IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.20, No.3, July 2026, pp. 307~318
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.117591

Comparative Performance Analysis of Xception and ResNet50 Architectures for Facial Expression Recognition Using the FER-2013 Dataset

Diah Aryani[*1], Habibullah Akbar[2], Ferdinand Defin Delio[3]
[1,3]Department of Informatics Engineering, Faculty of Computer Science, Universitas Esa Unggul, Jakarta, Indonesia.
[2]Department of Magister Computer Science, Faculty of Computer Science, Universitas Esa Unggul, Jakarta, Indonesia.
e-mail: [*1]**diah.aryani@esaunggul.ac.id**, [2]habibullah.akbar@esaunggul.ac.id, [3] ferdinanddefin16@student.esaunggul.ac.id

Abstrak

Studi ini bertujuan untuk mengevaluasi dan membandingkan kinerja dan efisiensi komputasi arsitektur Xception dan ResNet50 dalam tugas klasifikasi ekspresi wajah. Pengenalan Ekspresi Wajah (Facial Expression Recognition/FER) merupakan bidang penting dalam pengembangan sistem cerdas yang mampu memahami emosi manusia. Teknologi ini berpotensi diterapkan pada berbagai sektor, seperti pendidikan adaptif, layanan pelanggan berbasis emosi, serta dukungan kesehatan mental. Penelitian ini membandingkan dua arsitektur Convolutional Neural Network (CNN) populer, yaitu Xception dan ResNet50, untuk klasifikasi ekspresi wajah menggunakan dataset FER-2013. Dataset tersebut berisi 35.887 citra wajah grayscale beresolusi 48×48 piksel yang terbagi ke dalam tujuh kategori emosi dasar. Seluruh citra diubah ukurannya menjadi 224×224 piksel dan dikonversi ke format RGB agar sesuai dengan kebutuhan model pretrained ImageNet.

Kedua model dilatih menggunakan pendekatan transfer learning dengan fine-tuning selektif pada beberapa layer. Augmentasi data yang intensif diterapkan untuk meningkatkan variasi data serta meminimalkan overfitting. Evaluasi dilakukan menggunakan metrik accuracy, precision, recall, F1-score, dan confusion matrix. Hasil menunjukkan bahwa Xception memberikan performa lebih baik dibandingkan ResNet50, dengan akurasi validasi 70,69% dan weighted F1-score 0,71. Xception unggul dalam mengenali ekspresi bahagia dan terkejut, serta lebih konsisten pada kelas minoritas. Temuan ini menegaskan pentingnya pemilihan arsitektur dan strategi pelatihan yang tepat untuk meningkatkan kinerja sistem FER secara efektif dan skalabel.

Kata kunci*—Prediksi Kualitas Udara, Machine Learning, Ensemble Methods, XGBoost, SHAP, UCI Dataset*

Abstract

This study aims to evaluate and compare the performance and computational efficiency of Xception and ResNet50 architectures in facial expression classification tasks. Facial

Received March 12th,2026; Revised April 24th, 2026; Accepted July 27th, 2026

Expression Recognition (FER) plays an essential role in developing intelligent systems capable of interpreting human emotions. This technology has broad applications, including adaptive learning environments, emotion-sensitive customer service, and mental health assistance. This study compares two widely used Convolutional Neural Network (CNN) architectures, Xception and ResNet50, for facial expression classification using the FER-2013 dataset. The dataset contains 35,887 grayscale facial images with a resolution of 48×48 pixels, categorized into seven fundamental emotions. All images were resized to 224×224 pixels and converted into RGB format to align with the input specifications of pretrained ImageNet models.

Both architectures were trained using a transfer learning strategy with selective fine-tuning on specific layers. Extensive data augmentation techniques were applied to enhance dataset variability and mitigate overfitting. Model performance was assessed using accuracy, precision, recall, F1-score, and confusion matrix metrics. The results reveal that Xception surpassed ResNet50, achieving a validation accuracy of 70.69% and a weighted F1-score of

0.71. Xception demonstrated superior performance in recognizing happiness and surprise while maintaining greater stability across minority classes. These findings emphasize that selecting an appropriate architecture and structured training approach significantly improves FER system performance in scalable and practical implementations.

