Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3384~3393
e-ISSN: 2548-6861

Real-Time Facial Emotion Recognition Using Mini-Xception and EfficientNetB4

Gede Pradistya Evan Aryaputra [1], Christy Atika Sari [2]*, Eko Hari Rachmawanto [3]
Department of Informatics Engineering, Universitas Dian Nuswantoro, Semarang, Indonesia
111202314886@mhs.dinus.ac.id [1], christy.atika.sari@dsn.dinus.ac.id [2*], eko.hari@dsn.dinus.ac.id [3]

Article Info

Article history:

Received 2026-07-01
Revised 2026-07-24
Accepted 2026-08-08

Keyword:

Facial Emotion Recognition,
Mini-Xception,
EfficientNetB4,
RAF-DB,
Classifications Real-Time.

ABSTRACT

Facial Emotion Recognition (FER) is an important field within computer vision and human–computer interaction that focuses on the automatic recognition of human emotional expressions through facial images. This study presents a comparative analysis of two Convolutional Neural Network (CNN) architectures, namely Mini-Xception and EfficientNetB4, for real-time facial emotion classification using the RAF-DB (Real-world Affective Faces Database) dataset. Mini-Xception was employed as a lightweight model with lower computational requirements, whereas EfficientNetB4 utilized a transfer learning approach to achieve superior classification performance. The RAF-DB dataset consists of seven primary emotion categories: angry, disgust, fear, happy, neutral, sad, and surprise. The preprocessing stage included facial image resizing, grayscale conversion for Mini-Xception, RGB normalization for EfficientNetB4, and the application of data augmentation techniques to improve model generalization capability. Experimental results demonstrated that Mini-Xception achieved a validation accuracy of 52.12%, while EfficientNetB4 attained a validation accuracy of 86.02%. In real-time implementation using a webcam and OpenCV, Mini-Xception exhibited advantages in inference speed, whereas EfficientNetB4 produced more stable and accurate emotion predictions. The findings indicate a trade-off between computational efficiency and classification performance. Therefore, EfficientNetB4 is more suitable for systems requiring high classification accuracy, while Mini-Xception is more appropriate for real-time applications operating under limited computational resources.

I. Introduction

Facial Emotion Recognition (FER) is a rapidly growing research area in computer vision and artificial intelligence due to its wide range of applications in human–computer interaction systems [1], [10]. FER aims to identify facial expressions that represent basic human emotions, including happy, sad, angry, fear, surprise, disgust, and neutral. This technology has been widely applied in various fields, such as security systems, online learning, mental health assessment, human–computer interaction, and driver monitoring systems [4], [14]. Despite significant advancements, FER implementation still faces numerous challenges, particularly when deployed in uncontrolled real-world environments characterized by variations in lighting conditions, facial pose, image quality, occlusion, and similarities among emotional expressions [7], [20].

The development of deep learning methods, particularly Convolutional Neural Networks (CNNs), has demonstrated substantial performance improvements compared with traditional approaches that rely on handcrafted feature extraction [2], [16]. CNNs are capable of automatically extracting features and effectively identifying visual patterns in facial images. Various modern CNN architectures, including ResNet, MobileNet, EfficientNet, and Mini-Xception, have been extensively utilized in FER systems due to their superior classification capabilities [6], [11]. EfficientNetB4 is a state-of-the-art CNN architecture that implements a compound scaling strategy to balance network depth, width, and resolution, thereby improving classification accuracy in an efficient manner [6]. In contrast, Mini-Xception was designed as a lightweight model with lower computational complexity, making it particularly suitable for real-time FER applications [3], [5].

Several previous studies have demonstrated the successful application of CNNs in facial emotion recognition tasks. Research employing EfficientNet with transfer learning has reported improved facial emotion classification performance compared with conventional CNN architectures [1]. Other studies have developed lightweight CNN-based FER models capable of delivering satisfactory performance in real-time applications [5]. Furthermore, the implementation of FER in online learning systems using deep learning approaches has shown that emotion recognition can enhance user interaction effectiveness [18]. Research in the field of mental health has also reported promising classification results using EfficientNetB4 for facial expression analysis [4]. In addition, previous studies have emphasized the importance of balanced datasets and robust evaluation protocols to improve FER model performance under real-world conditions [17].

In FER research, the selection of an appropriate dataset plays a crucial role in evaluating model performance. One of the most widely used datasets in modern FER studies is the Real-world Affective Faces Database (RAF-DB) [7], [8]. The RAF-DB dataset contains facial images collected under real-world conditions, featuring diverse expressions, facial poses, lighting variations, and occlusions that are more complex than those found in controlled datasets. RAF-DB consists of seven basic emotion categories and is therefore considered more representative for evaluating FER model performance in real-world scenarios. However, the complexity of RAF-DB presents significant challenges, causing many FER models to struggle in maintaining high classification accuracy, particularly lightweight models designed for real-time applications [10], [12].

Although numerous previous studies have achieved high classification accuracy using modern CNN architectures, most have primarily focused on maximizing accuracy without adequately considering computational efficiency and real-time implementation [11], [13]. Conversely, lightweight models developed for real-time applications generally exhibit lower accuracy than more complex CNN architectures [3], [5]. This situation highlights a research gap regarding the performance comparison between lightweight models and high-accuracy modern CNN architectures in real-time FER systems using the RAF-DB dataset.

