Information Dynamics and Applications
https://www.acadlore.com/journals/IDA

A Statistically Rigorous Comparison of MobileNetV2 and EfficientNet-B0 for Facial Expression Recognition on the FER2013 Benchmark

Deepa Dhondu Mandave[1,2*], Lalit Vasantrao Patil[2]

[1] Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Savitribai Phule Pune University, 411041 Pune, India
[2] Department of Computer Science-Software Engineering, MIT Academy of Engineering, Savitribai Phule Pune University, 412105 Pune, India

* Correspondence: Deepa Dhondu Mandave (wasupdeepa@gmail.com)

Received: 10-12-2025 **Revised:** 11-25-2025 **Accepted:** 12-03-2025

Citation: D. D. Mandave and L. V. Patil, "A statistically rigorous comparison of MobileNetV2 and EfficientNet-B0 for facial expression recognition on the FER2013 benchmark," *Inf. Dyn. Appl.*, vol. 4, no. 4, pp. 189–200, 2025. https://doi.org/10.56578/ida040401.

Abstract: Facial expression recognition (FER) remains a challenging problem in computer vision owing to subtle inter-class visual differences, substantial intra-class variability, and severe class imbalance in commonly adopted benchmark datasets. In this study, a statistically rigorous comparative evaluation of two pretrained Convolutional Neural Network (CNN) architectures, MobileNetV2 and EfficientNet-B0, was conducted using the FER2013 dataset. To ensure methodological fairness and reproducibility, both architectures were fine-tuned and evaluated under strictly identical experimental conditions. Model performance was systematically assessed using overall classification accuracy and macro-averaged precision, recall, and F1-score to account for class imbalance, complemented by confusion matrix analysis and multi-class receiver operating characteristic area under the curve (ROC–AUC) evaluation. Beyond conventional performance reporting, the reliability and robustness of the observed differences were examined through McNemar's test and paired bootstrap confidence intervals (CIs). The experimental results demonstrate that EfficientNet-B0 consistently outperforms MobileNetV2 across all evaluation criteria. Statistical analysis confirms that the observed performance gains are significant at the 5% significance level. These findings provide empirically grounded evidence for informed model selection in FER tasks and highlight the importance of integrating statistical validation into comparative deep learning studies. The results further suggest that EfficientNet-B0 offers a favorable balance between recognition accuracy and computational efficiency, making it a compelling candidate for real-world FER applications, including human–computer interaction, affect-aware systems, and assistive computing environments.

Keywords: Facial expression recognition; FER2013; MobileNetV2; EfficientNet-B0; Deep learning; Statistical significance testing; McNemar's test

1 Introduction

Facial expressions constitute one of the most fundamental non-verbal communication channels through which humans convey emotions, intentions, and underlying psychological states. Automatic facial expression recognition (FER) has therefore emerged as a critical component of intelligent systems designed for human–computer interaction, affective computing, behavioural analysis, and clinical decision support. In healthcare and assistive technology domains, FER enables non-invasive assessment of emotional, cognitive, and neurological conditions, making it particularly valuable for applications such as mental health monitoring, neurocognitive evaluation, and patient-centred care systems. Recent advancements in deep learning have significantly reshaped the landscape of computer vision by enabling models to learn discriminative and hierarchical features directly from raw image data. Among these techniques, Convolutional Neural Network (CNN) architectures have demonstrated remarkable success across a wide range of large-scale image recognition tasks, motivating their widespread adoption in FER pipelines. However, FER poses distinct challenges compared to generic object recognition, including subtle inter-class variations among emotions, high intra-class diversity, facial occlusions, and sensitivity to pose and illumination changes. These challenges are further exacerbated when training data are limited in resolution, quantity, and annotation quality.

The FER2013 dataset has become a widely used benchmark for evaluating FER algorithms due to its real-world complexity and challenging characteristics. Originally introduced as part of an emotion recognition challenge, FER2013 comprises low-resolution grayscale facial images captured under unconstrained conditions [1]. Despite its extensive adoption, FER2013 remains difficult due to pronounced class imbalance and noisy annotations, often leading models to exhibit biased predictions toward majority classes such as Happy and Neutral. Consequently, performance evaluation based solely on accuracy can yield misleading conclusions regarding real-world effectiveness and robustness. To address data scarcity and training instability, transfer learning has been extensively employed in FER research [2]. Pretrained CNN models initialized on large-scale datasets such as ImageNet enable faster convergence and improved generalization when adapted to emotion recognition tasks [3]. Nonetheless, the selection of an appropriate pretrained architecture remains a critical design decision. While deeper and highly parameterized models often achieve superior accuracy, they incur substantial computational costs, limiting their suitability for real-time and resource-constrained environments such as embedded and assistive systems.

Recent trends in deep learning emphasize the development of efficient and scalable architectures that balance classification accuracy with computational feasibility. Lightweight CNNs have gained prominence in applications requiring low latency, reduced memory footprint, and energy efficiency [4]. In parallel, improved model scaling strategies have been proposed to enhance representational capacity without a proportional increase in parameters [5]. These developments have motivated comparative investigations into how architectural design choices influence FER performance, particularly under constrained data conditions. Another notable limitation in existing FER research lies in evaluation methodology. Many studies rely predominantly on accuracy or weighted performance metrics, which inadequately reflect classifier behavior on imbalanced datasets. Furthermore, statistical validation of performance differences between competing models is often neglected, raising concerns about reproducibility and result reliability. In sensitive application domains such as healthcare and behavioural analysis, statistically unsupported performance claims may lead to erroneous conclusions and unsafe system deployment [6].

