Journal of Soft Computing Exploration
Homepage: htttps://shmpublisher.com/index.php/joscex
p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Benchmarking mobileNetV3 and efficientNet-B0 for corn leaf disease classification with imbalanced dataset using stratified cross-validation

Muhammad Shandy Alfarizal[1], Muhamad Kelvin Saputra[2], Ade Fajar Kurniawan[3], Diki Wahyudi[4], Khanahaya Fadhil Adriano[5], Anindita Septiarini[6], Novianti Puspitasari[7]
[1,2,3,4,5,6,7]Department of Informatics, Universitas Mulawarman, Indonesia

Article Info

Article history:

Received Mar 26, 2026
Revised April 12, 2026
Accepted April 20, 2026

Keywords:

Corn leaf disease
Imbalanced dataset
Transfer learning
Computer vision
Deep learning

ABSTRACT

Corn leaf diseases pose a serious threat to crop productivity, yet most publicly available datasets for this task exhibit severe class imbalance that can mislead conventional accuracy-based evaluation. This study benchmarks two lightweight transfer learning architectures, MobileNetV3-Large and EfficientNet-B0, for multi-class corn leaf disease classification on the Seasonal Corn Leaf Disease Dataset from Mendeley Data 2025 containing 2,943 images across five imbalanced classes. Evaluation was conducted using Stratified 5-Fold Cross-Validation with Macro-F1 as the primary metric, complemented by per-class analysis through aggregated out-of-fold predictions. Class weights were applied to the CrossEntropyLoss function as a fixed experimental control for class imbalance, with the primary objective being the benchmarking of the two architectures rather than the comparison of imbalance-handling strategies. The experimental results revealed that EfficientNet-B0 consistently outperformed MobileNetV3, achieving a Macro-F1 of 0.9778 and an accuracy of 0.9796 with lower variance across folds. Error analysis through the OOF confusion matrix and a misclassification gallery confirmed that persistent errors predominantly occurred between Gray Leaf Spot and Healthy classes, particularly on early-symptom images captured under inconsistent lighting conditions.

Corresponding Author:

Anindita Septiarini,
Department of Informatics,
Faculty of Engineering, Universitas Mulawarman,
Kuaro Street, Gunung Kelua, Samarinda Ulu, Samarinda, East Kalimantan 75117, Indonesia
Email: anindita@unmul.ac.id
https://doi.org/10.52465/joscex.v7i2.30

1. INTRODUCTION

Indonesia ranks among the largest corn producers in Southeast Asia. According to Statistics Indonesia, national dry-shelled corn production reached 15.14 million tons from 2.55 million hectares of harvested area in 2024, reflecting a 2.47% year-on-year increase [1]. Despite this upward trend, productivity remains threatened by foliar diseases, including leaf blight, rust, and downy mildew that can reduce yields by 50 to 100% in susceptible varieties [2], [3]. Conventional disease identification relies on manual field inspection, a process that is inherently subjective, inconsistent, and heavily dependent on the observer's expertise [4], [5].

These limitations highlight the urgent need for automated classification systems that are faster, more objective, and reproducible.

Deep learning, particularly Convolutional Neural Networks, has emerged as a powerful approach for plant disease image classification [6], [7]. Transfer learning further accelerates this process by leveraging pretrained weights from large-scale datasets such as ImageNet, enabling strong feature extraction even on moderately sized domain-specific datasets [8], [9]. Among the available architectures, MobileNetV3 and EfficientNet-B0 stand out for their balance between accuracy and computational efficiency, making them suitable candidates for deployment on resource-constrained devices [10], [11]. Several prior studies have explored these and other architectures for corn leaf disease detection. A modified MobileNetV3 with attention modules achieved 98.23% accuracy on corn leaf data [12], while an enhanced EfficientNet-B0 incorporating CBAM and multi-scale fusion reached 98.32% accuracy [13]. Real-time maize disease detection using YOLOv8n reported 99.04% accuracy [14]. In related agricultural domains, a comparative study of YOLOv5 and YOLOv8 on rice leaf disease datasets demonstrated the superiority of YOLOv8 with mAP50 of 0.924 [15], and a benchmarking study of four transfer learning architectures on imbalanced skin cancer data using Focal Loss and Stratified Group-KFold cross-validation identified ConvNeXt-Tiny as the most stable model [16]. In the medical imaging domain, the integration of DenseNet-169 with CBAM for tuberculosis classification achieved 99.43% accuracy [17]; nevertheless, a common limitation across these studies is their reliance on single-split evaluation with accuracy as the primary metric, which reduces reliability when underlying data is imbalanced [18], [19].

Despite these advances, two research gaps remain. First, a direct head-to-head comparison between vanilla MobileNetV3 and EfficientNet-B0 on multi-class corn leaf disease data under rigorous cross-validation protocols has not been reported. Second, systematic error analysis that connects confusion matrix patterns to the visual characteristics of misclassified images is still rarely explored in the corn disease domain.