Several previous studies have applied CNN-based approaches for facial emotion recognition using different architectures. Akhand et al. [2] utilized transfer learning techniques in deep CNN models and reported promising classification performance. Na et al. [4] proposed FacialNet for emotion recognition in mental health analysis and demonstrated the effectiveness of transfer learning in improving model accuracy. Furthermore, Elsheikh et al. [11] and Lawpanom et al. [12] developed advanced CNN architectures that achieved competitive results on facial emotion recognition datasets. These findings indicate that deeper architectures generally provide higher classification accuracy but require greater computational resources. Therefore, this study compares Mini-Xception as a lightweight model with EfficientNetB4 as a high-performance model to investigate the trade-off between computational efficiency and classification performance in real-time facial emotion recognition systems.

Based on this research gap, the present study implements and compares Mini-Xception and EfficientNetB4 for facial emotion classification using the RAF-DB dataset. Both models were evaluated under similar experimental settings to ensure objective performance assessment. In addition, several evaluation metrics, including accuracy, precision, recall, F1-score, classification reports, and confusion matrices, were employed to evaluate the performance of each model. This study also implements a webcam-based facial emotion recognition system using OpenCV to assess model performance under real-time conditions. The findings are expected to provide insights into the trade-off between classification accuracy and computational efficiency in FER systems, thereby supporting the selection of appropriate models for real-time application requirements.

II. METHOD

A. Facial Emotion Recognition (FER)

Facial Emotion Recognition (FER) is a research area within computer vision that focuses on identifying human emotions through the analysis of facial expressions. FER systems have been widely applied in various domains, including human-computer interaction, education, mental health monitoring, security systems, and driver behavior analysis. With the advancement of deep learning techniques, FER systems have achieved significant improvements in performance compared to conventional approaches based on handcrafted feature extraction [4], [10], [17].

B. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is one of the most widely used deep learning methods for image classification tasks. CNNs are capable of automatically extracting hierarchical features from images through convolutional, pooling, and fully connected layers. This capability enables CNNs to effectively recognize visual patterns in facial images and perform accurate emotion classification. Previous studies have demonstrated the effectiveness of CNN-based approaches in facial expression recognition applications [2], [5], [18].

C. Mini-Xception

Mini-Xception is a lightweight CNN architecture developed for facial expression recognition tasks with low computational complexity. The model employs depthwise separable convolution to reduce the number of parameters and improve inference speed. Due to its compact architecture, Mini-Xception is particularly suitable for real-time applications and deployment on devices with limited computational resources. The architecture was derived from the Xception model proposed by Chollet and has been widely adopted in FER research because of its efficiency and practical implementation advantages [3], [5].

D. *EfficientNetB4*

EfficientNetB4 is a variant of the EfficientNet family that utilizes a compound scaling approach to balance network depth, width, and input resolution. By leveraging transfer learning from pre-trained models, EfficientNetB4 can achieve high classification accuracy across various image recognition tasks, including facial emotion recognition. Previous studies have reported that EfficientNet-based architectures outperform many conventional CNN models while maintaining efficient parameter utilization [1], [6], [9].

E. *Research Stages*

This study was conducted through several major stages to compare the performance of the Mini-Xception and EfficientNetB4 models in classifying facial emotions using the RAF-DB dataset. The research workflow consisted of dataset collection, data preprocessing, model training, performance evaluation, and the implementation of a real-time facial emotion recognition system. The overall research framework is illustrated in Figure 1. At the initial stage, the RAF-DB (Real-world Affective Faces Database) dataset was employed as the primary data source. This dataset was selected because it contains facial images captured under real-world conditions, including variations in lighting, facial pose, expression, and occlusion, which are more complex than those found in controlled datasets [7], [8]. The RAF-DB dataset comprises seven basic emotion categories: angry, disgust, fear, happy, neutral, sad, and surprise.

The next stage involved data preprocessing. All facial images were resized and normalized prior to model training. For the Mini-Xception model, the images were resized to 48 × 48 pixels and converted into grayscale format, whereas EfficientNetB4 utilized RGB images with a resolution of 380 × 380 pixels. In addition, data augmentation techniques, including rotation, horizontal flipping, and zooming, were applied to increase the diversity of the training data and reduce the risk of overfitting.

Following the preprocessing stage, model training was conducted using two different CNN architectures, namely Mini-Xception and EfficientNetB4. Mini-Xception was selected due to its lightweight architecture and low computational complexity, making it suitable for real-time applications [3], [5]. In contrast, EfficientNetB4 was employed because of its superior classification performance achieved through the compound scaling approach, which effectively improves both model efficiency and accuracy [1], [6].

The subsequent stage involved evaluating the models using several performance metrics, including accuracy, precision, recall, F1-score, classification reports, and confusion matrices. These evaluation metrics were utilized to assess the capability of each model in recognizing the individual facial emotion categories within the RAF-DB dataset.

The final stage of this study involved the implementation of a webcam-based facial emotion recognition system using OpenCV to evaluate model performance under real-time conditions. The system detects users' faces directly through a webcam and subsequently performs emotion prediction using the previously trained models. The predicted emotion labels are displayed in real time on the screen. This study is expected to provide valuable insights into the performance comparison between lightweight models and modern CNN architectures in real-time Facial Emotion Recognition (FER) systems.

