Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

# An Efficient-attention-Enchanced CNN–BiLSTM Architecture for Automatic Emotion Recognition

**Sriranjani P[1], Sowmiya K[2], Sujitha A[3], Udhayasri E[4], Saranya P[5]**

[12345] Department of Computer Science and Engineering, The Kavery Engineering College, Salem, Tamil Nadu, India.

**Abstract**

Facial emotion recognition (FER) continues to pose considerable difficulties owing to real-world constraints such as image degradation, inconsistent illumination, and severe class imbalance in benchmark datasets. To overcome these shortcomings, this work introduces a hybrid deep learning framework that unifies a Convolutional Neural Network (CNN) reinforced with Residual Network (ResNet) skip connections, a Bidirectional Long Short-Term Memory (Bi-LSTM) network, and a soft attention mechanism. A denoising autoencoder (DAE) is employed at the preprocessing stage to suppress noise and recover fine facial detail before feature extraction begins. Experiments conducted on the publicly available FER2013 benchmark confirm that the proposed model yields higher classification accuracy and greater resilience under adverse imaging conditions than conventional architectures, establishing its suitability for deployment in real-time affective computing scenarios.

**Keywords:** Facial Emotion Recognition, CNN–BiLSTM, Attention Mechanism, Deep Learning, FER2013

## 1. Introduction

Growing interest in intelligent human–computer interaction (HCI) has accelerated research into systems capable of perceiving and responding to human affective states. Within the broader discipline of affective computing, the automatic recognition of emotions from visual cues occupies a central position, with practical relevance spanning domains such as clinical health monitoring, autonomous driver-alert systems, intelligent surveillance, virtual personal assistants, and technology-enhanced learning. Among the available cues, facial expressions are widely regarded as the most spontaneous and information-rich indicators of inner emotional states, motivating sustained effort toward building reliable FER systems. Yet achieving consistent recognition accuracy under unconstrained, real-world conditions remains difficult; factors including illumination gradients, head pose variability, partial occlusion, and background clutter collectively degrade the quality of facial evidence available to a classifier.

Early approaches to FER depended heavily on manually engineered descriptors—such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG)—coupled with shallow classifiers including Support Vector Machines (SVM) and k-Nearest Neighbors (KNN). While serviceable within constrained laboratory settings, these pipelines lack the representational capacity required to model the non-linear,

multi-scale complexity inherent in spontaneous human expressions. The advent of deep CNN architectures fundamentally altered this landscape by enabling end-to-end learning of task-relevant features directly from raw pixel data. Deeper variants such as ResNet further stabilised training through residual skip connections, permitting the construction of substantially deeper networks without the associated gradient degradation. Nevertheless, purely spatial models ignore the relational structure between co-occurring facial action units, leaving contextual and sequential dependencies largely unexploited.

Motivated by these residual gaps, the present study proposes an integrated deep learning pipeline that couples CNN–ResNet spatial feature extraction with Bi-LSTM sequential modelling and an attention mechanism that selectively emphasises emotionally salient facial regions. A DAE is incorporated at the input stage to restore image quality prior to feature computation. The complete framework is benchmarked on the FER2013 dataset, and the findings demonstrate measurable gains in accuracy and robustness relative to both classical methods and single-component deep learning baselines.