This study addresses these gaps using the Seasonal Corn Leaf Disease Dataset from Mendeley Data 2025, comprising 2,943 images distributed across five classes: Healthy with 1,038 images, Bacterial Leaf Streak with 190, Common Rust with 129, Gray Leaf Spot with 1,497, and MCMV with 89 [2]. The extreme imbalance, Gray Leaf Spot alone accounts for over half the dataset while MCMV contributes only 3% — necessitates the adoption of Macro-F1 as the primary metric, which treats all classes equally regardless of their sample size [18], [20]. Stratified 5-Fold Cross-Validation ensures proportional class representation in every fold and reduces split-dependent bias [21], [22].

Specifically, this study offers three contributions: a head-to-head comparison of MobileNetV3 and EfficientNet-B0 under Stratified 5-Fold cross validation with a fixed seed for reproducibility; the adoption of Macro-F1 with mean and standard deviation reporting across folds to assess both performance level and consistency; and an error analysis grounded in aggregated OOF confusion matrices and a visual misclassification gallery that traces prediction errors back to data-level characteristics. This evaluation framework, in which class weighting serves as a fixed control rather than a subject of comparison, aligns with recent benchmarking practices in transfer learning [16] and plant disease detection studies emphasizing precision, recall, F1, and confusion matrix analysis [14], [15], [23].

2. METHOD

This section describes the complete experimental pipeline from data acquisition through model evaluation. The overall workflow is illustrated in Figure 1. After preprocessing, Stratified 5-Fold Cross-Validation splits the data while preserving class proportions in each fold. Both models are trained per fold, and the resulting validation predictions are concatenated into a single out-of-fold set so that every sample is evaluated exactly once in a held-out context [22].

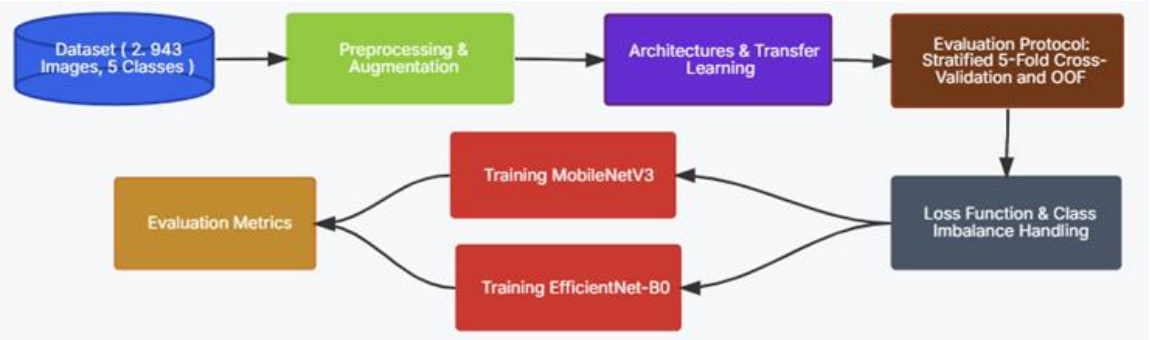

Figure 1. Workflow of corn leaf disease classification

Dataset

The Seasonal Corn Leaf Disease Dataset was collected from several agricultural fields in Gurudaspur, Natore, Rajshahi, Bangladesh, and published through the Mendeley Data repository in 2025 [2]. It contains 2,943 images distributed across five categories as shown in Table 1.

Table 1. Class distribution in the dataset

Class	Amount	Proportion
Healthy	1.038	35,27%
Bacterial Leaf Streak	190	6,46%
Common Rust	129	4,38%
Gray Leaf Spot	1.497	50,87%
MCMV	89	3,02%
Total	2.943	100%

Gray Leaf Spot dominates more than half the dataset while MCMV contributes only 89 images. This skewness motivates both the adoption of Macro-F1 as the primary metric and the application of class weights in the loss function [18], [20]. Representative raw samples from each class are shown in Figure 2. Several classes share overlapping symptom characteristics that make visual separation challenging, particularly at early stages [7], [14]. These similarities are discussed further in the error analysis section.

Figure 2. Raw image samples per class

Preprocessing and Augmentation

All images were resized to 224×224 pixels and normalized using ImageNet statistics [8], [24]. Three augmentation techniques were applied exclusively to the training subset: RandomHorizontalFlip with probability 0.5, RandomRotation within a 15-degree range, and ColorJitter with brightness and contrast variation of 0.15 each [25]. Validation data underwent only resizing and normalization. Rotation was bounded at 15 degrees to prevent unrealistic distortion of lesion morphology, and jitter parameters were kept low to simulate natural field lighting variation [7].