Keywords— *CNN, FER-2013, Xception, ResNet50, Facial Expression Recognition, Transfer Learning.*

1. INTRODUCTION

Facial Expression Recognition (FER) refers to the automated process of identifying and classifying human emotions—such as happiness, sadness, anger, fear, surprise, and disgust—through the analysis of static or dynamic facial images. This process typically employs computer vision and deep learning approaches (e.g., Convolutional Neural Networks (CNN), R-CNN, and Vision Transformer) and is grounded in the Facial Action Coding System (FACS), which detects facial muscle movements known as Action Units (AUs) [1]. By automatically interpreting facial expressions, FER systems enable the recognition of human emotions, thereby facilitating more natural and context-aware human–computer interactions [2]. Over the past two decades, FER has been extensively studied; however, significant progress has emerged only recently due to advancements in deep learning technologies, particularly with the development of Convolutional Neural Network (CNN) architectures such as ResNet and Xception [3][4].

According to Kaur et al., FER plays a crucial role in human-centered artificial intelligence systems, as it supports adaptive and interactive responses across various applications [5] Deep learning–based approaches, especially Convolutional Neural Networks (CNNs), have demonstrated superior performance in facial expression recognition tasks owing to their ability to automatically extract discriminative spatial features [6]. CNNs remain robust even in noisy data conditions, as they are capable of capturing subtle micro-patterns in human facial expressions that are difficult to detect using conventional feature-based methods [7].

CNN architectures have been shown to effectively support emotion classification and spatial feature extraction from facial images. Among various CNN models, ResNet50 and Xception are widely regarded as strong candidates due to their capacity to learn deep feature

representations in a stable and computationally efficient manner [8]. ResNet50 employs a residual learning framework to address the vanishing gradient problem in very deep networks, whereas Xception extends the Inception architecture by incorporating depthwise separable convolutions to reduce computational complexity [4].

Despite these advancements, facial expression recognition in real-world environments continues to present substantial challenges. These include variations in expression intensity, illumination conditions, facial pose, image resolution, and label noise within datasets. The complexity of these challenges increases in real-time scenarios, such as surveillance systems, intelligent interaction platforms, and health monitoring applications where models must achieve both high accuracy and computational efficiency [9][10].

FER-2013 is a widely used benchmark dataset for facial expression recognition tasks. It comprises 35,887 grayscale facial images with a resolution of 48 × 48 pixels, categorized into seven basic emotions [11]. The primary challenges associated with this dataset include low image resolution, high noise levels, subtle expression variations, and class imbalance. Previous studies have emphasized the importance of data augmentation techniques and appropriate training strategies to enhance classification accuracy on this dataset [12]. However, previous studies mostly evaluate CNN architectures independently or combine them with hybrid models, while a systematic comparison between Xception and ResNet50 under identical experimental configurations on the FER-2013 dataset remains limited.

Therefore, in this study a comparative performance evaluation was conducted between the Xception and ResNet50 models using a transfer learning approach. The experimental setup involved RGB image preprocessing with a resolution of 224 × 224 pixels, as well as identical augmentation and fine-tuning strategies for both models. Performance evaluation extended beyond overall accuracy to include multi-class metrics such as precision, recall, F1-score, and confusion matrix analysis, thereby providing a more comprehensive assessment. Experimental design considerations were informed by prior work utilizing Residual Network architectures and attention modules, as proposed by Siqueira et al. [13], as well as multi-scale CNN approaches applied by Zhao et al. [14], particularly in addressing the complexities of facial expression classification under challenging data conditions such as those presented by FER- 2013. This study provides a systematic comparison of two widely used CNN architectures, Xception and ResNet50, under identical preprocessing pipelines and training strategies to ensure a fair and reliable performance evaluation for facial expression recognition.

2. METHODS

2.1 System Architecture and Design Framework

The study commenced with the identification and formulation of the research problem, followed by a comprehensive literature review to assess contemporary approaches in CNN-based facial expression classification. The FER-2013 dataset was employed, comprising 35,887 grayscale facial images with a resolution of 48 × 48 pixels [15]. The dataset was partitioned into training and validation subsets.

