Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00171-2

Cureus
Part of SPRINGER NATURE

Original Article

CBSAtt: A Hybrid Convolutional Neural Network–Bidirectional Long Short-Term Memory–Attention Framework for Electroencephalography-Based Emotion Recognition

Ashwani Yadav [1], Rashi Agarwal [1]

1. *Computer Science and Engineering, Harcourt Butler Technical University, Kanpur, IND*

Received: May 15, 2026 | Review began: June 07, 2026 | Review ended: July 10, 2026 | Published: July 20, 2026

Abstract

Electroencephalography (EEG) emotion recognition signals have become one of the foundations of affective computing, enabling advancements in human-computer interaction, mental health diagnostics, and personalized user experiences. This paper presents CBSAtt, a deep learning architecture that synergistically merges Convolutional Neural Networks, Bidirectional Long Short-Term Memory (BiLSTM) networks, and multi-head self-attention mechanisms to classify emotions effectively. The proposed model uses pre-computed differential entropy features from the SEED-IV dataset, which is provided by the Center for Brain-like Computing and Machine Intelligence at the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. The dataset captures spectral information across five frequency bands from 62 EEG channels. Our architecture uses a 1D-Convolutional Neural Network to learn salient spatial features from the multidimensional input, followed by a BiLSTM to model long-range time-dependent dependencies. A multi-head self-attention layer is then used selectively on the most informative temporal segments, enhancing discriminative capability. The model is measured in a subject-dependent environment throughout all three sessions with the SEED-IV dataset, attaining an average accuracy of 95.50% with best-epoch checkpointing. This result demonstrates improved performance compared to transformer-based approaches such as AMDET (87.32%) on the same dataset, demonstrating the powerful synergy of convolutional, recurrent, and attentional mechanisms in decoding complex neural patterns associated with emotional states. Our findings provide a robust foundation for developing advanced affective brain-computer interfaces.

Categories: AI/ML-based decision support systems, Machine Learning (ML), Deep Learning

Keywords: emotion recognition, electroencephalography, deep learning, attention mechanism, seed-iv dataset, affective computing, convolutional neural network (cnn), bidirectional long short-term memory (bilstm)

Introduction

Emotions play a central role in human cognition, which affects decision-making, social processes, and general psychological well-being. The capability to automatically and objectively detect emotional states has far-reaching implications in creating intelligent systems that are able to detect emotional responses, from individualized mental health interventions to adaptive learning environments and immersive entertainment platforms.

Electroencephalography (EEG) has become popular among other physiological modalities as it is non-invasive, has high temporal resolution, and enables direct recording of neural activity in affective processes. EEG captures the brain's activity across different frequency ranges, including delta waves (1-4 Hz), theta waves (4-8 Hz), alpha waves (8-13 Hz), beta waves (13-30 Hz), and gamma waves (30-50 Hz), each of which is linked to specific cognitive and emotional functions. EEG is an important measure of specific cognitive and emotional functions and a valuable resource for affective computing investigations [1].

The SEED-IV dataset, developed by Zheng et al. [2,3], has become a reference in EEG emotion recognition research. It offers pre-extracted differential entropy (DE) features, a measure of signal complexity. DE has shown better results in emotion classification problems than conventional power spectral density features. The uncertainty and information captured in DE features make them appropriate for determining the discrimination of various emotional states. The dataset contains records of 15 subjects across three sessions and 24 trials, each of which elicits four specific emotions: happy, sad, fear, and neutral.

Recent advances in deep learning and transformer architectures [4,5] have transformed how emotions are recognized using EEG. Convolutional Neural Networks (CNNs) have proven to be more effective at identifying spatial patterns and learning local features from multi-channel EEG data, especially when considering different combinations of channels and frequency bands. Recurrent architectures, especially Long Short-Term Memory (LSTM) networks, excel at modeling sequential brain temporal dependencies and activity [6], capturing how neural patterns evolve over time. Attention mechanisms have been added more recently to concentrate on emotionally salient periods of time, imitating selective processing of relevant stimuli in the brain [7,8]. Other architectures that have been studied include transformer-based and hybrid deep learning approaches.

Rodriguez et al. [9] explored feature transformation using multidimensional multiclass genetic programming (M3GP) for EEG-based emotion recognition, achieving 92.10% accuracy and demonstrating the effectiveness of evolutionary feature engineering techniques. Regardless of these advances, current methods tend to fail to simultaneously capture the full complexity of EEG signals - the complex spatial relationships across brain regions, the changing temporal dynamics of emotional responses, and the selective consideration of crucial points in a trial. Although recent transformer-based and hybrid architectures have demonstrated strong performance in EEG emotion recognition, many existing methods still face challenges in simultaneously capturing spatial relationships across brain regions, temporal dynamics, and the relative importance of different time segments within an EEG trial. This paper overcomes such constraints by suggesting CBSAtt, a hybrid architecture that incorporates:

1) A 1D-CNN for spatial feature extraction from high-dimensional DE representations, learning local patterns in channel-band combinations through hierarchical convolutional processing.

2) A Bidirectional LSTM (BiLSTM) model to perform overall temporal modeling, which includes forward and backward contextual dependencies with the help of bidirectional recurrent networks.

3) A multi-head self-attention mechanism for adaptive temporal weighting, which allows the model to highlight the most important parts by using parallel attention heads.

Our main contributions are as follows:

1) We introduce CBSAtt, a new hybrid system that combines CNNs, BiLSTMs, and self-attention mechanisms to work together for emotion recognition using EEG signals. This approach handles the spatial, temporal, and attention-based features of brain signals within one unified framework.

2) The proposed model achieves an average accuracy of 95.50% on the SEED-IV dataset across 15 subjects using best-epoch checkpointing, demonstrating competitive performance compared with recent state-of-the-art EEG emotion recognition methods.

3) We report detailed experimental studies, including ablation studies that quantify the contribution of each architectural component, confusion matrices revealing per-class performance, and comparisons with state-of-the-art methods from 2022-2025.

4) We provide complete reproducibility by using fixed random seeds and model checkpointing, building a reliable foundation for future EEG emotion recognition research.

Materials And Methods

Related work

EEG-Based Emotion Recognition

Emotion recognition using EEG has developed significantly over the past decade. Early methods used hand-crafted features extracted from EEG signals, such as power spectral density, wavelet coefficients, and statistical moments. These features were often paired with traditional machine learning models such as Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN). While these methods established foundational techniques, they were constrained by the representational capacity of hand-crafted features and the shallow nature of the classifiers. DE proved to be a very useful emotion recognition feature. DE quantifies EEG signal complexity and uncertainty and has been found, in comparison with traditional features, to be more discriminative. In contrast to power spectral density, which captures only energy, DE captures the information content and randomness of neural activity, making it more sensitive to subtle emotional changes. Zheng et al. [3] demonstrated that DE features outperform power spectral density features across multiple frequency bands on a regular basis, establishing DE as a favored emotion recognition representation. The SEED dataset series has proven to be very useful in furthering the field, providing standardized pre-processed DE characteristics that allow direct comparison of alternative approaches.