Architectures and Transfer Learning

Two architectures pretrained on ImageNet were selected for direct comparison: MobileNetV3-Large and EfficientNet-B0, with their final classification layers replaced to produce five-class outputs. Both models have roughly comparable parameter counts, approximately 5.4 million for MobileNetV3 and 5.3 million for EfficientNet-B0, yet they differ fundamentally in design philosophy. MobileNetV3 relies on inverted residual

blocks combined with Squeeze-and-Excitation mechanisms and h-swish activation to minimize computational cost, an approach that has proven effective in various plant disease studies [10], [26]. EfficientNet-B0, by contrast, employs compound scaling to simultaneously balance network depth, width, and resolution, yielding richer feature representations without excessive parameter growth, and has been adopted in image-based detection systems across multiple domains [11], [27]. Both architectures were chosen because they remain competitive in plant disease image classification despite their relatively small footprint [12], [13], [28]. The detailed layer-by-layer structures of both architectures as configured in this study are illustrated in Figure 3(a) for MobileNetV3-Large and Figure 3(b) for EfficientNet-B0.

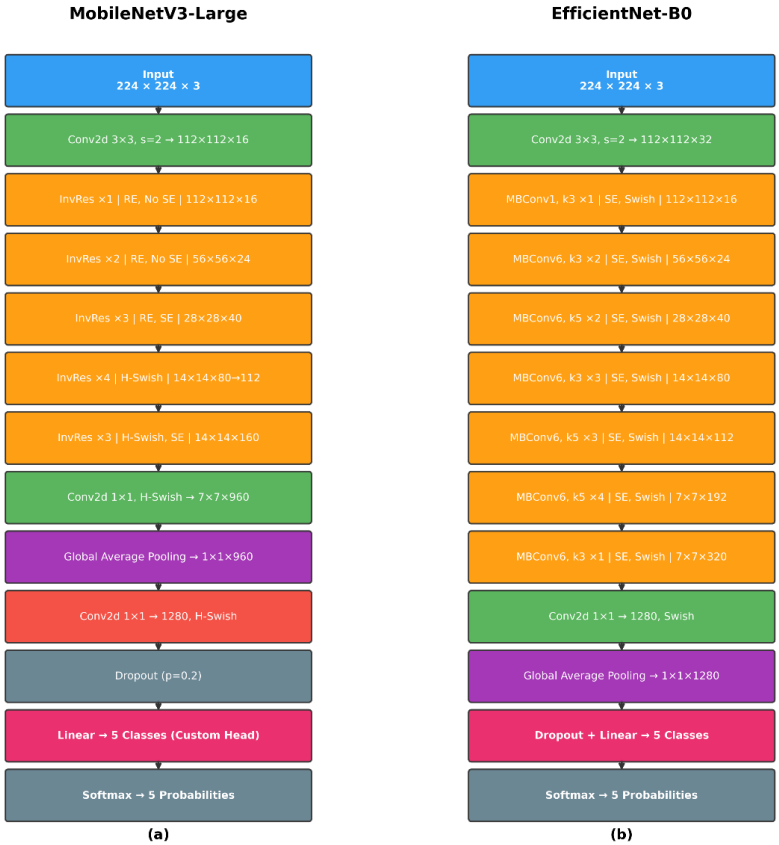

Figure 3. Model architectures: (a) MobileNetV3-Large, (b) EfficientNet-B0

As shown in Figure 3(a), MobileNetV3-Large begins with a standard 3×3 convolution with stride 2 that reduces the spatial resolution by half, followed by a series of inverted residual blocks organized into seven stages. The earlier stages employ the RE activation function without Squeeze-and-Excitation, while the deeper stages transition to H-Swish activation with SE modules enabled to enhance channel-wise feature recalibration. A 1×1 convolution expands the feature maps to 960 channels before global average pooling compresses the spatial dimensions. The original classifier was replaced with a custom head consisting of a 1×1 convolution projecting to 1,280 dimensions, a dropout layer with rate 0.2, and a fully connected layer producing five class outputs.

As shown in Figure 3(b), EfficientNet-B0 similarly starts with a 3×3 convolution with stride 2 but outputs 32 channels. The backbone consists of seven stages of Mobile Inverted Bottleneck Convolution blocks with expansion ratio 6, alternating between 3×3 and 5×5 kernel sizes. All MBConv blocks incorporate SE modules and Swish activation throughout the entire network, unlike MobileNetV3 which selectively enables these components. A 1×1 convolution expands the final features to 1,280 channels followed by global average pooling. The custom classification head consists of a dropout layer and a linear layer mapping to five output classes.

Both architectures share a comparable total parameter count of approximately 5.3 to 5.4 million, yet their structural differences in activation functions, SE module placement, and scaling strategy lead to distinct feature extraction behaviors that are reflected in the experimental results presented in Section 3.

Evaluation Protocol: Stratified 5-Fold Cross-Validation and OOF

Evaluation employed Stratified 5-Fold Cross-Validation with shuffling and a fixed seed of 42. Stratification ensures that each fold maintains approximately the same class proportions — critical for imbalanced datasets where a non-stratified fold might exclude minority classes entirely — while five folds were selected to balance estimation stability and computational cost, retaining approximately 17–18 samples per minority class in each validation fold [21], [22]. After all five folds complete training, validation predictions are aggregated into a single OOF prediction set, from which confusion matrices and all per-class metrics are computed [29].