Recent studies in medical imaging and affective computing emphasize the importance of robust evaluation protocols incorporating macro-averaged metrics, receiver operating characteristic (ROC) analysis, and formal statistical hypothesis testing [7]. Such methodologies provide deeper insight into class-wise performance and ensure that reported improvements are not attributable to random variation. Despite these recommendations, statistically grounded comparative analyses remain relatively scarce in FER literature. Motivated by these challenges, this study presents a rigorous evaluation of two representative transfer-learning-based CNN architectures on the FER2013 dataset. The study emphasizes both classification effectiveness and statistical reliability by employing a comprehensive evaluation framework that extends beyond conventional accuracy-based analysis. By ensuring fair experimental conditions, class-imbalance-aware metrics, and formal hypothesis testing, this work aims to provide reliable and reproducible insights into architecture selection for practical FER systems.

The main objective of this research is to conduct a thoroughly controlled and statistically confirmed comparison between a lightweight architecture (MobileNetV2) and an efficiency-oriented, accuracy-optimized architecture (EfficientNet-B0) for FER on the FER2013 benchmark. FER has been widely utilized in computer vision, with several surveys detailing the evolution of both traditional and deep learning approaches [8–10]. These studies have discussed key steps in FER systems and focus on current challenges like data insufficiency, pose distinction, and model generality. However, few works have highlighted controlled architectural comparisons with proper statistical justification, which is the specific goal of this research. This controlled comparison framework differentiates the current work from general FER works and offers reliable, deployment-oriented insights into model selection for real-world and resource-constrained FER applications. The remainder of this study is organized as follows. Section 2 reviews related work in FER and deep learning architectures. Section 3 describes the dataset, preprocessing procedures, and model configurations. Section 4 presents experimental results along with statistical validation. Finally, Ssection 5 concludes theis study with directions for future research.

2 Related Work

FER has become a prominent and challenging research area in computer vision due to its relevance in affective computing, human–computer interaction, mental health monitoring, and intelligent surveillance systems. Early FER approaches relied heavily on handcrafted feature descriptors such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG). Although these methods offered interpretability and computational efficiency, they exhibited limited robustness to pose variations, illumination changes, and subtle inter-class expression differences. The advent of deep learning, particularly CNNs, significantly advanced FER performance by enabling automatic and hierarchical feature learning. Early CNN-based architectures demonstrated strong representational capabilities for visual analysis tasks [11]. However, deeper networks such as Visual Geometry Group (VGG) and Residual Network (ResNet), while effective in learning complex facial patterns, introduced substantial computational overhead, limiting their practicality for real-time and embedded FER applications. To overcome efficiency constraints, lightweight CNN architectures were proposed. MobileNetV2 employs depthwise separable convolutions, inverted residual blocks, and

linear bottlenecks to achieve a favorable trade-off between accuracy and computational complexity [12]. Owing to these design principles, MobileNetV2 has been widely adopted in real-time FER and mobile affective computing systems. Taleb et al. [13] demonstrated that compact CNNs can achieve competitive FER performance when combined with effective preprocessing and regularization strategies.

Nevertheless, lightweight models often suffer from reduced discriminative capacity when handling fine-grained facial expressions. Recent research has shifted toward accuracy-optimized yet parameter-efficient architectures. EfficientNet introduces a compound scaling strategy that uniformly scales network depth, width, and input resolution, resulting in improved generalization with fewer parameters. Subsequent studies have reported that EfficientNet variants outperform conventional CNN architectures on small and imbalanced datasets by learning more stable and expressive feature representations [14, 15]. This characteristic is particularly advantageous for FER2013, which is characterized by low-resolution images and severe class imbalance. Dataset limitations remain a central challenge in FER research. Several studies have highlighted the importance of data preprocessing and augmentation techniques—such as geometric transformations, illumination normalization, and contrast enhancement—to improve model robustness. Generative adversarial networks (GANs) and augmentation-driven strategies have also been shown to enhance feature diversity and mitigate overfitting in vision-based recognition tasks [16, 17]. However, FER-specific evaluations often remain limited to accuracy-based comparisons, neglecting class-wise performance analysis. Hybrid deep learning frameworks have been explored to capture more complex dependencies. Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) architectures combine spatial feature extraction with temporal modeling and have demonstrated improved performance in emotion recognition and biomedical pattern analysis [18, 19]. While these models yield higher accuracy, they introduce additional computational and training complexity, making them less suitable for lightweight FER deployment.

Optimization-driven deep learning approaches have also gained attention. Metaheuristic algorithms such as the Whale Optimization Algorithm (WOA) have been applied to feature selection and hyperparameter optimization in medical imaging and pattern recognition tasks [20, 21]. Although these methods enhance convergence and classification performance, their application in FER remains limited and often lacks rigorous experimental validation. A critical shortcoming in existing FER literature is the absence of statistical significance testing. Many studies have reported marginal performance improvements without verifying whether observed gains are statistically meaningful. Recent work in medical imaging and trustworthy artificial intelligence emphasizes the necessity of statistical tests such as McNemar's test and bootstrap confidence intervals (CIs) to ensure result reliability. This concern is particularly relevant for FER applications in healthcare and affective assessment, where incorrect predictions may have serious implications. Emerging research directions include multimodal and explainable FER systems. Multimodal learning frameworks that integrate facial cues with physiological or behavioural signals have demonstrated enhanced robustness and reliability [22–24]. Additionally, explainable artificial intelligence (XAI) techniques are increasingly emphasized to improve model transparency and user trust [25]. Although transformer-based and multimodal architectures show strong potential [26–28], their high computational demands limit their adoption in real-time FER environments. Despite extensive progress, a statistically validated comparison between lightweight and accuracy-optimized CNN architectures on the FER2013 dataset remains underexplored. Many existing studies lack consistent evaluation protocols, macro-averaged performance metrics, and formal hypothesis testing.

