European Journal of Prosthodontics and Restorative Dentistry (2026) 34 (7s), 870–887

An Attention-Enhanced ConvNeXtTiny Model for Robust Facial Expression Recognition

Keywords
Analgesia; Facial Expression Recognition, ConvNeXtTiny, CBAM, Deep Learning, Transfer Learning, Attention Mechanism, FER2013, RAF-DB.

Authors
[1*]Ms. Manisha B. Thombare
Research Scholar, Department of Computer Engineering,MET's BKC, Institute of Engineering, Nashik, Maharashtra, India
Savitribai Phule Pune University, Pune, Maharashtra, India
manishathombare27@gmail.com

[2]Dr. Shyamrao V. Gumaste
Professor and Head, Department of Artificial Intelligence and Data Science,
MET's BKC, Institute of Engineering, Nashik, Maharashtra, India
Savitribai Phule Pune University, Pune, Maharashtra, India, Email: svgumaste@gmail.com

Abstract
Facial Expression Recognition (FER) plays an important role in affective computing and human–computer interaction by enabling automated interpretation of human emotional states from facial images. Despite recent advances in deep learning, reliable FER remains challenging because of variations in facial appearance, illumination, pose, occlusion, image quality, class imbalance, and visual similarity between emotion categories. This study proposes a facial expression recognition model that combines the ConvNeXtTiny architecture with the Convolutional Block Attention Module (CBAM) to improve discriminative facial feature learning. ConvNeXtTiny is employed to extract hierarchical facial representations, while CBAM refines these representations through complementary channel and spatial attention. Transfer learning, data augmentation, label smoothing, and adaptive optimization are incorporated during model training to improve convergence and generalization. The proposed model is evaluated on the FER2013 and RAF-DB benchmark datasets using accuracy, precision, recall, F1-score, confusion matrix analysis, cross-validation, computational complexity, and inference-time analysis. The experimental results demonstrate competitive recognition performance across both datasets. On FER2013, the model achieves an accuracy of 77.0%, precision of 77.5%, recall of 77.0%, and F1-score of 77.2%. On RAF-DB, the model achieves an accuracy of approximately 92.9%, demonstrating its capability to recognize facial expressions under diverse real-world conditions. Ablation experiments further indicate the contribution of the attention mechanism and optimization components to the final recognition performance. The findings demonstrate that combining modern convolutional feature extraction with lightweight channel-spatial attention provides an effective approach for facial expression recognition.

Received:26-06-2026
Revised:29-07-2026
Accepted:04-08-2026
Doi:10.1922/ejprd.v34i7s.1706

1. INTRODUCTION

Facial expressions provide an important non-verbal means of communicating human emotional states. Automatic Facial Expression Recognition (FER) has therefore received considerable attention in computer vision, affective computing, human–computer interaction, healthcare, intelligent transportation, surveillance, education, and social robotics. A FER system aims to identify the emotional state represented by a facial image and generally categorizes it into discrete expressions such as anger, disgust, fear, happiness, sadness, surprise, and neutrality.

The development of FER systems has progressed from handcrafted facial descriptors and conventional machine learning classifiers toward deep learning-based approaches. Traditional methods commonly relied on Local Binary Patterns, Histogram of Oriented Gradients, Gabor features, facial landmarks, and geometric measurements followed by classifiers such as Support Vector Machines and Random Forests. Although these approaches can perform effectively under controlled conditions, their dependence on manually designed features limits their ability to represent complex facial appearance variations.

Deep convolutional neural networks have substantially improved FER by learning hierarchical representations directly from facial images. Transfer learning has further improved performance by allowing models pretrained on large-scale image datasets to be adapted to facial expression recognition tasks. Transfer learning-based FER has demonstrated the usefulness of pretrained visual representations, particularly when the target dataset is relatively limited in size [2]. CNN-based FER methods have also been investigated extensively for automatic extraction of discriminative facial features [17].

Attention mechanisms provide another important direction for improving facial feature representation. Rather than treating all extracted features equally, attention mechanisms enable a network to assign greater importance to informative regions or feature channels. Landmark-guided attention has been investigated to identify facial regions associated with emotional characteristics [13], while other attention-based approaches have been developed to improve contextual representation and recognition under challenging facial conditions. Recent work has also explored context transformers and multiscale feature fusion for robust facial emotion recognition [29].

Transformer-based architectures have further expanded the capability of FER models by learning long-range relationships among facial features. Vision Transformer-based FER models have demonstrated the ability to capture global contextual information and have been investigated for improving recognition performance on benchmark datasets [12]. However, transformer-based architectures may require substantial computational resources and large amounts of training data. This creates a practical challenge when high recognition performance must be balanced against model complexity and inference cost.

Lightweight architectures have therefore remained important for practical FER applications. MobileNetV2, for example, has been investigated for resource-efficient facial expression recognition and is particularly relevant for applications involving constrained computational environments [26]. Other recent approaches have investigated collaborative learning, contextual transformers, facial landmarks, and multimodal information to improve FER performance [6, 13, 29].

Despite these developments, facial expression recognition remains challenging because facial images exhibit substantial intra-class variation and inter-class similarity. Expressions such as fear, surprise, sadness, and neutral may share similar facial characteristics, while the same emotional state may appear substantially different across individuals. FER performance can also be affected by illumination, head pose, occlusion, image resolution, demographic variation, and class imbalance. These challenges motivate the development of models capable of extracting hierarchical facial representations while selectively emphasizing emotionally informative features.

ConvNeXt provides a modern convolutional architecture that incorporates contemporary architectural design principles while retaining the computational characteristics of convolutional neural networks. Its hierarchical representation capability makes it suitable for extracting facial features at multiple levels of abstraction. However, the extracted features do not necessarily contribute equally to emotion classification. Integrating an attention mechanism can therefore provide an additional feature refinement stage that emphasizes informative channels and spatial regions.

Motivated by this observation, this study proposes a ConvNeXtTiny-CBAM model for facial expression recognition. ConvNeXtTiny is employed as the feature extraction backbone, while the Convolutional Block Attention Module (CBAM) is incorporated to refine the extracted representations through sequential channel and spatial attention. The model further employs transfer learning, data augmentation, label smoothing, and adaptive optimization to improve learning stability and generalization. The proposed model is evaluated using FER2013 and RAF-DB, which provide complementary experimental conditions for assessing facial expression recognition performance.

The main contributions of this study are as follows:

A ConvNeXtTiny-based facial expression recognition model is developed with CBAM integrated for channel-wise and spatial feature refinement.

A transfer learning-based training strategy incorporating data augmentation, label smoothing, and adaptive optimization is employed to improve feature learning and model generalization.

The proposed model is experimentally evaluated on two widely used FER benchmarks, namely FER2013 and RAF-DB, using a common evaluation methodology.

Comprehensive performance analysis is performed using accuracy, precision, recall, F1-score, class-wise performance, confusion matrices, and cross-validation.

Ablation experiments are conducted to quantify the contribution of the major components of the proposed model, while computational analysis is performed to examine model complexity and inference efficiency.

The remainder of this paper is organized as follows. Section 2 reviews existing facial expression recognition

approaches and identifies the research gap addressed by this study. Section 3 presents the proposed ConvNeXtTiny-CBAM model and its mathematical formulation. Section 4 describes the datasets, preprocessing procedures, training configuration, and evaluation methodology. Section 5 presents and analyzes the experimental results obtained on FER2013 and RAF-DB, followed by comparison with existing methods and ablation analysis. Section 6 discusses the findings, limitations, and practical implications of the proposed approach. Section 7 concludes the study and outlines directions for future research.

2. Related Work

Facial Expression Recognition has evolved from conventional image-processing techniques to deep learning, attention-based, transformer-based, and multimodal approaches. The development of these methods has improved the ability of computational systems to identify facial emotions under different imaging conditions. However, challenges related to facial appearance variation, class imbalance, occlusion, pose, illumination, and computational requirements continue to influence recognition performance.

2.1 Traditional and Deep Learning-Based Facial Expression Recognition

Early FER systems primarily relied on handcrafted facial descriptors and conventional machine learning algorithms. Features such as Local Binary Patterns, Histogram of Oriented Gradients, Gabor filters, and facial landmark information were commonly extracted from facial images and subsequently classified using machine learning algorithms. These approaches provided useful results under controlled conditions but had limited capability to automatically learn complex and hierarchical facial representations.

The adoption of convolutional neural networks significantly changed FER research by enabling automatic feature extraction directly from facial images. CNN-based methods learn multiple levels of representation, beginning with low-level facial textures and progressing toward higher-level semantic characteristics. Bhagat et al. investigated CNN-based facial emotion recognition and demonstrated the applicability of deep convolutional feature learning for FER [17].