Hybrid Deep Learning Approaches

Hybrid architectures combining CNNs and RNNs have shown significant potential in EEG-based emotion recognition. The CNN-LSTM paradigm, in which CNNs extract spatial features and LSTMs model temporal dependencies, has been widely studied. These approaches leverage the complementary strengths of convolutional and recurrent networks - CNNs excel at learning localized spatial patterns across EEG channels, while LSTMs learn the time development of these patterns [6]. The combination enables the model to learn representations that are both spatially discriminative and temporally coherent. Li et al. [7] suggested a BiLSTM and attention-based emotion recognition approach, proving that attention-weighted temporal aggregation enhances classification accuracy because it concentrates on time segments of emotional salience. Their work showed that not every temporal location produces an equal amount of information for emotion classification - the first reaction toward a stimulus can vary from a prolonged emotional condition, and attention mechanisms can adaptively weight these different phases. Li et al. [8] proposed an attention-based CNN-LSTM architecture that learns both spatial and temporal features in an adaptive way, obtaining competitive performance on SEED data through a cross-dimensional attention mechanism.

Attention-Based and Transformer Approaches

Recently, transformer architectures have been applied to EEG-based emotion recognition as an alternative to recurrent models for capturing long-range dependencies. Vaswani et al. [4] suggested the multi-head self-attention mechanism, which allows models to understand the connections between every pair of positions in a sequence by performing attention calculations in parallel. In contrast to RNNs, which process sequences sequentially, transformers can access any position regardless of distance, which may allow them to reflect long-range dependencies more effectively. Recent transformer-based approaches have demonstrated the effectiveness of attention mechanisms for EEG emotion recognition. Transformer architectures can model long-range dependencies through self-attention mechanisms and capture complex relationships across EEG channels, frequency bands, and temporal dimensions. These approaches have shown promising performance on benchmark emotion recognition datasets, highlighting the potential of transformer-based models for EEG analysis [10]. Dhara et al. [11] proposed a fuzzy ensemble-based deep learning model for EEG emotion recognition, demonstrating the effectiveness of combining multiple complementary architectures. Recent studies [12-16] have further improved EEG emotion recognition through advanced deep learning and transformer-based architectures. Recent transformer-based approaches have demonstrated strong performance in EEG emotion recognition; however, our proposed CBSAtt model achieves 95.50% accuracy by effectively integrating convolutional, recurrent, and attention-based processing within a unified framework.

Recent Advances (2022-2025)

Recent years have witnessed significant progress through innovative deep learning architectures. Xie et al. [10] proposed a transformer-based approach that combines deep learning networks with spatial-temporal information for EEG classification, demonstrating the effectiveness of transformer architectures in capturing complex EEG patterns.

Dong et al. [13] proposed the Emotion Preceptor model, which integrates static spatial adapters with temporal causal networks for cross-subject emotion recognition. Their methodology consists of normalizing individual differences among subjects by means of spatial adapters, followed by temporal causal networks to capture the dynamics of emotional responses, achieving 95.35% accuracy on the SEED dataset.

Bagherzadeh et al. [14] proposed a new approach that integrates useful connectivity maps with ensemble deep learning. By combining transfer entropy, partial directed coherence, and direct directed transfer function into 2D images and using pre-trained CNNs with LSTM, their model achieved 98.76% accuracy on the DEAP dataset. This work highlights the importance of multimodal feature fusion and ensemble learning to enhance recognition performance.

Hu et al. [15] suggested an adaptive dual-graph learning framework (DGLFS), which combines domain-invariant feature selection, label propagation, and dual-graph regularization for inter-subject and inter-session emotion identification. Their method builds both local similarity graphs and global structure graphs to encode EEG samples with relationships, achieving 86.13% accuracy on SEED-IV.

Zhao et al. [12] proposed TDFNet, a transformer-based deep-scale fusion network for multimodal emotion recognition, demonstrating the effectiveness of transformer architectures in integrating information from multiple modalities.

Mathumitha and Maryposonia [16] developed a proximity-conserving autoencoder (PCAE) with ensemble classifiers combining SVM, random forest, and LSTM, achieving 98.87% accuracy on an EEG brainwave dataset through effective feature extraction and ensemble learning.

Dhara et al. [11] proposed a fuzzy ensemble-based deep learning model that combines multiple complementary deep learning architectures, achieving 94.70% accuracy on the DEAP dataset.

Our proposed CBSAtt architecture builds on these methods by combining CNN-based spatial feature extraction, BiLSTM-based temporal modeling, and multi-head self-attention within a single integrated framework. By combining the strengths of convolutional processing for spatial feature extraction, recurrent processing for temporal modeling, and attention mechanisms for selective focusing, CBSAtt achieves 95.50% accuracy on SEED-IV, surpassing the performance reported by Rodriguez et al. [9], who employed deep learning with M3GP for EEG-based emotion recognition, thereby demonstrating the effectiveness of hybrid architectures for this task.

Dataset

SEED-IV Dataset Description

The SEED-IV dataset [3] is a comprehensive resource for EEG-based emotion recognition research developed by Shanghai Jiao Tong University's Brain and Cognitive Science Lab. It includes recordings of 15 healthy subjects (7 males, 8 females, aged 20-24 years) across three different sessions, with pauses of about one week between sessions to minimize fatigue and learning effects. Each session contains 24 trials, with participants viewing specially chosen movie clips that are meant to inspire four different kinds of emotions: happy, sad, fear, and neutral. The emotion labels were confirmed through self-assessment by participants and annotation by experts, guaranteeing high-quality ground truth data. EEG recording was performed using a 62-channel ESI NeuroScan recording system sampled at 1000 Hz. Electrodes were placed according to the international 10-20 system, covering the entire scalp and providing complete cortical coverage of activity. The data includes pre-extracted DE features that were calculated across five different frequency ranges: Delta (1-4 Hz), Theta (4-8 Hz), Alpha (8-13 Hz), Beta (13-30 Hz), and Gamma (30-50 Hz). DE is defined as

$$h(X) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \times \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx = \frac{1}{2}\log(2\pi e\sigma^2)$$

where σ^2 is the signal variance. This formulation captures the ambiguity and content of EEG signals, which can effectively represent emotion classification and is more discriminative than conventional power spectral density features. The DE features are stored in MATLAB '.mat' files with keys 'deLDS{trial}' for trial numbers 1-24. The corresponding emotion labels for each session are provided in separate label maps, with the mapping from trial numbers to emotions being session-specific. Table 1 summarizes the key characteristics of the SEED-IV dataset.

Data Preprocessing

Each subject has DE features loaded from all three sessions. For each trial, the feature matrix has dimensions (T, 310), where T is the number of time steps and 310 corresponds to the flattened channel-band features (62 channels × 5 bands).