Loss Function and Class Imbalance Handling

To prevent the model from defaulting to majority-class predictions during training, class weights were applied exclusively to the CrossEntropyLoss function using an inverse-frequency scheme, while validation folds retained their natural class distribution to reflect real-world evaluation conditions. The weight for each class is calculated as shown in Equation (1) [19], [30]:

$$w_c = \frac{N}{K \cdot n_c} \tag{1}$$

Where N denotes the total number of training samples in the current fold, K is the total number of classes (five in this study), and n_c is the number of training samples belonging to class c. The resulting weight w_c is inversely proportional to class frequency, meaning minority classes receive larger weights and thus contribute more to the loss during training.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log\left(p_{i,y_i}\right) \tag{2}$$

In Equation (2), w_{yi} is the weight assigned to the true class of sample i, and $p_{i,yi}$ is the predicted probability that the model assigns to that correct class. A lower predicted probability for the correct class yields a higher loss, and the class weight amplifies this penalty for underrepresented classes.

Training Configuration

Training used the AdamW optimizer with a weight decay of 1×10^{-4} and followed a two-stage scheme. In the first stage spanning three epochs, the backbone was frozen and only the classification head was trained at a learning rate of 1×10^{-3}, allowing the new head to stabilize without disrupting pretrained features [8]. In the second stage running for a maximum of ten epochs, the entire backbone was unfrozen for full fine-tuning at a reduced learning rate of 3×10^{-4}. All experiments used a batch size of 32 with 224×224 input resolution and mixed-precision training via AMP FP16. Early stopping monitored validation Macro-F1 with patience of three epochs [31]. The global random seed was fixed at 42.

Evaluation Metrics

Macro-F1 was selected as the primary metric because it assigns equal weight to all classes regardless of sample size, unlike accuracy which inherently favors the majority class [18], [20]. Per-class precision and recall are defined as shown in Equation (3) and Equation (4):

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{3}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{4}$$

These are combined via harmonic mean into per-class F1 as shown in Equation (5):

$$F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{5}$$

Macro-F1 is then the unweighted average across all classes as formulated in Equation (6):

$$\text{Macro-F1} = \frac{1}{K}\sum_{c=1}^{K} F\,1_c \tag{6}$$

Weighted-F1 is additionally reported as a supplementary metric for completeness [32].

$$\text{Accuracy} = \frac{\sum_{c=1}^{K} TP_c}{N} \tag{7}$$

In Equation (7), $\sum_{c=1}^{K} TP_c$ denotes the total number of correctly classified samples across all classes, and N is the total number of samples. Although accuracy provides an intuitive overall summary of model performance, it tends to be inflated by correct predictions on the majority class in imbalanced settings, making it an unreliable standalone metric for datasets with skewed class distributions [18], [32]. For this reason, accuracy is retained as a supplementary metric alongside Macro-F1 rather than as the basis for model selection.

3. RESULTS AND DISCUSSIONS

Augmentation Results

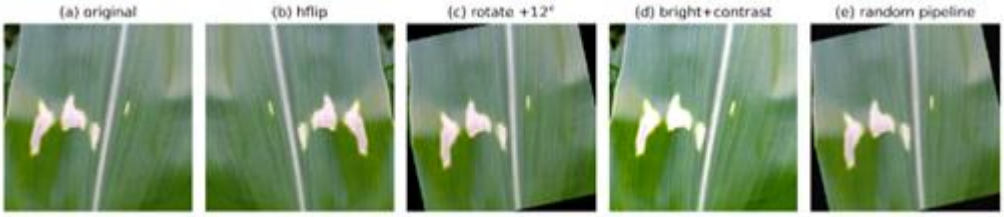

Figure 4. Examples of augmentation results

Representative augmentation results applied to the training subset are displayed in Figure 4. The transformations were deliberately kept mild to preserve the diagnostic features of each disease class while introducing sufficient variability to reduce overfitting.

Main Cross-Validation Results

EfficientNet-B0 achieved a Macro-F1 of 0.9778 with a standard deviation of 0.0085 and an accuracy of 0.9796 with a standard deviation of 0.0051, while MobileNetV3 obtained a Macro-F1 of 0.9671 with a standard deviation of 0.0215 and an accuracy of 0.9721 with a standard deviation of 0.0130. The gap in mean Macro-F1 is modest, but the more revealing difference lies in variance. EfficientNet-B0 exhibited substantially lower standard deviation across folds, indicating consistent performance regardless of how minority-class samples were distributed among folds [18], [22].

Architecturally, this advantage likely stems from its compound scaling strategy that jointly optimizes depth, width, and resolution [27], producing more diverse features compared to MobileNetV3 which prioritizes computational efficiency through depthwise separable convolutions [26]. A similar pattern was observed by Aufar *et al.* [16], where architectures with higher representational capacity demonstrated superior stability on imbalanced datasets.

