Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Hybrid Deep Learning Framework for efficient emotion detection on the facial image dataset

S. Hrushikesava Raju

Hrushikesava07@outlook.com

Koneru Lakshmaiah Education Foundation

S. Adinarayana

Anil Neerukonda Institute of Technology & Sciences

Kale Naga Venkata Srinivas

Aditya University

U. Sesadri

Vardhaman College of Engineering, Hyderabad

Vijaya Chandra Jadala

SR University

Nabanita Choudhury

Symbiosis Artificial Intelligence Institute, Symbiosis International University

Research Article

Keywords: Emotion recognition, Convolution neural network, Transformer, Attention feature measurement, Robustness, Performance, and Accuracy

Posted Date: July 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9924647/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Hybrid Deep Learning Framework for efficient emotion detection on the facial image dataset

S.Hrushikesava Raju[1]*, S.Adinarayana[2], Kale Naga Venkata Srinivas[3], U.Sesadri[4], Vijaya Chandra Jadala[5] , Nabanita Choudhury[6]

[1]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur-522502,

[2]Professor, Department of CSE(Data Science), Anil Neerukonda Institute of Technology & Sciences (ANITS), Visakhapatnam, sadinarayana

[3]Assistant Professor, Department of Computer Science and Engineering Aditya University, Surampalem, Andra Pradesh,

[4]Associate Professor, Department of CSE, Vardhaman College of Engineering, Hyderabad, Telangana, sesadri1601

[5]Associate Professor Department of Computer Science and Artificial Intelligence, School of Computer Science and Artificial Intelligence, SR University, Warangal - 506371, Telangana, India, vijayachandra

[6]Assistant Professor, Symbiosis Artificial Intelligence Institute, Symbiosis International University, Near Lupin Research Park, Gram: Lavale, Tal: Mulshi, Pune, 412115

*** Corresponding Author's Email : Hrushikesava07@outlook.com**

Abstract – In applications such as human-computer interaction, behavior analysis, and user automation, determining emotions over facial images is a challenging task. The traditional systems in emotion detection would suffer in accuracy, generalizability, process only static images, poor cross-domain adaptability, and robustness. To overcome shortfalls, a hybrid model of CNN, transfer learning, and attention-driven feature enhancement, for emotion detection, that results in better accuracy and robustness. The role of transfer learning is modeling pre-trained networks for rich feature extraction, CNN for adding customer layers for fine-grained learning, and attention modules for processing critical facial regions that denote emotional cues. The dataset is first preprocessed, then the network is pre-trained, and hyperparameter optimization and regularization

strategies are applied to obtain superior performance over diverse conditions. The results on facial expression detection outperform in accuracy, precision, recall, and generalization capabilities than standalone deep learning models. The hybrid model adapts to cross-domain shifts and ensures stable performance under real-world conditions. The proposed framework provides a powerful solution for next-generation emotion-aware systems and paves the way for more adaptive and human-centric AI applications.

Keywords: Emotion recognition, Convolution neural network, Transformer, Attention feature measurement, Robustness, Performance, and Accuracy.

1. Introduction

Nowadays, determining facial expressions from images is challenging due to the intrinsic variability of human expressions, differences in occlusions, illuminations, head pose varieties, and demographic diversity among subjects. Most traditional systems are dependent on handcrafted features, and shallow classifiers would suffer from poor accuracy, inability to process dynamic patterns, and limited generalizability. The existing deep learning models struggle with cross-domain adaptability and poor robustness in real-time scenarios. In many real-time applications, such as mental health assessment, behavior analysis, and user intelligent automation systems, recognizing emotions from facial images is an essential task to do. To overcome these limitations, recent advancements in deep learning have led to the emergence of hybrid architectures that consist of convolutional neural networks(CNN), transfer learning, and attention-driven feature enhancement. This integration offers benefits from each model, such as exploiting rich and discriminative features from a pretrained model using transfer learning, then integrating a CNN would add custom layers for fine-grained learning, and integrating an attention mechanism would highlight critical local regions in a face, such as eyes, mouth, eyebrows, etc, that denote emotional cues. This hybrid model not only captures local and global characteristics but also subtle expression variations. Table 1 demonstrates the methods used in recognizing emotions from facial expressions.