Subsequently, all images were resized to 224 × 224 pixels and converted into RGB format to ensure compatibility with the selected model architectures. Data augmentation techniques were

applied, including random rotation, translation, horizontal flipping, brightness shifting, shearing, zooming, and channel shifting, with the objective of enhancing data diversity and mitigating overfitting [4].

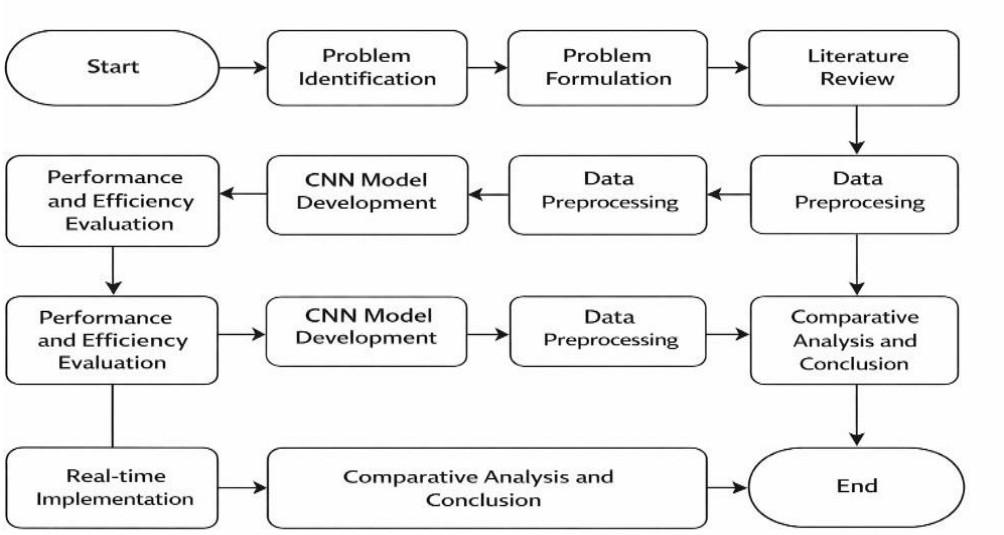


Figure 1. Research Workflow Diagram

As illustrated in Figure 1, the subsequent phase involved the development of CNN- based models employing two backbone architectures, Xception and ResNet50, both initialized with pretrained weights from ImageNet. A customized classification head was appended to each architecture, comprising a Global Average Pooling layer, Batch Normalization, a Dense layer with 512 units and ReLU activation, a Dropout layer with a rate of 0.5, and a final Dense layer with seven neurons using softmax activation. Model performance and computational efficiency were then assessed using evaluation metrics including accuracy, loss, and confusion matrix analysis. The results of this evaluation provided the foundation for real-time implementation, particularly to examine the models' effectiveness in more practical application scenarios. Finally, a comparative analysis was conducted to identify the most optimal architecture for facial expression classification, forming the basis for the study's conclusions.

2.2 Theoretical Framework

2.2.1 Xception

Xception, introduced by [16], represents an advancement of the Inception architecture through the implementation of depthwise separable convolution to enhance computational efficiency without compromising accuracy. This architecture decomposes the conventional convolution operation into two sequential stages: depthwise convolution, applied independently to each input channel, and pointwise convolution, which integrates the resulting feature maps across channels. By factorizing the convolution process in this manner, Xception reduces the number of parameters while maintaining strong generalization capability, particularly when applied to datasets such as FER-2013.

2.2.2 RestNet 50

ResNet-50, introduced by [17], employs a residual learning framework incorporating shortcut connections to address the vanishing gradient problem in very deep neural networks. The architecture consists of 50 layers and utilizes bottleneck blocks composed of a sequence of 1×1, 3×3, and 1×1 convolutional layers. This design enhances optimization efficiency while preserving representational capacity. ResNet-50 has demonstrated strong effectiveness in large-scale image classification and complex visual recognition tasks.