Per-Class Results from OOF Aggregated Predictions

Per-class performance computed from the aggregated OOF predictions is presented in Table 3. On MobileNetV3, the best F1 scores were observed on the two largest classes while MCMV recorded the lowest F1 among all five categories. Despite class weights and stratification, depthwise separable convolutions in MobileNetV3 are inherently constrained in building discriminative representations for visually ambiguous categories, and with only 89 MCMV training images, limited sample diversity compounded this architectural limitation [19], [26], [30].

Table 3. Per-class performance from OOF aggregated predictions

Class — Support	MobileNetV3 P	R	F1	EfficientNet-B0 P	R	F1
Bact. Leaf Streak — 190	0.9734	0.9632	0.9683	0.9742	0.9947	0.9844
Common Rust — 129	1.0000	0.9535	0.9762	0.9556	1.0000	0.9773
Gray Leaf Spot — 1,497	0.9804	0.9686	0.9745	0.9925	0.9713	0.9818
Healthy — 1,038	0.9622	0.9807	0.9714	0.9696	0.9846	0.9771
MCMV — 89	0.9158	0.9775	0.9457	0.9368	1.0000	0.9674

Macro Avg	**0.9672**		**0.9776**
Weighted Avg	**0.9722**		**0.9796**

EfficientNet-B0 showed a more uniform improvement, particularly on classes that lagged under MobileNetV3. Perfect recall was achieved on Common Rust and MCMV, meaning every sample from these classes was correctly identified. However, precision on Common Rust decreased compared to MobileNetV3, indicating a trade-off where some samples from other classes were incorrectly assigned to Common Rust [23], [32]. The gap between Macro Avg F1 and Weighted Avg F1 was notably smaller on EfficientNet-B0, confirming more equitable treatment of minority classes [20].

OOF Confusion Matrix and Error Patterns

Normalized OOF confusion matrices for both models are displayed in Figure 5. On MobileNetV3, off-diagonal cells reveal consistent confusion between Common Rust and Gray Leaf Spot, Gray Leaf Spot and Healthy, and Bacterial Leaf Streak and Gray Leaf Spot. All three pairs involve classes sharing lesion symptoms on the leaf surface, making visual separation difficult under suboptimal imaging conditions.

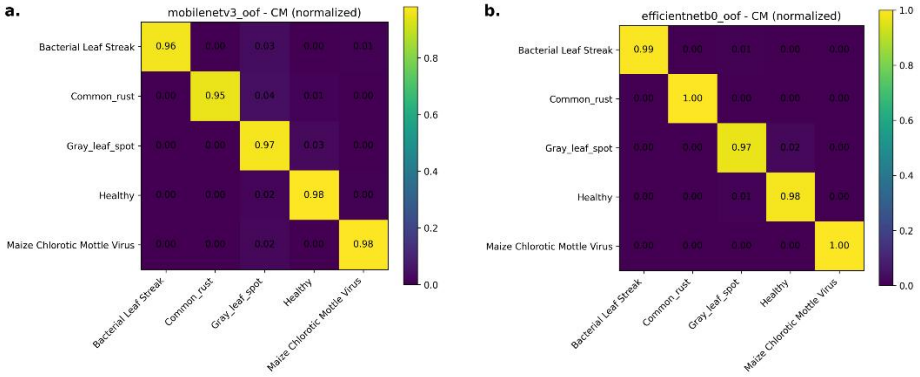

Figure 5. Normalized OOF confusion matrices: a. MobileNetV3, b. EfficientNet-B0

On EfficientNet-B0, off-diagonal cells shrank considerably, with the largest remaining confusion confined to the Gray Leaf Spot to Healthy pair. This reduction demonstrates that EfficientNet-B0 captures finer inter-class distinctions more effectively [7]. The persistence of this particular error is explained by the visual nature of early-stage symptoms: incipient Gray Leaf Spot lesions are small, faint, and low in contrast against the green leaf background, making them nearly indistinguishable from healthy tissue [7].

Training Dynamics

Training loss and validation Macro-F1 curves for representative folds are displayed in Figure 6. Both models exhibited a characteristic warmup-finetune pattern: rapid loss decrease and Macro-F1 climb during the frozen-backbone phase, followed by a temporary dip at the transition when the backbone was unfrozen due to gradient shock [8], [31]. After this transient phase, both models plateaued above 0.97. Loss decreased monotonically without upward rebound, indicating that overfitting was not a significant concern [31].