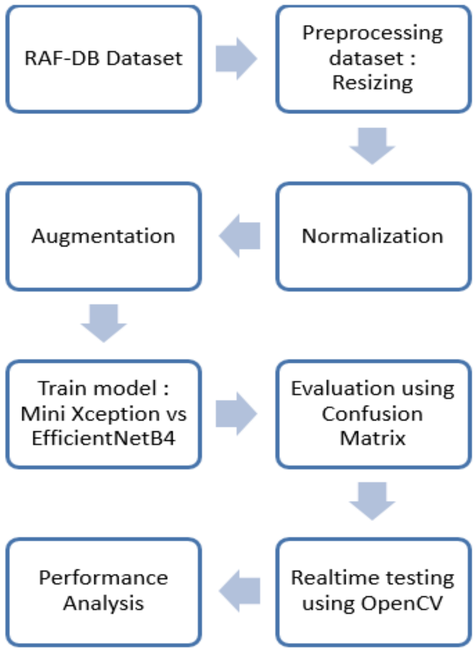

Figure 1. Research Stages of the Facial Emotion Recognition System

F. *Dataset and Preprocessing*

Here, the RAF-DB (Real-world Affective Faces Database) dataset was utilized as the primary dataset for training and evaluating facial emotion classification models. RAF-DB is recognized as one of the most widely used Facial Emotion Recognition (FER) datasets because it contains facial images captured under real-world (in-the-wild) conditions, including variations in lighting, facial pose, expression, and occlusion [7], [8]. The dataset comprises seven basic emotion categories: angry, disgust, fear, happy, neutral, sad, and surprise.

The RAF-DB dataset was divided into two subsets, namely the training set and the test set. Prior to model training, all images underwent a preprocessing stage. This process was performed to improve data quality while adapting the image characteristics to the requirements of each CNN architecture employed in this study. For the Mini-Xception model, facial images were resized to a resolution of 48 × 48 pixels and converted into grayscale format. The relatively small image size was selected to reduce computational costs, allowing the model to operate more efficiently in real-time environments. In contrast, EfficientNetB4 utilized RGB images with a resolution of 380 × 380 pixels to preserve richer and more complex visual information during the feature extraction process.

The preprocessing stage also included data normalization by scaling pixel values to a range between 0 and 1. This normalization process facilitated faster model convergence during training while improving the stability of the CNN learning process. In addition, data augmentation techniques were applied to increase the diversity of the training dataset and reduce the likelihood of overfitting. The augmentation methods included image rotation, horizontal flipping, zooming, and random image shifting. Through data augmentation, the models were expected to learn facial expression patterns more effectively and achieve better generalization under real-world conditions. Table 1 presents the emotion categories used in this study.

TABLE I
EMOTION CATEGORIES IN RAF-DB DATASET

No	Emotion Category	Description
1	Angry	Angry expression
2	Disgust	Disgusted expression
3	Fear	Fearful expression
4	Happy	Happy expression
5	Neutral	Neutral expression
6	Sad	Sad expression
7	Surprise	Surprised expression

The RAF-DB dataset used in this study consists of facial images representing seven basic emotion categories, which were divided into training and testing sets according to the official dataset protocol. The entire model training process was conducted using several configuration parameters, including input image size, optimizer, learning rate, batch size, and the number of training epochs, to accommodate the specific characteristics of each CNN architecture. Upon completion of the preprocessing stage, the processed facial image data were used to train the Mini-Xception and EfficientNetB4 models for facial emotion classification.

TABLE 2
TRAINING CONFIGURATION OF CNN MODELS

Parameter	Mini-Xception	EfficientNetB4
Input Size	48×48	380×380
Color Format	Grayscale	RGB
Optimizer	Adam	Adam
Learning Rate	0.001	0.0001
Batch Size	32	16
Epoch	50	50
Dataset	RAF-DB	RAF-DB

G. CNN Architecture

This study employed two Convolutional Neural Network (CNN) architectures, namely Mini-Xception and EfficientNetB4, to classify facial emotions using the RAF-DB dataset. These models were selected because they exhibit distinct characteristics in terms of computational complexity and classification performance, making them suitable for comparing the effectiveness of lightweight architectures and modern CNN models in Facial Emotion Recognition (FER) systems.

H. Mini-Xception

Mini-Xception is a lightweight variant of the Xception architecture that was specifically designed to reduce computational requirements, making it suitable for real-time applications [3], [5]. The architecture employs depthwise separable convolution to reduce the number of trainable parameters while improving computational efficiency without significantly compromising the model's feature extraction capability. In this study, Mini-Xception utilized grayscale facial images with a resolution of 48 × 48 pixels as input. The model architecture consists of several convolutional layers, batch normalization layers, activation functions, and a global average pooling layer before generating emotion classification outputs through a softmax activation function. The selection of Mini-Xception was motivated by its relatively lightweight architecture compared with many modern CNN models. This characteristic makes the model suitable for webcam-based facial emotion recognition systems and other real-time applications. Furthermore, its lower memory requirements and faster inference time enable efficient deployment on devices with moderate computational resources.

I. EfficientNetB4

EfficientNetB4 is a modern CNN architecture that employs a compound scaling approach to achieve an optimal balance among network depth, width, and input resolution [1], [6]. This approach enables the model to achieve superior classification performance while utilizing parameters more efficiently than conventional CNN architectures. In this study, EfficientNetB4 received RGB images with a resolution of 380 × 380 pixels as input. The model was implemented using a transfer learning strategy with pretrained ImageNet weights to accelerate the training process and enhance the model's generalization capability. The final classification layer was subsequently modified to match the number of emotion categories contained in the RAF-DB dataset. In addition to transfer learning, fine-tuning was applied to several of the final layers of EfficientNetB4. This process was intended to improve the model's ability to learn the facial expression characteristics present in the RAF-DB dataset, thereby enabling more effective extraction and representation of emotion-related features.