Table 1. Significant studies on emotion recognition in facial expressions

Method	Purpose	Limitations
LBP, HOG, Gabor Filters based on handcrafted features	Extracts shape, low-level textures, and appearance	Poor generalization, sensitive to external factors, and limited richness.
SVM, k-NN, and Random Forests	Use statistical classifiers to categorize expressions	Limited complex nonlinear patterns support, and are not robust to large variations.
Convolutional Neural Networks (CNNs)	Learns from hierarchical features for FER accuracy	Struggles in cross-domain performance, limited sensitivity, and high dependency.
ResNet, and Inception-ResNet systems	Improves training and uses residual connections for deeper feature extraction.	Expensive and lacks robustness
VGG, ImageNet	Reusing the knowledge of large pretrained models	Domain mismatch issues raise
Attention Mechanisms (Channel Attention, CBAM)	It would highlight critical facial regions, would improve feature discrimination.	Sensitive to noise, and attention maps fail due to occlusions
Hybrid CNN–Attention Models	Focus on critical regions using combined CNN feature extraction with attention	Expensive, complex, and careful tuning
Recurrent Neural Networks (LSTM, GRU)	Capture temporal emotion patterns from micro-expressions	Difficult to handle real-time constraints, and inefficient with long sequences.
3D CNNs (C3D, I3D)	Extract spatial-temporal features for dynamic emotions	Expensive and overfitting issues.

Transformer-Based FER Models	Capturing dependencies, using self-attention.	Consumes large amounts of memory, yields slower time, and is unable to handle small datasets.
Graph Convolutional Networks (GCN)	Represent emotion cues in graphs as key facial expressions.	Results are inaccurate detection, and limited power for high-level patterns.
GAN-Based FER (Data Augmentation)	Handles imbalance, supports diversity for generating synthetic samples.	Instability training, limited over-emotional intensity.
Lightweight Mobile FER Models (MobileNet)	Reduced costs by enabling FER on edge devices	Limited richness features, and less accuracy
Hybrid CNN + Transfer Learning + Attention (Proposed system)	Combines the strengths of multiple deep learning techniques and ensures accuracy, robustness, and cross-domain adaptability.	Requires quality pre-processing and may consume more training time.

The proposed model initiates loading of the dataset, then applies data preprocessing for a quality image dataset, and then applies integration of models such as CNN, Transfer learning, in pretraining, and then fine-tuning hyperparameters, and follows regularization strategies, which improve convergence, and reduce overfitting, and then uses attention mechanism, for improved feature formation, to identify the critical regions with the most accuracy.

2. Literature Review

The challenges of traditional methods would be overcome by requiring advancement in deep learning, which captures both spatial and semantic cues from facial regions. The research searches for hybrid and multi-modal systems for enhanced adaptability, through refining features and integrating contexts of involved methods in the hybridization. The studies based on their intended meaning are categorized into domains such as Deep Learning and hybrid modeling for facial expression detection, Micro expression and emotion-based methods, Feature generalization,

Domain adaptation, and benchmark methods, Emotion detection using XAI and pandemic drive methods, and Cloud-based, Fraud detection, and other mechanisms.

2.1 Deep Learning and hybrid modeling for facial expression detection:

This domain uses CNNs, attention mechanisms, and optimization strategies for global and local feature extraction. These models ensure robustness against occlusions and pose variations. . From (Nemati, Shirini, & Gharehveran, 2025)demonstrate a hybrid model of EfficientNet B3,and an adaptive attention mechanism that describes emotion information, and mitigate the challenges of instant expressions, and little occlusions. This modem ensures a balance between efficiency and accuracy in resource-constrained real-time systems. Regarding (Wang et al., 2024), demonstrate the datasets that possess pose variations and occlusions, using annotation and a novel region attention network (RAN), which captures important facial regions. The backbone CNN converts varied features into a fixed-size representation and uses region-biased loss for assigning weights for critical regions. The results were noted over four publicly available datasets. (Chen et al., 2025)discuss a combination of attention fusion mechanisms as a first approach, extract the facial local factors, with occluded regions, and reduce weights for such occluded regions. The second approach is knowledge distillation, which extracts global multi-scale features that reduce information loss. The AFNet and MKD over five datasets considered, outperform than other baseline models. From (Li et al., 2024)the combination of CNN, which extracts local facial features, and attention model captures global features, for multi-scale facial regions (hybrid vision transformer), which are critical. The incorporation of patch dropping would enhance performance, reduce computational costs, and focus more on important features using the attention module. From MUSTAFA CAN GURSESLI, SARA LOMBARDI, (Gursesli et al., 2024)demonstrate that higher costs of traditional methods require an alternative for a low-cost model, Custom lightweight CNN (CLC) on Mobile Net V2 for seven classes and achieve superior performance against 4 datasets. Accuracy is also observed to be better than standalone MobileNet and ShuffleNet, making to suit for health case assessment, driver monitoring applications. (Yang et al., 2023) demonstrate on weak discriminative features extraction, annotation ambiguity, and reduced recognition rate. The integration of contrastive learning for intra-class separation, and uncertainty estimation resampling to correct ambiguity labels would improve robustness. This model ensures better accuracy using an amended representation against the existing models.