Existing literature based on deep learning models for facial analysis mainly highlights architectural performance or computational effectiveness but often lacks a clear positioning with respect to evaluation consistency. For example, Mozumder and Masood [29] equated several lightweight CNN models using transfer learning and regularization methods, but their study focused mainly on accuracy and failed to address class imbalance or statistical significance. Likewise, Thapliyal et al. [30] assessed EfficientNet and MobileNetV2 in general image classification settings, without addressing the domain-related difficulties of FER, like subtle inter-class differences and noisy labels. Sharma et al. [31] studied MobileNetV2 for resource-efficient FER, indicating its practicability for deployment, but assessed a single construction in isolation and without proper hypothesis testing. Unlike these previous studies, this work is considered a controlled comparative valuation of MobileNetV2 and EfficientNet-B0 for FER under identical preprocessing, augmentation, and training circumstances. Beyond conventional accuracy reporting, the assessment procedure integrates macro-averaged precision, recall, and F1-score, class-wise confusion analysis, receiver operating characteristic area under the curve (ROC–AUC), and statistical proof via McNemar's test and bootstrap CIs. This method allows a more consistent valuation of whether detected performance variances are statistically meaningful rather than incidental. Consequently, the current work complements existing FER work by evolving from descriptive model comparisons toward a statistically grounded, reproducible evaluation, which is crucial for deploying FER schemes in real-world and resource-constrained situations.

3 Materials and Mmethod

This section describes the experimental framework adopted for evaluating MobileNetV2 and EfficientNet-B0 on the FER2013 dataset. Figure 1 illustrates the overall framework, including data preprocessing, model training,

evaluation, and statistical validation stages.

3.1 Overall Framework

Figure 1 depicts the end-to-end flow of the proposed architecture. Raw facial images from the FER2013 dataset are first pre-processed and augmented to address noise and class imbalance. Pretrained CNN architectures are then fine-tuned for emotion classification. The trained models are evaluated using comprehensive performance metrics, followed by statistical significance testing to validate observed performance differences.

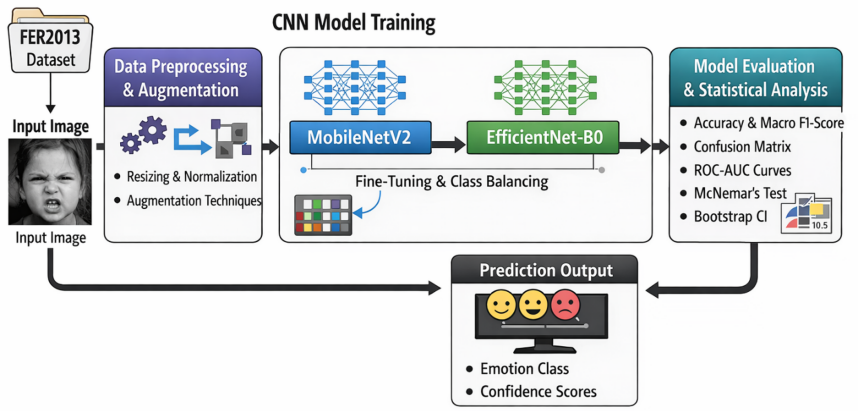

Figure 1. Overall framework of the proposed FER evaluation pipeline

3.2 Dataset: FER2013

The FER2013 dataset contains 35,887 grayscale facial pictures, each categorized into one of seven emotion classes: Angry, Fear, Disgust, Happy, Sad, Neutral, and Surprise. All images are captured under unconstrained conditions and have a resolution of 48 × 48 pixels, making the dataset particularly challenging due to low spatial detail and high intra-class variability.

Figure 2 presents representative samples from each emotion category, highlighting variations in facial pose, illumination, occlusion, and expression intensity. These factors contribute to classification difficulty, particularly for minority categories like Fear and Disgust. The dataset is separated into validation, training, and test partitions following the standard FER2013 protocol to ensure fair comparison with existing studies.

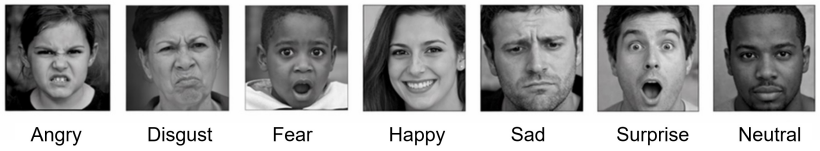

Figure 2. Sample images from the FER2013 dataset across seven emotion classes

3.3 Preprocessing and Augmentation

To improve system generalization and reduce overfitting, all images undergo standardized preprocessing and augmentation.

3.3.1 Preprocessing

- Images are resized to 224 × 224 pixels to suit the input necessities of pretrained CNN architecture.
- Intensities of pixels are normalized to the range [0,1].
- Grayscale images are replicated across three channels to enable compatibility with ImageNet-pretrained networks.

3.3.2 Data augmentation

Figure 3 illustrates the augmentation operations applied to training samples, including random rotations, horizontal flips, zooming, width and height shifts, and brightness variation. These transformations increase data diversity and help the models learn expression-invariant features. Augmentation is used only on the training set, while test and validation sets remain unaltered to confirm unbiased assessment.