J. Model Evaluation

Model performance was evaluated using several evaluation metrics, including accuracy, precision, recall, and F1-score. In addition, the performance of each model across individual emotion categories was analyzed using classification reports and confusion matrices. Accuracy was used to measure the overall correctness of the model in classifying all testing samples. Precision was employed to assess the accuracy of the model's predictions for each emotion category, whereas recall was used to evaluate the model's ability to correctly identify all samples belonging to a specific category. Furthermore, the F1-score was utilized as a combined metric that balances in this study are presented in Equations (1), (2), (3), and (4), where TP is True Positive, TN is True Negative, FP is False Positive, FN is False Negative. The evaluation results of both models were subsequently compared to determine the most effective model for real-time facial emotion classification.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$Precision = TP / (TP + FP) \quad (2)$$

$$Recall = TP / (TP + FN) \quad (3)$$

$$F1 - Score = (2 \times Precision \times Recall) / (Precision + Recall) \quad (4)$$

III. Result and Discussion

A. Mini-Xception Model

In this study, the Mini-Xception model was employed as a lightweight architecture for real-time facial emotion classification. The model was trained using the RAF-DB dataset with grayscale input images of 48 × 48 pixels. The use of a smaller image resolution was intended to reduce computational complexity, thereby enabling faster inference during real-time operation. Based on the training results, Mini-Xception achieved a best validation accuracy of 52.12%. The progression of accuracy and loss values throughout the training process is presented in Figure 2. As illustrated in the figure, both training and validation accuracy gradually increased as the number of epochs increased. In addition, the loss value consistently decreased, indicating that the model was able to learn the underlying data patterns effectively.

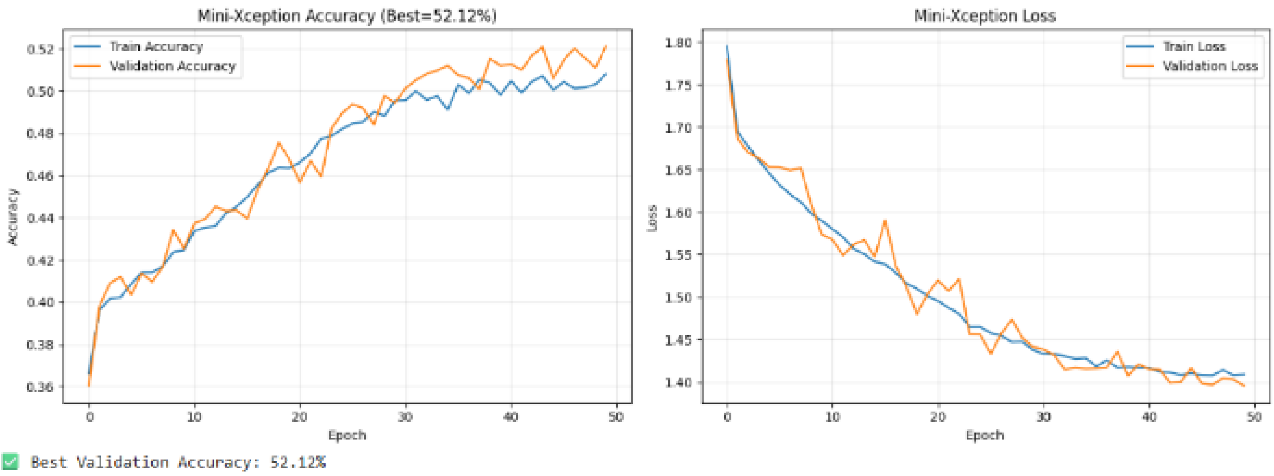

Figure 2. Training Accuracy and Loss of Mini-Xception Model

Based on the classification report results, the Mini-Xception model exhibited varying levels of performance across different emotion categories. The happy category achieved the highest precision and recall values among all emotion classes. In contrast, categories such as fear and disgust demonstrated lower performance due to the similarity of facial expression patterns and the relatively limited number of training samples available for these categories.As illustrated in the confusion matrix shown in Figure 3, the model still produced misclassifications among several emotion categories with similar visual characteristics. Most prediction errors occurred between the sad, neutral, and fear categories. These findings indicate that lightweight architectures such as Mini-Xception still face challenges in distinguishing complex facial expressions under real-world conditions. The confusion matrix of the Mini-Xception model is presented in Figure 3. Although it achieved lower classification accuracy compared with other modern CNN architectures, Mini-Xception demonstrated notable advantages in terms of computational efficiency and inference speed. The model was capable of operating in real-time using a webcam and OpenCV while maintaining relatively low memory consumption. Therefore, Mini-Xception remains a suitable choice for deployment on devices with moderate hardware specifications and applications that require rapid response times.

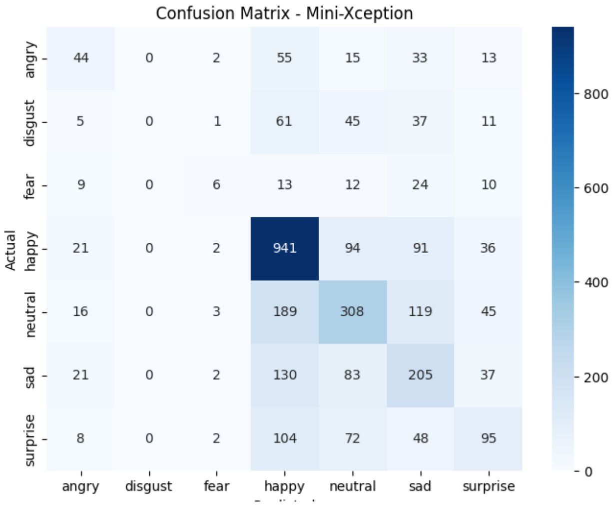

Figure 3. Confusion Matrix of Mini-Xception Model

The classification report results indicate that the performance of Mini-Xception varied across different emotion categories. The happy category achieved a precision of 0.63, a recall of 0.79, and an F1-score of 0.70, representing the best performance among all emotion classes. These results suggest that the model was able to recognize happy expressions more effectively due to their more distinctive and dominant visual patterns within the RAF-DB dataset.