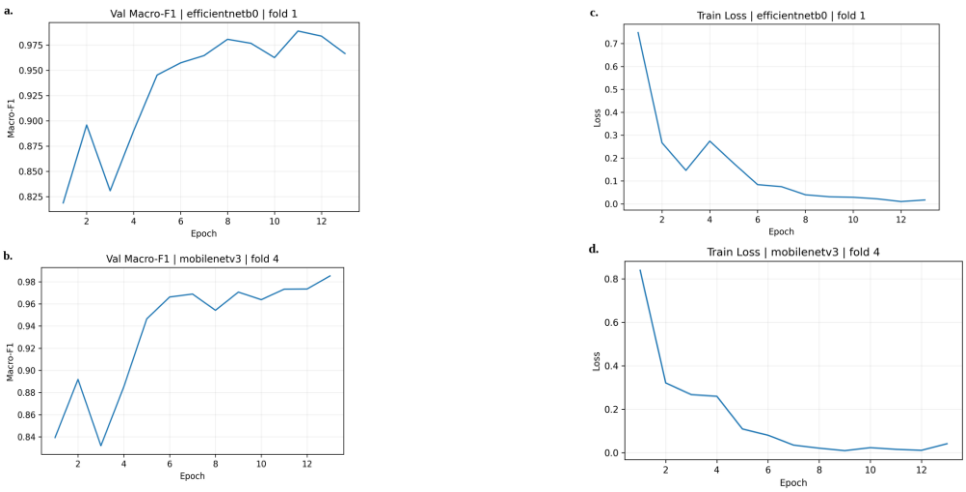

Figure 6. Training curves: a. validation Macro-F1 EfficientNet-B0 fold 1, b. validation Macro-F1 MobileNetV3 fold 4, c. training loss EfficientNet-B0 fold 1, d. training loss MobileNetV3 fold 4

Visual Error Analysis

To understand why models erred on specific class pairs, misclassified images from the dominant Gray Leaf Spot to Healthy pair are presented in Figure 7. Three recurring patterns were identified:

Figure 7. Misclassification gallery: Gray Leaf Spot samples predicted as healthy

Early-stage symptoms with small faint lesions leaving overall leaf appearance dominated by green coloration; overexposure causing whitish spots to blend with natural light reflections; and leaves not filling the frame with distracting backgrounds. These findings confirm that prediction errors stem partly from data characteristics rather than solely from architectural limitations. Improvements in image acquisition protocols could meaningfully reduce this error type [7], [25].

Comparison with Previous Studies

A comparison with related work is presented in Table 4. This comparison highlights the differences in evaluation protocols, architectures, and metrics adopted across recent studies in the plant disease classification domain.

Table 4. Comparison with previous studies

Study	Object	Architecture	Metric	Result	Evaluation
[28]	Rice leaf, 6 classes	MobileNetV3 vs EfficientNet-B0	Macro-F1	95.80% vs 93.02%	Single split
[12]	Corn leaf	Modified MobileNetV3	F1-score	98.26%	Single split
[13]	Corn leaf	Improved EfficientNet-B0	Accuracy	98.32%	Single split
[14]	Corn leaf	YOLOv8n	Accuracy	99.04%	Single split
[15]	Rice leaf	YOLOv8	mAP50	92.4%	Multi-split
This study	Corn leaf, 5 classes	EfficientNet-B0	Macro-F1	97.78%	5-Fold CV
This study	Corn leaf, 5 classes	MobileNetV3	Macro-F1	96.71%	5-Fold CV

The results remain competitive despite employing a more rigorous 5-Fold CV protocol compared to single-split designs commonly used in prior studies [18], [22]. Notably, the head-to-head outcome between MobileNetV3 and EfficientNet-B0 is not universal across domains, on rice leaf disease, MobileNetV3 was found to outperform EfficientNet-B0, whereas the reverse holds in this study, suggesting that architectural superiority is task-dependent [9], [28]. Studies reporting above 98% employed architecturally modified models

with attention modules and multi-scale fusion on different datasets [12], [13], while this study deliberately used vanilla architectures to establish pure baseline performance. The evaluation approach aligns with recent benchmarking practices in transfer learning and the emphasis on F1 and confusion matrix analysis in related fields [15], [16], [23].

4. CONCLUSION

This study benchmarked MobileNetV3 and EfficientNet-B0 for five-class corn leaf disease classification on an imbalanced dataset of 2,943 images using Stratified 5-Fold Cross-Validation with Macro-F1 as the primary metric. EfficientNet-B0 achieved a Macro-F1 of 0.9778 ($\sigma = 0.0085$) and accuracy of 0.9796 ($\sigma = 0.0051$), outperforming MobileNetV3 which obtained a Macro-F1 of 0.9671 ($\sigma = 0.0215$) and accuracy of 0.9721 ($\sigma = 0.0130$), with the notably lower variance of EfficientNet-B0 indicating more consistent behavior across folds. This advantage was consistently observed in both the aggregate metrics and the per-class OOF evaluation. Class weights proved effective in supporting minority-class recognition, although errors persisted on class pairs sharing visual symptom overlap, most notably Gray Leaf Spot misclassified as Healthy on early-symptom images. Comparison with prior studies demonstrated that these findings remain competitive despite the more stringent evaluation protocol employed. Limitations include the absence of external validation on datasets from different geographic locations or growing seasons, the lack of visual interpretability techniques such as Grad-CAM, and the use of only three basic augmentation transforms. Future work should explore cross-dataset validation, integrate Grad-CAM for model transparency, and evaluate mobile deployment feasibility to support real-time corn leaf disease detection in the field.