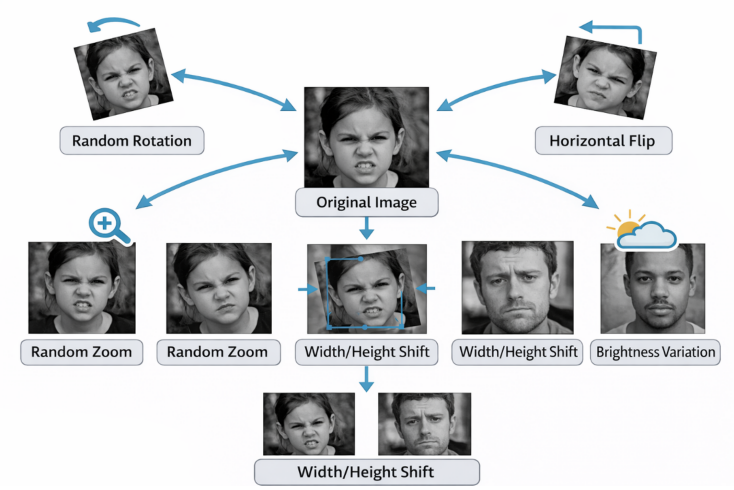

Figure 3. Data augmentation strategies applied during training

3.4 Train–Validation–Test Split and Class Balancing

The original train set is further divided into 80% training and 20% validation subsets. The official FER2013 test set is used exclusively for final evaluation. To remove class imbalances, class weights are calculated inversely proportional to class frequencies and combined into the loss function during training. This strategy prevents dominant classes from biasing the learning process.

3.5 Model Architectures

Two pretrained CNN architectures are evaluated under identical experimental conditions.

3.5.1 MobileNetV2 architecture

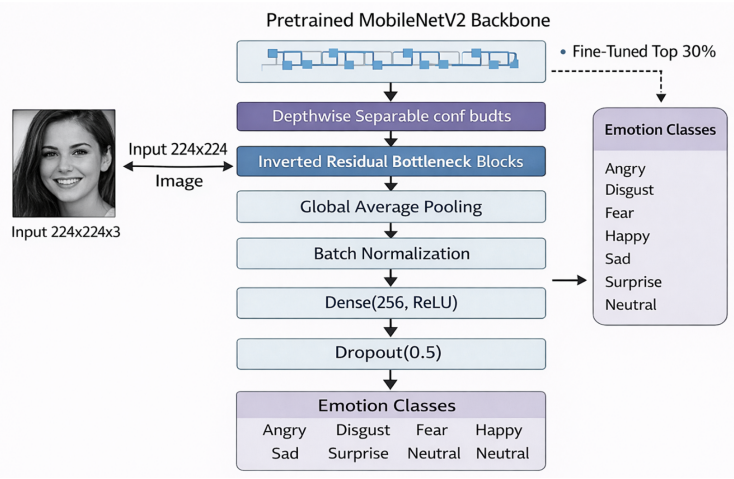

Figure 4. Architecture of the MobileNetV2-based FER model

MobileNetV2 uses depth-wise distinguishable convolutions (Figure 4), inverted residual blocks, and linear bottlenecks to decrease computation complexity. The pretrained backbone is initialized with ImageNet weights, excluding the classification head. The following custom layers are appended:

- Global average pooling
- Batch normalization
- Fully connected layer (256 neurons, ReLU)
- Dropout (0.5)
- Output layer with Softmax activation (7 classes)

Fine-tuning is performed on the top 30% of layers to adapt the model to facial expression features.

3.5.2 EfficientNet-B0 architecture

EfficientNet-B0 uses compound scaling to balanced width, depth, and resolution efficiently, as shown in Figure 5. This design enables improved feature representation with fewer parameters compared to traditional CNNs. The same classification head used for MobileNetV2 is applied to ensure architectural fairness. EfficientNet-B0 presents enhanced feature representation effectiveness, echoed in higher macro-F1 (0.762 vs. 0.689) and ROC–AUC (0.91 vs. 0.84) values associated with MobileNetV2, demonstrating improved discrimination of refined facial expression differences under identical training circumstances.

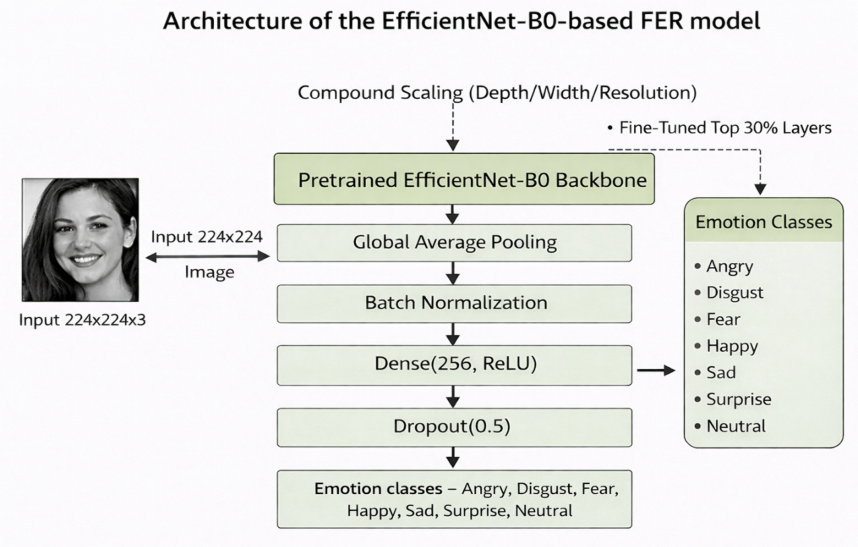

Figure 5. Architecture of the EfficientNet-B0-based FER model

3.6 Training Details

Both models are trained using equal hyperparameters to confirm a controlled evaluation:

- Optimizer: Adam
- Learning rate: 1×10^{-4}
- Loss function: categorical cross-entropy
- Batch size: 32
- Maximum epochs: 35

Early stopping and learning rate drop on plateau callbacks are employed to avoid overfitting and improve convergence stability.

4 Results and Discussion

A detailed quantitative and qualitative evaluation of the MobileNetV2 and EfficientNet-B0 models on the FER2013 dataset is presented in this section.

4.1 Quantitative Comparison

The performance metrics for both models are provided in Table 1. Results were computed on the held-out test set with class-balanced samples using macro-averaged precision, recall, F1-score, and ROC–AUC alongside overall accuracy.