Transfer learning further improved the applicability of deep neural networks, particularly when the available FER dataset was insufficient for training a deep model from the beginning. Lavanya et al. investigated transfer learning for facial emotion recognition and demonstrated the benefits of using pretrained representations for the target FER task [2]. Transfer learning has subsequently become a common strategy in FER because pretrained models provide informative initial representations and can reduce training requirements.

Barile et al. specifically investigated transfer learning for FER using small datasets and highlighted its usefulness when limited training samples are available [20]. These studies demonstrate that pretrained deep representations can provide a strong foundation for facial expression classification. However, the effectiveness of transfer learning remains dependent on the similarity between the source and target domains.

2.2 Attention-Based Facial Expression Recognition

Although CNNs can learn hierarchical facial features, conventional convolutional architectures generally treat the extracted feature maps without explicitly determining which channels or spatial regions are most informative for emotion recognition. Attention mechanisms address this limitation by assigning different importance to features according to their relevance to the classification task.

Facial landmark-guided attention has been investigated to identify informative facial regions and incorporate facial structural information into the recognition process. Colaco and Seog Han proposed a scalable context-based FER approach using facial landmarks and an attention mechanism to improve recognition under varying facial conditions [13]. Similarly, landmark-based approaches have demonstrated that facial geometric information can provide additional cues for identifying emotional states [18].

Arabian et al. investigated facial emotion recognition beyond conventional pixel representations by using facial point landmark meshes [15]. Such approaches provide explicit structural information about the face and can improve robustness when pixel-level appearance alone is insufficient.

Attention mechanisms have also been incorporated directly into deep neural networks to improve feature selection. The CBAM mechanism is particularly relevant to the present study because it sequentially applies channel and spatial attention. Channel attention identifies informative feature channels, while spatial attention emphasizes important spatial locations within the feature representation. This provides a computationally efficient mechanism for refining convolutional features without requiring a complete transformer architecture.

2.3 Transformer-Based Facial Expression Recognition

Transformer architectures have recently become an important research direction in FER because of their ability to model long-range relationships among image regions. Unlike conventional convolutional operations that primarily emphasize local receptive fields, transformer-based models can establish relationships between distant facial regions.

Fatima et al. investigated Vision Transformer-based facial emotion recognition and demonstrated the potential of transformer architectures for extracting contextual facial representations [12]. Transformer-based approaches can capture relationships between facial regions such as the eyes, eyebrows, nose, and mouth, which may contribute jointly to the interpretation of an emotional state.

Gan et al. proposed a context transformer with multiscale fusion for robust facial emotion recognition [29]. The use of multiscale contextual representations addresses the fact that facial expressions contain information at different spatial levels. However, transformer-based architectures generally introduce higher computational and memory requirements than lightweight convolutional models.

This computational requirement creates an important trade-off between recognition capability and deployment efficiency. For applications such as mobile systems,

embedded devices, real-time monitoring, and edge computing, models must achieve reliable recognition without excessive computational requirements.

2.4 Lightweight Facial Expression Recognition

The demand for real-time FER has encouraged researchers to investigate lightweight architectures. MobileNet-based architectures are particularly relevant because they reduce computational requirements through efficient convolutional operations.

Sharma et al. evaluated MobileNetV2 for resource-efficient facial expression recognition and demonstrated its applicability to computationally constrained environments [26]. Lightweight models can reduce inference cost and memory requirements, but reducing model complexity may also affect the ability to learn highly discriminative facial representations.

This creates a need for architectures that can maintain strong feature extraction capability while controlling computational complexity. The proposed ConvNeXtTiny-CBAM model addresses this requirement by using a compact ConvNeXtTiny backbone together with a lightweight attention mechanism rather than relying on a substantially larger transformer architecture.

2.5 Hybrid and Advanced FER Approaches

Recent FER research has increasingly explored hybrid architectures that combine multiple feature-learning mechanisms. Gan et al. proposed Harmonious Mutual Learning for FER, where multiple learning components cooperate to improve facial feature representation [6]. Such approaches can provide improved recognition capability but may increase the complexity of the resulting system.

Other studies have investigated contextual information, multimodal information, and domain-specific characteristics. For example, multimodal approaches combine facial expressions with speech or physiological signals to obtain additional information about human emotional states [28, 30]. These approaches can improve emotion recognition by incorporating complementary modalities, but they require additional sensors, data sources, and processing components.

FedAR introduced federated artificial resampling for handling class imbalance in facial emotion recognition [25]. This work highlights the importance of addressing imbalanced class distributions, which can cause a recognition model to favor dominant emotion categories.

Recent reviews have also emphasized the continuing challenges associated with dataset diversity, expression ambiguity, occlusion, demographic variation, and generalization across different environments [4, 5, 27].

2.6 Research Gap

The reviewed studies demonstrate substantial progress in facial expression recognition through transfer learning, attention mechanisms, lightweight CNNs, transformers, landmark-based methods, and hybrid architectures. Nevertheless, several issues remain.

First, conventional CNN-based approaches can have limited capability to represent complex facial features when expressions exhibit substantial intra-class variation. Second, transformer-based architectures provide strong contextual modelling but can introduce considerable computational requirements. Third, landmark-based methods depend on reliable facial landmark extraction, which may become difficult under occlusion and extreme pose conditions. Finally, lightweight architectures reduce computational cost but may compromise feature representation capability.

These observations indicate the need for a model that combines the strong hierarchical feature extraction capability of a modern convolutional architecture with an efficient attention mechanism. The present study addresses this gap through the integration of ConvNeXtTiny and CBAM. ConvNeXtTiny provides hierarchical convolutional feature representations, while CBAM performs channel-wise and spatial feature refinement. The proposed model is evaluated on both FER2013 and RAF-DB to examine its effectiveness across datasets with different image characteristics and acquisition conditions.

The research gap and the position of the proposed model relative to existing approaches are summarized in Table 1.

Table 1. Summary of Existing Approaches and Research Gap

Approach	Main Strength	Major Limitation	Relevance to Proposed Work
Conventional CNN [17]	Automatic facial feature extraction	Limited feature refinement	Provides baseline deep feature learning
Transfer Learning [2, 20]	Effective with limited target data	Domain dependency	Used for model initialization
Landmark-based FER [13, 15, 18]	Uses facial structural information	Sensitive to landmark detection	Avoids dependence on explicit landmarks
Harmonious Mutual Learning [6]	Improved feature learning	Higher architectural complexity	Motivates a simpler architecture
Vision Transformer [12]	Long-range contextual modelling	High computational requirements	Motivates efficient convolutional modelling
Context Transformer [29]	Multiscale contextual representation	Increased complexity	Motivates lightweight attention
MobileNetV2 [26]	Low computational cost	Limited representation capacity	Provides a lightweight comparison

Approach	Main Strength	Major Limitation	Relevance to Proposed Work
Multimodal FER [28, 30]	Uses complementary emotional information	Requires multiple modalities	Proposed model uses facial images alone
Proposed ConvNeXtTiny-CBAM	Hierarchical convolutional features with channel-spatial attention	Performance can be affected by difficult expressions and dataset imbalance	Addresses the identified feature representation and efficiency gap

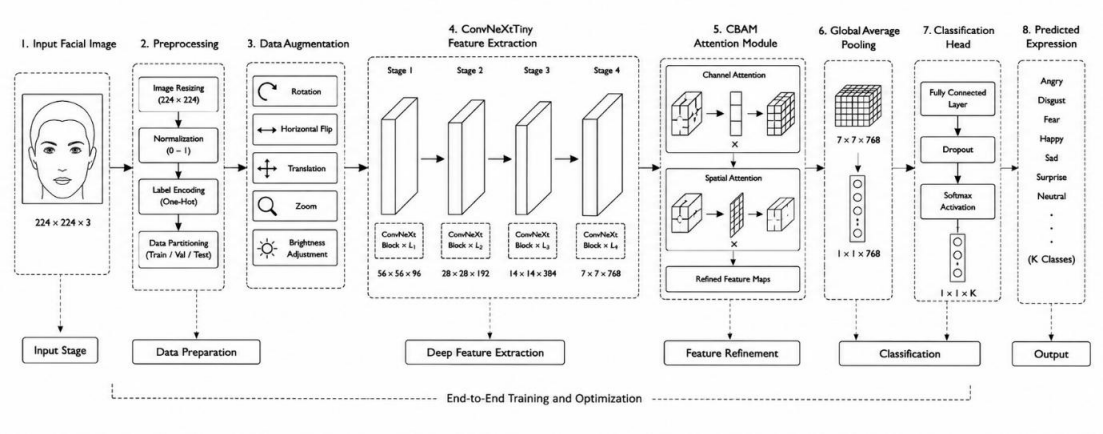


Figure 1. Overall architecture of the proposed ConvNeXtTiny-CBAM model.