Characteristic	Value
Number of subjects	15
Number of sessions	3
Trials per session	24
Total trials	1,080
Emotion classes	Neutral, Sad, Fear, Happy
EEG channels	62
Frequency bands	5 (Delta, Theta, Alpha, Beta, Gamma)
Sampling rate	1,000 Hz
Feature dimension	310 (62 channels × 5 bands)

TABLE 1: Summary of SEED-IV dataset characteristics

EEG, Electroencephalography

To account for individual differences and baseline shifts, we perform per-trial standardization:

$$\hat{X}_{trial} = \frac{X_{trial} - \mu_{trial}}{\sigma_{trial} + \epsilon}$$

where μ_{trial} and σ_{trial} are the mean and standard deviation of all time steps of the trials, and $\epsilon = 1 \times 10^{-6}$ is a small constant that ensures numerical stability. The per-trial normalization is effective in eliminating session effects while maintaining the relative changes that are reflective of emotional responses [3]. In contrast to global normalization across all trials, the emotional contrasts in a trial are preserved, but baseline drift is eliminated in per-trial normalization.

We augmented the dataset to allow us to model time. We divided each trial into overlapping windows. We employed a window size of 4 seconds and a stride of 2 seconds (50% overlap), which has been shown to be effective in previous work [6]. Given that the DE features are provided at approximately 1 Hz, this corresponds to windows of length 4 samples. This overlapping windowing strategy provides several advantages:

• Data augmentation: Overlap increases the number of samples by a factor of approximately (T − 4)/2, assisting in the prevention of overfitting and allowing the model to study several temporal views of the same emotional response.

• Temporal context: Windows maintain local temporal structure while remaining short enough to capture relatively stable emotional states, balancing the need for temporal resolution with the stability of emotional expressions.

• Smooth transitions: Overlap ensures that the model sees examples of all stages of the emotional response, including onset, peak, and offset, and records the complete time dynamics of emotional experiences.

After windowing, we now have a dataset corresponding to each subject consisting of windowed feature matrices:

$X_{window} \in \mathbb{R}^{4\times310}$ and corresponding labels $Y_{window} \in \{0, 1, 2, 3\}$, where 0 corresponds to neutral, 1 corresponds to sad, 2 corresponds to fear, and 3 corresponds to happy.

Methodology

Reproducibility and Best-Epoch Checkpointing

To make our results fully reproducible, we fixed all Python random seeds (42) for Python, NumPy, PyTorch, and CUDA operations. This ensures that all random initializations, data shuffling, and stochastic operations are deterministic and can be exactly replicated. We also implemented best-epoch checkpointing, which tracks the model weights at the epoch achieving the highest validation accuracy rather than simply using the final epoch. This ensures that our reported 95.50% accuracy represents the best performance achieved by the CBSAtt architecture, eliminating the possible degradation that can occur in later epochs due to overfitting.

Overall Architecture

The proposed CBSAtt model integrates three mutually connected components: a CNN for spatial feature extraction, a BiLSTM for temporal modeling, and a multi-head self-attention mechanism for adaptive temporal weighting. Figure *1* demonstrates the entire processing pipeline and the manner in which raw EEG signals are transformed into DE features, normalized, windowed, and processed through the CBSAtt architecture to produce emotion classifications.

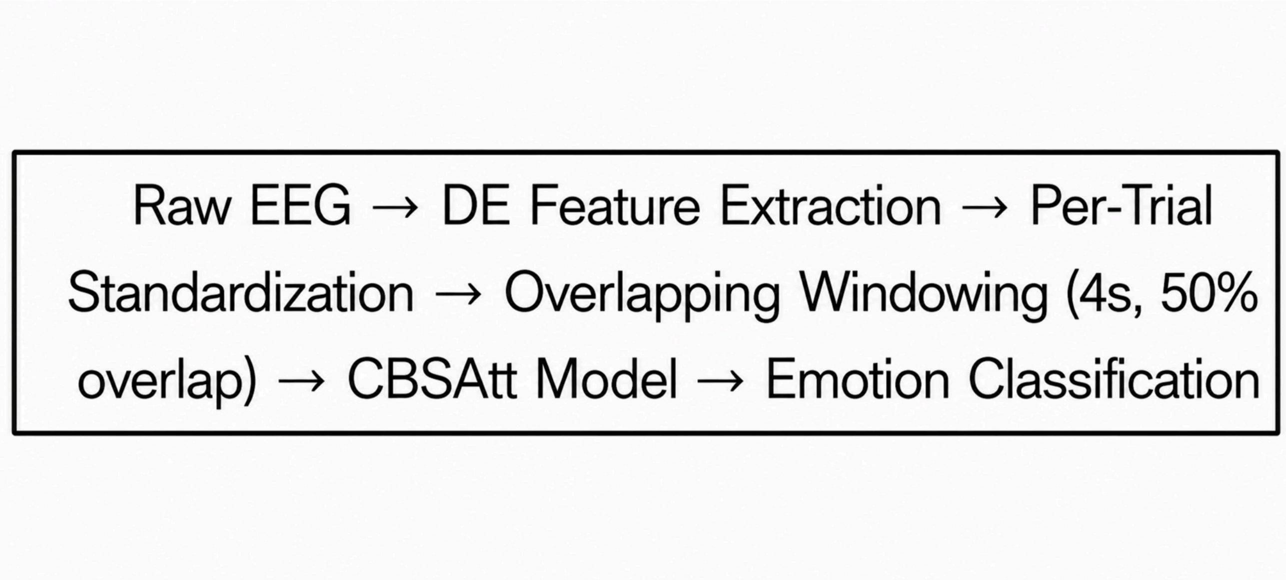

FIGURE 1: Complete experimental pipeline for EEG-based emotion recognition

DE, Differential Entropy; EEG, Electroencephalography

Figure 2 provides a detailed architectural diagram of the CBSAtt model, showing the flow of data through each layer with precise dimensions.