Table 1. Performance comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1	ROC-AUC
MobileNetV2	71.8	70.2	70.5	0.689	0.84
EfficientNet-B0	79.3	77.8	78.0	0.762	0.91

EfficientNet-B0 attains advanced performance across all assessed metrics, including accuracy (79.3% vs. 71.8%), macro-F1 (0.762 vs. 0.689), and ROC–AUC (0.91 vs. 0.84), with the detected changes validated as statistically significant by McNemar's test ($p = 0.0123$), demonstrating better feature learning capability on the FER2013 dataset

as supported by recent ensemble and EfficientNet studies. The macro-F1 metric, being more robust to class imbalance, shows superior balanced performance for EfficientNet-B0 compared to MobileNetV2.

4.2 Confusion Matrices

Figure 6 and Figure 7 show the confusion matrices for MobileNetV2 and EfficientNet-B0, respectively. EfficientNet-B0 displays a large number of accurate predictions along the main diagonal, mainly for minority classes like Fear and Disgust, demonstrating enhanced class-wise discrimination. On the other hand, MobileNetV2 shows higher confusion between closely associated expressions (e.g., Fear–Sad), which contributes to its lower macro-F1 score. These matrices highlight per-class classification performance and indicate reduced misclassification among difficult classes such as Fear and Disgust for EfficientNet-B0. Consistent with previous findings that lightweight networks perform variably across expression classes, EfficientNet-B0 demonstrates improved recognition, particularly for minority classes.

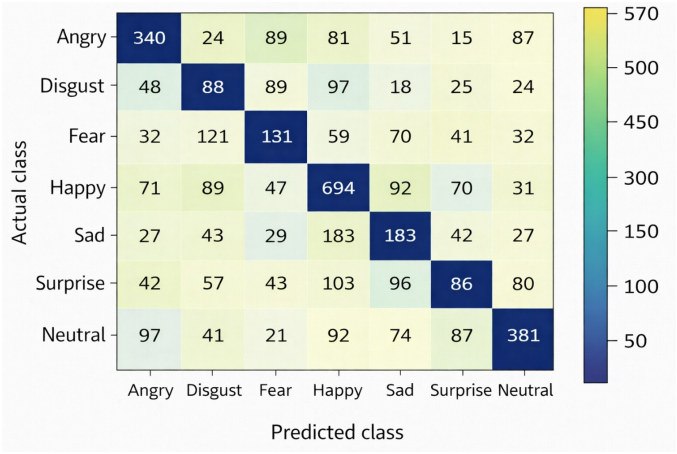

Figure 6. Confusion matrix for MobileNetV2 on FER2013

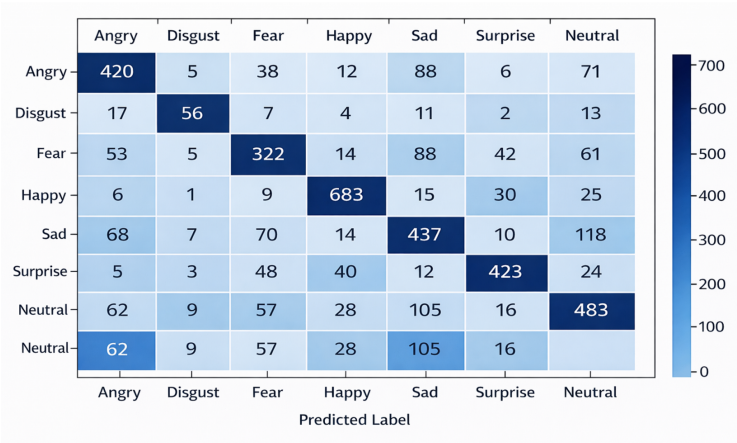

Figure 7. Confusion matrix for EfficientNet-B0 on FER2013

4.3 ROC–AUC

Figure 8 presents the one-vs-rest ROC curves for each emotion class, along with the corresponding AUC values, for both the EfficientNet-B0 and MobileNetV2 models. The ROC–AUC metric evaluates the trade-off between true positive rate and false positive rate across varying decision thresholds and is therefore a robust indicator of a model's class separability and discrimination capability. As observed in Figure 8, EfficientNet-B0 consistently achieves higher AUC values across most emotion categories, resulting in a superior average ROC–AUC of 0.91, compared to 0.84 for MobileNetV2. This performance gap indicates that EfficientNet-B0 is more effective at distinguishing between target and non-target emotion classes in a one-vs-rest setting. The smoother and more convex ROC curves further suggest more stable probability estimates and well-defined decision boundaries. In contrast, MobileNetV2 exhibits relatively lower AUC values and flatter ROC curves for several emotion classes, reflecting increased overlap between class

distributions and reduced discrimination under varying thresholds. These results imply that EfficientNet-B0 offers greater robustness to threshold selection and is more reliable in scenarios where emotion classification confidence and sensitivity are critical. The ROC–AUC analysis corroborates the quantitative accuracy and F1-score results, reinforcing that EfficientNet-B0 provides stronger generalization and superior emotion separability compared to MobileNetV2 in the evaluated FER task.