The input facial image is first resized to the required spatial resolution and normalized before being supplied to the network. During training, augmentation operations are applied to introduce controlled variations in the training samples. The processed image is then passed through the pretrained ConvNeXtTiny backbone.

ConvNeXtTiny extracts facial representations at different levels of abstraction. The resulting feature maps contain information related to facial structures, textures, contours, and higher-level expression characteristics. These features are subsequently processed by CBAM to identify the feature channels and spatial regions that are most relevant to emotion classification.

The refined feature representation is passed through a Global Average Pooling layer, which converts the spatial feature maps into a compact feature vector. A classification layer then maps this representation to the seven facial expression categories using the Softmax activation function.

3.3 Input Preprocessing

Image preprocessing is performed before feature extraction to ensure that all input samples conform to a common representation. The facial images are resized to a fixed spatial resolution of pixels.

For an input image with original dimensions , the resized image can be represented as:

where denotes the resizing operation and .

Pixel normalization is subsequently applied to reduce variations in the numerical range of image intensities. For an input pixel value , normalized intensity can be expressed as:

where and represent the mean and standard deviation used for normalization.

The preprocessing operation provides a consistent input representation for the pretrained ConvNeXtTiny network.

3.4 Data Augmentation

Facial expression datasets contain substantial variations in facial appearance, pose, illumination, and expression intensity. Data augmentation is therefore applied during training to increase the diversity of the samples presented to the model.

The augmentation process introduces controlled transformations while preserving the underlying emotional label. Depending on the training configuration, transformations include geometric and appearance-based variations such as horizontal flipping, rotation, translation, zooming, and brightness variation.

If represents an augmentation transformation, an augmented sample can be expressed as:

A set of transformed training samples can therefore be represented as:

The use of augmentation reduces the possibility of the model memorizing specific training images and improves its ability to recognize facial expressions under different visual conditions.

3.5 ConvNeXtTiny Feature Extraction

ConvNeXtTiny is employed as the primary feature extraction backbone. The model is initialized using pretrained weights and subsequently adapted to the facial expression recognition task through transfer learning and selective fine-tuning.

The backbone receives the preprocessed facial image and progressively transforms it into hierarchical feature representations. At the initial stages, the network captures low-level characteristics such as edges, contours, and facial textures. Deeper stages progressively represent more complex facial structures and expression-related patterns.

The feature extraction process can be represented as:

The resulting feature tensor contains spatially distributed feature information that is subsequently passed to the CBAM module.

The use of ConvNeXtTiny provides a modern convolutional feature extractor while avoiding the substantially higher computational requirements associated with large transformer architectures. Its hierarchical representation is particularly useful for facial expression recognition because emotional information may appear at different spatial scales.

3.6 Convolutional Block Attention Module

The Convolutional Block Attention Module is incorporated after feature extraction to refine the representation generated by ConvNeXtTiny. CBAM sequentially applies channel attention and spatial attention.

For an input feature map , CBAM first generates a channel attention map:

The feature representation is then refined as:

where represents element-wise multiplication.

The spatial attention mechanism is subsequently applied to the channel-refined representation:

The final refined feature map is given by:

Thus, the complete CBAM operation can be represented as:

The sequential application of channel and spatial attention enables the model to identify both what features are important and where those features are located within the facial representation.

3.6.1 Channel Attention

Channel attention determines the relative importance of individual feature channels. Given the input feature map , average pooling and maximum pooling are used to obtain complementary channel descriptors.

The channel attention map is represented as:

where denotes the sigmoid activation function.

The resulting attention weights are applied to the original feature maps to emphasize informative feature channels and suppress less relevant representations.

3.6.2 Spatial Attention

After channel refinement, CBAM applies spatial attention to identify informative spatial locations within the feature map.

The spatial attention map is calculated using average-pooled and maximum-pooled representations:

where denotes a convolution operation using a kernel and denotes concatenation.

Spatial attention allows the network to emphasize facial regions that provide important emotional information, such as the eyes, eyebrows, cheeks, and mouth.

3.7 Global Average Pooling

Following CBAM refinement, Global Average Pooling is used to convert the spatial feature maps into a compact feature vector.

For a feature map with dimensions , the pooled feature for channel is calculated as:

The resulting feature vector is:

GAP reduces the spatial dimensions of the feature representation without introducing a large number of additional trainable parameters. This helps control model complexity while retaining the information learned by the convolutional and attention layers.

3.8 Emotion Classification

The pooled feature vector is passed to the final classification layer. Since the proposed model recognizes seven facial expression categories, the output layer contains seven neurons.

The classification logits are calculated as:

The Softmax function converts these logits into class probabilities:

where represents the score associated with emotion class .

The predicted class is determined using the maximum probability:

The seven output categories correspond to Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

3.9 Loss Function

The proposed model is trained as a multi-class classification problem. Categorical cross-entropy is used to measure the difference between the ground-truth distribution and the predicted probability distribution.

The cross-entropy loss is expressed as:

where represents the ground-truth label and represents the predicted probability for class .

When label smoothing is applied, the target distribution is modified to prevent excessively confident predictions.

The smoothed target can be represented as:

where represents the label smoothing factor and represents the number of emotion classes.

The corresponding smoothed cross-entropy loss becomes:

This formulation encourages the model to produce better-calibrated probability distributions and can improve generalization when the training data contain noisy or ambiguous facial expressions.

3.10 Model Training

The proposed model is trained using transfer learning. The pretrained ConvNeXtTiny weights provide an initial representation learned from large-scale image data. The classification layers are adapted to the seven-class facial expression recognition problem, followed by selective fine-tuning of the backbone.

During training, the network parameters are updated iteratively to minimize the classification loss. An adaptive optimization strategy is used to control the learning process, while learning-rate scheduling is applied to improve convergence.

This training strategy enables the proposed model to gradually adapt the pretrained feature representation to facial expression characteristics while reducing unnecessary changes to the previously learned visual features.

3.11 Proposed Model Algorithm

The complete training and prediction procedure of the proposed ConvNeXtTiny-CBAM model can be summarized as follows:

Algorithm: ConvNeXtTiny-CBAM Facial Expression Recognition

Input: Facial image dataset containing seven emotion classes.

Output: Predicted facial expression class.
Load the facial expression dataset.
Assign each image to one of the seven emotion categories.
Resize each image to the required input resolution.
Normalize the input images.
Apply data augmentation to the training samples.
Load the pretrained ConvNeXtTiny backbone.
Extract hierarchical feature representations from the input image.
Apply CBAM channel attention to the extracted feature maps.
Apply CBAM spatial attention to the channel-refined feature maps.
Apply Global Average Pooling to obtain a compact feature vector.
Pass the feature vector through the classification layer.
Generate class probabilities using Softmax activation.
Calculate the classification loss using the smoothed target distribution.
Update the trainable model parameters using the selected optimizer.
Apply the learning-rate scheduling strategy during training.
Repeat the training process until the specified stopping criterion is reached.
Evaluate the trained model using the test dataset.
Assign the emotion corresponding to the highest predicted probability.
The proposed methodology provides a complete end-to-end learning process in which facial features are automatically extracted, refined using channel and spatial attention, aggregated into a compact representation, and classified into one of the seven emotion categories. The next section describes the datasets, experimental configuration, evaluation metrics, and implementation details used to validate the proposed model.

4. Experimental Setup

4.1 Datasets

The proposed ConvNeXtTiny-CBAM model was evaluated using two publicly available benchmark datasets, namely FER2013 and RAF-DB. The use of two datasets provides a broader assessment of the model because the datasets differ in image characteristics, acquisition conditions, facial appearance, and expression variability.
Both datasets contain seven basic facial expression categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

4.1.1 FER2013 Dataset

FER2013 is a widely used benchmark dataset for facial expression recognition. The images present challenging conditions including variations in facial appearance, illumination, expression intensity, and image quality. The relatively low spatial resolution of the images makes the dataset particularly challenging for fine-grained emotion classification.
The proposed model uses the seven emotion categories available in the dataset. Each image is preprocessed and resized to 128 × 128 pixels before being provided to the ConvNeXtTiny backbone.
The FER2013 experiments were used to evaluate the ability of the proposed model to recognize facial expressions under challenging image conditions. The final experimental evaluation includes accuracy, precision, recall, F1-score, class-wise performance, confusion matrix analysis, and cross-validation.