CLASSIFICATION REPORT

	precision	recall	f1-score	support
angry	0.35	0.27	0.31	162
disgust	0.00	0.00	0.00	160
fear	0.33	0.08	0.13	74
happy	0.63	0.79	0.70	1185
neutral	0.49	0.45	0.47	680
sad	0.37	0.43	0.40	478
surprise	0.38	0.29	0.33	329
accuracy			0.52	3068
macro avg	0.37	0.33	0.33	3068
weighted avg	0.48	0.52	0.49	3068

Figure 4. Classification Report of Mini-Xception Model

In contrast, the fear and disgust categories exhibited relatively poor classification performance, with F1-scores of 0.13 and 0.00, respectively. These low scores indicate that the model was not able to effectively distinguish facial expressions with visual characteristics that closely resemble those of other emotion categories, particularly sad and neutral. Another factor contributing to the model's performance was the imbalanced distribution of the RAF-DB dataset, resulting in certain emotion categories being underrepresented during the training process. Meanwhile, the neutral and sad categories achieved relatively satisfactory results, with F1-scores of 0.47 and 0.40, respectively. These findings suggest that the model was still capable of recognizing the fundamental patterns of several facial expressions, although its performance remained suboptimal under real-world conditions characterized by variations in lighting, facial pose, and occlusion. Overall, Mini-Xception achieved an accuracy of 52% with a weighted average F1-score of 0.49. These results indicate that lightweight CNN architectures still face limitations in handling complex facial expression classification tasks on in-the-wild datasets. Nevertheless, the model offers advantages in terms of computational efficiency and inference speed, making it more suitable for real-time Facial Emotion Recognition (FER) systems operating under limited computational resources.

B. EfficientNetB4 Model

EfficientNetB4 was employed in this study to enhance facial emotion classification performance on the RAF-DB dataset. The model incorporates transfer learning and fine-tuning techniques to optimize the extraction of visual features from facial images. Unlike Mini-Xception, which prioritizes a lightweight architecture, EfficientNetB4 utilizes a compound scaling approach to effectively balance network depth, width, and input resolution, thereby achieving superior classification performance.

During the training process, EfficientNetB4 received RGB images with a resolution of 380 × 380 pixels as input. The training procedure was conducted in two stages, namely the frozen stage and the fine-tuning stage. In the initial stage, all layers of the base model were frozen, and only the classification layers were trained for 20 epochs. Subsequently, several layers of the base model were unfrozen and retrained using a lower learning rate, enabling the model to learn dataset-specific features from RAF-DB more effectively. Based on the training results, EfficientNetB4 achieved a best validation accuracy of 86.02%. The validation accuracy curve is presented in Figure 5. As shown in the figure, the fine-tuning process initiated at epoch 20 resulted in a substantial improvement in model performance, as indicated by the increase in accuracy from approximately 48% to over 80% within only a few epochs. Following this stage, the model's accuracy continued to improve gradually until it reached its highest validation accuracy of 86.02%. EfficientNetB4 demonstrated superior classification performance compared with Mini-Xception due to its ability to extract visual features more effectively through a deeper and more sophisticated network architecture. In addition, the model exhibited strong generalization capability when applied to real-world facial images characterized by variations in pose, illumination, and occlusion. Nevertheless, the increased architectural complexity of EfficientNetB4 resulted in higher computational requirements than those of lightweight models. Consequently, EfficientNetB4 is more suitable for deployment on systems equipped with greater computational resources.

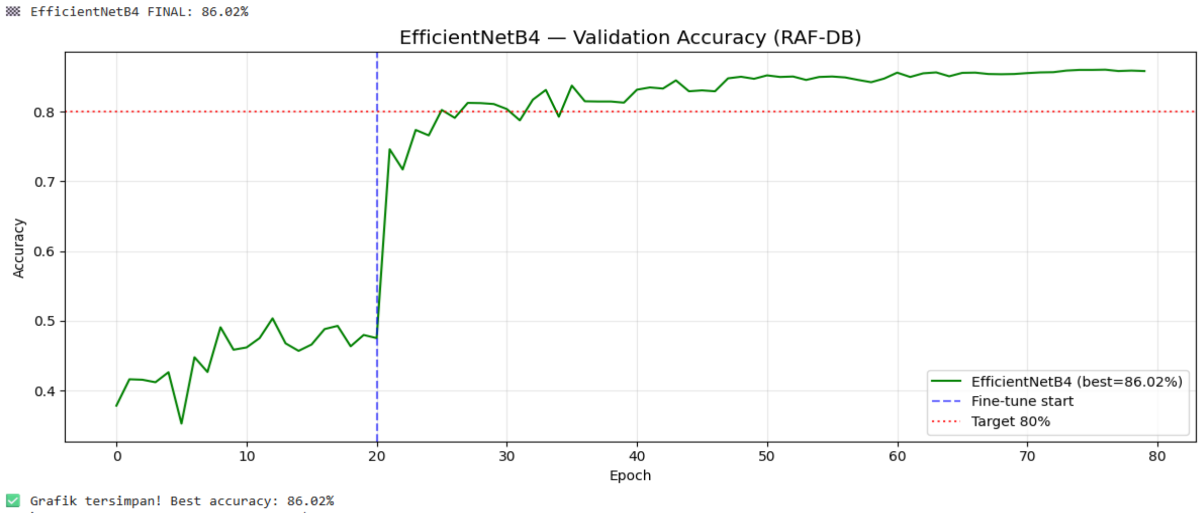

Figure 5. Validation Accuracy of EfficientNetB4 Model

The confusion matrix analysis revealed that EfficientNetB4 achieved superior classification performance compared with Mini-Xception across nearly all emotion categories. Most predictions were concentrated along the main diagonal of the confusion matrix, indicating that the model was able to recognize facial expressions with a high degree of accuracy.