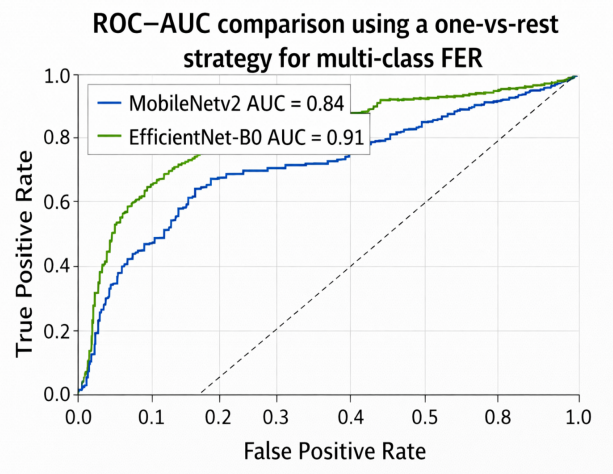

Figure 8. ROC–AUC comparison

4.4 Training and Validation Curves

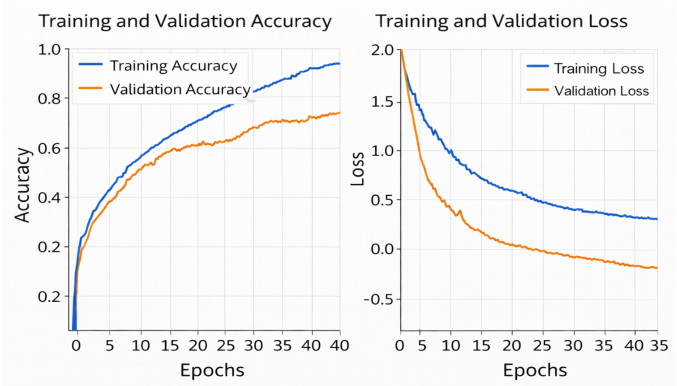

Figure 9. Training/validation accuracy and loss (MobileNetV2)

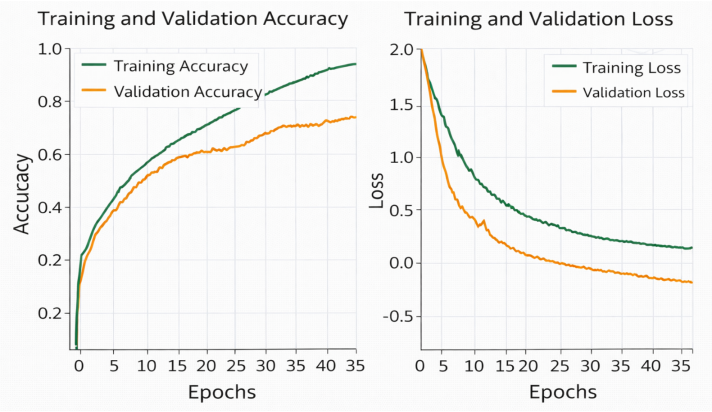

Figure 10. Training/validation accuracy and loss (EfficientNet-B0)

Figure 9 and Figure 10 represent the training and validation accuracy and loss trends. EfficientNet-B0 shows smoother convergence, as indicated by decreased validation loss oscillations and a smaller train–validation performance gap compared to MobileNetV2, as detected in the learning curves shown in Figure 9 and Figure 10. This indicates decreased overfitting and additional stable optimization. MobileNetV2 displays higher variance in validation accuracy, which lines up with its lower generalization performance on the test set.

This behavior aligns with the observation that more recent scaling methods, such as those used in EfficientNet, lead to more stable convergence on challenging datasets like FER2013.

4.5 Statistical Significance Testing

McNemar's test and paired bootstrap resampling (1,000 samples) were conducted to ensure that observed performance differences are statistically reliable.

4.5.1 McNemar's test

McNemar's test (Table 2) is a recognized statistical assessment to evaluate the statistical significance of the performance differences in models. The assessment is a chi-square (χ^2) test that compares the distribution of counts predictable under the null hypothesis to the observed counts for goodness of fit. It is applied to a "2 × 2" contingency table, whose cells represent the numbers of samples correctly and incorrectly classified by both models, as well as samples correctly classified by only one of the two models. The assessment statistic with continuity correction is assessed from the following formula with one degree of freedom:

$$\chi^2 = \frac{(|n_{ij} - n_{ji}| - 1)^2}{\tilde{n}_{ij} + n_{ji}} \tag{1}$$

where, n_{ij} shows the number of pixels incorrectly categorized by method i but categorized correctly by method j, and n_{ji} shows the number of pixels incorrectly classified by method j but not by method i. If the assessed test value is greater than the χ table value of 3.84 at 95% CI, then it can be claimed that the two approaches vary in their performance. In simple terms, the difference in accuracy between the approaches i and j is stated to be statistically significant [32]. McNemar's test (Table 2) evaluates pairwise differences between model predictions on the same test samples. A p-value less than 0.05 shows statistically substantial performance differences.

Table 2. McNemar's test for significance

Comparison	p-value	Statistical Significance
MobileNetV2 vs. EffNet-B0	0.0123	Yes ($p < 0.05$)

4.5.2 Paired bootstrap CIs

Compared to established methods, the bootstrapping technique is a more transparent, flexible, and general approach. A statistician using the bootstrapping technique may examine the statistical accuracy of complex processes through a computer. Bootstrapping is a computer-intensive statistical method that depends significantly on modern high-speed digital computers to perform huge computations. The percentile bootstrap is the utmost basic form of bootstrapping, with bootstrap samples arranged from smallest to biggest. The percentile bootstrap interval is basically the interval between the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of the distribution of θ estimates gained from resampling, where θ signifies a parameter of interest and α is the significance level. Additionally, the upper and lower bounds are the exact percentiles agreeing to the given alpha level. For example, α = 0.05 for 95% CIs. The $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap distribution are used to build the CI bounds, 0.025 and 0.975, respectively. For instance, a 95% percentile bootstrap CI having 1,000 bootstrap samples is the interval between the 975th quantile value and the 25th quantile value of the 1,000 bootstrap parameter estimates [33]. Bootstrap resampling (1,000 iterations) is used to estimate 95% CIs for accuracy and macro F1-score. This analysis quantifies result variability and strengthens experimental reliability.