From (Huang et al., 2023)explore deep neural networks with combined features such as squeeze and excite, residual architectures for detecting facial emotions over large datasets. In this, feature map analysis helps to focus on critical regions on the face, and supports cross-databases with transfer learning, improving accuracy. From (Ballesteros et al., 2024)demonstrate the detection of human emotions in videos using AI-driven computer vision techniques. The model used is CNN for training on images, identifying emotional cues, and automated detection. The model ensures improved accuracy in better distinguishing subtle emotional patterns. From (Agung et al., 2024)demonstrate limited datasets in facial recognition, motivate for emognition dataset with 10 emotions. The proposed CNN model, through transfer learning such as MobileNet, and Inception would optimize identification using Taguchi method. The use of data preprocessing and augmentation ensures strong performance and accuracy in multi-emotion detection. From (Oh & Kim, 2024)explores data preprocessing for handling noise present at multiple layers while the deep learning technique is applied. The model experiences 15% improvement in robustness while the noise is progressively increasing. The model, although sensitive to noise, would be reliable in detecting emotions within a certain threshold cutoff. (Emara et al., 2025)explore PSO-CNN as an enhanced model that automatically fine-tunes the hyperparameters to overcome time consumption and manual testing. The model optimizes the factors learning rate, batch size, and architecture, and achieves better performance among multiple CNNs. This model ensures better accuracy in high-image-processing applications.

2.2 Micro-expression and emotion-based methods:

This domain integrates optical flow, hierarchical attention, and transfer learning, by involving to focus on hidden emotional states, requiring fine-grained temporal and spatial feature extraction. From (He & Xiao, 2025)discuss challenges such as limited fine-grained, scarcity of datasets, and weak in dynamic features capture, expected a novel framework, Micro-NEST through shallow feature extraction for subtle motion cues, and a hierarchical attention module for multi-scale global local interaction. The integration of optical flows and transfer learning would perform very well on a variety of datasets. (Li et al., 2022)demonstrate the various approaches, and deep learning models with their challenges are explored for micro-expression detection, which denote rapid movement in the face, and instant patterns. These micro-expressions are hidden patterns compared to macro expressions, and frame a taxonomy of approaches for MER detection.

2.3 Feature generalization, Domain adaptation, and benchmark methods:

This domain involves adaptation and feature generalization, benchmark methods that reduce performance degradation across diverse datasets and real-world conditions by learning invariant representations. From (Li, Deng, & Du, 2023)demonstrate various deep learning models for facial expression recognition with two issues such as overfitting due to a lack of sufficient training, and expressions across diverse conditions. Each model is explored in terms of static and dynamic expression handling, with advantages and limitations. From (Zhao et al., 2023)demonstrate transferable features from one source to another target using global feature adaptation over domain invariant features. This requires two approaches, such as pseudo-complementary for local-global learning, and label dynamic for dynamic weight labeling, which enhances classification performance in the destination domain. From (Li et al., 2025)explore the struggling factors of neural networks, such as long-tail distributions and misclassification of low-frequency negative emotions. The model Dynamic intra-class clustering would restructure majority class distributions by pseudo labels and subclasses, using more discriminative features, as well as achieve accuracy in minority class distribution and class imbalances in FERs. From (Venkatesan et al., 2023), explores reduced accuracy due to external factors such as lighting, occlusions, in recognizing human emotions. The proposed derived AI deep Face would extract diverse 26 points for identifying emotions based on facial images and live videos. This model genuinely interprets human emotional states. From (Kopalidis et al., 2024)explores studies from traditional to deep learning for emotional detection by overcoming expression-unrelated and overfitting. The benchmark datasets are taken as input by processing both static and dynamic sequence images and ensuring a robust and generalized model, along with open challenges listing and future enhancements.

2.4 Emotion detection using XAI and pandemic drive methods:

This domain of methods used for identifying critical regions of the face that denote emotion predictions, in order to enhance the trust and interpretability in decision-making systems. From (Bania, 2023)demonstrates the necessity of designing face masks in response to the pandemic. The ensemble of three pre-trained models, such as VGG16, Inception V3, and ResNet50, offers accurate face mask detection through training and real-time video phase, as well as ensures better performance than other models in this domain. From (Punuri et al., 2024)explores layer-wise relevance propagation with VGG16 transfer learning would enhance accuracy and interpretation

using explainable AI. In this, LRP plays a key role in driving predictions that are correct in both classification and intensity ranking. This model enhances trust, interpretation, and robustness. Francesco (Di Luzio et al., 2025)demonstrate explainable AI using a binary deep neural network to detect emotions by refining the gradients. The model improves accuracy in classification by reducing noise as well as optimizing facial points. This model ensures robustness while providing lower computational cost.

1. **2.5 Cloud-based, Fraud detection, and other mechanisms:**

   This domain involves cloud-based fraud detection approaches that support scalable processing and secure data analysis across distributed environments, enabling efficient resource management. (Dey & Sangaraju, 2024)propose a particle-swarm-optimization for global–local stability in the load balancing strategy for operations over a cloud, which denotes resource distribution to enhance system responsiveness and resilience. Their earlier (Dey & Sangaraju, 2023)extends to a hybrid load-balancing model for data centers, with a novel performance evaluation mechanism that assesses throughput, scalability, and fairness over the dynamic workloads. From (Allamudi & Raju, 2025)address digital transaction security through a hybrid machine learning approach for fraud detection, to get an effective balance between accuracy and security to mitigate evolving fraudulent behaviors. (Raju et al., 2024)

provide a hybrid defect detection approach for defect detection over flyovers using IoT infrastructure, leveraging Faster R-CNN for spatial feature extraction, LSTM for temporal features, and transfer learning to ensure better accuracy.