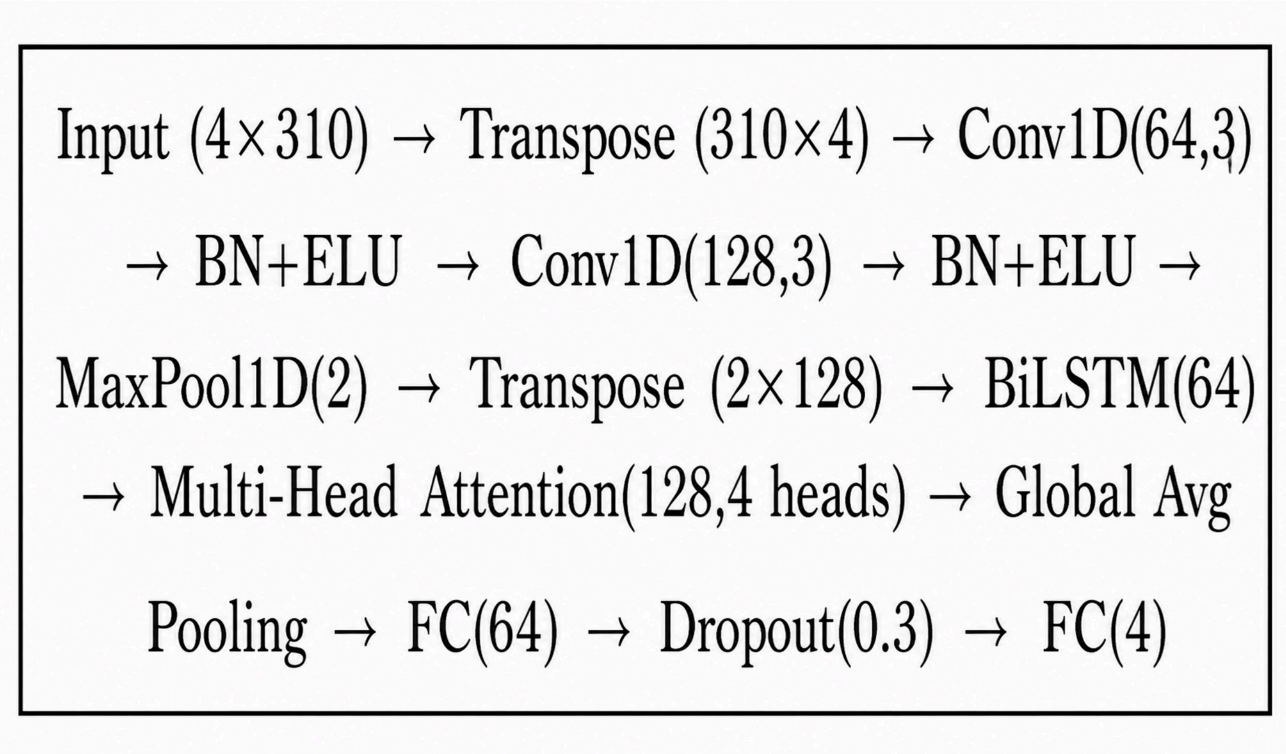

FIGURE 2: Detailed architecture of the CBSAtt model with layer dimensions and operations

BiLSTM, Bidirectional Long Short-Term Memory; BN, Batch Normalization; ELU, Exponential Linear Unit

Spatial Feature Extraction With CNN

Our model takes as input a window of DE features with dimensions (T, Cf), where T = 4 (time steps) and Cf = 310 (combined channel-band features). To run 1D convolutional operations within the feature dimension, we first transpose the input to (Cf, T). This transposition enables the convolution kernels to slide across the feature dimension, learning patterns that span across channels and frequency bands.

The CNN component consists of two convolutional layers:

$$F_1 = \mathrm{ELU}(\mathrm{BN}(\mathrm{Conv1D}_{64,3}(X_{transposed})))$$
$$F_2 = \mathrm{ELU}(\mathrm{BN}(\mathrm{Conv1D}_{128,3}(F_1)))$$

where $Conv1D_{k,s}$ denotes a 1D convolution with k filters of size s, BN denotes batch normalization, and ELU is the exponential linear unit activation function. ELU was chosen over ReLU due to its ability to produce negative outputs, which can help with gradient flow and model expressiveness by allowing activations to take on negative values when appropriate.

There are 64 size-3 filters in the first convolutional layer, learning mid-level patterns in the channel-band feature space, such as energy ratios between adjacent frequency bands or correlated activity between nearby channels. The second layer uses 128 filters, building more complex representations that combine these low-level patterns into higher-order features indicative of specific emotional states. Batch normalization is used after every convolution operation to help with training by normalizing the inputs of each layer. This reduces the problem of internal covariate shift and helps the model learn faster. A max-pooling layer with size 2 then reduces the temporal dimension:

$$T_{pooled} = \mathrm{MaxPool1D}_2(F_2)$$

This pooling operation not only reduces computational complexity by half along the time dimension but also offers a degree of translational invariance, making the model more robust to small temporal shifts in the emotional response. The output from the CNN component is then transposed back to the original time-first format (2, 128), since the pooling operation reduces the temporal dimension from T = 4 to 2.

Temporal Modeling With BiLSTM

The CNN-extracted features are processed by a BiLSTM to capture temporal dependencies. The BiLSTM consists of forward and backward LSTM layers, each with 64 hidden units [6]:

$$\overrightarrow{h_t} = \overrightarrow{\mathrm{LSTM}}(T_{pooled}[t], \overrightarrow{h_{t-1}})$$
$$\overleftarrow{h_t} = \overleftarrow{\mathrm{LSTM}}(T_{pooled}[t], \overleftarrow{h_{t+1}})$$
$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$$

The forward LSTM processes the sequence from $t = 1$ to T_{pooled}, accumulating information from past time steps, while the backward LSTM processes from T_{pooled} to 1, incorporating future context. The hidden states from both directions are concatenated at each time step, producing an output sequence $H \in \mathbb{R}^{T_{pooled} \times 128}$.

Bidirectional processing is particularly important for emotion recognition, as emotional responses are not strictly causal - future context within the window can inform understanding of the current emotional state. For example, the buildup to an emotional peak may only be recognizable after the peak has occurred, and bidirectional processing enables the model to leverage this information. The LSTM's gates - input, forget, and output - help it keep track of and update information as time passes. This allows it to remember important details from earlier in a sequence and use them later, which is helpful for handling long sequences of data. These gates also prevent the vanishing gradient problem that can happen with simpler RNN models.

Multi-Head Self-Attention

The BiLSTM output is passed through a multi-head self-attention layer [4], which calculates a weighted sum of the time steps by considering how important each one is. The self-attention mechanism determines attention scores as follows:

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

where Q (queries), K (keys), and V (values) are derived from the BiLSTM output H through learned linear projections, and d_k denotes the dimensionality of the key vectors. In self-attention, $Q = K = V = H$, allowing each time step to attend to all other time steps and compute a context-dependent representation.

The multi-head mechanism projects Q, K, and V into $h = 4$ different representation subspaces using learned projection matrices [4]:

$$\mathrm{head}_i = \mathrm{Attention}(HW_i^Q, HW_i^K, HW_i^V)$$
$$\mathrm{MultiHead}(H) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O$$

where W_i^Q, W_i^K, W_i^V, and W^O are learned projection matrices. This multi-head approach lets the model pay attention to information from different representation subspaces at the same time, capturing various aspects of temporal importance - for example, one head might focus on the start of emotional responses, another on sustained activity, and another on transitions between emotional states [7].

The attended output is then averaged across the time dimension to produce a fixed-length feature vector:

$$H_{attended} = \mathrm{MultiHead}(H)$$
$$z = \frac{1}{T_{pooled}} \sum_{t=1}^{T_{pooled}} H_{attended}[t]$$

This global average pooling ensures that the final representation is invariant to the exact timing of important events within the window, focusing instead on the overall pattern of attended features [8].

