. They found that these analyses, to a significant extent, failed to reveal subtle spatial correlations among brain areas during emotion processing. Their main improvement was the use of a CNN architecture on Delta, using gamma-band CNN features and differential entropy (DE) features, followed by spatial-frequency analysis to encode electrode connectivity patterns, reducing 62-channel SEED data to 3D feature maps to enable deeper hierarchical learning. In subject-dependent testing of the SEED three-class task (positive, neutral, negative), this achieved 94.8% accuracy across sessions, higher than previous CNNs (e.g., 90-92), and highlighted intra- and inter-hemispheric asymmetries using custom convolutional kernels. To address the limitations of unimodal emotion recognition, Bhatlawande *et al*. [22] proposed a multimodal emotion recognition framework that integrates EEG, ECG, EMG, and facial expression signals to identify the full range of emo-

tions. The data were gathered using the Affective Video Response System and handcrafted time- and frequency-based methods. Wavelet-domain features were extracted from the physiological signals, and the facial expression was described in terms of such descriptors as LBP, LTP, HOG, and Gabor features. In terms of classification, the optimized Random Forest achieved 84% accuracy on physiological signals, and KNN achieved the highest accuracy in face recognition at 84.6%. Still, the last emotion was determined by fusing the modality-specific outputs. A notable strength of this work is its multimodal design, which improves robustness and supports real-time deployment on embedded hardware. Building on these developments, our work integrates a hybrid CNN-LSTM-attention architecture with a broad set of well-established EEG features to capture the complex spatio-temporal dynamics of emotion more effectively.

3. EEG SIGNAL FEATURES AND DATASET

3.1. SEED DATASET

A popular EEG-based emotion recognition benchmark is the SEED (SJTU Emotion EEG Dataset). It contains EEG samples from 15 participants (7 men and 8 women), all of whom participated in three experimental sessions. During the sessions, the subjects were shown 15 film clips expected to elicit positive, neutral, or negative feelings, with 45 trials per participant across all sessions [23], [24]. The ESI NeuroScan system was used to record signals from 62 channels, positioned according to the international 10-20 system. In this research, each trial was divided into 4-second windows, with the extracted features and trained models not overlapping. This window length offers a useful compromise between time resolution and the quasi-stationary assumption often used in EEG analysis.

3.2. EEG PREPROCESSING

EEG signals were bandpass filtered between 1 and 75Hz, and a 50Hz notch filter was used to remove power-line interference. Simple artifact handling and normalization were then performed, followed by feature extraction. The non-overlapping epochs of 4 seconds were then extracted from the continuous recordings, which maintain adequate temporal context and allow efficient downstream processing of the results.

3.3. FEATURE EXTRACTION FRAMEWORK

The proposed model includes five feature extraction methods that represent different neurophysiological processes involved in emotion processing.

3.3.1. Power spectral Density (PSD)

A Welch periodogram with a Hamming window was used to compute the PSD and measure the distribution of spectral power across frequency bands. The spectrum was divided into the five standard EEG bands: Delta (1-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-50 Hz). For each electrode and frequency band, a PSD value was obtained [25]:

$$S_{xx}(f) = (1 / (N * f_s)) * | \Sigma_{\{n=0..N-1\}} x[n] * exp(-j * 2 * pi * f * n / f_s) |^2 \quad (1)$$

where f stands for frequency, f_s for sampling rate, N for window length, and $x[n]$ for EEG samples from that electrode.

3.3.2 Differential Entropy (DE)

Differential Entropy (DE) captures the complexity and uncertainty of EEG signals and is particularly sensitive to nonlinear dynamics associated with emotional states [26]. For a signal segment assumed to follow a Gaussian distribution $N(\mu, \sigma^2)$, the differential entropy is computed as:

$$DE_b = (1/2) * ln (2 * \pi * e * \sigma_b^2) \quad (2)$$

where b denotes the frequency-band index. Higher DE values indicate greater signal complexity and provide an efficient measure of randomness within each frequency range.

3.3.3 Asymmetry Features (RASM, DASM, ASM)

Asymmetry features quantify hemispheric differences in brain activity and are widely used to analyze valence-related emotional processing. These features are computed from symmetric (left–right) electrode pairs, referred to as homologous electrode pairs, which correspond to identical anatomical regions across the two hemispheres. In this study, five homologous electrode pairs are considered: Fp1–Fp2, F3–F4, C3–C4, P3–P4, and O1–O2, representing frontal pole, frontal, central, parietal, and occipital regions, respectively.

The asymmetry measures are extracted across five standard EEG frequency bands: Delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–50 Hz). Let F_r^L and F_r^R denote the feature values (e.g., Differential Entropy (DE) or Power Spectral Density (PSD)) for rhythm r at the left and right hemispheres, respectively.

Three asymmetry measures are defined as follows:

1. **Differential Asymmetry (DASM):**

- DASM captures the absolute difference between left and right hemispheric activity:

$$DASM_r = F_r^L - F_r^R \quad (3)$$

2. **Rational Asymmetry (RASM):**

- RASM represents the relative hemispheric difference:

$$RASM_r = (F_r^L) / (F_r^R) \quad (4)$$

3. **Asymmetry (ASM):**

- ASM is defined as a normalized asymmetry index to account for scale variations:

$$ASM_r = (F_r^L - F_r^R) / (F_r^L + F_r^R) \quad (5)$$

These definitions are consistent with prior studies on EEG-based emotion recognition [27].