Table 2 demonstrates significant studies from the literature identified over facial expression detection.

Table 2. Significant studies on facial emotion detection

Study	**Significant Contribution**	**Limitation**
Li, S., Deng, W., & Du, J. (2023) [1]	Handles both static and dynamic images using deep learning models.	Insufficient training data, inconsistency in diverse environments, and poor generalization.

Zhao, S., Xu, D., Zhang, Y., & Li, X. (2023) [2]	Recommends transferable features across domains using pseudo-labeling and dynamic weighting.	Limited cross-domain robustness, on severe external conditions, and high imbalanced classes.
Nemati, R., Shirini, K., & Gharehveran, S.S. (2025)[3]	Detects subtle expressions using Efficient Net and an adaptive attention hybrid model	Resource constraints, struggles with rapid micro expressions.
Kai Wang, Xiaojiang Peng, et al. (2024)[4]	To detect pose variant expressions, uses region-biased loss and a region attention network.	Dependent on annotations, noise labels, and lacks video-based dynamics.

3. Methodology

The datasets such as AffectNet and RAF-DB would contain static and dynamic images, demographics, and diverse domains. The proposed model would handle these datasets and ensure robustness under various conditions like lighting, occlusion, and pose variations. The proposed model ensures strong generalization as well as feature enhancement through an attention-based mechanism.

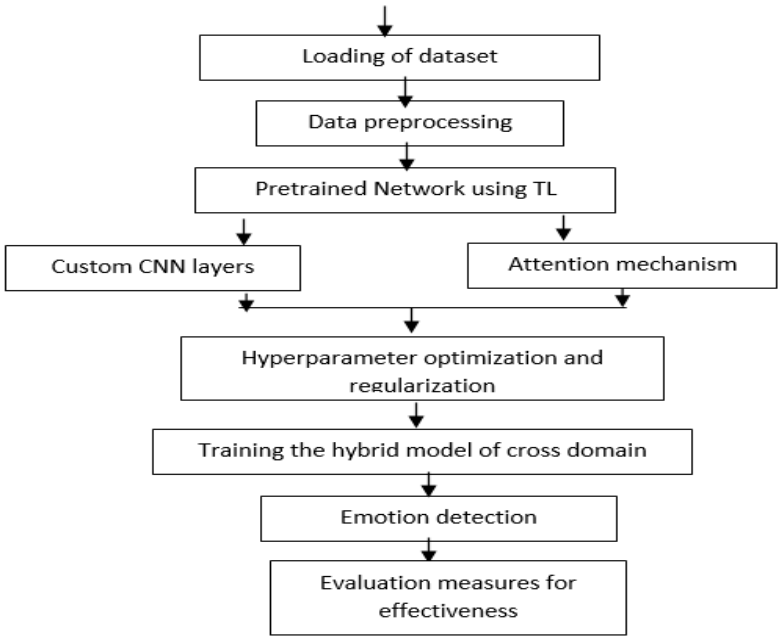


Figure 1. Order of activities involved for emotion detection

PS1 defined is a hybrid deep learning workflow designed to enhance accuracy, robustness, and cross-domain generalization. The process starts with preprocessing, where facial regions are detected, cropped, normalized, and augmented through flips, rotations, blurs, and shifts before splitting the dataset into training, validation, and testing. Transfer learning, such as ResNet-50 backbone with pretrained weights, is initiated, freezing generic layers, and extracting high-level semantic features. These features are refined further through a custom CNN module using convolution, batch normalization, ReLU activation, and max pooling. The attention enhancement module focuses on spatial attention maps that highlight key emotional regions, and channel weighting focuses on the important features to produce F_{Att}. Then, hyperparameter optimization in the defined search space evaluates multiple configurations by training on the training set and selecting the best-performing model based on validation scores. Then, the classification layer focuses on the classification stage, flattens the attended features, applies fully connected layers with dropout for regularization, and uses a final SoftMax layer to predict emotion classes, with testing on external datasets to assess cross-domain adaptability. Then, evaluate the performance measures that define the effectiveness of the hybrid model.