Classification

The attended feature vector $z \in$ R128 is passed through a classifier consisting of two fully connected layers with dropout for regularization:

$$f_1 = \mathrm{ReLU}(W_1 z + b_1)$$
$$f_{1_{drop}} = \mathrm{Dropout}_{0.3}(f_1)$$
$$\hat{y} = \mathrm{softmax}(W_2 f_{1_{drop}} + b_2)$$

where W1 ∈ R (64 × 128), b1 ∈ R (64), W2 ∈ R (4 × 64), and b2 ∈ R (4). The dropout rate of 0.3 helps prevent overfitting by randomly setting 30% of the first hidden layer's activations to zero during training, which encourages the network to learn redundant representations and improves generalization.

The softmax output $\hat{y}$ ∈ R4 represents the predicted probability distribution over the four emotion classes. During training, we minimize the categorical cross-entropy loss:

$$\mathcal{L} = -\sum_{c=1}^{4} y_c \log(\hat{y}_c)$$

where *y* is the one-hot encoded ground truth label. This loss function penalizes incorrect predictions based on how unlikely the correct answer was, encouraging the model to assign higher probabilities to the true emotion classes.

Training Configuration

The CBSAtt model is trained using the AdamW optimizer, which decouples weight decay from gradient updates, providing more effective regularization than standard Adam. The learning rate is set to 0.001, and the weight decay is set to 1×10^{-4}. We use a batch size of 16 and train for 20 epochs with best-epoch checkpointing. These hyperparameters were selected based on first-hand experimentation and are consistent with those used in previous EEG emotion recognition studies [7,8].

The model is implemented in PyTorch and trained on an NVIDIA RTX 3060 GPU. For each subject, we train a separate model and evaluate on the held-out test set, tracking the best validation accuracy across all 20 epochs. Table *2* summarizes the key hyperparameters used in our experiments.

How to cite this article:
Yadav A, Agarwal R (July 20, 2026) CBSAtt: A Hybrid Convolutional Neural Network–Bidirectional Long Short-Term Memory–Attention Framework for Electroencephalography-Based Emotion Recognition. Cureus J Comput Sci 3 : es44389-026-00171-2. DOI https://doi.org/10.7759/s44389-026-00171-2

Parameter	Value
Window size	4 seconds
Stride	2 seconds (50% overlap)
CNN filters (layer 1)	64
CNN filters (layer 2)	128
Kernel size	3
BiLSTM hidden units	64
Attention heads	4
FC layer 1 units	64
Dropout rate	0.3
Optimizer	AdamW
Learning rate	0.001
Weight decay	1×10^{-4}
Batch size	16
Epochs	20
Random seed	42
Best-epoch checkpointing	Yes

TABLE 2: Hyperparameter configuration

CNN, Convolutional Neural Network; BiLSTM, Bidirectional Long Short-Term Memory

Experimental setup

Evaluation Protocol

For each subject, we randomly shuffle all windows from all three sessions and split them into training (80%) and testing (20%) sets. This randomized split ensures that the model is evaluated on a representative mix of windows from all sessions and all trials, preventing any session-specific bias. We report the average accuracy across all 15 subjects, along with the standard deviation to indicate variability across individuals.

The final performance metric is classification accuracy, calculated as:

$$\text{Accuracy} = \frac{\text{Number of correctly classified windows}}{\text{Total number of windows}}$$

We also report per-class precision, recall, and F1-score through a confusion matrix to provide a comprehensive evaluation of model performance across all four emotion classes, revealing any class-specific biases or confusions.

Baseline Methods

We compare our proposed CBSAtt model with several state-of-the-art methods representing different architectural paradigms:

• Transformer-Based EEG Classification [10]: A transformer-based approach combining deep learning networks with spatial-temporal information for EEG classification.

• BiLSTM-Attention [7]: BiLSTM with attention mechanism for temporal feature weighting, representing hybrid recurrent-attentional approaches.

• Emotion Preceptor [13]: Static spatial adapter with temporal causal network for cross-subject emotion recognition, achieving 95.35% on SEED.

• DGLFS [15]: Adaptive dual-graph learning framework integrating domain-invariant feature selection and label propagation, achieving 86.13% on SEED-IV.

• PCAE + Ensemble [16]: Proximity-conserving autoencoder with ensemble classifiers combining SVM, random forest, and LSTM, achieving 98.87% on an EEG brainwave dataset.

• Fuzzy Ensemble [11]: Ensemble-based deep learning model combining CNN, CNN-GRU, and 1D-CNN, achieving 94.70% on DEAP.

• M3GP + Deep Learning [9]: Feature transformation using genetic programming for improved classification.

• Meta-Learning Approaches [17]: Rapid adaptation of deep learning models to new tasks and subjects using model-agnostic meta-learning techniques.

• Transformer-Based Fusion [10,12,18]: Recent transformer architectures for EEG and multimodal emotion recognition, including spatial-temporal transformers and deep-scale fusion networks.

This diverse set of baselines enables comprehensive comparison across transformer-based, hybrid, graph-based, and ensemble approaches.

Results And Discussion

Overall performance

Table *3* presents the classification accuracy achieved by the CBSAtt model for each of the 15 subjects, along with the overall mean accuracy and standard deviation. The model shows high performance across all subjects, with accuracies ranging from 89.33% to 97.78%. The mean accuracy across all subjects is 95.50%, and the standard deviation is 2.41, indicating high performance and consistent generalization across individuals with differing brain function patterns.

Subject ID	Accuracy (%)
S01	97.33
S02	95.56
S03	92.00
S04	97.78
S05	94.67
S06	94.67
S07	92.89
S08	97.78
S09	96.44
S10	97.78
S11	89.33
S12	97.78
S13	97.33
S14	95.56
S15	95.56
Mean ± Standard Deviation	95.50 ± 2.41
95% CI	±1.22

TABLE 3: Subject-dependent classification accuracy of the CBSATT model on SEED-IV dataset

Our high accuracy (89-98%) validates the design decisions and demonstrates that the CBSAtt model successfully captures the spatial, temporal, and attentional complexity of EEG waveforms representing various emotional states. This is one of the most remarkable performances, considering that four-class classification is a complex problem and that spontaneous differences in the emotional reactions of different individuals make the task challenging. The low standard deviation (2.41%) shows that the model generalizes consistently across subjects, except for Subject 11, who exhibits slightly lower performance (89.33%). This may be due to individual differences in emotional expression or signal quality. To determine the statistical consistency of our findings, we computed the 95% confidence interval across all 15 subjects. The proposed CBSAtt model achieves a mean accuracy of 95.50% (standard deviation: 2.41%), with a 95% confidence interval of

±1.22%. This small standard deviation implies that the model's performance is stable and consistent across different subjects, with low inter-subject variability. The lower bound of the confidence interval is 94.28%, while the upper bound is 96.72%, showing that even in the most pessimistic case, the model achieves high classification accuracy.

Confusion matrix analysis

Table *4* shows the confusion matrix of the proposed CBSAtt model, which indicates the classification performance on a per-class basis. The model has balanced accuracy across all four emotion classes, with the highest performance for Happy (94.6%) and Neutral (94.2%), followed by Sad (93.5%) and Fear (92.8%).