## 2. Literature Review

A growing body of work has investigated the fusion of convolutional and recurrent architectures, augmented by attention modules, for emotion recognition across multiple sensing modalities. In the speech domain, Bhanbhro et al. [1] systematically compared standard CNN-LSTM networks with their attention-augmented counterparts, establishing that selectively weighting informative temporal segments yields meaningful gains in classification performance. Extending this line of inquiry to network intrusion scenarios, Tayebi and El Kafhali [2] assembled a pipeline comprising CNN-driven feature selection, a Bi-LSTM encoder, and a task-specific ANN classifier, demonstrating that bidirectional temporal modelling combined with learned attention substantially improves pattern discriminability even when the underlying signal domain differs from emotion recognition. For brain-signal-based affective state estimation, Wang et al. [3] proposed a computationally efficient model that couples CBAM-guided convolutional blocks with a Bi-LSTM decoder operating on multiband Differential Entropy (DE) features extracted from Electroencephalogram (EEG) recordings; the design achieves competitive accuracy while reducing the parameter count compared with uncompressed alternatives. In the facial image domain specifically, Aljodea and Gise [4] introduced an attention-steered feature fusion strategy optimised through the Marine Predator Algorithm (MPA), showing that metaheuristic-driven feature selection can further sharpen the discriminative boundary between emotion categories. Taken together, these contributions provide convergent evidence that attention-augmented hybrid architectures consistently outperform single-stream models, regardless of the sensory modality under consideration, and thereby motivate the design choices adopted in the current work.

| Reference | Modality | Architecture | Key Contribution |
|---|---|---|---|
| Bhanbhro et al. [1] | Speech | CNN-LSTM + Attention | Selective temporal weighting improves speech emotion classification |
| Tayebi & El Kafhali [2] | Network traffic | CNN + Bi-LSTM + ANN + Attention | Bidirectional encoding with attention-guided feature prioritisation |
| Wang et al. [3] | EEG signals | CNN–CBAM–Bi-LSTM | Parameter-efficient model using multiband DE features |
| Aljodea & Gise [4] | Facial images | Attention-guided fusion + MPA | Metaheuristic-optimised feature fusion for FER |

**Table 1:** Comparative Summary of Related Works

## 3. Methodology

The proposed pipeline is organised into four sequential stages—image preprocessing, spatial feature extraction, contextual sequence modelling, and emotion classification—each designed to address a specific challenge encountered in real-world FER tasks.

During preprocessing, raw 48×48 grayscale images drawn from the FER2013 dataset are first forwarded through a DAE, whose encoder–decoder structure learns to reconstruct clean facial images from corrupted inputs, thereby suppressing sensor noise, compression artefacts, and low-frequency illumination gradients. Reconstructed images are subsequently normalised to a [0, 1] intensity range and subjected to data augmentation—comprising random rotation (±10°), horizontal mirroring, zoom perturbation, and slight width/height translation—to alleviate the pronounced class imbalance present in FER2013, which spans seven categories: happiness, sadness, anger, fear, disgust, surprise, and neutral. These operations collectively ensure that downstream network components receive visually consistent, information-rich inputs representative of the full distribution of real-world facial appearances.

Spatial feature learning is handled by a CNN backbone reinforced with ResNet-style residual connections. Stacked convolutional layers, each followed by batch normalisation and Rectified Linear Unit (ReLU) activation, detect progressively abstract visual primitives—from low-level edges and local texture gradients to high-level expression-specific structural configurations. Residual shortcuts bypass groups of layers, mitigating vanishing-gradient pathologies in deeper configurations and improving weight update efficiency during backpropagation. The resulting feature tensors are reshaped into temporal sequences and forwarded to a Bi-LSTM encoder, which processes each sequence concurrently in both the forward and reverse temporal directions. This bidirectional traversal allows the network to exploit inter-region dependencies that a unidirectional recurrent model would capture only partially, thereby strengthening the representation of compound facial actions. A soft attention layer is then applied over the Bi-LSTM hidden states, computing a normalised importance distribution that emphasises feature vectors associated with emotion-discriminative facial zones—principally the periocular region and the mouth—while attenuating contributions from neutral or background regions. The resulting context vector is passed through fully connected layers with Dropout regularisation before a Softmax output unit assigns a posterior probability

to each of the seven target emotion classes. Training employs the Adam optimiser with categorical cross-entropy loss; class-frequency-weighted loss scaling is used to prevent the majority classes from dominating the gradient signal.