The feature dimensions of PSD and DE are 310 each (62 electrodes × 5 frequency bands) [28]. Since asymmetry features are computed from five homologous electrode pairs across five frequency bands, each of DASM, RASM, and ASM contributes 5×5=25 features. These features are concatenated in a fixed order to form the final representation used by the proposed deep learning model.

3.3.4. Feature Fusion and Tensor Construction

For each 4-second EEG window, PSD and DE are computed for all 62 channels across the five standard frequency bands, yielding two channel-wise feature matrices of size 62 × 5. The asymmetry descriptors ASM, DASM, and RASM are computed over five homologous left-right electrode pairs across the same five bands, producing 25 values for each asymmetry descriptor. These outputs are flattened and concatenated in the fixed order [PSD | DE | ASM | DASM | RASM] to form a fused feature vector of length F = 695 for each time window. Consecutive windows from a trial are then stacked over time to form a sequence of length T. For the CNN block, each window is represented as a channel-by-feature matrix, with C = 62, following the fixed SEED/NeuroScan channel order used during acquisition; no 2-D topographic remapping is applied. The 1-D convolution operates along the channel axis to capture local inter-channel dependencies, while the feature axis retains the concatenated spectral and asymmetry descriptors. Because ASM, DASM, and RASM are defined at the electrode-pair level rather than per channel, these descriptors are broadcast across the channel dimension during tensor construction, ensuring each window has a consistent C × F representation. Fig. 2 illustrates the proposed EEG-based emotion recognition methodology.

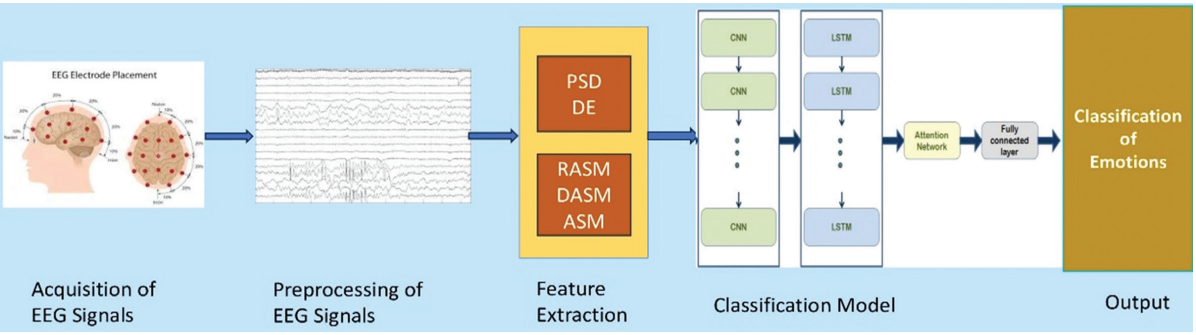


Fig. 2. Overview of the proposed EEG-based emotion recognition framework, illustrating feature extraction, feature fusion, and the attention-augmented CNN–LSTM architecture for spatiotemporal modeling

4. A HYBRID CNN-LSTM ARCHITECTURE FOR EEG ANALYSIS USING BAHDANAU TEMPORAL ATTENTION

This section describes the proposed hybrid CNN-LSTM-attention framework for EEG-based emotion recognition. Each EEG sample is represented as a three-dimensional tensor of size T × C × F, where T denotes the number of consecutive 4-second windows, C = 62 denotes the EEG channels, and F = 695 denotes the fused feature dimension. The network first extracts spatial representations within each window, then models temporal dependencies across the sequence, and finally applies additive attention to emphasize the most informative temporal segments before classification.

4.1. SPATIAL FEATURE ENCODING: TIMEDISTRIBUTED CNN BLOCK

A TimeDistributed 1-D CNN is applied independently to each time window to learn spatial dependencies across EEG channels. For each window, use two Conv1D layers with 32 and 64 filters, respectively, with kernel size 3, ReLU activation, and same padding. The kernel size of 3 was chosen as a compact receptive field for modeling short-range local interactions among neighboring channels while keeping the number of trainable parameters manageable. Each convolutional layer is followed by batch normalization and max pooling with a pool size of 2 to stabilize optimization and reduce redundancy. Global average pooling produces a compact fixed-length representation for each window, and a Dense projection maps this representation to the latent embedding passed to the temporal module. This design preserves channel-wise structure while reducing parameter count and supporting generalization.

4.2. TEMPORAL MODELING: STACKED LONG SHORT-TERM MEMORY LAYERS

The sequence of spatial embeddings is processed by two stacked LSTM layers with 128 and 64 hidden units, respectively, to capture temporal evolution across consecutive EEG windows. Both layers return hidden states for all time steps, enabling hierarchical modeling of short-, mid-, and longer-range dependencies. Dropout (0.3) is applied between the recurrent layers to reduce overfitting. The final hidden state of the second LSTM is used as the query vector q for the attention mechanism. In contrast, the full set of hidden states {h_t} provides the temporal memory over which relevance weighting is performed.