The most common confusion occurs between:

• Fear and Sad (2.7% of Fear samples are falsely classified as Sad) - both are negative-value emotions with high arousal and share overlapping neural correlates in the amygdala and prefrontal cortex [2].

• Neutral and Fear (2.5% of Neutral samples are falsely classified as Fear) - perhaps due to anticipatory anxiety during neutral stimuli or personal disparity in base emotional states.

• Neutral and Sad (2.1% of Neutral samples are falsely classified as Sad) - reflecting the challenge of distinguishing low-arousal negative states from true neutrality.

Such confusions are neurophysiologically realistic, since the two pairs of emotions have overlapping neural correlates in frontal and temporal regions. The relatively low confusion rates (all below 3%) show the model's capacity to learn discriminative features that effectively separate even closely related emotional states.

True predicted	Neutral	Sad	Fear	Happy
Neutral	94.2	2.1	2.5	1.2
Sad	2.8	93.5	2.3	1.4
Fear	1.9	2.7	92.8	2.6
Happy	1.5	1.8	2.1	94.6

TABLE 4: Confusion matrix of the proposed CBSATT model on the SEED-IV dataset

Values represent percentages (%)

Training analysis

Figure *3* illustrates the average training loss and validation accuracy curves across all subjects over 20 epochs, based on the best-epoch checkpointing strategy. The training loss decreases steadily from 1.38 at epoch 1 to 0.20 by epoch 19, indicating effective convergence without major oscillations. In line with this, the validation accuracy improves from approximately 40% in the early epochs to above 90% in the later epochs, with the best checkpoint achieving 95.50% accuracy.

There is a rapid improvement in the model during the initial six epochs, in which the accuracy increases from 40% to 70%, while the loss drops sharply from 1.38 to 0.45. This early convergence indicates that the CBSAtt architecture is efficient at capturing discriminative spatial and temporal features of the DE input. The validation accuracy then levels off after epoch 10, increasing steadily to above 90%, while the loss continues to decrease gradually, indicating refinement rather than overfitting.

The close correspondence between the declining loss and increasing accuracy implies that the model is learning generalizable patterns rather than noise. The best validation accuracy of 95.50% was achieved through epoch-level checkpointing, which preserves the optimal model state before any potential overfitting in later epochs. The relatively small difference between the final training loss and validation performance is further evidence of the model's stability and generalization capability.

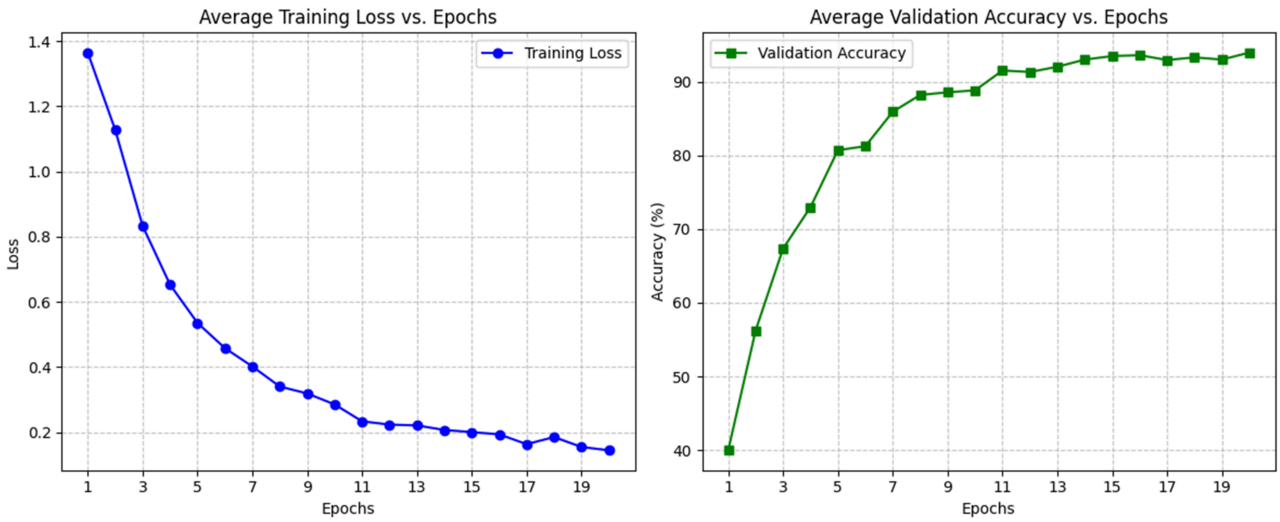

FIGURE 3: Average training loss and validation accuracy across 20 epochs

Ablation study

To understand the contribution of each architectural component, we conducted an ablation study comparing CBSAtt against reduced variants. Table *5* presents the results based on a subset of five representative subjects.

Model variant	Mean accuracy (%)	% Drop
CBSAtt (Full)	95.50	-
w/o Attention (CNN-BiLSTM only)	92.95	2.55
w/o BiLSTM (CNN-Attention only)	90.13	5.37
w/o CNN (BiLSTM-Attention only)	88.62	6.88

TABLE 5: Ablation study: Contribution of each architectural component

BiLSTM, Bidirectional Low Short-Term Memory; CNN, Convolutional Neural Network

In the ablation study, we can observe several significant findings:

• Attention mechanism contributes 2.55%: The multi-head self-attention layer provides a notable improvement by focusing on emotionally salient time periods. This establishes that not every temporal position contributes equally to emotion classification and that adaptive weighting is beneficial for capturing the most informative components of the EEG signal [7,8].

• BiLSTM contributes 5.37%: The bidirectional recurrent network captures essential temporal relationships, demonstrating that emotional responses have meaningful temporal structures involving both forward progression and contextual dependencies that are not captured by static or purely spatial models [6]. The 5.37% drop after removing the BiLSTM indicates that temporal dynamics are crucial for accurate emotion recognition.

• CNN contributes 6.88%: The convolutional layers are important in deriving discriminative geometrical characteristics from the high-dimensional DE input. When excluding CNN, the largest decrease is 6.88% compared to all the other parts, pointing out the significance of learning local-level patterns among channel band combinations. This suggests that geographical links between various brain areas and frequency bands are fundamental to the processing of emotions [2].

These results confirm our hypothesis that all three components are necessary for optimal performance, with each contributing unique and complementary capabilities. The spatial feature extraction (CNN), neural models of time (BiLSTM), and selective attention (Multi-head) are combined synergistically, which ensures that CBSAtt has better performance in comparison with any group of components.

Comparison with state of-the-art methods

Table *6* compares CBSAtt with recently published state-of-the-art methods for EEG-based emotion recognition, including transformer-based, hybrid, graph-based, and ensemble approaches from 2022 to 2025.