2.3 Research Gap Analysis

Based on the literature review conducted, several limitations remain insufficiently explored in prior studies. A primary gap is the lack of comprehensive evaluations that directly compare the performance of Xception and ResNet-50 architectures in the context of Facial Expression Recognition (FER) under real-world (in-the-wild) conditions. Although previous research has examined the performance of each architecture individually [3][4], only a limited number of studies have provided a direct comparison using equivalent experimental methodologies and comprehensive evaluation metrics, including accuracy, inference time (latency), and runtime efficiency. A thorough understanding of the trade-off between accuracy and computational efficiency is essential for determining the most appropriate architecture for FER systems deployed in practical environments, particularly in mobile applications or real- time systems.

2.4 State of the Art Review

Recent studies have addressed various challenges in Facial Expression Recognition (FER). Study [3] compared traditional CNNs with ResNet-50 on the FER-2013 dataset and reported that ResNet-50 achieved an accuracy of 85.75%, outperforming the standard CNN at 74%, although with higher training and computational costs, indicating a notable trade-off. Sunil and Hariprasad (2023) [4] proposed a hybrid architecture combining Xception and ResNet50V2, demonstrating that the integrated model improved accuracy while enhancing computational efficiency compared to individual architectures. Similarly, Shanimol and Charles (2024) [8] combined ResNet-50 with a Gated Recurrent Unit (GRU), achieving 95.56% accuracy on the CK+ dataset. While the CNN–RNN integration effectively captured temporal dynamics in video-based facial expressions, the model remained computationally intensive and less suitable for real-time image-based applications.

Li and Deng (2019) [2] introduced DLP-CNN (Deep Locality-Preserving CNN), incorporating a locality-preserving loss function to enhance discriminative feature learning, particularly for mixed emotions in real-world scenarios. Furthermore, Sheng and Lau (2024) [9] evaluated ResNet18, ResNet34, and ResNet-50, concluding that ResNet18 provided a more balanced trade-off between accuracy and processing efficiency. Collectively, these findings indicate that no consensus has yet been established regarding the most effective CNN architecture for FER in real-world applications. Therefore, a comprehensive quantitative and qualitative comparison between Xception and ResNet-50 remains essential, given their widespread adoption in practical FER systems

2.5 Experimental Configuration

The training process was conducted on Google Colab utilizing GPU acceleration. Both models were compiled using the Adam optimizer with an initial learning rate of 1e-5. To ensure a fair comparison, identical training parameters were applied to both architectures, including a batch size of 32, 20–30 epochs, and consistent preprocessing procedures. Callbacks such as EarlyStopping, ReduceLROnPlateau, and ModelCheckpoint were implemented to improve training efficiency and mitigate overfitting.

2.6 *Model Evaluation Framework*

BModel evaluation was performed on the validation set using accuracy, precision, recall, F1-score, and confusion matrix as performance metrics. These metrics were employed to assess the predictive performance of the classification model. Accuracy, precision, recall, and F1-score quantify overall and class-level prediction capability, while the confusion matrix provides a detailed representation of classification outcomes based on true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions across classes, as formulated in [5,6,7,8]. In addition, training and validation accuracy curves were analyzed to evaluate model stability throughout the training process [3][18].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

3. RESULTS AND DISCUSSION

3.1 Data Preparation and Preprocessing

FER-2013, which comprises 35,887 grayscale images with a resolution of 48 × 48 pixels. These images are categorized into seven fundamental emotion classes: angry, disgust, fear, happy, sad, surprise, and neutral. Prior to model training, all images were converted to

RGB format and resized to 224 × 224 pixels to meet the input requirements of the Xception and ResNet50 architectures.

Subsequent preprocessing involved image augmentation using the ImageDataGenerator utility from Keras. Distinct augmentation strategies were applied to each model. The Xception model utilized more aggressive augmentation techniques, including rotations of up to 45 degrees, brightness adjustments, horizontal and vertical shifts, zooming, shearing, and horizontal flipping. In contrast, ResNet50 employed a lighter augmentation scheme to maintain training stability. Pixel values were normalized by dividing each value by 255 (rescale = 1./255) to ensure consistent input scaling across all images.

3.2 Model Training and Evaluation

The training process was conducted separately for the Xception and ResNet50 models using consistent configurations, including the Adam optimizer, the sparse_categorical_crossentropy loss function, and a batch size of 64. Training was performed incrementally over 20, 25, and 30 epochs. During the initial training phase of 20 epochs, ResNet50 achieved a higher validation accuracy of 54.55%, whereas Xception attained only 43.55%.