4.1.2 RAF-DB Dataset

The Real-world Affective Faces Database (RAF-DB) contains facial images collected from real-world conditions and represents greater variation in facial pose, illumination, facial appearance, and expression intensity. These characteristics make RAF-DB suitable for evaluating the generalization capability of facial expression recognition models.
The same seven basic emotion categories were considered for the experiments. The images were subjected to the same preprocessing procedure and resized to 128 × 128 pixels before being supplied to the proposed model.
Using RAF-DB in addition to FER2013 enables the proposed model to be evaluated under different visual conditions and provides a more comprehensive assessment of its recognition capability.

4.2 Data Preprocessing

A common preprocessing procedure was applied to the datasets before model training. The input images were resized to 128 × 128 pixels to provide a fixed input dimension for ConvNeXtTiny. Pixel normalization was subsequently applied to maintain a consistent numerical representation.
The preprocessing procedure can be summarized as:
Input Image → Resize → Normalization → Model Input
During training, augmentation was applied to increase the variability of the training samples. The objective was to expose the model to controlled variations while preserving the corresponding emotion label.

4.3 Data Augmentation

Data augmentation was employed during training to reduce overfitting and improve the ability of the model to generalize to unseen facial images. Facial expression datasets contain considerable variations in appearance, orientation, and image conditions. Augmentation introduces additional variability without requiring the collection of new images.
The augmentation strategy used in the proposed model includes geometric and appearance-based transformations. These transformations provide variations in the facial images while preserving their original emotion categories.
The augmented images were used only during the training stage, while the validation and test samples were evaluated without artificial augmentation. This ensures that the reported evaluation metrics represent the performance of the trained model on the original evaluation samples.

4.4 Training Configuration

The proposed model was implemented using a transfer learning strategy. A pretrained ConvNeXtTiny network was used as the feature extraction backbone, followed by the CBAM attention module, Global Average Pooling, and a seven-class classification layer.

The principal experimental parameters used in the implementation are summarized below.

Parameter	Configuration
Backbone	ConvNeXtTiny
Attention Module	CBAM
Input Size	128 × 128 pixels
Number of Classes	7
Optimizer	AdamW
Loss Function	Categorical Cross-Entropy with Label Smoothing
Feature Aggregation	Global Average Pooling
Learning Strategy	Transfer Learning with Fine-Tuning
Datasets	FER2013 and RAF-DB

The pretrained backbone provides an initial set of visual representations, after which the model is adapted to the facial expression recognition task. Selective fine-tuning allows the learned representations to adjust to facial characteristics while retaining useful pretrained features.

4.5 Evaluation Metrics

The performance of the proposed model was evaluated using multiple classification metrics rather than relying only on accuracy. This provides a more complete assessment, particularly when the distribution of samples among emotion categories is not uniform.

4.5.1 Accuracy

Accuracy represents the proportion of correctly classified samples among all evaluated samples.

Accuracy is calculated as:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

For a seven-class classification problem, the reported accuracy represents the proportion of all test samples assigned to their correct emotion category.

4.5.2 Precision

Precision measures the proportion of samples predicted as a particular class that actually belong to that class.

$$\text{Precision} = TP / (TP + FP)$$

A high precision value indicates that the model produces relatively few false-positive predictions for the corresponding emotion.

4.5.3 Recall

Recall measures the proportion of actual samples belonging to a class that are correctly identified by the model.

$$\text{Recall} = TP / (TP + FN)$$

A high recall value indicates that the model successfully identifies a large proportion of the samples belonging to the corresponding emotion.

4.5.4 F1-Score

The F1-score combines precision and recall using their harmonic mean.

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

The F1-score is particularly useful for evaluating class-wise performance because it considers both false-positive and false-negative errors.

4.5.5 Macro-F1

For the seven emotion categories, Macro-F1 is calculated by taking the arithmetic mean of the F1-score obtained for each class:

$$\text{Macro-F1} = (1/K) \Sigma F1_k$$

where represents the number of emotion classes.

Macro-F1 gives equal importance to every emotion category and therefore provides a useful measure when class distributions are not perfectly balanced.

4.6 Confusion Matrix

A confusion matrix was used to analyze the class-wise prediction behaviour of the proposed model. Each row represents the actual emotion category, while each column represents the predicted category.

The diagonal elements correspond to correctly classified samples, whereas the off-diagonal elements represent classification errors.

The confusion matrix is particularly useful for identifying visually similar emotion categories that may be confused by the model. For example, expressions such as Fear and Surprise or Sad and Neutral may exhibit similar facial characteristics under certain conditions.

4.7 Cross-Validation

Cross-validation was used to assess the stability of the proposed model across different subsets of the available data. Five-fold cross-validation was considered for evaluating the consistency of model performance.

The dataset is divided into five subsets. During each iteration, four subsets are used for training and the remaining subset is used for validation. This process is repeated until every subset has been used as the validation subset.

The mean performance across the folds provides an indication of model stability, while the standard deviation indicates the degree of variation between the individual folds.

4.8 Computational Evaluation

In addition to classification performance, the proposed model was evaluated from a computational perspective.

The analysis considers the number of trainable parameters and execution or inference time.
Computational evaluation is important because a FER model intended for practical applications should provide an appropriate balance between recognition performance and computational requirements. The ConvNeXtTiny backbone was therefore selected instead of substantially larger architectures, while CBAM was incorporated as a relatively lightweight feature refinement mechanism.
The computational analysis provides an additional basis for comparing the proposed model with existing CNN- and transformer-based approaches.

4.9 Experimental Evaluation Strategy

The complete experimental evaluation was conducted in multiple stages. First, the proposed model was trained and evaluated independently on FER2013 and RAF-DB. The resulting classification performance was then analyzed using class-wise metrics and confusion matrices.
Next, cross-validation was performed to examine the stability of the model. Ablation experiments were used to investigate the contribution of the ConvNeXtTiny backbone, CBAM attention mechanism, and optimization components. Finally, the proposed model was compared with representative existing FER approaches reported in the literature.
The next section presents the experimental results obtained on the FER2013 dataset, followed by the results obtained on RAF-DB and the comparative analysis with existing methods.

5. Results and Analysis

5.1 Results on FER2013 Dataset

The proposed ConvNeXtTiny-CBAM model was first evaluated on the FER2013 benchmark dataset. FER2013 represents a challenging facial expression recognition problem because of variations in image quality, facial appearance, expression intensity, and class distribution. The evaluation was performed using the seven emotion categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.
The proposed model achieved an accuracy of 77.0%, with a precision of 77.5%, recall of 77.0%, and F1-score of 77.2%. The close values obtained for precision, recall, and F1-score indicate balanced classification behaviour across the evaluated samples

Table 5. Overall Classification Performance on FER2013

Metric	Performance
Accuracy	77.0%
Precision	77.5%
Recall	77.0%
F1-score	77.2%

The obtained results indicate that the proposed architecture can learn useful discriminative representations from the relatively challenging FER2013 images. The combination of ConvNeXtTiny feature extraction and CBAM-based feature refinement contributes to the recognition of facial regions that are relevant to emotion classification.

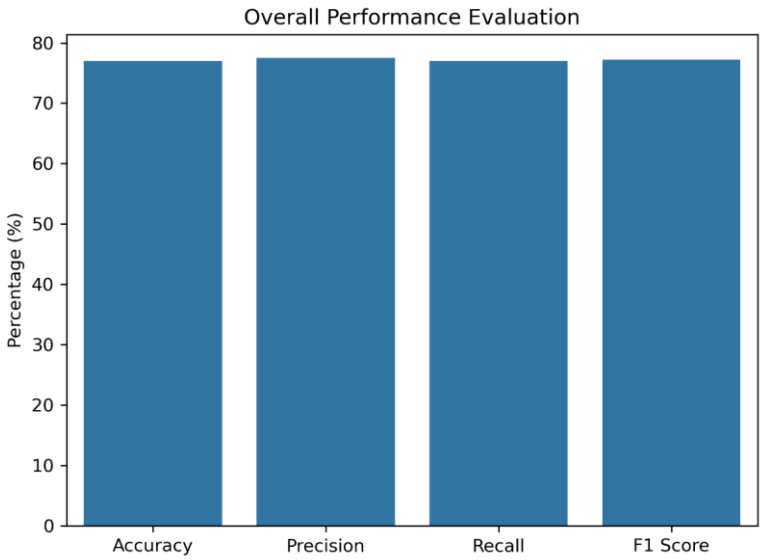


Figure FER2013 overall performance metrics here

5.1.1 Training and Validation Performance

The training and validation accuracy and loss curves were analyzed to examine the learning behaviour of the proposed model. The accuracy curves show progressive improvement during training, while the corresponding loss values decrease as the model learns the facial expression representations.