4.3. TEMPORAL RELEVANCE WEIGHTING: BAHDANAU ADDITIVE ATTENTION

Because different temporal segments do not contribute equally to emotion recognition, a Bahdanau additive attention mechanism is applied over the LSTM outputs [29]. For each time step t, an alignment score is computed as

$$e_t = v_a{}^T tanh\,(W_h h_t + WqhT + b_a)\,\alpha_t \qquad (6)$$

where h_t is the hidden state at time step t, and q is the query vector obtained from the final state of the second LSTM. The attention weights are then obtained by $\alpha t = softmax(e_t)$, and the context vector is computed as $c = \Sigma_t\, \alpha_t\, h_t$. This mechanism allows the model to focus on the most discriminative temporal segments and provides interpretable weights for the sequential representations.

The context vector c is forwarded to a Dense layer with ReLU activation, followed by batch normalization and dropout. A final softmax layer outputs the posterior probabilities for the three emotion classes. This stage integrates the spatially encoded, temporally modeled, and attention-weighted information into the final class prediction.

The overall framework is fully trainable end-to-end and jointly learns local spatial patterns, sequential dependencies, and temporal relevance weighting. The design novelty does not lie in introducing a new attention equation, but in the task-specific integration of complementary handcrafted EEG descriptors with a compact spatiotemporal deep architecture tailored to the SEED setting. Fig. 3 presents the architecture of the proposed model.

Pipeline Summary

1. Each EEG trial is organized as a $T \times C \times F$ input tensor.
2. A TimeDistributed 1-D CNN extracts window-level spatial features across channels.
3. Pooling and Dense projection generate compact spatial embeddings for each window
4. Two stacked LSTM layers model temporal dependencies across the embedding sequence.
5. Bahdanau attention assigns importance weights to the hidden states.
6. The weighted hidden states are aggregated into a context vector.
7. A Dense-softmax classifier predicts positive, neutral, or negative emotion.

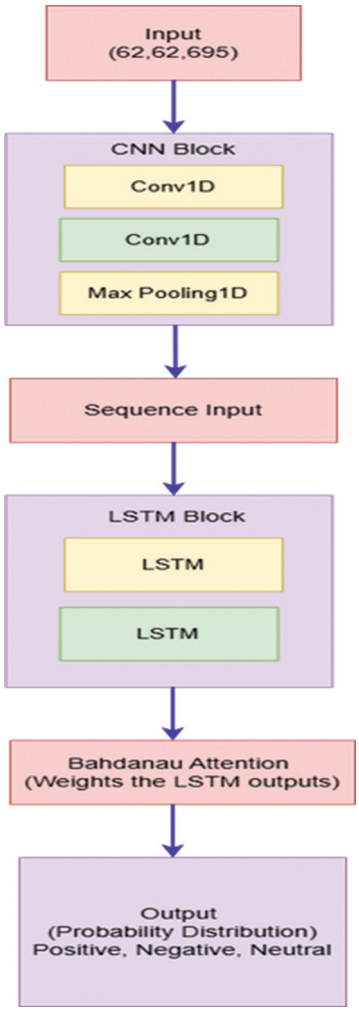


Fig. 3. Detailed architecture of the proposed attention-augmented CNN–LSTM model

5. EXPERIMENTS

5.1. VALIDATION PROTOCOL AND GENERALIZATION SCOPE

The model was evaluated using subject-dependent 5-fold cross-validation on pooled EEG samples derived from all available subjects and sessions. Under this protocol, the folds are disjoint at the sample level, but samples from the same participant may appear in both the training and test partitions across different folds. This setting measures within-dataset discrimination performance in the presence of subject-specific characteristics and therefore typically yields higher accuracy than a strict subject-independent protocol. Accordingly, the reported 98.63% accuracy should be interpreted as strong performance under a subject-dependent validation setting, not as a direct measure of full cross-subject generalization.

5.2. IMPLEMENTATION AND REPRODUCIBILITY DETAILS

The model was implemented in Python using TensorFlow/Keras, while NumPy, SciPy, and scikit-learn were used for preprocessing and evaluation. Each input sample was represented as a $T \times 62 \times 695$ tensor. The spatial encoder used two TimeDistributed Conv1D layers with 32 and 64 filters, kernel size 3, same-padding, ReLU activation, batch normalization, and max pooling (pool

size 2), followed by global average pooling and a Dense projection. The temporal encoder consisted of stacked LSTM layers with 128 and 64 hidden units, respectively, with dropout of 0.3 between recurrent layers. Bahdanau additive attention was applied over the second LSTM outputs, and the resulting context vector was passed to a Dense-ReLU layer and a three-class softmax classifier. To improve generalization and reduce overfitting, batch normalization, dropout (0.3), and L2 weight decay (1×10^{-4}) were applied throughout the network. The Adam optimizer was used with a learning rate of 1×10^{-4}, and categorical cross-entropy served as the loss function. Training was performed for up to 50 epochs with a batch size of 64, and early stopping based on validation loss was employed to avoid overfitting. The experiments were conducted on a workstation equipped with an NVIDIA RTX 3080 GPU (10 GB VRAM), an Intel i7-10700K CPU at 3.8 GHz, and 32 GB DDR4 RAM.

To evaluate the proposed model, a 5-fold subject-dependent cross-validation strategy was adopted, which provides a reliable estimate of performance within the same Dataset while not claiming strict cross-subject generalization. Performance was assessed using accuracy, precision, recall, and F1-score. Fig. 4 presents the corresponding confusion matrices.