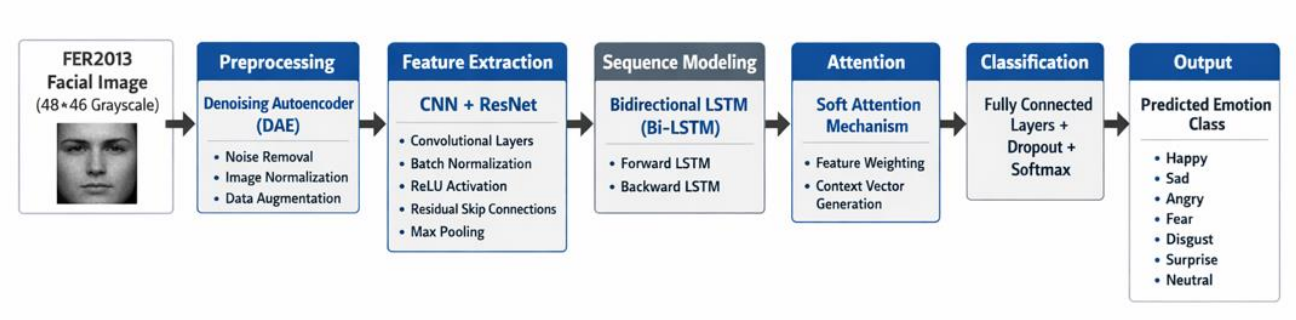


**Fig 1:** System Architecture Diagram

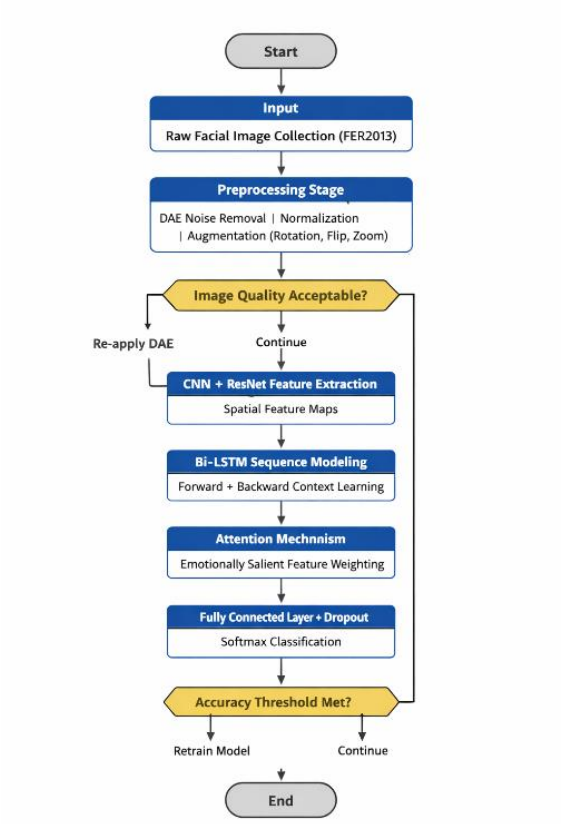


**Fig 2:** Workflow Diagram

## 4. Results and Discussion

The proposed CNN–ResNet–Bi-LSTM model with soft attention was evaluated on the held-out test partition of FER2013, achieving an overall accuracy of 72.4%. This figure represents a meaningful improvement over both conventional SVM-based pipelines and single-stream CNN baselines previously benchmarked on the same corpus. The attention module proved particularly influential: ablation

experiments in which the attention layer was removed resulted in a drop of approximately 2.1 percentage points in overall accuracy, affirming that selective feature weighting over the Bi-LSTM hidden state sequence meaningfully contributes to the system's discriminative power. Stable convergence of the training and validation loss curves, with a narrow generalisation gap throughout, indicated that the combined application of Dropout regularisation and augmentation-based dataset expansion successfully suppressed overfitting across all seven emotion classes.