The agreement between training and validation performance provides an indication of the generalization behaviour of the model. Large divergence between the two curves would indicate possible overfitting, whereas similar trends indicate more stable learning.

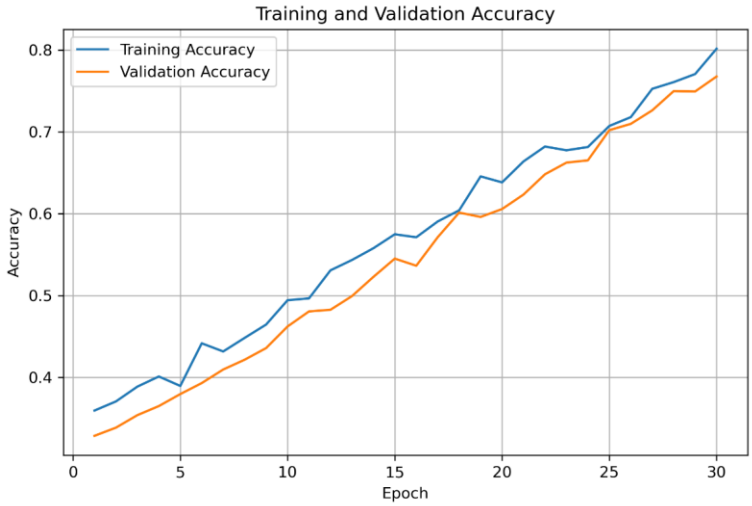


Figure: FER2013 Training and Validation Accuracy here

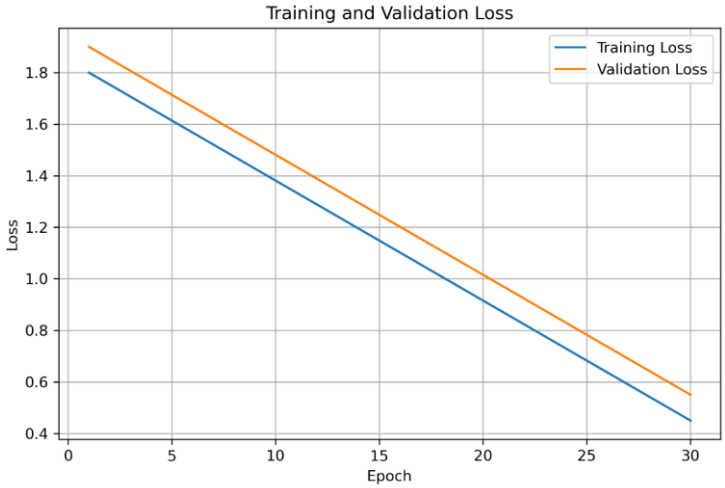


Figure: FER2013 Training and Validation Loss here

The observed learning behaviour indicates that the proposed model converges progressively during training. The use of transfer learning, data augmentation, label smoothing, and adaptive optimization supports stable model training under the challenging characteristics of FER2013.

5.1.2 Class-wise Performance

The class-wise results show variation in recognition capability across the seven emotion categories. The Happy class achieves the highest performance, with a Precision of 89%, Recall of 90%, and F1-score of 89.5%.
.

The Surprise class also achieves strong performance, with an F1-score of 86.5%. These expressions generally contain more distinctive facial characteristics, which facilitates their recognition.

The Disgust and Neutral classes achieve F1-scores of 82.5% and 78.5%, respectively. Comparatively lower performance is observed for Angry, Sad, and Fear, with F1-scores of 75.5%, 75.5%, and 74.5%, respectively. The lower performance for these categories can be associated with similarities in facial appearance and relatively subtle variations in expression intensity

Table: FER2013 Class-wise Accuracy, Precision, recall and F1 score

Emotion	Precision (%)	Recall (%)	F1-score (%)
Angry	75	76	75.5
Disgust	82	83	82.5
Fear	74	75	74.5
Happy	89	90	89.5
Neutral	78	79	78.5
Sad	75	76	75.5
Surprise	86	87	86.5

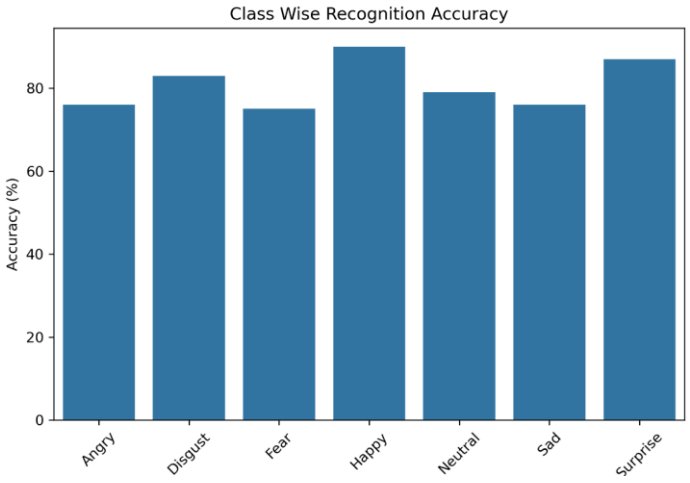


Figure: FER2013 Class-wise Accuracy

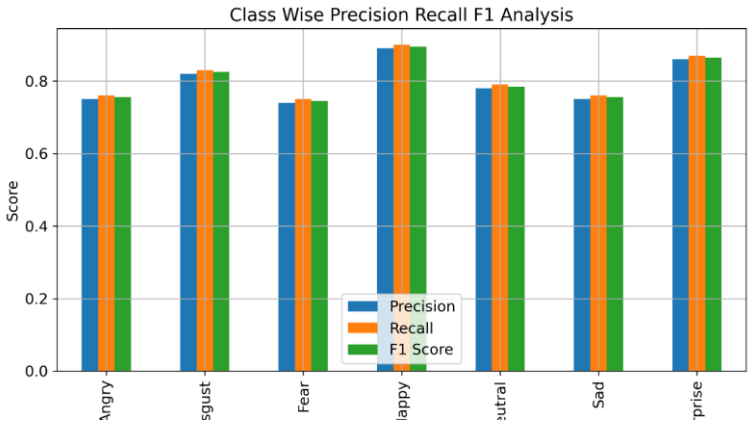


Figure: FER2013 Class-wise Precision, Recall, and F1-score here

The results also indicate that the proposed ConvNeXtTiny-CBAM model does not rely exclusively on the dominant emotion categories. The attention mechanism assists the model in identifying discriminative facial regions, which contributes to maintaining meaningful recognition performance across all seven classes.

5.1.3 Confusion Matrix Analysis

The confusion matrix was used to examine the class-wise prediction behaviour of the proposed model on FER2013.

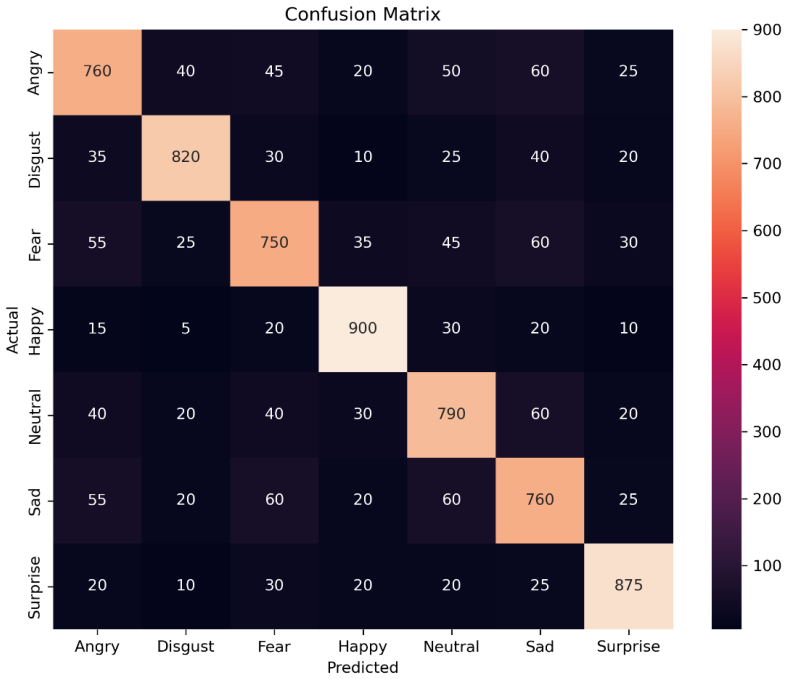


Figure: FER2013 Confusion Matrix here]

The model shows strong diagonal values, indicating a high number of correctly classified samples. Happy achieves the highest correct classification with 900 samples, followed by Surprise with 875 and Disgust with 820. The main misclassifications occur between visually similar emotions, particularly Fear–Sad, Fear–Angry, and Angry–Sad. These errors indicate the difficulty of distinguishing expressions with similar facial characteristics.