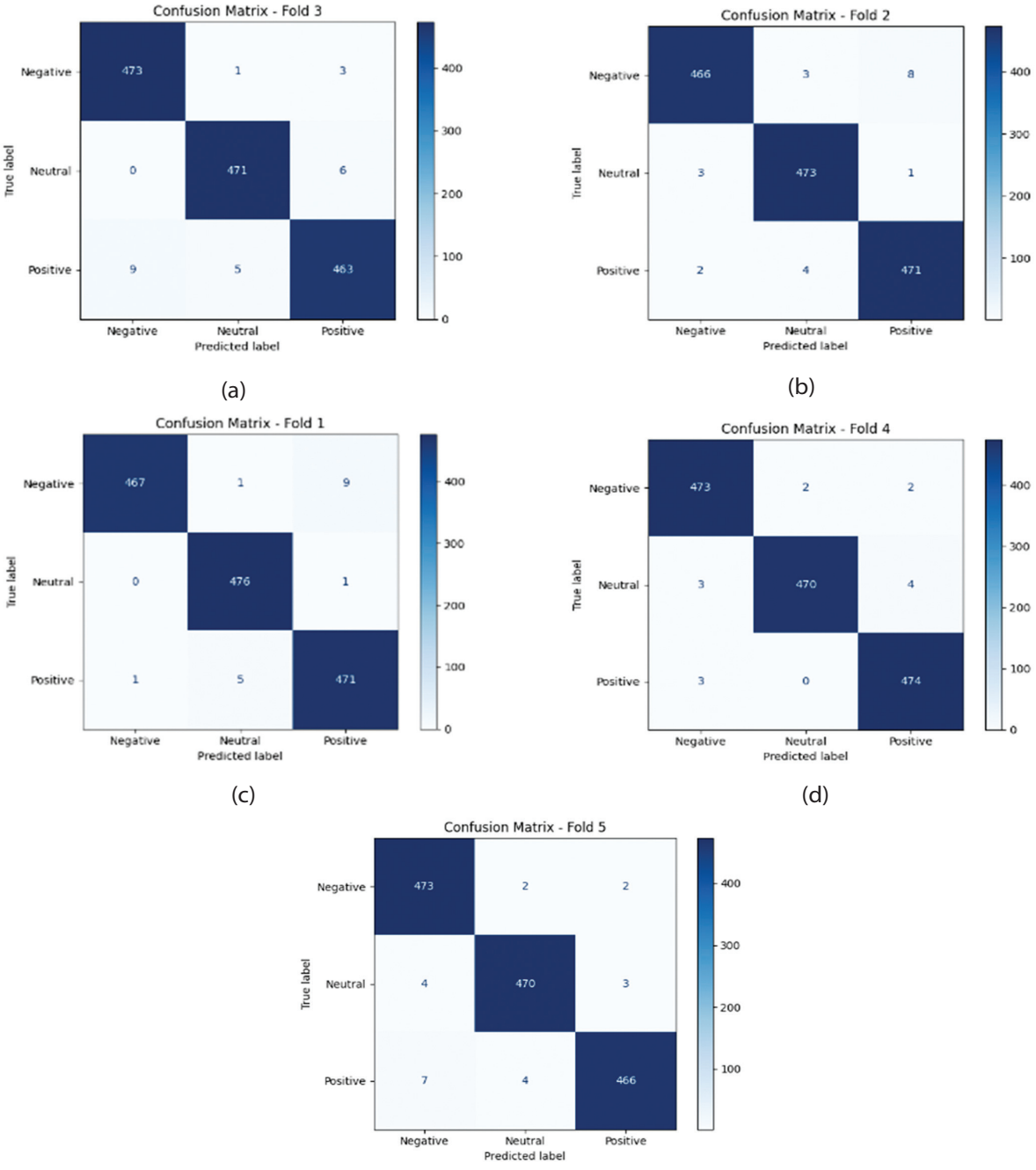


Fig. 4. Confusion matrices obtained from the proposed model under 5-fold cross-validation: (**a**) Fold 1, (**b**) Fold 2, (**c**) Fold 3, (**d**) Fold 4, and (**e**) Fold 5, illustrating classification performance across negative, neutral, and positive emotion classes

Table 1 compiles representative results from the EEG emotion-recognition literature to contextualize the proposed model, and each reported accuracy is interpreted with respect to the corresponding study's dataset and evaluation protocol.

Table 1. Evaluation of Proposed Model Versus Existing Methods on EEG Dataset

Authors and Year	Model Architecture	Datasets	Best Accuracy
Xu *et al.*[12] (2023)	GRU-Conv	DEAP, SEED	87.04% (SEED)
Iyer *et al.* [14] (2023)	CNN-LSTM Ensemble	DEAP, SEED	97.16% (SEED)
Li *et al.* [15] (2020)	CNN-LSTM + Attention	DEAP, SEED	93%
Lu *et al.* [30] (2021)	Multimodal DCCA	SEED	94.6%
Yu *et al.* [17] (2023)	LSTM + Additive Attention	SEED, SEED-IV	85.40% (SEED)
J.Zhang *et al.* [31] (2024)	Band frequency + connectivity	SEED	92.29%
Liu *et al.* [19] (2024)	ERTNet (Transformer)	DEAP, SEED-V	74.23% (DEAP)
Hu. F *et al.* [32] (2024)	GCN + adaptive multi feature	SEED	97.89%
M. Li *et al.* [33] (2025)	Spatial temporal Transformer	SEED	92.67%
Ma. Y *et al.* [34] (2025)	Multi scale CNN + Bi LSTM + Attention	SEED	96.67%
Our Proposed Model	CNN-LSTM + additive attention	SEED	98.63% (subject-dependent)