ACKNOWLEDGEMENTS

The authors would like to thank the Department of Informatics, Faculty of Engineering, Universitas Mulawarman for providing research facilities.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Muhammad Shandy Alfarizal: Conceptualization, Methodology, Software, Writing – original draft. **Muhamad Kelvin Saputra**: Data curation, Software. **Ade Fajar Kurniawan**: Data curation, Visualization. **Diki Wahyudi** Validation, Visualization. **Khanahaya Fadhil Adriano**: Validation, Visualization. **Anindita Septiarini**: Supervision, Methodology, Writing – review & editing. **Novianti Puspitasari**: Methodology, Writing – review & editing.

DECLARATION OF COMPETING INTERESTS

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

DATA AVAILABILITY

Data will be made available on request.

REFERENCES

[1] B.-S. Indonesia, "Luas Panen dan Produksi Jagung di Indonesia 2024 (Angka Tetap) [Harvested Area and Corn Production in Indonesia 2024 (Final Figures)," Berita Resmi Statistik.

[2] M. R. Islam, M. A. Hossain, and M. S. Islam, "Seasonal Corn Leaf Disease Dataset: A Multi-Year Collection for Robust Analysis," *Mendeley Data*, 2025, doi: 10.17632/vy629dngm8.1.

[3] N. F. Sulfitri, A. K. Parawansa, and M. S. Gani, "ISOLASI DAN INTENSITAS SERANGAN PENYAKIT BULAI (Peronosclerospora philippinensis Shaw) PADA TANAMAN JAGUNG (Zea mays L.) DI KABUPATEN MAROS," *AGrotekMAS J. Indones. J. Ilmu Peranian*, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:273687240

[4] M. S. Krishna, P. Machado, R. I. Otuka, S. W. Yahaya, F. Neves dos Santos, and I. K. Ihianle, "Plant Leaf Disease Detection Using Deep Learning: A Multi-Dataset Approach," 2025. doi: 10.3390/j8010004.

[5] M. K. A. Mazumder, M. F. Mridha, S. Alfarhood, M. Safran, M. Abdullah-Al-Jubair, and D. Che, "A robust and light-weight transfer learning-based architecture for accurate detection of leaf diseases across multiple plants using less amount of images," *Front. Plant Sci.*, vol. Volume 14-2023, 2024, doi: 10.3389/fpls.2023.1321877.

[6] B. Tugrul, E. Elfatimi, and R. Eryigit, "Convolutional Neural Networks in Detection of Plant Leaf Diseases: A Review," *Agriculture*, vol. 12, no. 8, 2022, doi: 10.3390/agriculture12081192.

[7] J. Zhao *et al.*, "A review of plant leaf disease identification by deep learning algorithms," *Front. Plant Sci.*, vol. Volume 16-2025, 2025, doi: 10.3389/fpls.2025.1637241.

[8] W. Shafik, A. Tufail, C. De Silva Liyanage, and R. A. A. H. M. Apong, "Using transfer learning-based plant disease classification and detection for sustainable agriculture," *BMC Plant Biol.*, vol. 24, no. 1, p. 136, 2024, doi: 10.1186/s12870-024-04825-y.

[9] B. Hanh, H. Manh, and N.-V. Nguyen, "Enhancing the performance of transferred efficientnet models in leaf image-based plant disease classification," *J. Plant Dis. Prot.*, vol. 129, Apr. 2022, doi: 10.1007/s41348-022-00601-y.

[10] B. Khomkham and Y. Pankaseam, "Lightweight Convolutional Neural Network Model Based on Modified MobileNetV3 for Plant Disease Classification," *SN Comput. Sci.*, vol. 6, Dec. 2025, doi: 10.1007/s42979-025-04607-9.

[11] F. Rajeena P. P., A. S. U., M. A. Moustafa, and M. A. S. Ali, "Detecting Plant Disease in Corn Leaf Using EfficientNet Architecture—An Analytical Approach," *Electronics*, vol. 12, no. 8, 2023, doi: 10.3390/electronics12081938.

[12] C. Bi, S. Xu, N. Hu, S. Zhang, Z. Zhu, and H. Yu, "Identification Method of Corn Leaf Disease Based on Improved Mobilenetv3 Model," *Agronomy*, vol. 13, no. 2, 2023, doi: 10.3390/agronomy13020300.