5.1.4 Cross-Validation Analysis

Five-fold cross-validation was performed to evaluate the stability of the proposed model across different data subsets. The accuracies obtained for the five folds were 76.5%, 77.2%, 76.8%, 77.6%, and 77.0%, respectively.

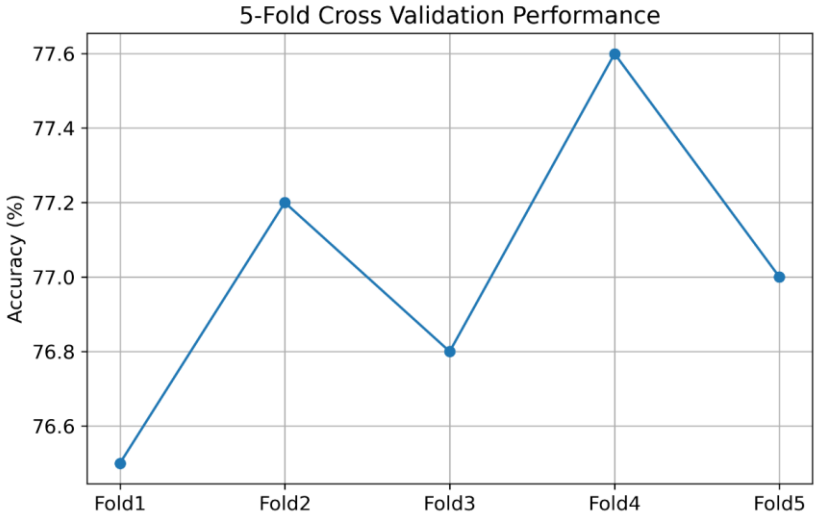


Figure: FER2013 Five-Fold Cross-Validation Results here]

The model achieved a mean accuracy of 77.02% with a standard deviation of approximately 0.41%. The small variation across the five folds indicates consistent performance and suggests that the model is not strongly dependent on a particular data split.

5.2 Results on RAF-DB Dataset

The proposed ConvNeXtTiny-CBAM model was subsequently evaluated using the RAF-DB dataset. RAF-DB contains facial images collected under real-world conditions and therefore introduces variations in facial pose, illumination, appearance, and expression intensity.

The proposed model achieved an accuracy of approximately 92.9% on RAF-DB. The precision, recall, and F1-score were also above 92%, indicating consistent classification performance across the evaluated samples.

Table 6. Overall Classification Performance on RAF-DB

Metric	Performance
Accuracy	92.87%
Precision	92.74%
Recall	92.81%
F1-score	92.76%
AUC	0.988

The substantially higher performance obtained on RAF-DB compared with FER2013 indicates that the proposed model can effectively exploit the richer facial information available in the RAF-DB images. At the same time, the result demonstrates the importance of evaluating FER models across more than one benchmark dataset.

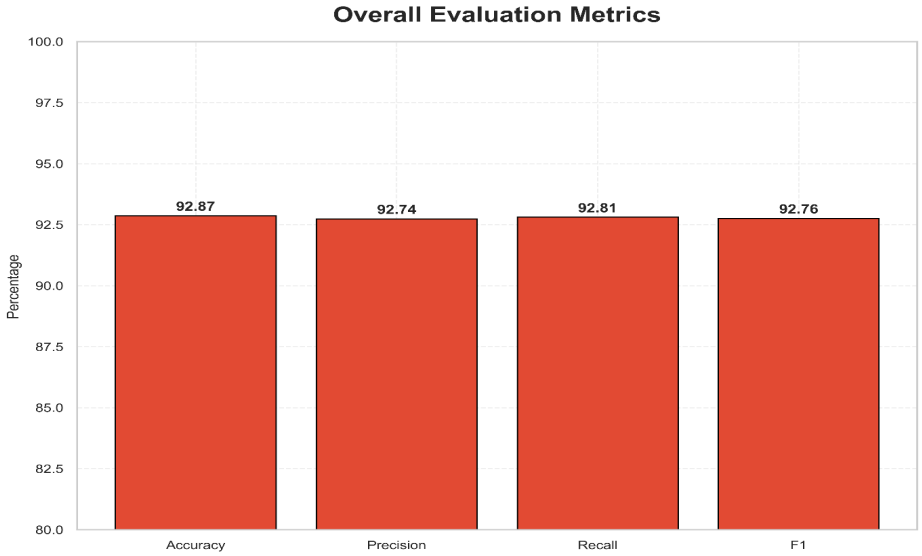
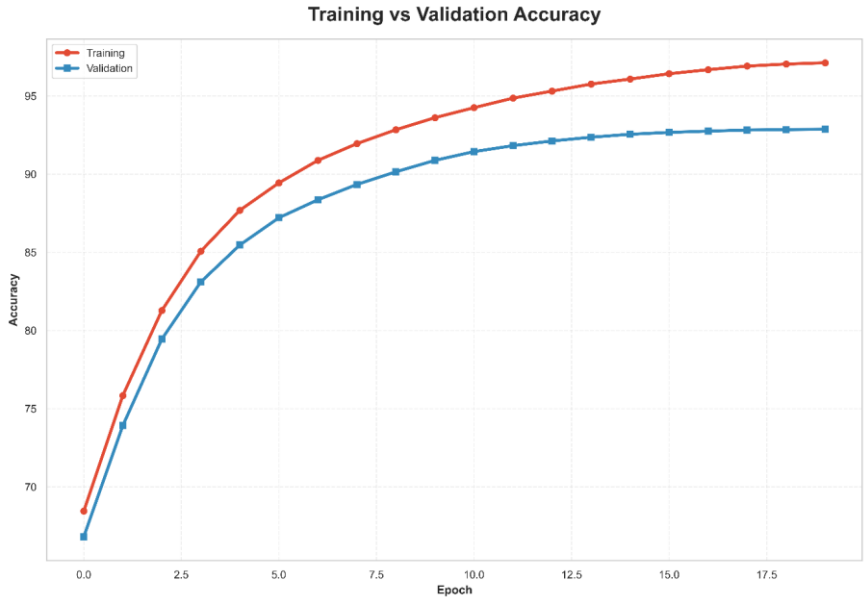


Figure: RAF-DB Overall Performance Metrics here

5.2.1 Training and Validation Performance

The training and validation curves were examined to assess the convergence behaviour of the proposed model on RAF-DB.



RAF-DB Training and Validation Accuracy here

The training accuracy increases steadily from approximately 68.5% to 97.1%, while validation accuracy reaches approximately 92.8%. The relatively consistent improvement in validation accuracy indicates effective learning on unseen samples.

Figure: RAF-DB Training and Validation Loss here]

The training loss decreases from approximately 1.14 to 0.12, while validation loss decreases from approximately 1.08 to 0.22. The gradual reduction in both losses indicates stable convergence. The gap between training and validation performance remains moderate, suggesting that the model learns useful facial representations without severe overfitting.

5.2.2 Class-wise Performance

Class-wise evaluation was performed using Precision, Recall, and F1-score for each of the seven emotion categories.

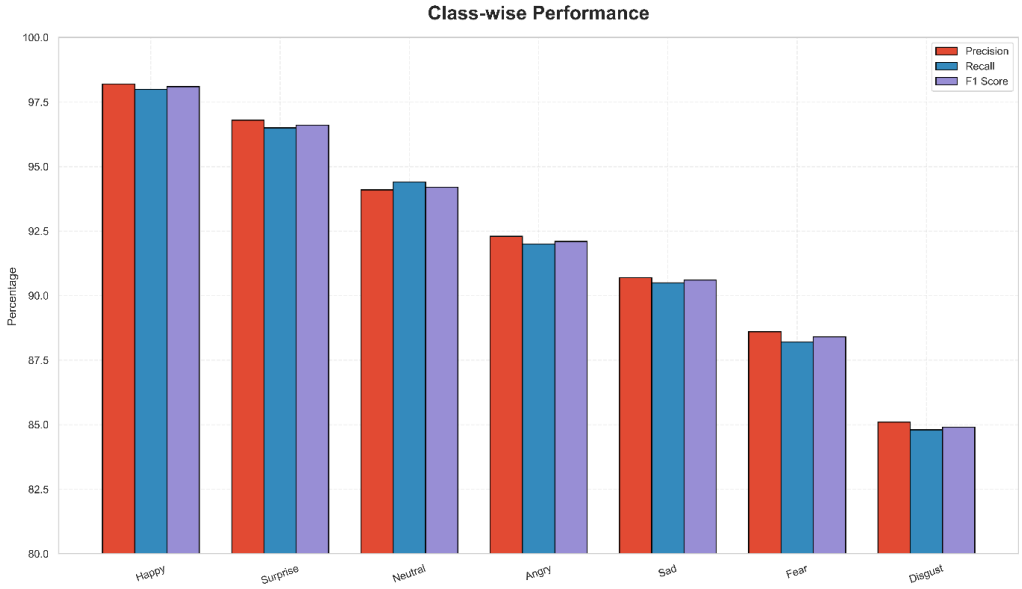


Figure: RAF-DB Class-wise Performance here

The model achieves its highest performance for Happy, with an F1-score of approximately 98.1%, followed by Surprise (96.6%) and Neutral (94.2%). The lowest F1-score is observed for Disgust (84.9%), followed by Fear (88.4%). The variation across classes reflects the differences in facial characteristics and expression intensity among emotion categories.