6. RESULTS AND DISCUSSION

The proposed hybrid model was evaluated under the subject-dependent 5-fold protocol described in Section 5.1, and the results are summarised in Table 2. The model achieved consistently high performance across all folds, with an average classification accuracy of 98.63%. In Fold 1, for example, 467 of 477 negative samples were classified correctly. At the same time, the neutral and positive classes also showed strong precision and recall with only a small number of confusions between adjacent emotional states. Fold 2 displayed a similar pattern, with 466 negative, 473 neutral, and 471 positive samples correctly classified. The limited number of errors occurred mainly between neighboring affective categories, such as negative-neutral and neutral-positive, which is expected because these states can exhibit partially overlapping EEG patterns. The strong overall performance can be attributed to three factors. First, the feature set combines frequency-domain descriptors (PSD and DE) with hemispheric asymmetry measures (RASM, DASM, and ASM), providing a rich representation of emotion-related brain dynamics. Second, the CNN-LSTM architecture captures the dual nature of EEG data by modeling spatial dependencies across channels and temporal dependencies across successive windows. Third, the attention mechanism enables the network to emphasize the most informative temporal segments for classification. At the same time, the high accuracy should be interpreted in light of the adopted subject-dependent protocol, which is less stringent than a fully subject-independent evaluation and may therefore produce more optimistic estimates of generalisability.

Table 2. Classification performance for emotion classes under 5-fold cross-validation

Fold	Labels	TP	FP	FN	Precision	Recall	F1-score
1	Negative	467	2	10	99.57%	97.90%	98.73%
	Neutral	476	6	1	98.75%	99.79%	99.27%
	Positive	471	10	6	97.92%	98.74%	98.33%
2	Negative	466	5	11	98.94%	97.70%	98.31%
	Neutral	473	7	4	98.54%	99.16%	98.85%
	Positive	471	9	6	98.13%	98.74%	98.43%
3	Negative	473	9	4	98.13%	99.2%	98.66%
	Neutral	471	6	6	98.74%	98.74%	98.74%
	Positive	463	9	14	98.10%	97.1%	97.6%
4	Negative	473	6	4	98.75%	99.2%	98.97%
	Neutral	470	2	7	99.58%	98.53%	99.05%
	Positive	474	6	3	98.75%	99.37%	99.06%
5	Negative	473	11	4	97.67%	99.2%	98.43%
	Neutral	470	6	7	98.74%	98.53%	98.63%
	Positive	466	5	11	98.9%	97.7%	98.3%

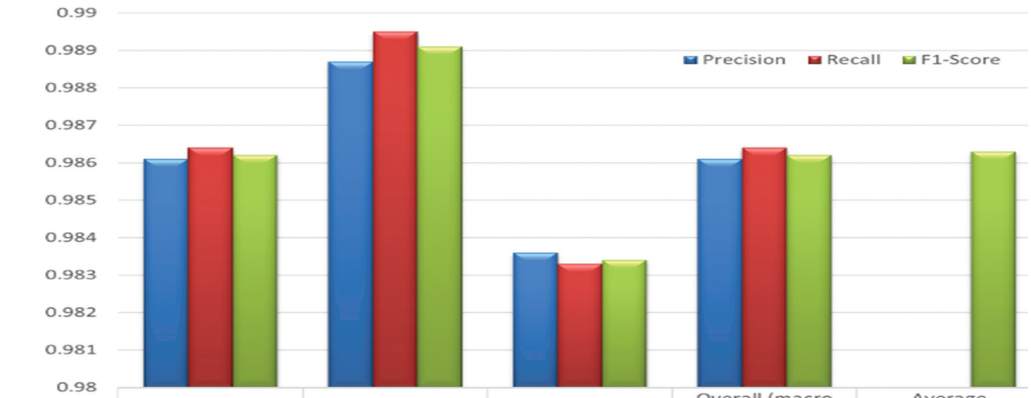
6.1. COMPARATIVE ANALYSIS OF MODEL COMPONENTS AND FEATURE REPRESENTATIONS

The comparative results provide clear evidence that both model architecture and feature representation substantially influence EEG-based emotion recognition performance. Among the model components, the spatial-only configuration achieved 86.95% accuracy, whereas the temporal-only configuration improved to 90.20%. This implies that temporal dependency modeling provides a greater contribution than spatial extraction in the current scenario, with an absolute gain of 3.25 percentage points. But the hybrid spatial-temporal structure achieved the best accuracy of 94.98, respectively beating the spatial-only and temporal-only counterparts by 8.03 and 4.78 percentage points, respectively. These advantages support the idea that neither spatial nor temporal modeling alone is adequate to summarize the full discriminative structure of emotional EEG signals. Instead, the better performance of the hybrid CNN-LSTM attention model indicates that emotion-related neural patterns are simultaneously defined by spatially distributed channel-level representations and temporally dynamic dependencies, both of which need to be learned together.