This result demonstrates that EfficientNetB4 was able to learn more discriminative facial representations through its deeper architecture and transfer learning strategy. Consequently, the model exhibited greater robustness in distinguishing subtle facial variations, even under challenging real-world conditions characterized by illumination changes, pose variations, and partial occlusions.

The happy category achieved the highest number of correct predictions, as reflected by a recall of 93.00% and a precision of 96.24%. These results suggest that happy expressions possess more distinctive visual characteristics, making them easier for the model to recognize. Furthermore, the expression patterns associated with this category appear to be more consistent throughout the RAF-DB dataset. In addition, the neutral, sad, and surprise categories also demonstrated strong classification performance, achieving F1-scores above 84%. The architectural complexity of EfficientNetB4 enables the extraction of more discriminative facial features, thereby improving the model's generalization capability when applied to real-world facial images. Despite its strong overall performance, the model still exhibited misclassification errors in the fear and disgust categories. In particular, the fear category achieved a recall of only 50.00%, indicating that the model continued to experience difficulties in distinguishing facial expressions with similar visual characteristics. Moreover, factors such as variations in lighting conditions, facial pose, occlusion, and less pronounced facial expressions may have contributed to the reduced classification performance observed in certain emotion categories. Overall, the confusion matrix analysis demonstrated that EfficientNetB4 achieved more accurate and stable classification performance than Mini-Xception. These findings indicate that the combination of transfer learning and fine-tuning within a modern CNN architecture is effective in enhancing the performance of Facial Emotion Recognition (FER) systems when applied to challenging in-the-wild datasets such as RAF-DB.

The classification report results indicate that EfficientNetB4 achieved excellent classification performance across the majority of facial emotion categories. Overall, the model attained an accuracy of 86.02% and a weighted average F1-score of 85.95%, demonstrating stable and reliable classification performance on the RAF-DB dataset. The happy category achieved the best performance, with a precision of 96.24%, a recall of 93.00%, and an F1-score of 94.59%. The outstanding performance in this category suggests that the model was highly effective in recognizing happy facial expressions, primarily due to their more distinctive visual patterns and the availability of more representative training samples within the dataset.

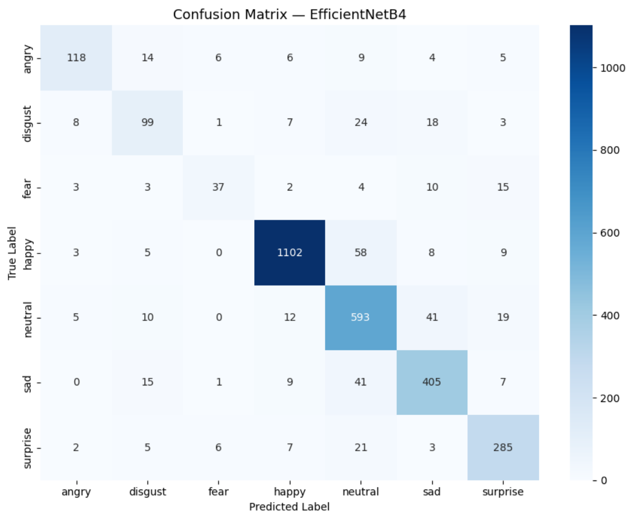

Figure 6. Confusion Matrix of EfficientNetB4 Model

Furthermore, the neutral, sad, and surprise categories also demonstrated strong classification performance, achieving F1-scores of 82.94%, 83.76%, and 84.82%, respectively. These results indicate that EfficientNetB4 was able to extract facial features more effectively through the combination of a modern CNN architecture and transfer learning techniques. In contrast, the fear and disgust categories continued to exhibit lower classification performance compared with the other emotion classes. The relatively low recall values observed in these categories suggest that the model was not yet fully effective in distinguishing facial expressions with visual characteristics similar to those of other emotions, particularly sad and neutral. Furthermore, the imbalanced distribution of the RAF-DB dataset may have contributed to the model's reduced ability to accurately recognize certain emotion categories. Nevertheless, the evaluation results demonstrate that EfficientNetB4 consistently outperformed lightweight models such as Mini-Xception in terms of classification performance. Therefore, EfficientNetB4 is more suitable for Facial Emotion Recognition (FER) systems that require high classification accuracy, despite its higher computational complexity and resource requirements.