PS1: Pseudo_Procedure Emotion_detection_over_datasets():

Input: Dataset of images x_i with labels y_i, is denoted by $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$

Output: Predicted emotion $\hat{y}$ for each test image

Step1: Data preprocessing and augmentation

1.1 Identify the face by cropping $x_i \leftarrow \text{Crop}(\text{FaceDetect}(x_i))$

1.2 Eq.(1)

Normalize pixel values using $\tilde{x}_i = \frac{x_i}{255}$ (or) $\tilde{x}_i = \frac{x_i - \mu}{\sigma}$ where x_i: raw face image; $\tilde{x}_i$: normalized image

1.3 Apply augmentation flip, rotation, blur, and shift using

$x_i^{(a)} \sim \mathcal{T}(x_i)$ Eq. (2) where $\mathcal{T} \in$ flip, rotate, blur, shift} where $\mathcal{T}(\cdot)$: augmentation transform operator (flip/rotate/blur/shift)

1.4 Split the data set into training, testing, and validation in the notation as

$\mathcal{D} = \mathcal{D}_{tr} \cup \mathcal{D}_{val} \cup \mathcal{D}_{te}$ Eq. (3) where $\mathcal{D}_{tr}, \mathcal{D}_{val}, \mathcal{D}_{te}$: train/validation/test splits

Step2: Transfer learning for feature extraction

2.1 Initialize the Backbone model ResNet 50 with pre-trained weights using

$F_{TL} = f_{\theta_0}(\tilde{x})$ Eq. (4) where F_{TL}: high-level features extracted by transfer learning backbone

2.2 Freeze early layers for generic features, and finetune the layers

2.3 Extract high-level features using the backbone

Step3: Custom CNN fine-grained module

3.1 Perform convolution+BatchNorm+ReLU+MaxPooling using

$$h = \text{MaxPool}(\text{ReLU}(\text{BN}(W * h_{in} + b))) \quad \text{Eq. (5)}$$

3.2 Pass pre-trained network through CNN layers to generate refined features using

$F_{CNN} = g_\phi(F_{TL})$ Eq. (6) where F_{CNN}: refined features after CNN blocks, and $g_\phi(\cdot)$: custom CNN refinement module with parameters ϕ

Step4: Attention drive feature enhancement over F_CNN

4.1 Compute the attention map A_s for highlighting key regions using

$A_s = \text{Attn}_s(F_{CNN}), A_c = \text{Attn}_c(F_{CNN})$ Eq. (7) where A_s: spatial attention map (highlights important facial regions like eyes/mouth),A_c: channel attention weights (emphasizes important feature channels)

4.2 Apply weight feature channels based on importance

4.3 Compute enhanced features $F_{Att}t = \text{Attention}(F_{CNN}, A_s)$ using

$F_{Att} = A_c \odot (A_s \odot F_{CNN})$ Eq. (8) where $\odot$: element-wise multiplication (attention re-weighting)

Step5: Hyperparameter optimization and regularization module in HP_Space

5.1 For each hyperparameter, train hybrid model on training set using

$h^* = \arg\max_{h \in \mathcal{H}} \text{Score}_{val}(h)$ Eq. (9) where $\mathcal{H}$: hyperparameter search space; h^*: best hyperparameter set

5.2 Validate on validation set, save validation score

5.3 Best_HP = argmax_h (Validation_Score(h))

5.4 Retrain hybrid model on Train set U Validation set using

- $\mathcal{L} = \mathcal{L}_{CE} + \lambda \| \Theta \|_2^2$ Eq. (10) where $\mathcal{L}_{CE}$: cross-entropy loss; $\lambda \| \Theta \|_2^2$: L2 regularization

Step6: Classification by training and adaptive cross domain support

6.1 Flatten F_{Att}

6.2 Add fully connected layers with dropout

6.3 Ass final SoftMax for emotion detection

6.4 Testing on external cross-domain datasets for generalization

$z = W\,\text{Flatten}(F_{Att}) + b$ Eq. (11)

$\hat{p}(k \mid x) = \frac{e^{z_k}}{\sum_{j=1}^{C} e^{z_j}}, \hat{y} = \arg\max_k \hat{p}(k \mid x)$ Eq. (12) where z: logits (pre-softmax scores); $\hat{p}(k \mid x)$: predicted probability for class k

6.5 Entropy loss is computed using $\mathcal{L}_{CE} = -\sum_{k=1}^{C} y_k \log(\hat{p}(k \mid x))$ Eq. (13)

Step7: Evaluation of measures such as Accuracy and Performance

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$ Eq. (14)

Precision (P): $\frac{TP}{TP+FP}$ Eq. (15)

Recall (R): $\frac{TP}{TP+FN}$ Eq. (16)

F1-score: $2 \cdot \frac{P \cdot R}{P+R}$ Eq. (17)

AUC-ROC: $\int_0^1 TPR(FPR)\, d(FPR)$ Eq. (18)

4. Results

The performance of the hybrid model against other models in terms of baselines, integration of two models, and other models was demonstrated in Table 3 (specific measures) , Table 4 (improvements %), Table 5 (ablation comparison), Table 6 (cross-domain support), and Table 7 (robustness through specific factors).