The complementary nature of the mechanisms that comprise the hybrid model can explain the model's stronger performance. Spatial learning via CNNs is useful for deriving localized inter-channel regularities and detecting salient topological differences across electrodes in the EEG, which are commonly linked to emotion-specific cortical activity. By contrast, LSTM-based temporal modeling can capture sequential dependencies and dynamic changes in neural activity over time, which are important because emotional responses are non-instantaneous and nonstationary. A combination of these two mechanisms provides a more detailed representation space in which spatial correlations can be maintained and temporal evolution modeled simultaneously. The resulting enhancement of about 9.23 percent relative to the spatial-only proxy and 5.30 percent relative to the temporal-only proxy further supports the importance of spatiotemporal fusion in creating powerful emotion classifications.

The same tendency is evident in the feature-level comparison. Differential Entropy (DE) scored highest at 90.37, followed by Power Spectral Density (PSD) at 83.17, and asymmetry-based features at 77.19. DE thus outperformed PSD by 7.20 percentage points and the lack of asymmetry features by 13.18 percentage points, indicating a clear advantage in representational efficiency. The fact that this result is technically plausible is that DE is more sensitive to the statistical distribution and the variance of EEG energy, and can therefore detect small changes in signal complexity, which are strongly linked to affective brain states. Even though PSD is a well-defined measure of spectral power, it is more of a measure of frequency-domain energy. It may not necessarily capture the higher-order distributional properties inherent in emotional EEG activity. The less favorable results for asymmetry features indicate that hemispheric imbalance alone, despite its neurophysiological significance, can provide only partial affective information when applied separately. Taken together, these results suggest that effective EEG emotion recognition must have expressive signal descriptors and architectures that can represent spatiotemporal dependencies. In that regard, DE provides a more discriminative feature space.

In contrast, the hybrid CNNLSTM approach offers the most appropriate modeling strategy for leveraging the complexity, dynamism, and dispersion of emotional EEG signals. Specifically, the 98.63% accuracy in the SEED dataset with the 5-fold protocol demonstrates the efficiency of the offered combination of spectral and asymmetry characteristics with CNN-based spatial representations, LSTM-based temporal learning, and attention-based feature redefinition. Fig. 6 represents the EEG Emotion Recognition experimental comparison results.



	Negative	Neutral	Positive	Overall (macro avg)	Average Accuracy
Precision	0.9861	0.9887	0.9836	0.9861	
Recall	0.9864	0.9895	0.9833	0.9864	
F1-Score	0.9862	0.9891	0.9834	0.9862	0.9863

Fig. 5. Performance evaluation of the proposed model showing accuracy, precision, recall, and F1-score across different validation folds

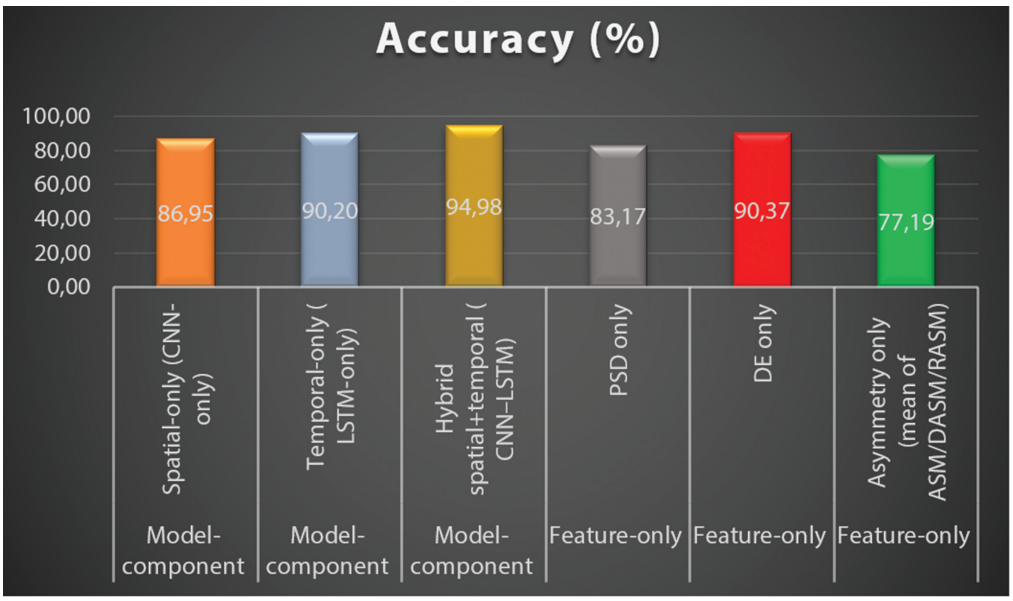


Fig. 6. EEG Emotion Recognition Experimental Comparison Results Accuracy (%)

7. CONCLUSION AND FUTURE WORK

The paper introduced a hybrid CNN-LSTM-attention model to EEG emotion recognition using the SEED dataset. The proposed framework, which incorporates a variety of handcrafted properties, such as PSD, DE, and various asymmetry measures, along with an attention-enhanced deep learning model, achieved an average classification accuracy of 98.63% across three emotional classes under a subject-dependent 5-fold cross-validation scheme. Despite overall good performance, the remaining misclassifications are predominantly between emotionally related states, indicating that there remains a challenge in affective neuroscience. Future research will take four directions, namely: the integration of explainable AI (XAI) methods to enhance interpretability; a comprehensive quantitative ablation study of groups of features and model components; testing the framework in stricter subject-independent and cross-dataset conditions to understand generalisability better; and applying the method to more realistic continuous emotion recognition tasks.