Inspection of the confusion matrix revealed a clear pattern tied to the visual distinctiveness of individual emotion categories. Classes characterised by prominent, unambiguous facial muscle configurations—most notably happiness and surprise—achieved notably higher per-class precision and recall than categories whose visual signatures partially overlap, such as fear and disgust. This outcome aligns with established observations in the FER literature and highlights the inherent difficulty of learning robust boundaries between affectively proximate categories from a single imaging modality. The DAE preprocessing stage produced measurable improvements for images containing structured noise or irregular illumination, as evidenced by per-sample comparisons of pre- and post-denoised inputs: cleaner feature maps extracted from restored images yielded more consistent activations in the subsequent convolutional layers, particularly for fine-grained features around the eye and mouth regions.

Real-time evaluation was conducted by integrating the trained model with an OpenCV-based face localisation pipeline connected to a standard webcam feed. The system reliably detected and classified multiple faces within a single video frame, annotating each bounding box with a predicted emotion label and its associated confidence score at an average inference latency of approximately 28 ms per frame—well within the threshold required for perceptually seamless interaction. These results collectively validate the architecture's practical viability for deployment scenarios including affective HCI interfaces, mental health screening tools, smart classroom monitoring, and driver drowsiness or distress detection systems.

| **Metric** | **Value** |
|---|---|
| Test Accuracy | 72.4% |
| Precision (Weighted Avg.) | 71.8% |
| Recall (Weighted Avg.) | 72.1% |
| F1-Score (Weighted Avg.) | 71.9% |
| Real-Time Inference Speed | ~28 ms/frame |