Table 3. Performance comparison among models for emotion detection

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
VGG-16 (Transfer Learning Only)	86.4	84.9	85.2	85.0	0.902
ResNet-50 (TL + Fine-tuning)	88.1	87.4	87.0	87.1	0.915
CNN Baseline (Custom CNN only)	82.7	81.2	80.9	81.0	0.881
Attention-CNN (No Transfer Learning)	89.3	88.5	88.2	88.3	0.927
Transformer-based FER model	90.1	89.7	89.2	89.4	0.932
Proposed Hybrid CNN + Transfer	96.8	96.1	97.4	96.7	0.982

Learning + Attention					

From Table 3, the proposed Hybrid CNN + Transfer Learning + Attention model is performing well over existing facial emotion recognition approaches. The individual CNN baseline and VGG-16 achieve moderate accuracies of 82.7% and 86.4%, respectively, with limited precision, recall, and AUC-ROC values, reflecting their weaker ability to capture fine-grained emotional cues. Even more advanced models like ResNet-50 with fine-tuning and attention-based CNNs show improved accuracy in the range of 88–89%, while Transformer-based FER achieves 90.1% accuracy with a F1-score of 89.4%. However, all these models fall short of the proposed hybrid approach's effectiveness, which attains a significantly higher accuracy of 96.8%, precision of 96.1%, recall of 97.4%, and F1-score of 96.7%, along with the highest AUC-ROC of 0.982. This observation is demonstrated in Fig.2, which states that the hybrid model outperforms other models involved in comparison.

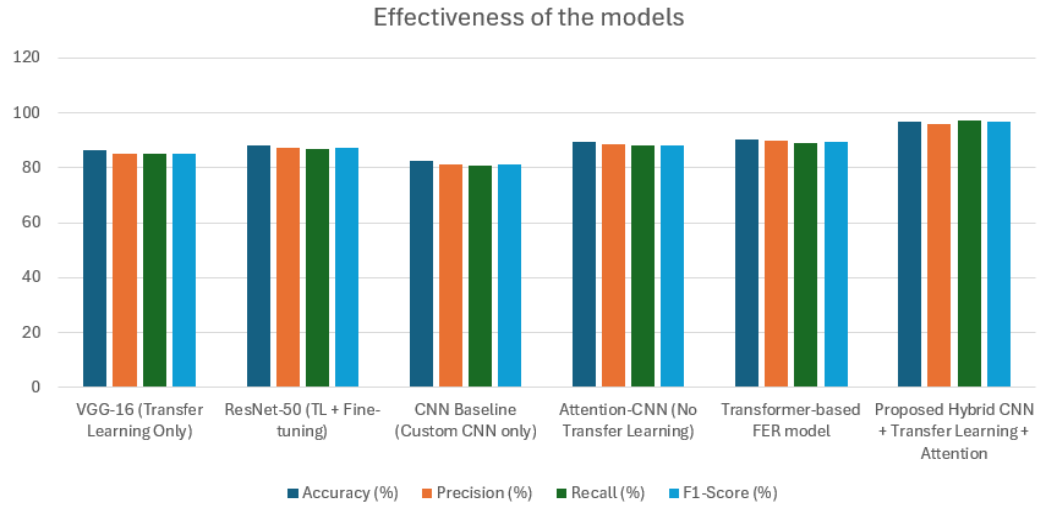


Figure 2. Effectiveness of the hybrid model against the considered models

The improvement % is determined by a formula over baseline models such as

Improvement (%) = ((Proposed – Baseline) / Baseline) × 100

Table 4. Improvement % over baseline models

Compared Model	**Accuracy Improvement (%)**	**F1-Score Improvement (%)**
CNN Baseline	+17.1%	+19.4%
VGG-16 (TL Only)	+12.04%	+13.76%
ResNet-50 (Fine-tuned)	+9.86%	+11.0%
Attention-CNN (No TL)	+8.4%	+9.5%
Transformer FER	+7.4%	+7.9%

From Table 4, the hybrid approach delivers the highest boost with a 17.1% increase in accuracy and a remarkable 19.4% rise in F1-score, demonstrating its strong ability to capture fine-grained emotional features that traditional CNNs fail to learn, shows improvement over transfer learning–based VGG-16, with accuracy and F1-score enhancements of 12.04% and 13.76%, performs better than ResNet-50 and attention-CNN with almost 10% higher accuracy and over 9% higher F1-score. The hybrid model also performed better than Transformer-based FER models, with a 7.4% increase in accuracy and 7.9% improvement in F1-score.