DATA AVAILABILITY STATEMENT:

The SEED Dataset used in the present paper is publicly accessible at https://bcmi.sjtu.cn/home/seed/seed.html.

DECLARATION OF THE USE OF AI TOOLS:

AI tools are not used.

REFERENCES

[1] S. Poria, E. Cambria, R. Bajpai, A. Hussain, "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion", Information Fusion, Vol. 37, 2017, pp. 98-125. doi: 10.1016/j.inffus.2017.02.003

[2] H. Liu, Y. Zhang, Y. Li, X. Kong, "Review on Emotion Recognition Based on Electroencephalography", Frontiers in Computational Neuroscience, Vol. 15, 2021, p. 758212. doi: 10.3389/fncom.2021.758212

[3] E. Dritsas, M. Trigka, "Application of Deep Learning for Heart Attack Prediction with Explainable Artificial Intelligence", Computers, Vol. 13, 2024, p. 244. doi: 10.3390/computers13100244

[4] A. Kotwal, V. Gupta, M. Verma, J. Manhas, V. Sharma, "EEG-based Emotion Recognition: Leveraging CNNs for Precision", Proceedings of the 2nd International Conference on Next Generation Computing and Information Systems, Jammu, India, 18-19 December 2023, p. 151. doi: 10.1201/9781003466383

[5] R. J. Davidson, "Emotion and Affective Style: Hemispheric Substrates", Psychological Science, Vol. 3, No. 1, 1992, pp. 39-43. doi: 10.1111/j.1467-9280.1992.tb00254.x

[6] S. Zhang, C. Chu, X. Zhang, X. Zhang, "EEG Emotion Recognition Using AttGraph: A Multi-Dimensional Attention-Based Dynamic Graph Convolutional Network", Brain Sciences, Vol. 15, No. 6, 2025, p. 615. doi: 10.3390/brainsci15060615

[7] F. I. Mouri, C. E. Valderrama, S. G. Camorlinga, "Identifying Relevant Asymmetry Features of EEG for Emotion Processing", Frontiers in Psychology, Vol. 14, 2023, p. 1217178. doi: 10.3389/fpsyg.2023.1217178

[8] M. A. Ozdemir, M. Degirmenci, E. Izci, A. Akan, "EEG-based Emotion Recognition with Deep Convolutional Neural Networks", Biomedical Engineering / Biomedizinische Technik, Vol. 66, No. 1, 2021, pp. 43-57. doi: 10.1515/bmt-2019-0306

[9] B. Chakravarthi, S.-C. Ng, M. R. Ezilarasan, M.-F. Leung, "EEG-based Emotion Recognition Using Hybrid CNN and LSTM Classification", Frontiers in Computational Neuroscience, Vol. 16, 2022, p. 1019776. doi: 10.3389/fncom.2022.1019776

[10] B. Chakravarthi, S.-C. Ng, M. R. Ezilarasan, M.-F. Leung, "EEG-based Emotion Recognition Using Hybrid CNN and LSTM Classification", Frontiers in Computational Neuroscience, Vol. 16, 2022, p. 1019776. doi: 10.3389/fncom.2022.1019776

[11] A. Kotwal, V. Sharma, J. Manhas, "Deep Neural Based Learning of EEG Features Using Spatial, Temporal and Spectral Dimensions Across Different Cognitive Workload of Human Brain: Dimensions, Methodologies, Research Challenges and Future Scope", Proceedings of the International Conference on Emerging Trends in Expert Applications and Security, Jaipur, India, 17-19 February 2023. doi: 10.1007/978-981-99-1946-8_7

[12] X. Shi, Q. She, F. Fang, M. Meng, T. Tan, Y. Zhang, "Enhancing Cross-Subject EEG Emotion Recognition Through Multi-Source Manifold Metric Transfer Learning", Computers in Biology and Medicine, Vol. 174, 2024, p. 108445. doi: 10.1016/j.compbiomed.2024.108445

[13] M. Zuo, F. Wang, "Application of Stock Trading-Related Emotion Recognition from EEG Signals Using Deep Learning EEGNet", Proceedings of the 4th International Conference on Artificial Intelligence and Computer Engineering, Dalian, China, 17-19 November 2023. doi: 10.1145/3652628.3652734

[14] A. Iyer, S. S. Das, R. Teotia, S. Maheshwari, R. R. Sharma, "CNN and LSTM Based Ensemble Learning for Human Emotion Recognition Using EEG Recordings", Multimedia Tools and Applications, Vol. 82, No. 4, 2023, pp. 4883-4896. doi: 10.1007/s11042-02212310-7

[15] N. Piedrabuena, W. M. Elmannai, Y. Wang, "Emotion Recognition From EEG Signals Using a Hybrid Deep Learning Approach", IEEE Access, Vol. 14, 2026, pp. 11035-11047. doi: 10.1109/ACCESS.2025.3642249

[16] M. R. Yousefi, A. Dehghani, H. Taghaavifar, "Enhancing the Accuracy of Electroencephalogram-Based Emotion Recognition Through Long Short-Term Memory Recurrent Deep Neural Networks", Frontiers in Human Neuroscience, Vol. 17, 2023, p. 1174104. doi: 10.3389/fnhum.2023.1174104