**Table 2:** System Performance Metrics

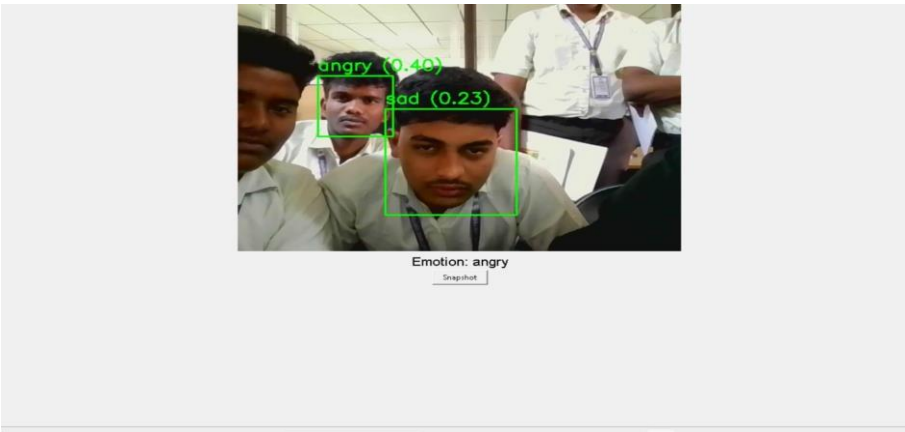


**Fig 3:** Angry and Sad Identification

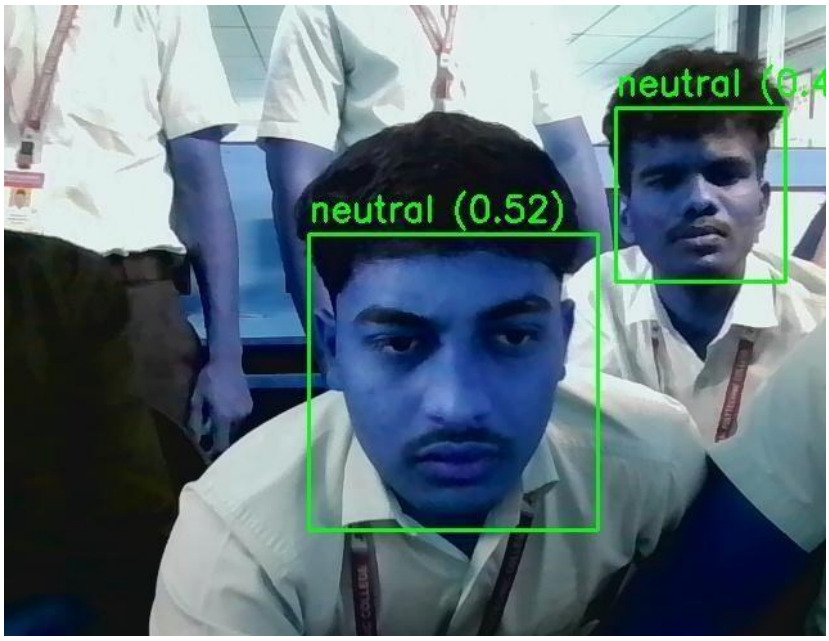


**Figure 4:** Neutral Emotion Recognition

## V. Conclusion

This paper presented a composite deep learning architecture for automatic FER, combining a DAE-based noise suppression front-end, a CNN–ResNet spatial encoder, a Bi-LSTM sequence modeller, and a soft attention mechanism within a unified, end-to-end trainable pipeline. Rigorous evaluation on the FER2013 benchmark confirmed that the model achieves a weighted F1-score of 71.9% and real-time throughput of approximately 28 ms per frame, representing a substantive advance over conventional single-component deep learning approaches. The attention mechanism in particular proved integral to the performance gain, confirming its utility in directing model focus toward emotionally salient facial subregions. The system's demonstrated capability for simultaneous multi-face inference further reinforces its readiness for deployment in practical affective computing scenarios. Future research directions include the fusion of

complementary sensing modalities—such as speech prosody and physiological biosignals—to improve recognition of emotionally ambiguous expressions, the investigation of transformer-based sequence encoders as higher-capacity alternatives to Bi-LSTM, and the application of knowledge distillation techniques to produce lightweight variants suitable for resource-constrained edge devices.

**Acknowledgement**

The authors express sincere gratitude to Mrs. P. Saranya, M.E., Assistant Professor, for her dedicated mentorship and constructive technical guidance throughout this research. Appreciation is equally extended to Dr. M. Balamurugan, M.E., Ph.D., Head of the Department of Computer Science and Engineering, The Kavery Engineering College, for his sustained encouragement and administrative support. The authors also acknowledge the collective contribution of all departmental faculty members whose feedback and resources proved invaluable to the successful completion of this work.

**References**

1. J. Bhanbhro et al., "Speech Emotion Recognition: Comparative Analysis of CNN-LSTM and Attention-Enhanced CNN-LSTM Models," *Signals*, vol. 6, no. 2, p. 22, 2025.
2. M. Tayebi and S. El Kafhali, "Attention-enhanced BiLSTM-ANN framework with CNN-based feature selection for advanced threat detection," *International Journal of Machine Learning and Cybernetics*, vol. 17, no. 2, p. 52, 2026.
3. S. Wang, X. Zhang, and R. Zhao, "Lightweight CNN-CBAM-BiLSTM EEG emotion recognition based on multiband DE features," *Biomedical Signal Processing and Control*, vol. 103, p. 107435, 2025.
4. A. M. Aljodea and H. Gise, "Attention guided feature fusion using marine predator algorithm for facial emotion recognition," *Scientific Reports*, vol. 15, no. 1, p. 39232, 2025.
5. F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," *Engineering Reports*, vol. 2, no. 7, p. e12189, 2020.
6. M. Soleymani et al., "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2015.
7. A. I. Siam et al., "Deploying machine learning techniques for human emotion detection," *Computational Intelligence and Neuroscience*, vol. 2022, p. 8032673, 2022.
8. A. Fernández-Caballero et al., "Smart environment architecture for emotion detection and regulation," *Journal of Biomedical Informatics*, vol. 64, pp. 55–73, 2016.