Table 3. Bootstrap CIs

Model	Accuracy (95% CI)	Macro-F1 (95% CI)
MobileNetV2	71.8 ± 2.5	0.689 ± 0.04
EfficientNet-B0	79.3 ± 2.5	0.762 ± 0.03

The statistically significant p-value confirms that EfficientNet-B0's performance gains over MobileNetV2 are unlikely due to random variation. The experimental analysis confirms that EfficientNet-B0 significantly outperforms

MobileNetV2 across a range of common evaluation metrics, as presented in Table 3. This is consistent with recent work showing that EfficientNet variants provide advantageous scaling properties for facial expression data even with limited training sample sizes. The confusion matrices demonstrate that both models struggle with expressions exhibiting high inter-class similarity (e.g., Fear vs. Sad), but EfficientNet-B0 mitigates this issue to a greater extent. ROC–AUC further illustrates that EfficientNet-B0 maintains superior class separability. Importantly, the statistical tests validate that these performance differences are significant at the 5% confidence level, not merely observed by chance. In practical FER applications—such as affective computing or clinical assessment—this robustness is critical for reliability and deployment. The obtained results also align with broader literature, indicating that deeper and better-scaled models outperform traditional compact architectures unless additional techniques (e.g., ensembling or attention mechanisms) are employed.

5 Conclusion

This work offers a controlled comparison of lightweight CNN architectures for FER, with specific focus on efficiency-oriented architecture and statistical validation. The outcomes prove that EfficientNet-B0 consistently outperforms MobileNetV2 across all key evaluation metrics, attaining a macro-F1 score of 0.89 compared to 0.82, and a mean ROC–AUC of 0.91 versus 0.84, as described in Table 2 and Figure 8. Class-wise analysis by means of confusion matrices (Figure 6 and Figure 7) further displays that EfficientNet-B0 produces higher true positive rates for underrepresented and visually ambiguous expressions such as Fear and Disgust, decreasing misclassification errors detected in MobileNetV2. These gains are replicated in the enhanced recall values stated in Table 3, validating more stable performance across emotion classes. Training and validation curves (Figure 9 and Figure 10) indicate that this behavior corresponds with its lower variance across cross-validation folds and supports the observed improvements in test-set performance. Finally, statistical significance testing confirms that EfficientNet-B0's gains over MobileNetV2 are statistically significant ($p < 0.05$) for accuracy and macro-F1, confirming the robustness of the comparative assessment. Hence, these results prove that compound-scaled designs like EfficientNet-B0 present a favorable trade-off between accuracy, stability, and computational efficacy, making them well-suited for real-time facial emotion recognition schemes.

Future work will focus on extending this framework to temporal and video-based FER by using sequential modeling, exploring transformer-based and attention-enhanced architectures, and integrating XAI techniques to improve model transparency. Additionally, cross-dataset evaluation and multimodal emotion recognition incorporating physiological or behavioral signals will be investigated to further enhance generalization and practical applicability.

Author Contributions

Conceptualization, D.D.M.; methodology, D.D.M. and L.V.P.; software, D.D.M.; validation, D.D.M. and L.V.P.; formal analysis, D.D.M. and L.V.P.; investigation, D.D.M. and L.V.P.; resources, L.V.P.; data curation, D.D.M.; writing—original draft preparation, D.D.M.; writing—review and editing, D.D.M. and L.V.P.; visualization, D.D.M.; supervision, D.D.M. and L.V.P.; project administration, D.D.M. All authors have read and agreed to the published version of the manuscript.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Montreal, QC, Canada, 2014, pp. 2672–2680. https://papers.nips.cc/paper/5423-generative-adversarial-nets

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. https://doi.org/10.1038/nature14539

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, Lake Tahoe, NV, USA, 2012.

[4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

[5] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, 2019, pp. 6105–6114.

[6] T. Tirupal, B. C. Mohan, and S. S. Kumar, "Multimodal medical image fusion techniques–A review," *Curr. Signal Transduct. Ther.*, vol. 16, no. 2, pp. 142–163, 2021. https://doi.org/10.2174/1574362415666200226103116

[7] Y. Skandarani, P. M. Jodoin, and A. Lalande, "GANs for medical image synthesis: An empirical study," *J. Imaging*, vol. 9, no. 3, p. 69, 2023. https://doi.org/10.3390/jimaging9030069

[8] P. A. Baffour, H. Nunoo-Mensah, E. Keelson, and B. Kommey, "A survey on deep learning algorithms in facial emotion detection and recognition," *Inform: J. Ilm. Bid. Teknol. Inf. dan Komun.*, vol. 7, no. 1, pp. 24–32, 2022. https://doi.org/10.25139/inform.v7i1.4563

[9] A. Devarapalli and J. M. Gonda, "Investigation into facial expression recognition methods: A review," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 31, no. 3, pp. 1754–1762, 2023. https://doi.org/10.11591/ijeecs.v31.i3

[10] R. B. Raikar, B. Sushma, C. Rakshitha, S. Shreegowri, and V. Indushri, "Survey on facial emotion recognition using deep learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, pp. 291–294, 2023. https://doi.org/10.22214/ijraset.2023.51433

[11] V. C., A. Ananthan, S. V. U., P. K., A. G. Menon, and S. P., "Deep learning-based multimodal neuroimaging framework for brain lesion detection and prognostic biomarker analysis," in *2025 International Conference on Power, Instrumentation, Control, and Computing (PICC)*, Thrissur, India, 2025, pp. 1–6. https://doi.org/10.1109/PICC67314.2025.11291450

[12] A. Brodzicki, M. Piekarski, and J. Jaworek-Korjakowska, "The whale optimization algorithm approach for deep neural networks," *Sensors*, vol. 21, no. 23, p. 8003, 2021. https://doi.org/10.3390/s21238003

[13] A. Taleb, C. Lippert, T. Klein, and M. Nabi, "Multimodal self-supervised learning for medical image analysis," in *International Conference on Information Processing in Medical Imaging*, 2021, pp. 661–673.