[13] X. Sun and H. Huo, "Corn leaf disease recognition based on improved EfficientNet," *IET Image Process.*, vol. 19, Jan. 2025, doi: 10.1049/ipr2.13288.
[14] F. Khan, N. Zafar, M. N. Tahir, M. Aqib, H. Waheed, and Z. Haroon, "A mobile-based system for maize plant leaf disease detection and classification using deep learning," *Front. Plant Sci.*, vol. Volume 14-2023, 2023, doi: 10.3389/fpls.2023.1079366.
[15] M. N. Fadhillah, A. Septiarini, H. Hamdani, R. Rajiansyah, and A. Tejawati, "Comparative of YOLOv5 and YOLOv8 for rice leaf disease detection on diverse image datasets," *J. Soft Comput. Explor.*, vol. 7, no. 1 SE-Articles, pp. 31–42, doi: 10.52465/joscex.v7i1.19.
[16] Y. Aufar, M. D. A. Rahman, and M. F. Ridhani, "Benchmarking deep transfer learning for imbalanced skin cancer classification: Integrating focal loss, explainable AI, and web deployment," *J. Soft Comput. Explor.*, vol. 7, no. 1 SE-Articles, pp. 55–65, doi: 10.52465/joscex.v7i1.20.
[17] M. Izzulhaq and E. Sugiharti, "Tuberculosis classification on chest x-ray images using DenseNet-169 and convolutional block attention module," *J. Soft Comput. Explor.*, vol. 7, pp. 19–30, Mar. 2026, doi: 10.52465/joscex.v7i1.14.
[18] P. Thölke *et al.*, "Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data," *Neuroimage*, vol. 277, p. 120253, 2023, doi: https://doi.org/10.1016/j.neuroimage.2023.120253.
[19] M. Ulum and J. Unjung, "Enhancing diabetes classification performance using XGBoost integrated with SMOTE and bayesian hyperparameter optimization," *J. Soft Comput. Explor.*, vol. 7, pp. 9–18, Mar. 2026, doi: 10.52465/joscex.v7i1.3.
[20] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, "Confidence interval for micro-averaged F1 and macro-averaged F1 scores," *Appl. Intell.*, vol. 52, no. 5, pp. 4961–4972, Mar. 2022, doi: 10.1007/s10489-021-02635-5.
[21] H.-W. Zhang, R.-F. Wang, Z. Wang, and W.-H. Su, "DLCPD-25: A Large-Scale and Diverse Dataset for Crop Disease and Pest Recognition," *Sensors*, vol. 25, no. 22, 2025, doi: 10.3390/s25227098.
[22] J. White and S. D. Power, "k-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation," 2023. doi: 10.3390/s23136077.
[23] I. Bakti, H. Jati, N. Nurkhamid, and Y. Bunda, "A comparative analysis of five textual similarity methods for automatic short answer grading," *J. Soft Comput. Explor.*, vol. 7, pp. 43–54, Mar. 2026, doi: 10.52465/joscex.v7i1.11.
[24] S. Saponara and A. Elhanashi, "Impact of Image Resizing on Deep Learning Detectors for Training Time and Model Performance," in *Lecture Notes in Electrical Engineering*, Springer Science and Business Media Deutschland GmbH, 2022, pp. 10–17. doi: 10.1007/978-3-030-95498-7_2.
[25] K. Alomar, H. I. Aysel, and X. Cai, "Data Augmentation in Classification and Segmentation: A Survey and New Strategies," 2023. doi: 10.3390/jimaging9020046.
[26] A. Ardiansyah and N. F. Hasan, "Deteksi dan Klasifikasi Penyakit Pada Daun Kopi Menggunakan Yolov7," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 12, no. 1 SE-Articles, pp. 30–35, Mar. 2023, doi: 10.32736/sisfokom.v12i1.1545.
[27] Y. Pamungkas and D. S. Eljatin, "Hyperparameter Tuning of EfficientNet Method for Optimization of Malaria Detection System Based on Red Blood Cell Image," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 13, no. 3 SE-Articles, pp. 360–368, Nov. 2024, doi: 10.32736/sisfokom.v13i3.2257.
[28] A. N. Abiyyu and M. Rahardi, "Comparison of Transfer learning Models MobileNetV3-Large and EfficientNet-B0 for Rice Leaf Disease Classification," *J. Appl. Informatics Comput.*, vol. 10, no. 1 SE-Articles, pp. 818–828, Feb. 2026, doi: 10.30871/jaic.v10i1.12033.
[29] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Min.*, vol. 16, no. 1, p. 4, 2023, doi: 10.1186/s13040-023-00322-4.
[30] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif. Intell. Rev.*, vol. 57, no. 10, p. 273, 2024, doi: 10.1007/s10462-024-10884-2.
[31] E. Hassan, M. Y. Shams, N. A. Hikal, and S. Elmougy, "The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study," *Multimed. Tools Appl.*, vol. 82, no. 11, pp. 16591–16633, May 2023, doi: 10.1007/s11042-022-13820-0.
[32] K. M. Sujon, R. Hassan, K. Choi, and M. A. Samad, "Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *J. Big Data*, vol. 12, no. 1, p. 268, 2025, doi: 10.1186/s40537-025-01313-4.

.