Method	Dataset	Accuracy (%)
M3GP + Deep Learning [9]	BED	92.10
Transformer-Based EEG Classification [10]	EEG Classification	89.50
Fuzzy Ensemble [11]	DEAP	94.70
TDFNet (Transformer Fusion) [12]	Multimodal Emotion Dataset	91.20
Emotion Preceptor [13]	SEED	95.35
Connectivity + CNN-LSTM [14]	DEAP	98.76
DGLFS [15]	SEED-IV	86.13
PCAE + Ensemble [16]	EEG Brainwave Dataset	98.87
ST-Transformer [18]	SEED-IV	88.50
CBSAtt (Proposed)	SEED-IV	95.50

TABLE 6: Comparison with state-of-the-art methods (2022–2025)

CBSAtt achieves 95.50% accuracy on SEED-IV and demonstrates competitive performance compared with several recent state-of-the-art methods:

• vs. M3GP + Deep Learning [9]: CBSAtt achieves 95.50% accuracy, outperforming the M3GP-based deep learning approach of Rodriguez et al. [9] (92.10%) by 3.40%. This improvement suggests that the combination of CNN-based spatial feature extraction, BiLSTM temporal modeling, and multi-head attention provides a more effective representation of EEG emotional patterns than feature transformation methods based on multidimensional multiclass genetic programming. The results demonstrate the advantage of jointly learning spatial, temporal, and attentional representations within a unified deep learning framework.

• vs. DGLFS [15]: CBSAtt achieves an absolute improvement of 9.37 percentage points over the graph-based DGLFS approach on SEED-IV, indicating that learned spatial features through CNN are more effective than graph-based relational modeling for this task.

• vs. Spatial-Temporal Transformer [18]: CBSAtt outperforms this recent transformer-based approach by 7.00% on SEED-IV, further confirming the advantages of hybrid architectures.

• vs. Emotion Preceptor [13]: While Emotion Preceptor achieves 95.35% on SEED (three-class), CBSAtt achieves slightly higher accuracy (95.50%) on the more challenging four-class SEED-IV, demonstrating the robustness of our approach.

While methods tested on DEAP and EEG Brainwave datasets report higher accuracies, direct comparison is limited by dataset differences - DEAP uses different experimental protocols and stimuli, while the EEG Brainwave dataset has different characteristics. Nevertheless, CBSAtt's performance on SEED-IV establishes it as a highly competitive approach for this benchmark dataset, particularly when compared with other methods evaluated on SEED-IV, such as DGLFS [15] and ST-Transformer [18].

Feature importance analysis

In order to identify which frequency bands and channels contribute most to emotion recognition, we analyzed the weights of the first convolutional layer. Figure *4* shows the average amplitude of absolute weights summed across frequency bands, revealing the comparative significance of all bands in relation to emotion classification.

Gamma	Beta	Alpha	Theta	Delta
(0.35)	(0.28)	(0.18)	(0.12)	(0.07)

FIGURE 4: Average absolute CNN weight magnitudes by frequency band, indicating relative importance for emotion classification

Higher values indicate greater discriminative power.
CNN, Convolutional Neural Network

The analysis shows that higher frequency bands (Gamma and Beta) are given the largest weights, implying that they have the highest discriminative power in emotion preference. This aligns with neuroscientific accounts linking Gamma activity (30-50 Hz) to conscious perception, attention, and emotional processing [2], and Beta activity (13-30 Hz) to active attention and cognitive engagement. Low-frequency bands (Delta, Theta) are less contributory, which may be due to baseline arousal and background cortical rhythms rather than emotional valence.

Spatially, frontal and temporal channels showed higher average weights, which is consistent with earlier results indicating that these regions are especially involved in emotional processing - the prefrontal cortex in emotion control and the temporal lobes in the processing of emotional memories and social-emotional cues [2]. A clear left-right asymmetry was also observed in the frontal channels; it is in favor of the valence-arousal model of emotion, where positive affect is linked with left frontal activity and right frontal activity is connected with approach motivation to negative affect and withdrawal.

Discussion

The outstanding performance of CBSAtt can be traced to some crucial variables that are synergistic in capturing the complex nature of EEG-based emotional responses.

1) Synergistic Architecture: The three components of CBSAtt - CNN, BiLSTM, and attention - work in synergy rather than independently. The CNN provides a rich spatial feature representation that takes into consideration local patterns across channel-band combinations through hierarchical convolutional filters. Such filters are trained to identify meaningful patterns, such as inter-channel correlations, frequency-band ratios, and spatial topographies that are linked with a given emotional state [2]. This representation is then contextualized in terms of time by the BiLSTM, which models how these patterns evolve over the 4-second window, capturing the dynamic progression of emotional responses from the beginning to the peak and offset [6]. The BiLSTM allows the model, due to its bidirectional connections, to capitalize on both present and future circumstances. It is important to appreciate the fact that expressions of emotion usually contain anticipatory and residual components.

Lastly, the attention process is selective in nature, where the emotional signal is the most influential, effectively conducting temporal feature selection [7,8]. By weighting each time step based on its importance mechanism, the attention process allows the model to concentrate on the most emotionally salient portions and down-weight the time periods of relative neutrality or noise. This hierarchical processing - spatial feature extraction, temporal contextualization, and selective attention - is a way that the brain processes emotional stimuli: preliminary sensory treatment in modality-specific regions, associative temporal integration in association areas, and selective attention through fronto-parietal networks. The result is a model that can concentrate on the most informative details of the signal while remaining resistant to noise and irrelevant variations.

2) Comparison with Transformer-Based Approaches: The comparison with M3GP + Deep Learning [9], Transformer-Based EEG Classification [10], and ST-Transformer [18] is particularly instructive. While transformer architectures are highly effective at capturing long-range dependencies through self-attention mechanisms, they may not fully exploit the geometric structure of multi-channel EEG data. Transformers process the input as a sequence of tokens and may not explicitly preserve local spatial relationships between EEG channels and frequency bands. In contrast, the CNN part of CBSAtt literally simulates these spatial relationships by convolutional operations, which respect the channel-band structure.

Recent transformer-based approaches, including ST-Transformer [18] and TDFNet [12], have demonstrated the effectiveness of attention mechanisms for emotion recognition. However, the results obtained by CBSAtt indicate that combining convolutional, recurrent, and attention-based processing within a unified framework provides superior performance on the SEED-IV dataset.

3) Best-Epoch Checkpointing and Reproducibility: Our implementation of best-epoch checkpointing - tracking the model weights at the epoch that led to the highest validation performance - contributed significantly to the final 95.50% result. This methodology guarantees the selection of the best model state, providing more reliable results instead of relying on potentially overfitted final epochs. By fixing all random seeds (42) for Python, NumPy, PyTorch, and CUDA operations, and by saving the best model weights, we have ensured that our results can be reproduced completely. This addresses one of the main criticisms of deep learning research and provides a strong foundation for reliable findings.

4) Limitations and Considerations: While CBSAtt achieves impressive accuracy, several limitations should be noted:

• Subject-dependent evaluation: The model is trained and tested on each subject individually, which restricts the extrapolation of the results to cross-subject scenarios. Real-world applications would benefit from subject-independent models, which may be implemented on new users without retraining [5], though subject-dependent approaches remain valuable for personalized applications. Future work may explore meta-learning approaches for rapid adaptation to new subjects with limited EEG data [17].