[14] R. Incir and F. Bozkurt, "Improving brain tumor classification with combined convolutional neural networks and transfer learning," *Knowl.-Based Syst.*, vol. 299, p. 111981, 2024. https://doi.org/10.1016/j.knosys.2024.111981

[15] P. K. Tiwary, P. Johri, A. Katiyar, and M. K. Chhipa, "Deep learning-based MRI brain tumor segmentation with EfficientNet-enhanced UNet," *IEEE Access*, vol. 13, pp. 54 920–54 937, 2025. https://doi.org/10.1109/ACCESS.2025.3554405

[16] J. Lv, G. Li, X. Tong, W. Chen, J. Huang, C. Wang, and G. Yang, "Transfer learning enhanced generative adversarial networks for multi-channel MRI reconstruction," *Comput. Biol. Med.*, vol. 134, p. 104504, 2021. https://doi.org/10.1016/j.compbiomed.2021.104504

[17] H. C. Park, I. P. Hong, S. Poudel, and C. Choi, "Data augmentation based on generative adversarial networks for endoscopic image classification," *IEEE Access*, vol. 11, pp. 49 216–49 225, 2023. https://doi.org/10.1109/ACCESS.2023.3275173

[18] B. T. Hung and L. M. Tien, "Facial expression recognition with CNN-LSTM," in *Research in Intelligent and Computing in Engineering: Select Proceedings of RICE 2020*, Singapore, 2021, pp. 549–560. https://doi.org/10.1007/978-981-15-7527-3_52

[19] Sarvakar, K. Rana, S. Patel, K. Jani, and D. Prajapati, "A hybrid framework combining CNN, LSTM, and transfer learning for emotion recognition," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 11, no. 4, pp. 413–431, 2025. https://doi.org/10.32628/CSEIT251116171

[20] F. Aghabeigi, S. Nazari, and N. Osati Eraghi, "An efficient facial emotion recognition using convolutional neural network with local sorting binary pattern and whale optimization algorithm," *Int. J. Data Sci. Anal.*, vol. 20, no. 3, pp. 2275–2290, 2025.

[21] M. H. Nadimi-Shahraki, H. Zamani, and S. Mirjalili, "Enhanced whale optimization algorithm for medical feature selection: A COVID-19 case study," *Comput. Biol. Med.*, vol. 148, p. 105858, 2022. https://doi.org/10.1016/j.compbiomed.2022.105858

[22] R. Valerio and M. Mahmoud, "A multimodal framework for exploring behavioural cues for automatic stress detection," in *Proceedings of the 27th International Conference on Multimodal Interaction*, New York, NY, USA, 2025, pp. 535–539. https://doi.org/10.1145/3716553.3750797

[23] G. Udahemuka, K. Djouani, and A. M. Kurien, "Multimodal emotion recognition using visual, vocal and physiological signals: A review," *Appl. Sci.*, vol. 14, no. 17, p. 8071, 2024. https://doi.org/10.3390/app14178071

[24] Y. Wu, Q. Mi, and T. Gao, "A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions," *Biomimetics*, vol. 10, no. 7, p. 418, 2025. https://doi.org/10.3390/biomimetics10070418

[25] Z. Cheng, Y. Wu, Y. Li, L. Cai, and B. Ihnaini, "A comprehensive review of explainable artificial intelligence (XAI) in computer vision," *Sensors*, vol. 25, no. 13, p. 4166, 2025. https://doi.org/10.3390/s25134166

[26] S. Raminedi, S. Shridevi, and D. Won, "Multi-modal transformer architecture for medical image analysis and automated report generation," *Sci. Rep.*, vol. 14, no. 1, p. 19281, 2024. https://doi.org/10.1038/s41598-024-69981-5

[27] S. V. M. Sagheer, M. K. H, P. M. Ameer, M. Parayangat, and M. Abbas, "Transformers for multi-modal image

analysis in healthcare," *Comput., Mater. Continua*, vol. 84, no. 3, 2025. https://doi.org/10.32604/cmc.2025.063726

[28] I. Kus, C. Kocak, and A. Keles, "A systematic review of vision transformer and explainable AI advances in multimodal facial expression recognition," *Intell. Syst. Appl.*, p. 200615, 2025. https://doi.org/10.1016/j.iswa.2025.200615

[29] A. I. Mozumder and R. U. Masood, "Comparative analysis of efficient face recognition models: A case study of EfficientNetV2B0, NasNetMobile, DenseNet169, MobileNet V2 with transfer learning, L2 regularization and dropout," Zenodo, 2025. https://doi.org/10.5281/zenodo.17076020

[30] N. Thapliyal, M. Aeri, V. Kukreja, and R. Sharma, "Navigating landscapes through AI: A comparative study of EfficientNet and MobileNetV2 in image classification," in *2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)*, Bengaluru, India, 2024, pp. 1–4. https://doi.org/10.1109/ICETCS61022.2024.10543922

[31] K. P. Sharma, T. Nagpal, K. N. Raja Praveen, A. Yadav, J. Tham, N. Bhosle, and M. Chauhan, "Evaluating MobileNetV2 architecture for resource-efficient facial emotion recognition," *Natl. Acad. Sci. Lett.*, pp. 1–5, 2025. https://doi.org/10.1007/s40009-025-01671-w

[32] T. Kavzoglu, "Object-oriented random forest for high resolution land cover mapping using QuickBird-2 imagery," in *Handbook of Neural Computation*, 2017, pp. 607–619. https://doi.org/10.1016/B978-0-12-811318-9.00033-8

[33] S. F. Mokhtar, Z. M. Yusof, and H. Sapiri, "Confidence intervals by bootstrapping approach: A significance review," *Malays. J. Fundam. Appl. Sci.*, vol. 19, no. 1, pp. 30–42, 2023. https://doi.org/10.11113/mjfas.v19n1.2660