Table 1 Perband Performance Comparison of Xception and ResNet50 Models at 20

Parameter	**Xception**	**ResNet50**
Learning Rate	1e-5	1e-5
Dropout	0.5	0.5
Fine-Tuning	Semua layer freeze	Semua layer freeze
Augmentasi	sedang	sedang
Akurasi Validasi	43.55%	47.55%
Epoch	20	20

Based on Table 1, both models were trained under identical hyperparameter settings, including the same learning rate (1e-5), dropout rate (0.5), fully frozen layers, and moderate data augmentation. Despite these equivalent configurations, ResNet50 demonstrated superior validation performance compared to Xception at 20 epochs, achieving 47.55% versus 43.55%. This result indicates that, under fully frozen conditions and limited training duration, ResNet50 exhibited better initial generalization capability than Xception

Table 2. Performance Comparison of Xception and ResNet50 Models at 25 Epochs

Parameter	**Xception**	**ResNet50**
Learning Rate	1e-4 → 1e-5	1e-4 → 5e-5
Dropout	0.5	0.5
Fine-Tuning	Unfreeze > layer 50	Unfreeze > layer 50
Augmentasi	Menengah	Menengah
Akurasi Validasi	68.72%	63.55%
Epoch	25	25

Based on Table 2, This discrepancy can be attributed to the fact that, at this stage, all layers of the Xception model remained frozen, whereas ResNet50 had already unfrozen several upper layers to enable fine-tuning. However, with the increase in the number of training epochs and the implementation of a more aggressive fine-tuning strategy for Xception, its performance improved substantially. By the 25th epoch, Xception surpassed ResNet50, achieving a validation accuracy of 68.72%, compared to 63.55% for ResNet50.

Table 3. Performance Comparison of Xception and ResNet50 Models at 30 Epochs

Parameter	**Xception**	**ResNet50**
Learning Rate	1e-5 → 1e-6	1e-5 → 5e-6
Dropout	0.5	0.5
Fine-Tuning	Hampir semua layer	Hampir semua layer
Augmentasi	Agresif	Agresif
Akurasi Validasi	70.69%	67.62%
Epoch	30	30

Based on Table 3, summarizes the comparative configuration and performance of the Xception and ResNet50 models after 30 training epochs. Both architectures were trained using a progressively reduced learning rate (1e-5 → 1e-6 for Xception and 1e-5 → 5e-6 for ResNet50), a dropout rate of 0.5, and extensive fine-tuning in which nearly all layers were unfrozen. An aggressive data augmentation strategy was applied to both models.

Based on Table 3, Xception achieved a validation accuracy of 70.69%, outperforming ResNet50, which obtained 67.62% under the same number of epochs. These results indicate

that, with extended training, lower learning rates, and near-complete fine-tuning, Xception demonstrated superior performance compared to ResNet50.

3.3 Training Performance Visualization

This section presents a visual analysis of the training performance for both models across different epochs. The visualization aims to illustrate trends in accuracy and loss during the training and validation phases, thereby providing insight into model convergence behavior, stability, and potential overfitting or underfitting.