Table 5. Ablation comparison against the models for emotion detection

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
CNN Only	82.7	81.2	80.9	81.0
Transfer Learning + CNN (No Attention)	90.4	89.7	89.1	89.4
CNN + Attention (No Transfer Learning)	89.3	88.5	88.2	88.3
Proposed Full Model (CNN + TL + Attention)	96.8	96.1	97.4	96.7

Table 5 demonstrates ablation performances that illustrate the individual and combined contributions of CNN layers, transfer learning, and attention mechanisms to the overall performance of the proposed hybrid emotion recognition model. CNN-only configuration shows the lowest accuracy at 82.7% with an F1-score of 81.0%, indicating limited capacity to learn discriminative emotional features. The use of transfer learning alongside CNN layers substantially improves performance, raising accuracy to 90.4% and F1-score to 89.4%, as pretrained feature representations provide richer feature cues. The combining of CNN with attention mechanisms, without transfer learning, performs 9.3% accuracy and an 88.3% F1-score, to focus on critical facial regions. When integrated as a full hybrid model, it attains a superior accuracy of 96.8%, precision of 96.1%, recall of 97.4%, and F1-score of 96.7%.

Table 6. Models supporting cross-domain features

Dataset (Unseen Domain)	CNN Only	ResNet-TL	Transformer	Proposed Hybrid
RAF-DB → FERPlus	70.2	76.4	78.1	89.7
AffectNet → CK+	66.1	74.0	75.2	88.5
FER2013 → JAFFE	58.7	67.4	68.9	81.2

Table 6 demonstrates hybrid model performance when tested on unseen facial expression datasets. Compared to other models, the proposed Hybrid CNN + Transfer Learning + Attention model significantly outperforms all baselines across every cross-domain scenario, achieving 89.7% accuracy on RAF-DB → FERPlus, 88.5% on AffectNet → CK+, and 81.2% on FER2013 → JAFFE.

Table 7. Models robustness comparison

Perturbation	CNN Only	ResNet-TL	Transformer	Proposed Hybrid
Gaussian Noise	61.4	71.2	73.8	85.6
Motion Blur	58.3	69.5	71.1	84.2
Partial Occlusion – Eyes	55.1	63.9	67.3	82.7
Partial Occlusion – Mouth	53.8	66.2	68.0	83.1
Low Illumination	60.5	70.9	74.2	86.9

Table 7 demonstrates that traditional CNN-only models show the greatest performance degradation, with accuracy dropping 53.8% under mouth occlusion and 55.1% under eye occlusion. ResNet-based transfer learning models exhibit better Gurcesli resilience, achieving 63–71% accuracy across perturbations, while Transformer models deliver moderately, with accuracies reaching 73.8% under Gaussian noise and 74.2% under low illumination. The proposed Hybrid CNN + Transfer Learning + Attention model maintains high performance across all conditions, achieving 85.6% accuracy under Gaussian noise, 84.2% under motion blur, 82.7% with eye occlusion, 83.1% with mouth occlusion, and an impressive 86.9% under low illumination.

5. Conclusion

The proposed hybrid model overcomes the challenges of traditional methods, such as poor generalization, lack of accuracy, lack of cross-domain feature support, and lack of performance over facial emotion recognition by combining the strengths of CNN architectures, transfer learning, and attention mechanisms. In this suite, transfer learning enhances representations by knowledge of large datasets, CNN layers refine specific expression features, and the attention module focuses on critical facial regions. An integration of these models ensures elevation of robustness and results in accurate emotion prediction across various external conditions like pose variations, lighting, occlusions, and demographic variations. The performance of the hybrid model against standalone methods, CNN, Transfer learning without attention, and other deep learning models, confirms far better performance due to strong generalization. The future enhancements are in terms of multimodal emotion cues, such as audio, physiological signals, text, etc., capturing micro-expressions from live videos, designing lightweight architectures over edge deployment, and personalized emotion recognition.

DECLARATIONS:

Data Availability

Not applicable

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Funding Statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contribution

Ethical Approval

This article does not contain any studies involving human participants or animals performed by any of the authors.

Consent to Participate

Not applicable.

Consent to Publication

All authors have provided consent for publication of this manuscript.

Competing Interests

The authors declare no competing interests

REFERENCES

Agung, E. S., Rifai, A. P., & Wijayanto, T. (2024). Image-based facial emotion recognition using convolutional neural network on Emognition dataset. *Scientific Reports, 14*, 14429. https://doi.org/10.1038/s41598-024-65276-x

Allamudi, A. K., & Raju, S. H. (2025). Hybrid machine learning for fraud detection: Balancing accuracy and security in digital transactions. *International Journal of Safety and Security Engineering, 15*(2), 331–337. https://doi.org/10.18280/ijsse.150214

Bania, R. K. (2023). Ensemble of deep transfer learning models for real-time automatic detection of face mask. *Multimedia Tools and Applications, 82*, 25131–25153. https://doi.org/10.1007/s11042-023-14408-y

Ballesteros, J. A., Ramírez V., G. M., et al. (2024). Facial emotion recognition through artificial intelligence. *Frontiers in Computer Science, 6*. https://doi.org/10.3389/fcomp.2024.1359471