Classification Report – EfficientNetB4

	precision	recall	f1-score	support
angry	0.8489	0.7284	0.7841	162
disgust	0.6556	0.6188	0.6367	160
fear	0.7255	0.5000	0.5920	74
happy	0.9624	0.9300	0.9459	1185
neutral	0.7907	0.8721	0.8294	680
sad	0.8282	0.8473	0.8376	478
surprise	0.8309	0.8663	0.8482	329
accuracy			0.8602	3068
macro avg	0.8060	0.7661	0.7820	3068
weighted avg	0.8616	0.8602	0.8595	3068

Figure 7. Classification Report of EfficientNetB4 Model

C. Real-Time Implementation

During the implementation stage, the facial emotion recognition system was evaluated in real time using a webcam and the OpenCV library. The system performed face detection directly from the camera feed, after which the captured facial images were processed using the previously trained CNN models to classify facial emotions. The system was developed using the Python programming language with the support of the TensorFlow and Keras frameworks. OpenCV was utilized for real-time image acquisition and facial region detection prior to the emotion classification process.The implementation results demonstrated that both models were capable of operating effectively in a real-time environment. Mini-Xception achieved faster inference times due to its lightweight architecture, resulting in more stable system responsiveness on devices with moderate hardware specifications. In contrast, EfficientNetB4 produced more accurate emotion predictions but required slightly longer processing times than Mini-Xception. The system was able to recognize several basic emotion categories, including happy, sad, neutral, angry, fear, disgust, and surprise, directly through webcam input. The predicted emotions were displayed on the screen in the form of emotion labels corresponding to the classification results generated by the model.

Based on the implementation results presented in Figure 8, the system was able to detect the user's face in real time while simultaneously displaying the predicted emotion. The model successfully classified the happy emotion with a confidence score of 53.5%. These findings demonstrate that the CNN-based approach can be effectively implemented in real-time human–computer interaction systems.

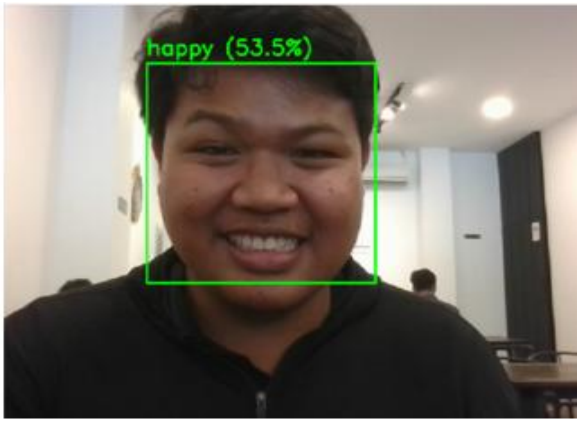

Figure 8. Real-Time Facial Emotion Recognition Using Webcam and OpenCV

D. Discussion

The results of this study demonstrate that different CNN architectures produce varying levels of performance in real-time Facial Emotion Recognition (FER) systems. Mini-Xception and EfficientNetB4 exhibit distinct characteristics in terms of classification accuracy, computational efficiency, and effectiveness under real-world conditions.

Mini-Xception achieved a validation accuracy of 52.12%, with its primary advantages being faster inference speed and lower memory consumption. The lightweight architecture enabled the model to operate efficiently in real-time environments using a webcam and OpenCV while maintaining rapid system responsiveness. Therefore, Mini-Xception is more suitable for deployment on devices with moderate hardware specifications and applications that require low-latency performance.

In contrast, EfficientNetB4 achieved a substantially higher validation accuracy of 86.02%. This result indicates that the application of transfer learning and fine-tuning significantly enhanced the model's ability to extract discriminative visual features from facial expressions. Furthermore, EfficientNetB4 demonstrate d superior generalization capability on the RAF-DB dataset, which contains complex variations in facial pose, illumination, expression, and occlusion.

TABLE III
COMPARISON OF MODEL PERFORMANCE USING RAF-DB DATASET

Model	Accuracy	Real-Time Performance	Complexity
Mini-Xception	52.12%	Fast	Low
EfficientNetB4	86.02%	Medium	High

Based on Table 3, EfficientNetB4 achieved superior classification performance compared with Mini-Xception. However, the model requires greater computational complexity, resulting in longer inference times. In contrast, Mini-Xception demonstrated better real-time performance despite achieving lower classification accuracy. The findings of this study were also compared with those reported in previous research, as presented in Table 4.

TABLE IV
COMPARISON WITH PREVIOUS STUDIES

Research	Model	Dataset	Accuracy
Singh et al. [1]	EfficientNet	FER2013	72.4%
Akhand et al. [2]	Deep CNN	CK+	84.6%
Proposed Method	EfficientNetB4	RAF-DB	86.02%

Based on the comparison presented in Table 4, the EfficientNetB4 model developed in this study demonstrated competitive performance relative to those reported in previous research. Although it did not achieve the highest classification accuracy among all compared studies, this research successfully implemented a real-time facial emotion recognition system using a webcam and OpenCV, thereby providing additional value from a practical system implementation perspective. Furthermore, the use of the RAF-DB dataset introduced greater challenges than controlled datasets such as CK+ and FER2013, as it contains a wider range of real-world variations in facial pose, illumination, expression, and occlusion. Therefore, the achieved accuracy of 86.02% indicates that EfficientNetB4 is capable of effectively recognizing and classifying facial emotions under in-the-wild conditions. The findings of this study highlight the trade-off between classification accuracy and computational efficiency in Facial Emotion Recognition (FER) systems. EfficientNetB4 outperformed Mini-Xception in terms of classification performance, whereas Mini-Xception proved more suitable for real-time implementation due to its lower computational complexity and faster inference speed.