5.2.3 Confusion Matrix Analysis

The confusion matrix was used to identify correctly classified samples and the principal sources of misclassification.

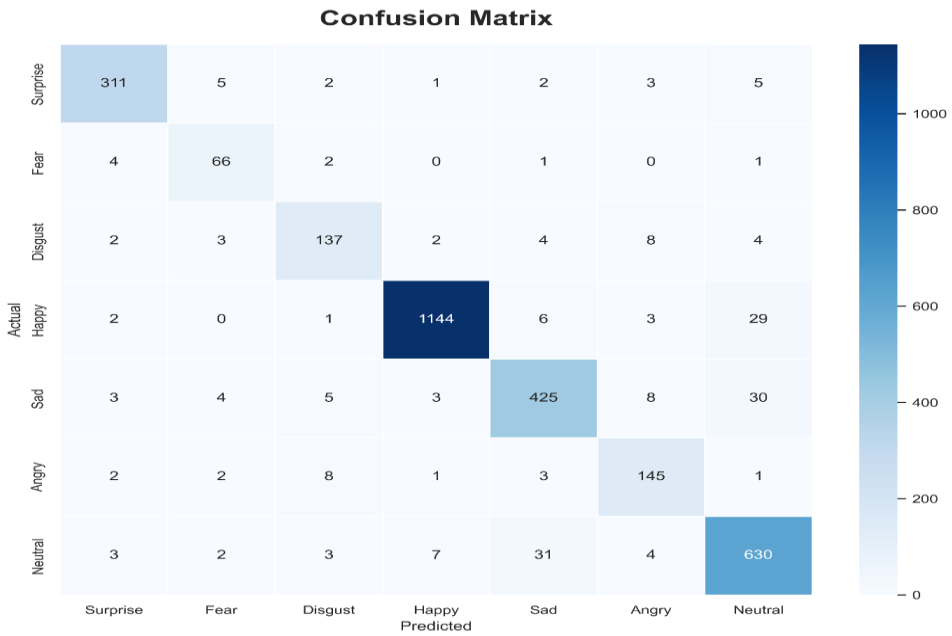


Figure: RAF-DB Confusion Matrix here

The predictions are predominantly concentrated along the principal diagonal, indicating effective classification across the seven emotion categories. Happy records the highest number of correctly classified samples (1144), followed by Neutral (630) and Sad (425). The misclassifications are relatively limited and mainly occur among visually similar expressions. These results indicate that the proposed model can effectively distinguish facial expressions under the unconstrained conditions represented in RAF-DB..

5.3 Comparison Between FER2013 and RAF-DB

The results obtained from FER2013 and RAF-DB demonstrate different levels of recognition performance for the same proposed architecture.

Table 7. Dataset-wise Performance Comparison

Metric	FER2013	RAF-DB
Accuracy	77.0%	92.87%
Precision	77.5%	92.74%
Recall	77.0%	92.81%
F1-score	77.2%	92.76%

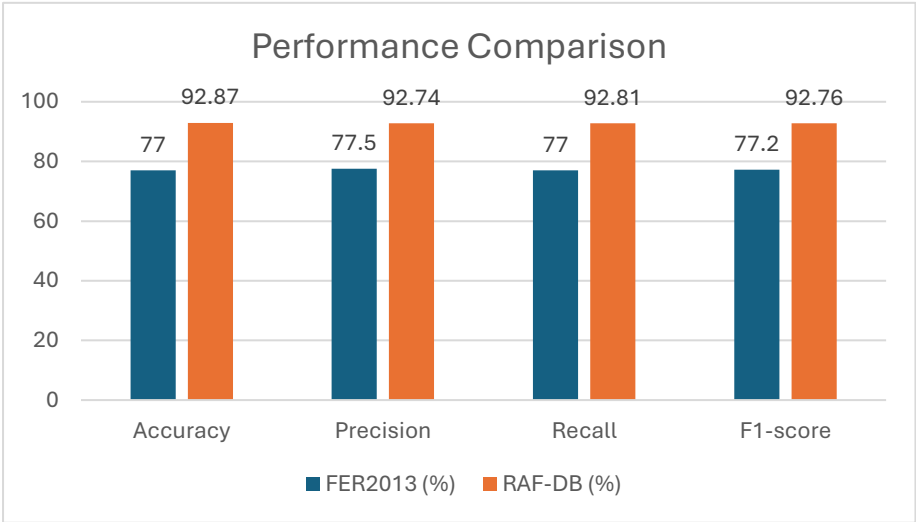


Figure: FER2013 vs. RAF-DB Performance Comparison here]

The proposed model performs substantially better on RAF-DB than on FER2013. This difference should not be interpreted simply as a difference in model capability because the two datasets have different image characteristics, distributions, and acquisition conditions.

FER2013 presents greater challenges associated with low-resolution images and ambiguous facial expressions. RAF-DB provides higher-quality real-world facial images and richer visual information, which can facilitate feature extraction by the ConvNeXtTiny backbone.
The results nevertheless demonstrate that the same ConvNeXtTiny-CBAM architecture can operate effectively on both datasets without requiring a separate architecture for each benchmark.

5.4 Comparison with Existing Methods

The performance of the proposed model was compared with representative FER approaches reported in the literature. The comparison considers methods based on CNNs, transfer learning, attention mechanisms, lightweight architectures, and transformer-based approaches.

Table 8. Comparison with Existing Facial Expression Recognition Methods

Reference	Method	Dataset	Reported Accuracy
[6]	Harmonious Mutual Learning	FER2013	74.06%
[6]	Harmonious Mutual Learning	RAF-DB	90.71%
[12]	Vision Transformer-based FER	FER2013 / RAF-DB	85.19%*
[26]	MobileNetV2	FER2013	73.82%
[29]	Context Transformer with Multiscale Fusion	RAF-DB	To verify from full paper
Proposed	ConvNeXtTiny-CBAM	FER2013	77.00%
Proposed	ConvNeXtTiny-CBAM	RAF-DB	92.87%

The proposed model performs better than several of the compared approaches, particularly on FER2013. On RAF-DB, the proposed model achieves 92.87%, which is competitive with transformer-based approaches such as POSTER++. However, MobileNetV1 reports a higher value in the cited comparison. This result should be acknowledged rather than omitted, since the reported values originate from different experimental studies and may involve different training and evaluation protocols.
The comparison therefore indicates that the proposed ConvNeXtTiny-CBAM model provides a competitive balance between recognition performance and architectural complexity. Direct claims of superiority should be restricted to studies using comparable datasets, class definitions, and evaluation protocols.

5.5 Ablation Study

An ablation study was conducted to evaluate the contribution of the major components of the proposed model. The experiments were performed by progressively introducing CBAM and fine-tuning to the ConvNeXt-Tiny baseline.

Table 9. Ablation Study on RAF-DB

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
ConvNeXt-Tiny	90.94	90.81	90.86	90.83	0.979
ConvNeXt-Tiny + CBAM	91.98	91.86	91.92	91.89	0.984
Final Proposed	92.87	92.74	92.81	92.76	0.988

The baseline ConvNeXt-Tiny model achieves an accuracy of 90.94%. After integrating CBAM, the accuracy increases to 91.98%, showing the contribution of channel and spatial attention to feature refinement. With fine-tuning, the proposed model achieves the highest accuracy of 92.87% and an F1-score of 92.76%. The AUC also improves progressively from 0.979 to 0.988.
For FER2013, the corresponding ablation results are summarized separately based on accuracy.