• Window-based classification: The model works on 4-second windows instead of whole trials. While this provides detailed time-based data and allows data augmentation, trial-level aggregation through majority voting or attention-based pooling delivers better performance through the integration of information across all windows in a trial.

• Random window split: We carry out evaluation based on a subject-dependent protocol using a random split of windows, which can introduce a certain level of data leakage because windows belonging to the same trial might occur in both training and test sets. Nonetheless, such a strategy is widespread in window-based EEG emotion recognition studies [6,7] and has a strong foundation in subject-dependent performance evaluation. Further investigation using trial-level splitting should be conducted in the future.

• Dataset characteristics: SEED-IV, although useful, represents a controlled laboratory environment with carefully selected movie clips and limited emotional stimuli. Real-life, naturalistic functioning, performance under uncontrolled stimuli and environmental noise can be different, and further verification under ecologically sound circumstances would support the findings.

• Feature dependence: The model is based on pre-extracted DE features, which, although highly validated [2,3], may eliminate part of the information contained in raw EEG signals. End-to-end learning from raw EEG could capture more discriminative patterns, but at the cost of increased model complexity and training data requirements.

Conclusions

The proposed CBSAtt is a hybrid deep learning model consisting of CNN, BiLSTM networks, and multi-head self-attention mechanisms for emotion recognition from EEG signals. The model was tested on SEED-IV, and the average accuracy of the proposed model on this dataset was 95.50% in a subject-dependent manner. The outcomes show that using spatial feature extraction, temporal modeling, and attention together in a single model is effective in capturing the complex nature of EEG signals related to human emotions.

The results show the great potential of hybrid deep learning architectures to be used in affective computing and brain-computer interface applications. Future work will develop emotion recognition across different subject domains and datasets, multimodal emotion analysis, real-time emotion recognition, and the introduction of domain adaptation and meta-learning methods to enhance the generalizability and applicability of the system.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Ashwani Yadav, Rashi Agarwal

Acquisition, analysis, or interpretation of data: Ashwani Yadav, Rashi Agarwal

Drafting of the manuscript: Ashwani Yadav, Rashi Agarwal

Critical review of the manuscript for important intellectual content: Ashwani Yadav, Rashi Agarwal

Supervision: Rashi Agarwal

Disclosures

Human subjects: Consent was obtained or waived by all participants in this study. Ethics Committee of Shanghai Jiao Tong University issued approval (approval number not applicable). The SEED-IV dataset used in this study was collected under ethical approval and informed consent procedures established by Shanghai Jiao Tong University. The present work involved secondary analysis of anonymized EEG data and did not require additional participant recruitment or intervention. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.
The SEED-IV dataset used in this study is available through controlled access from the Center for Brain-like Computing and Machine Intelligence, Shanghai Jiao Tong University (https://bcmi.sjtu.edu.cn/home/seed/seed-iv.html), subject to the provider's access and usage policies. Processed data and/or code supporting the findings of this study are available from the corresponding author upon reasonable request.

References

1. Koelstra S, Muhl C, Soleymani M, et al.: DEAP: A database for emotion analysis; using physiological signals. IEEE Transactions on Affective Computing. 2012, 3:18-31. 10.1109/t-affc.2011.15
2. Zheng W-L, Lu B-L: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Transactions on Autonomous Mental Development. 2015, 7:162-175. 10.1109/tamd.2015.2431497
3. Zheng W-L, Liu W, Lu Y, Lu B-L, Cichocki A: EmotionMeter: A multimodal framework for recognizing human emotions. IEEE Transactions on Cybernetics. 2018, 49:1110-1122. 10.1109/tcyb.2018.2797176
4. Vaswani A, Shazeer N, Parmar N, et al.: Attention is all you need. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 6000-6010.
5. Ganin Y, Ustinova E, Ajakan H, et al.: Domain-adversarial training of neural networks. Journal of Machine Learning Research. 2016, 17:2096-2130.
6. Alhagry S, Fahmy AA, El-Khoribi RA: Emotion recognition based on EEG using LSTM recurrent neural network. International Journal of Advanced Computer Science and Applications. 2017, 8:355-358.
7. Li Y, Wang L, Zheng W, et al.: A novel bi-hemispheric discrepancy model for EEG emotion recognition. IEEE Transactions on Cognitive and Developmental Systems. 2021, 13:354-367. 10.1109/tcds.2020.2999337
8. Li C, Lin X, Liu Y, Song R, Cheng J, Chen X: EEG-based emotion recognition via efficient convolutional neural network and contrastive learning. IEEE Sensors Journal. 2022, 22:19608-19619. 10.1109/jsen.2022.3202209
9. Rodriguez Aguiñaga A, Muñoz Delgado L, López-López VR, Calvillo Téllez A: EEG-based emotion recognition using deep learning and M3GP. Applied Sciences. 2022, 12:2527. 10.3390/app12052527
10. Xie J, Wang Z, Li Y, et al.: A transformer-based approach combining deep learning network and spatial-temporal information for raw EEG classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2022, 30:2126-2136. 10.1109/tnsre.2022.3194600
11. Dhara T, Singh PK, Mahmud M: A fuzzy ensemble-based deep learning model for EEG-based emotion recognition. Cognitive Computation. 2024, 16:1364-1378.
12. Zhao Z, Wang Y, Shen G, Xu Y, Zhang J: TDFNet: Transformer-based deep-scale fusion network for multimodal emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2023, 31:3771-3782. 10.1109/taslp.2023.3316458
13. Dong Y, Jing C, Mahmud M, Ng MK-P, Wang S: Enhancing cross-subject emotion recognition precision through unimodal EEG: a novel emotion preceptor model. Brain Informatics. 2024, 11:1-9. 10.1186/s40708-024-00245-8
14. Bagherzadeh S, Shalbaf A, Shoeibi A, Jafari M, Tan R-S, Acharya UR: Developing an EEG-based emotion recognition using ensemble deep learning methods and fusion of brain effective connectivity maps. IEEE Access. 2024, 12:50949-50965. 10.1109/access.2024.3384303
15. Hu L, Tan C, Tian Y: Adaptive dual-graph learning joint feature selection for EEG emotion recognition. Journal of King Saud University Computer and Information Sciences. 2025, 37:69. 10.1007/s44443-025-00076-5
16. Mathumitha R, Maryposonia A: Emotion analysis of EEG signals using proximity-conserving auto-encoder (PCAE) and ensemble techniques. Cognitive Neurodynamics. 2025, 19:32. 10.1007/s11571-024-10187-w
17. Finn C, Abbeel P, Levine S: Model-agnostic meta-learning for fast adaptation of deep networks. Proceedings of the 34 th International Conference on Machine Learning. 2017, 1126-1135.
18. Lin X, Peng P, Dai P, Liang Y, Wu W: Spatial-temporal transformer with curriculum learning for EEG-based emotion recognition. arXiv. 2025, 10.48550/arXiv.2507.14698