IV. CONCLUSION

This study implemented and compared two CNN architectures, namely Mini-Xception and EfficientNetB4, for real-time facial emotion recognition using the RAF-DB dataset, which consists of seven primary emotion categories: angry, disgust, fear, happy, neutral, sad, and surprise. The experimental results revealed substantial differences in classification performance between the two models, highlighting the trade-off between classification accuracy and computational efficiency in Facial Emotion Recognition (FER) systems. Mini-Xception achieved a validation accuracy of 52.12% with a weighted average F1-score of 0.49. Although its classification accuracy was relatively lower, Mini-Xception demonstrated advantages in inference speed and memory efficiency due to its lightweight architecture. These characteristics make the model more suitable for deployment on devices with limited computational resources or applications that require real-time responses with low latency. The model achieved its best performance in the happy category, whereas fear and disgust yielded the lowest F1-scores of 0.13 and 0.00, respectively. This limitation can be attributed to the visual similarity between these emotions and other negative emotion categories, as well as the limited number of representative samples available in the RAF-DB dataset. EfficientNetB4 achieved a substantially higher validation accuracy of 86.02% with a weighted average F1-score of 85.95%, demonstrating the effectiveness of transfer learning and fine-tuning in optimizing feature extraction capabilities for in-the-wild facial expression datasets. The two-stage training strategy, consisting of classifier training during the initial phase followed by fine-tuning of selected base model layers from epoch 20 onward, contributed significantly to the model’s performance improvement. The happy category once again achieved the highest F1-score of 94.59%, while fear and disgust remained the most challenging categories, with recall values of 50.00% and 61.88%, respectively. The real-time implementation using OpenCV and a webcam confirmed that both models can operate effectively in live environments. EfficientNetB4 produced more accurate emotion predictions, whereas Mini-Xception provided faster response times. Despite these promising results, this study has several limitations, including the relatively low classification accuracy of Mini-Xception and the higher computational requirements of EfficientNetB4. Future research may explore model compression techniques, such as knowledge distillation, pruning, and quantization, to achieve a better balance between classification accuracy and inference efficiency. Furthermore, extending the evaluation to video-based temporal emotion recognition and validating the models on additional real-world datasets represent promising directions for future investigation.

REFERENCES

[1] M. Sigala, A. Beer, L. Hodgson, and A. O’Connor, *Big Data for Measuring the Impact of Tourism Economic Development Programmes: A Process* Singh, R., Sharma, H., & Mehta, N. K. (2022). EfficientNet for Human FER Using Transfer Learning. ICTACT Journal on Soft Computing, 13(1), 2792–2797. https://doi.org/10.21917/ijsc.2022.0397

[2] Akhand, M. A. H., et al. (2021). Facial emotion recognition using transfer learning in the deep CNN. Electronics, 10(9), 1036. https://doi.org/10.3390/electronics10091036

[3] Arriaga, O., Valdenegro-Toro, M., & Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*. https://doi.org/10.48550/arXiv.1710.07557

[4] Na, I., et al. (2024). FacialNet: facial emotion recognition for mental health analysis using UNet segmentation with transfer learning model.

Frontiers in Computational Neuroscience, 18, 1485121. https://doi.org/10.3389/fncom.2024.1485121
[5] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1251–1258. https://doi.org/10.1109/CVPR.2017.195
[6] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 6105–6114. https://doi.org/10.48550/arXiv.1905.11946
[7] Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2584–2593. https://doi.org/10.1109/CVPR.2017.277
[8] Li, S., & Deng, W. (2019). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing, 28*(1), 356–370. https://doi.org/10.1109/TIP.2018.2868382
[9] Handayani, N. M. K. K., et al. (2023). Pengenalan ekspresi wajah menggunakan transfer learning MobileNetV2 dan EfficientNet-B0. *Jurnal Media Informatika Budidarma, 8*(1). https://doi.org/10.30865/mib.v8i1.7048
[10] Kopalidis, T., Solachidis, V., Vretos, N., & Daras, P. (2024). Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets. *Information, 15*(3), 135. https://doi.org/10.3390/info15030135
[11] Elsheikh, R. A., Mohamed, M. A., Abou-Taleb, A. M., & Ata, M. M. (2024). Improved facial emotion recognition model based on a novel deep convolutional structure. *Scientific Reports, 14*, 29050. https://doi.org/10.1038/s41598-024-79167-8
[12] Lawpanom, R., Songpan, W., & Kaewyotha, J. (2024). Advancing facial expression recognition using homogeneous ensemble CNN. *Applied Sciences, 14*(3), 1156. https://doi.org/10.3390/app14031156
[13] Raj, R., & Demirkol, I. (2025). An improved facial emotion recognition system using convolutional neural network for the optimization of human robot interaction. *Scientific Reports, 15*, 38940. https://doi.org/10.1038/s41598-025-22835-0
[14] Qian, C., et al. (2025). Real-time emotion recognition based on facial expressions. *Multidisciplinary Reviews, 6*(3). https://doi.org/10.61838/kman.mr.6.3.25
[15] Özdemir, M. A., et al. (2019). Real time emotion recognition from facial expressions using CNN architecture. *2019 27th Signal Processing and Communications Applications Conference (SIU)*. https://doi.org/10.1109/SIU.2019.8806250
[16] Siqueira, H., Magg, S., & Wermter, S. (2020). Efficient facial feature learning with wide ensemble-based convolutional neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(04), 5800–5809. https://doi.org/10.1609/aaai.v34i04.6037
[17] Khaireddin, Y., & Chen, Z. (2021). Facial emotion recognition: State of the art performance on FER2013. *arXiv preprint arXiv:2105.03588*. https://doi.org/10.48550/arXiv.2105.03588
[18] Zhang, J., Li, S., & Wang, Y. (2023). Facial expression recognition based on improved convolutional neural networks. *Scientific Reports, 13*, 17654. https://doi.org/10.1038/s41598-023-44893-6
[19] Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 279–283. https://doi.org/10.1145/2993148.2993165
[20] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923