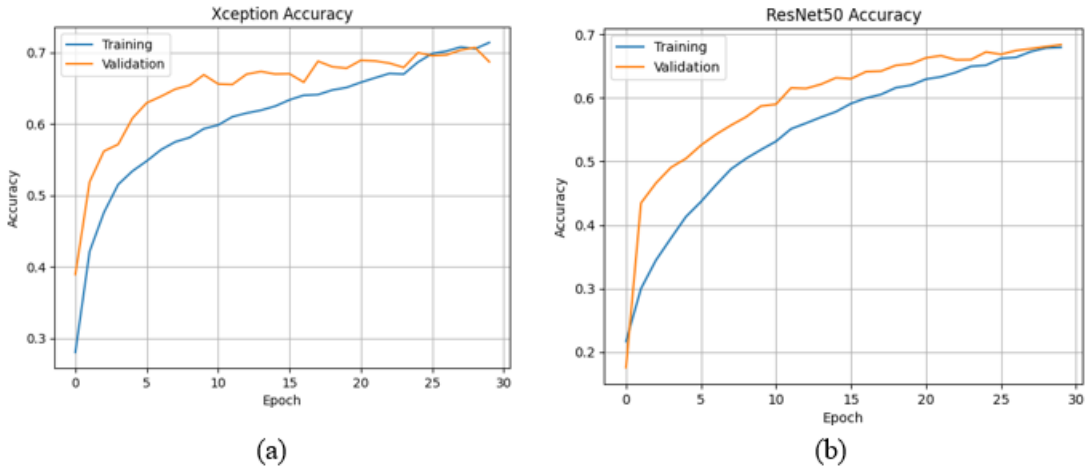


Figure 2 Accuracy Performance Visualization of Xception (a) and ResNet50 (b)

The visualization of the training and validation accuracy curves for both models reveals distinct performance patterns. For the Xception model, the accuracy curve demonstrates a rapid increase up to approximately the 10th epoch, followed by stabilization around 70% after the 20th epoch. No evidence of overfitting is observed, as the validation accuracy consistently remains above or closely aligned with the training accuracy. This pattern indicates that the model effectively learns the underlying data representation while maintaining good generalization capability.

In contrast, ResNet50 also exhibits a sharp improvement in accuracy during the early stages of training; however, its validation accuracy becomes slightly lower than that of Xception beyond the 20th epoch. Although both models display convergence behavior, Xception shows greater stability in sustaining higher performance levels through the final stages of training

3.4 Evaluation Based on Classification Metrics

The models were evaluated using multi-class classification metrics, including accuracy, precision, recall, F1-score, and the confusion matrix. While overall accuracy has been discussed previously, this section emphasizes class-wise performance and the consistency of predictions across models.

The confusion matrix for Xception indicates that the model effectively recognizes the happy class, with approximately 88.8% of images correctly classified. Nevertheless, misclassifications remain evident in visually and emotionally similar categories, particularly fear, which is frequently predicted as sad or neutral. Despite these challenges, Xception demonstrates a well-balanced performance across classes, including minority categories such as disgust. This robustness can be

attributed to the implementation of aggressive data augmentation and extensive fine-tuning strategies.

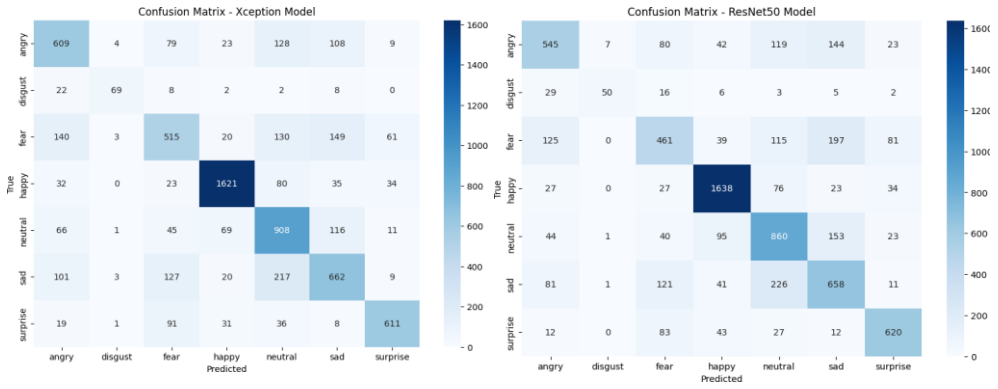


Figure 3. *Confussion Matrix Xception Model (a) dan ResNet50 Model (b)*

ResNet50 likewise achieves strong performance in identifying the happy expression, attaining a prediction accuracy of 89.7%. However, it exhibits a higher rate of misclassification in minority classes, notably disgust and fear, which are more often incorrectly assigned to majority classes such as neutral or sad. Compared to Xception, ResNet50 shows less balanced performance across categories, suggesting greater sensitivity to class imbalance within the dataset