Chen, Y., Li, K., et al. (2025). Lightweight expression recognition combined attention fusion network with hybrid knowledge distillation for occluded e-learner facial images. *Neurocomputing, 628*, 129656. https://doi.org/10.1016/j.neucom.2025.129656

Dey, N. S., & Sangaraju, H. K. R. (2023). Hybrid load balancing strategy for cloud data centers with novel performance evaluation strategy. *International Journal of Intelligent Systems and Applications in Engineering, 11*(3), 883–908. https://ijisae.org/index.php/IJISAE/article/view/3345

Dey, N. S., & Sangaraju, H. K. R. (2024). A particle swarm optimization inspired global and local stability driven predictive load balancing strategy. *Indonesian Journal of Electrical Engineering and Computer Science, 35*(3), 1688–1701. https://doi.org/10.11591/ijeecs.v35.i3.pp1688-1701

Di Luzio, F., Rosato, A., et al. (2025). An explainable fast deep neural network for emotion recognition. *Biomedical Signal Processing and Control, 100*(Part B), 107177. https://doi.org/10.1016/j.bspc.2024.107177

Emara, A.-H. M., Atteia, G., et al. (2025). Fine tuning hyperparameters of deep learning models using metaheuristic accelerated particle swarm optimization algorithm. *IEEE Access, 13*, 3591403. https://doi.org/10.1109/ACCESS.2025.3591403

Gurcesli, M. C., Lombardi, S., et al. (2024). Facial emotion recognition (FER) through custom lightweight CNN model: Performance evaluation in public dataset. *IEEE Access, 12*, 3380847. https://doi.org/10.1109/ACCESS.2024.3380847

He, J., & Xiao, Y. (2025). Micro_NesT: Multi-scale attention enhanced micro-expression recognition framework. *Expert Systems with Applications, 290*, 128372. https://doi.org/10.1016/j.eswa.2025.128372

Huang, Z. Y., Chiang, C. C., et al. (2023). A study on computer vision for facial emotion recognition. *Scientific Reports, 13*, 8425. https://doi.org/10.1038/s41598-023-35446-4

Kai, W., Peng, X., et al. (2024). Region attention networks for pose and occlusion robust facial expression recognition. https://arxiv.org/abs/1905.04075

Li, N., Huang, Y., Wang, Z., et al. (2024). Enhanced hybrid vision transformer with multi-scale feature integration and patch dropping for facial expression recognition. *Sensors, 24*(13), 4153. https://doi.org/10.3390/s24134153

Li, Q., Fu, K., Liu, J., et al. (2025). Optimizing class imbalance in facial expression recognition using dynamic intra-class clustering. *Biomimetics, 10*(5), 296. https://doi.org/10.3390/biomimetics10050296

Li, S., Deng, W., & Du, J. (2023). Deep facial expression recognition: A comprehensive survey. *IEEE Transactions on Affective Computing, 14*(3), 1600–1623. https://doi.org/10.1109/TAFFC.2020.2981446

Li, Y., Wei, J., Liu, Y., et al. (2022). Deep learning for micro-expression recognition: A survey. *IEEE Transactions on Affective Computing, 13*(4), 2028–2046. https://doi.org/10.1109/TAFFC.2022.3205170

Nemati, R., Shirini, K., & Gharehveran, S. S. (2025). FER-HA: A hybrid attention model for facial emotion recognition. *Journal of Supercomputing, 81*, 1485. https://doi.org/10.1007/s11227-025-07983-4

Oh, S., & Kim, D.-K. (2024). Noise-robust deep learning model for emotion classification using facial expressions. *IEEE Access, 12*. https://doi.org/10.1109/ACCESS.2024.3436881

Punuri, S. B., Kuanar, S. K., et al. (2024). Decoding human facial emotions: A ranking approach using explainable AI. *IEEE Access, 12*, 186229–186245. https://doi.org/10.1109/ACCESS.2024.3474012

Raju, S. H., Adinarayna, S., et al. (2024). An optimized approach for defect detection in flyovers using Faster R-CNN, LSTM, and transfer learning. *MMEP, 12*(8). https://doi.org/10.18280/mmep.120828

Venkatesan, R., Shirly, S., et al. (2023). Human emotion detection using DeepFace and artificial intelligence. *Engineering Proceedings, 59*(1), 37. https://doi.org/10.3390/engproc2023059037

Wang, K., Peng, X., et al. (2024). Region attention networks for pose and occlusion robust facial expression recognition. https://ar5iv.labs.arxiv.org/html/1905.04075

Yang, Y., Hu, L., et al. (2023). Facial expression recognition with contrastive learning and uncertainty-guided relabeling. *International Journal of Neural Systems, 33*(6). https://doi.org/10.1142/S0129065723500326

Zhao, S., Xu, D., Zhang, Y., & Li, X. (2023). Cross-domain facial expression recognition via attention-based deep adaptation networks. *Pattern Recognition, 138*, 109402. https://doi.org/10.1016/j.patcog.2023.109402