[17] X. Yu, Z. Li, Z. Zang, Y. Liu, "Real-Time EEG-Based Emotion Recognition", Sensors, Vol. 23, No. 18, 2023, p. 7853. doi: 10.3390/s23187853.https://doi.org/10.3390/s23187853

[18] R. Du, S. Zhu, H. Ni, T. Mao, J. Li, R. Wei, "Valence-Arousal Classification of Emotion Evoked by Chinese Ancient-Style Music Using 1D-CNN-BiLSTM Model on EEG Signals for College Students", Multimedia Tools and Applications, Vol. 82, No. 10, 2023, pp. 15439-15456. doi: 10.1007/s11042-022-14011-7

[19] 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, Vol. 18, 2024, p. 1320645. doi: 10.3389/fnins.2024.1320645

[20] Z. Cheng *et al*. "EEG-based Emotion Recognition Using Multi-Scale Dynamic CNN and Gated Transformer", Scientific Reports, Vol. 14, 2024, p. 31319. doi: 10.1038/s41598-024-82705-z

[21] A. Kotwal, M. Verma, J. Manhas, V. Sharma, "Robust EEG-Based Emotion Recognition Using CNN: A High-Accuracy Approach with Differential Entropy Features and Spatial-Frequency Domain Analysis on the SEED Dataset", Journal of Scientific Research, Vol. 17, No. 3, 2025, pp. 777-787. doi: 10.3329/jsr.v17i3.78907

[22] S. Bhatlawande, S. Shilaskar, S. Pramanik, S. Sole, "Multimodal Emotion Recognition Based on the Fusion of Vision, EEG, ECG, and EMG Signals", International Journal of Electrical and Computer Engineering Systems, Vol. 15, No. 1, 2024, pp. 41-58. doi: 10.32985/ijeces.15.1.5

[23] R.-N. Duan, J.-Y. Zhu, B -L. Lu, "Differential entropy feature for EEG-based emotion classification", Proceedings of the 6th International IEEE/EMBS Conference on Neural Engineering, San Diego, CA,

USA, 6-8 November 2013, pp. 81-84. doi: 10.1109/NER.2013.6695876

[24] 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, Vol. 7, No. 3, 2015, pp. 162-175. doi: 10.1109/TAMD.2015.2431497

[25] S. Hwang *et al*. "Learning CNN Features from DE Features for EEG-Based Emotion Recognition", Pattern Analysis and Applications, Vol. 23, 2020, pp. 1323-1335. doi: 10.1007/s10044-019-00860-w

[26] M. Alsolamy, A. Fattouh, "Emotion Estimation from EEG Signals During Listening to Quran Using PSD Features", Proceedings of the 7th International Conference on Computer Science and Information Technology, Amman, Jordan, 13-14 July 2016, pp. 1-5. doi: 10.1109/CSIT.2016.7549457

[27] G. Zhang, M. Yu, Y.-J. Liu, G. Zhao, D. Zhang, W. Zheng, "SparseDGCNN: Recognizing Emotion From Multichannel EEG Signals", IEEE Transactions on Affective Computing, Vol. 14, No. 1, 2023, pp. 537-548. doi: 10.1109/TAFFC.2021.3051332

[28] W.-L. Zheng, J.-Y. Zhu, Y. Peng, B.-L. Lu, "EEG-based Emotion Classification Using Deep Belief Networks", Proceedings of the IEEE International Conference on Multimedia and Expo, Chengdu, China, 14-18 July 2014, pp. 1-6. doi: 10.1109/ICME.2014.6890166

[29] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate", arXiv:1409.0473, 2014

[30] J. Li, S. Li, J. Pan, F. Wang, "Cross-Subject EEG Emotion Recognition With Self-Organized Graph Neural Network", Frontiers in Neuroscience, Vol. 15, 2021, p. 611653. doi: 10.3389/fnins.2021.611653

[31] J. Zhang, Y. Hao, X. Wen, C. Zhang, H. Deng, J. Zhao, R. Cao, "Subject-Independent Emotion Recognition Based on EEG Frequency Band Features and Self-Adaptive Graph Construction", Brain Sciences, Vol. 14, No. 3, 2024, p. 271. doi: 10.3390/brainsci14030271

[32] F. Hu, K. He, M. Qian, X. Liu, Z. Qiao, L. Zhang, J. Xiong, "STAFNet: An Adaptive Multi-Feature Learning Network via Spatiotemporal Fusion for EEG-Based Emotion Recognition", Frontiers in Neuroscience, Vol. 18, 2024, p. 1519970. doi: 10.3389/fnins.2024.1519970

[33] M. Li, P. Yu, Y. Shen, "A Spatial and Temporal Transformer-Based EEG Emotion Recognition in VR Environment", Frontiers in Human Neuroscience, Vol. 19, 2025, p. 1517273. doi: 10.3389/fnhum.2025.1517273

[34] Y. Yang, Z. Chen, Q. Dong, S. Zhang, Y. Li, "MSBiLSTM-Attention: EEG Emotion Recognition Model Based on Spatiotemporal Feature Fusion", Biomimetics, Vol. 10, No. 3, 2025, p. 178. doi: 10.3390/biomimetics10030178