Classification Report (Precision, Recall, F1-Score):

	precision	recall	f1-score	support
angry	0.62	0.63	0.62	960
disgust	0.85	0.62	0.72	111
fear	0.58	0.51	0.54	1018
happy	0.91	0.89	0.90	1825
neutral	0.60	0.75	0.67	1216
sad	0.61	0.58	0.60	1139
surprise	0.83	0.77	0.80	797
accuracy			0.71	7066
macro avg	0.71	0.68	0.69	7066
weighted avg	0.71	0.71	0.71	7066

(a)

Classification Report (ResNet50):

	precision	recall	f1-score	support
angry	0.63	0.57	0.60	960
disgust	0.85	0.45	0.59	111
fear	0.56	0.45	0.50	1018
happy	0.86	0.90	0.88	1825
neutral	0.60	0.71	0.65	1216
sad	0.55	0.58	0.56	1139
surprise	0.78	0.78	0.78	797
accuracy			0.68	7066
macro avg	0.69	0.63	0.65	7066
weighted avg	0.68	0.68	0.68	7066

(b)

Figure 4 Classification Report Xception Model

The class-wise F1-score analysis for the Xception model indicates that the happy class achieved the highest score of 0.90, followed by surprise and neutral. In contrast, the fear class recorded the lowest F1-score at 0.54, suggesting that the model continues to encounter difficulty in consistently recognizing this expression. The macro-average F1-score of 0.69 and weighted-average F1-score of 0.71 demonstrate that Xception delivers stable and balanced performance across most classes, without exhibiting excessive bias toward any particular category.

For the ResNet50 model, the F1-score results reveal strong performance in the happy class, with the highest score of 0.88. However, the F1-scores for the fear and disgust classes are comparatively low, each reaching only 0.45. These findings indicate that despite achieving relatively high overall accuracy, the class-wise performance remains uneven. The macro- average F1-score of 0.63 and weighted-average F1-score of 0.68 are lower than those obtained by Xception, suggesting that ResNet50 is more susceptible to misclassification in minority classes or in expressions that are more subtle and difficult to distinguish

3.5 Latency and Computational Efficiency Evaluation

In addition to evaluating CNN model accuracy, this study also assessed runtime efficiency and inference latency to determine the feasibility of deploying the models in real- time applications, particularly on resource-constrained devices such as mobile or embedded systems. Both Xception and ResNet50 architectures were measured under identical configurations over 30 training epochs, yielding the following average per-epoch times.

Table 4. Training Time Comparison of Xception and ResNet50 over 30 Epochs

Arsitektur	**Rata-rata Waktu per Epoch**	Total Waktu Pelatihan 30 Epoch
Xception	±392 detik	±11.760 detik (~3 jam 16 menit)
ResNet50	±460–500 detik	±13.800–15.000 detik (~3 jam 50 menit – 4 jam 10 menit)

The results indicate that Xception required an average of approximately 392 seconds per epoch (totaling around 3 hours and 16 minutes), whereas ResNet50 required approximately 460–500 seconds per epoch (totaling between 3 hours and 50 minutes and 4 hours and 10 minutes). The training efficiency of Xception is attributed to its use of depthwise separable convolutions, which reduce the number of parameters and computational complexity, in contrast to ResNet50's more complex architecture with multiple residual blocks.

Inference time per image further highlights Xception's advantage, averaging approximately 20–25 milliseconds, compared to 30–35 milliseconds for ResNet50. This discrepancy arises from the higher computational load in ResNet50 due to its larger number of layers and parameters, resulting in slower predictions, particularly for high-resolution images or large batch sizes.

Considering training efficiency, inference latency, and validation accuracy (70.69% for Xception versus 68.38% for ResNet50), Xception demonstrates superior overall performance Consequently, this architecture is more suitable for developing real-time Facial Expression Recognition systems on mobile devices, where computational efficiency is a critical factor.

4. CONCLUSIONS

This study aimed to compare the performance of two Convolutional Neural Network (CNN) architectures, Xception and ResNet50, in the task of human facial expression classification using the FER-2013 dataset. Both models were evaluated under multiple training configurations employing transfer learning, fine-tuning, and structured image augmentation techniques.

Experimental results indicate that the Xception model outperformed ResNet50 in classification tasks. Xception achieved a validation accuracy of 70.69% and a weighted F1-score of 0.71, whereas ResNet50 attained a validation accuracy of 67.62% and a weighted F1-score of 0.68. Moreover, Xception demonstrated greater stability in predictions, particularly for dominant classes such as happy, neutral, and surprise, while also providing relatively accurate predictions for minority classes, including disgust.