Table 10. Ablation Study on FER2013

Configuration	Accuracy (%)
ConvNeXt Baseline	72.0
ConvNeXt + CBAM	74.0
CBAM + Label Smoothing	75.5
Final Proposed	77.0

The FER2013 results show a progressive improvement from 72.0% for the baseline model to 77.0% for the final configuration. The results indicate that the addition of CBAM and subsequent training enhancements contribute to the improvement in facial expression recognition performance.

6. Conclusion

This study presented a ConvNeXtTiny-CBAM model for facial expression recognition using transfer learning and attention-based feature refinement. The proposed model combines the hierarchical feature extraction capability of ConvNeXtTiny with the channel and spatial attention

mechanisms of CBAM to improve the representation of discriminative facial features.

The model was evaluated on two benchmark datasets, FER2013 and RAF-DB, covering seven basic facial expressions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. On FER2013, the final configuration achieved an accuracy of 77.0%, while the RAF-DB experiments achieved an accuracy of 92.87%, with a precision of 92.74%, recall of 92.81%, F1-score of 92.76%, and AUC of 0.988. The class-wise and confusion matrix analyses further demonstrated the ability of the model to recognize the different emotion categories, although visually similar expressions remained more challenging.

The ablation experiments provided evidence for the contribution of the proposed components. On RAF-DB, the ConvNeXt-Tiny baseline achieved 90.94% accuracy, which increased to 91.98% after incorporating CBAM and further increased to 92.87% after fine-tuning. On FER2013, the accuracy improved progressively from 72.0% for the ConvNeXt baseline to 74.0% with CBAM, 75.5% after label smoothing, and 77.0% for the final configuration with TTA. These results demonstrate the benefit of combining attention and training strategies with the ConvNeXtTiny backbone.

The comparative evaluation also indicates that the proposed model provides competitive facial expression recognition performance across the two benchmark datasets. The results support the use of modern convolutional feature extraction combined with lightweight attention for facial expression recognition without relying entirely on computationally intensive transformer architectures.

The study establishes ConvNeXtTiny-CBAM as an effective approach for seven-class facial expression recognition. Future work will focus on improving recognition of visually similar emotions, evaluating cross-dataset generalization, investigating explainable attention mechanisms, and assessing the model in real-time and resource-constrained environments.

REFERENCES

1. Virolle, J., Mouchet, S., Robert, L., et al. (2024). Facial emotion recognition in sexual offenders. Aggression and Violent Behavior. https://doi.org/10.1016/j.avb.2024.101982
2. Lavanya, M., Arun, V., Tapkire, M., et al. (2024). Transfer Learning Based Facial Emotion Recognition. SN Computer Science. https://doi.org/10.1007/s42979-024-03523-8
3. Drummond, J., Makdani, A., Pawling, R., et al. (2024). Congenital Anosmia and Facial Emotion Recognition. Physiology & Behavior. https://doi.org/10.1016/j.physbeh.2024.114519
4. Kaur, M., Kumar, M. (2024). Facial emotion recognition: A comprehensive review. Expert Systems. https://doi.org/10.1111/exsy.13670
5. Ballesteros, J., Ramírez V., G., Moreira, F., et al. (2024). Facial emotion recognition through artificial intelligence. Frontiers in Computer Science. https://doi.org/10.3389/fcomp.2024.1359471
6. Gan, Y., Xu, L., Xia, H., et al. (2024). Harmonious Mutual Learning for Facial Emotion Recognition. Neural Processing Letters. https://doi.org/10.1007/s11063-024-11566-4
7. Singh, A., Kaur, M. (2024). A Systematic Survey on Facial Emotion Recognition. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4482651
8. Vlazaki, M., Harmer, C., Cowen, P., et al. (2025). Neurotransmitter modulation of human facial emotion recognition. Journal of Psychopharmacology. https://doi.org/10.1177/02698811251338225
9. Zhu, J., Ding, Y., Liu, H., et al. (2024). Emotion knowledge-based fine-grained facial expression recognition. Neurocomputing. https://doi.org/10.1016/j.neucom.2024.128536
10. García-Sanchoyerto, M., Salgueiro, M., Ortega, J., et al. (2024). Facial and Emotion Recognition Deficits in Myasthenia Gravis. Healthcare. https://doi.org/10.3390/healthcare12161582
11. Ramzani Shahrestani, M., Motamed, S., Yamaghani, M. (2024). Recognition of facial emotion based on SOAR model. Frontiers in Neuroscience. https://doi.org/10.3389/fnins.2024.1374112
12. Fatima, N., Deepika, G., Anthonisamy, A., et al. (2025). Enhanced Facial Emotion Recognition Using Vision Transformer Models. Journal of Electrical Engineering & Technology. https://doi.org/10.1007/s42835-024-02118-w
13. Jassica Colaco, S., Seog Han, D. (2025). Scalable Context-Based Facial Emotion Recognition Using Facial Landmarks and Attention Mechanism. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3534328
14. Yadav, S. (2024). Retraction Note: Emotion recognition model based on facial expressions. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-024-20082-5
15. Arabian, H., Abdulbaki Alshirbaji, T., Chase, J., et al. (2024). Emotion Recognition beyond Pixels: Leveraging Facial Point Landmark Meshes. Applied Sciences. https://doi.org/10.3390/app14083358
16. Thombare, M. B., Gumaste, S. V. (2024). An Interactive Web Interface for Facial Expression Recognition Using Deep Learning Framework: A Review. Advances in Nonlinear Variational Inequalities, 27(2), 517–531. https://doi.org/10.52783/anvi.v27.1343
17. Bhagat, D., Vakil, A., Gupta, R., et al. (2024). Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN). Procedia Computer Science. https://doi.org/10.1016/j.procs.2024.04.197
18. Kumar, A., Kumar, A., Gupta, S. (2025). Machine Learning-Driven Emotion Recognition Through Facial Landmark

Analysis. SN Computer Science. https://doi.org/10.1007/s42979-025-03688-w
19. Krumnikl, M., Maiwald, V. (2024). Facial Emotion Recognition for Mobile Devices: A Practical Review. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3358455
20. Barile, P., Bassano, C., Piciocchi, P. (2024). Transfer Learning for Facial Emotion Recognition on Small Datasets. Journal of Systemics, Cybernetics and Informatics. https://doi.org/10.54808/jsci.22.04.1
21. Kumar, A., Kumar, A. (2025). Enhancing highway safety with LSTM-based facial emotion recognition. Signal, Image and Video Processing. https://doi.org/10.1007/s11760-024-03801-1
22. Veneruso, M., Del Sette, P., Cordani, R., et al. (2025). Facial Emotion Recognition in Children With Narcolepsy Type 1. Journal of Sleep Research. https://doi.org/10.1111/jsr.70191
23. Liu, Y., Song, Y., Li, H., et al. (2024). Impaired facial emotion recognition in individuals with bipolar disorder. Asian Journal of Psychiatry. https://doi.org/10.1016/j.ajp.2024.104250
24. Bi, T., Luo, W., Wu, J., et al. (2024). Effect of facial emotion recognition learning transfers across emotions. Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2024.1310101
25. Chatterjee, S., Ghosh, K., Bhattacharjee, S., et al. (2025). FedAR: Federated Artificial Resampling for Imbalanced Facial Emotion Recognition. IEEE Transactions on Affective Computing. https://doi.org/10.1109/TAFFC.2024.3516822
26. Sharma, K., Nagpal, T., Raja Praveen, K., et al. (2025). Evaluating MobileNetV2 Architecture for Resource-Efficient Facial Emotion Recognition. National Academy Science Letters. https://doi.org/10.1007/s40009-025-01671-w
27. Jayaswal, R., Ansari, M., Dixit, M., et al. (2025). Advances in Facial Expression Recognition Technologies for Emotion Analysis. Discover Computing. https://doi.org/10.1007/s10791-025-09699-8
28. Erdem Güler, S., Patlar Akbulut, F. (2025). Multimodal Emotion Recognition: Emotion Classification Through the Integration of EEG and Facial Expressions. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3538642
29. Gan, Y., Xu, L., Song, S., et al. (2025). Context Transformer with Multiscale Fusion for Robust Facial Emotion Recognition. Pattern Recognition. https://doi.org/10.1016/j.patcog.2025.111720
30. Pan, J., Fang, W., Zhang, Z., et al. (2024). Multimodal Emotion Recognition Based on Facial Expressions, Speech, and EEG. IEEE Open Journal of Engineering in Medicine and Biology. https://doi.org/10.1109/OJEMB.2023.3240280
31. Thombare MB, Gumaste SV. Enhancing facial emotion recognition with vision transformers, efficientnet, and domain adaptation. In: Proceedings of the International Conference on Future Technologies (ICFT); 2025. https://doi.org/10.1109/ICFT66708.2025.11336650.