The superior performance of Xception can be attributed to its architectural efficiency in spatial feature extraction through depthwise separable convolutions, complemented by comprehensive fine-tuning and controlled yet aggressive image augmentation strategies. In contrast, ResNet50 exhibited strength during early-stage training but proved more sensitive to data variability and class imbalance.

Therefore, Xception is recommended as the preferred model for CNN-based facial expression recognition on the FER-2013 dataset. These findings offer practical implications for the development of efficient and accurate FER systems, particularly for deployment in resource-

constrained environments such as edge computing devices, emotion-aware service systems, and interactive educational or psychological applications.

REFERENCES

[1] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.

[2] R. K. Kumar and S. Vuppala, "Robust Facial Expression Recognition using CNN- based Feature Fusion," *IEEE Access*, vol. 9, pp. 49385–49395, 2021, doi: 10.1109/ACCESS.2021.3069164.

[3] J. He, "Face Expression Recognition (FER) Dataset." 2019. [Online]. Available: https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset

[4] Y. Gao, Y. Cai, X. Bi, B. Li, S. Li, and W. Zheng, "Cross-Domain Facial Expression Recognition through Reliable Global–Local Representation Learning and Dynamic Label Weighting," *Electronics*, vol. 12, no. 21, 2023, doi: 10.3390/electronics12214553.

[5] Z. Zhao and X. Zhang, "Static Facial Expression Recognition via Multi-scale CNN," *J. Vis. Commun. Image Represent.*, vol. 76, 2021, doi: 10.1016/j.jvcir.2021.103115.

[6] A. Dapogny, K. Bailly, and S. Adam, "Confidence-based Adversarial Training for Robust Facial Expression Recognition," *IEEE Trans. Affect. Comput.*, 2021.

[7] S. M. J. S. Priya and S. Umamaheswari, "Deep Convolutional Neural Network-Based Facial Expression Recognition," *Expert Syst. Appl.*, vol. 197, 2022, doi: 10.1016/j.eswa.2022.116696.

[8] H. Sheng and M. Lau, "Optimising Facial Expression Recognition: Comparing ResNet Architectures for Enhanced Performance," in *International Conference on Control, Dynamic Systems and Robotics*, 2024. doi: 10.11159/cdsr24.123.

[9] M. Kaur, "Facial Emotion Recognition: A Comprehensive Review," *Expert Syst.*, vol. 41, no. 10, pp. 105–112, 2024, doi: 10.1111/exsy.13670.

[10] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, 2019, doi: 10.1109/TAFFC.2017.2740923.

[11] J. Yu, Y. Liu, and R. Fan, "MixCut: A Data Augmentation Method for Facial Expression Recognition." 2024. [Online]. Available: http://arxiv.org/abs/2405.10489

[12] S. L. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, 2022, doi: 10.1109/TAFFC.2020.2981446.

[13] W. Li and J. He, "Cross-dataset Facial Expression Recognition via CNN Feature Alignment," *Pattern Recognit. Lett.*, vol. 125, pp. 166–171, 2019, doi: 10.1016/j.patrec.2019.01.009.

[14] A. Shanimol and J. Charles, "ResNet50 and GRU: A Synergistic Model for Accurate Facial Emotion Recognition," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 8, pp. 611– 620, 2024, doi: 10.14569/IJACSA.2024.0150861.

[15] L. Liao, S. Wu, C. Song, and J. Fu, "RS-Xception: A Lightweight Network for Facial Expression Recognition," *Electronics*, vol. 13, no. 16, 2024, doi: 10.3390/electronics13163217.

[16] J. Yang, L. Tang, Z. Chen, X. Zhu, and X. Lai, "Facial Emotion Recognition Based on Optimized Xception Training," in *ACM International Conference Proceeding Series*, 2024. doi: 10.1145/3697355.3697358.

[17] C. W. Wee and S. Budianto, "Improving Facial Expression Recognition with Attention-guided CNN," *Image Vis. Comput.*, vol. 103, 2020

[18] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798– 1828, 2013