

Comparative Analysis of Decision-Level Ensemble Voting Strategies for Tomato Leaf Disease Classification Using EfficientNet-B0 and MobileNetV2

Hani Nafis Muzaki¹, Yahya Nur Ifriza^{2*}

^{1,2}Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract. Tomato leaf diseases threaten agricultural productivity, making accurate and timely disease classification essential for supporting precision agriculture. Recent advances in convolutional neural networks have significantly improved automated plant disease recognition, and ensemble learning has further enhanced performance by combining multiple models. Among ensemble approaches, decision-level voting strategies, including hard voting, soft voting, and weighted soft voting, are widely used for their simplicity and effectiveness. To address the limited systematic comparisons of these strategies using lightweight CNNs, this study evaluates their performance to identify the optimal ensemble approach.

Purpose: This study systematically compares hard voting, soft voting, and weighted soft voting for combining EfficientNet-B0 and MobileNetV2 to identify the most effective decision-level ensemble strategy for tomato leaf disease classification.

Methods/Study design/approach: EfficientNet-B0 and MobileNetV2 were trained independently and combined using hard voting with confidence-based tie breaking, soft voting, and weighted soft voting, with voting weights optimized through grid search. Performance was evaluated using accuracy, macro-F1 score, confusion matrices, and McNemar's test.

Result/Findings: EfficientNet-B0 outperformed MobileNetV2 as the best single model (94.05% accuracy, 0.9266 macro-F1 score vs. 90.64% accuracy, 0.8781 macro-F1 score). All three ensemble strategies further improved classification performance, with weighted soft voting achieving the best results (95.15% accuracy, 0.9417 macro-F1). McNemar's test confirmed that soft voting and weighted soft voting significantly outperformed the best single model, while hard voting did not. No significant differences were found among the three ensemble strategies.

Novelty: This study provides a systematic comparison of hard voting, soft voting, and weighted soft voting for decision-level ensemble using EfficientNet-B0 and MobileNetV2 in tomato leaf disease classification, incorporating McNemar's test to statistically validate performance differences and offer empirical guidance for selecting appropriate voting strategies.

Keywords: Tomato Leaf Disease, Decision-Level Voting, Ensemble Learning, EfficientNet-B0, MobileNetV2

Received August 2026 / **Revised** August 2026 / **Accepted** September 2026

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



INTRODUCTION

Tomato is an important horticultural crop worldwide, but its productivity and quality are significantly affected by various leaf diseases. Conventional visual diagnosis by farmers is often subjective and prone to error, leading to delayed treatment and reduced crop yield. Consequently, automated disease identification has become an important application of deep learning (DL). Among DL techniques, Convolutional Neural Networks (CNNs) have demonstrated excellent performance by automatically extracting distinct visual patterns from leaf images. However, practical deployment requires models that provide strong classification performance while remaining suitable for platforms with limited computational power, such as mobile and edge devices.

Recent studies have enhanced the classification of tomato leaf diseases primarily via architectural refinement and feature-level enhancement. Several works have modified or combined network architectures to enrich feature representation, such as employing DenseNet for mobile-based classification [1] and

*Corresponding author.

Email addresses: yahyanurifriza@mail.unnes.ac.id (Ifriza)

DOI: 10.15294/rji.v4i2.61961

integrating Dense Inception with MobileNetV2 through an attention mechanism [2]. Other studies pursued a complementary direction by enriching feature representation prior to classification, combining deep features with handcrafted descriptors [3], fusing feature maps from multiple CNNs [4], [5], or concatenating feature maps from ResNet50 and MobileNetV2 [6]. Although these directions achieve competitive performance, they share a common limitation: reliance on a single prediction model or feature-level fusion, which increases architectural complexity and often requires additional preprocessing, thereby constraining their scalability for lightweight, resource-constrained deployment [7].

Decision-level ensemble learning offers an alternative that preserves the independence of individual models while improving robustness. Existing studies in this direction have primarily focused on isolated voting strategy, such as averaging-based ensembles of heterogeneous CNNs [8], [9], majority voting ensembles [10], [11], and weighted voting schemes combining two lightweight architectures [12]. Despite differing implementations, these studies share two important gaps. First, none systematically compares multiple decision-level voting methods (e.g., hard voting, soft voting, and weighted soft voting) under identical conditions, making it unclear which strategy is most effective for a given architecture combination. Second, statistical significance testing is rarely conducted, so it remains uncertain whether reported performance gains reflect genuine methodological improvement or random sampling variation. Furthermore, several of these ensembles incorporate computationally demanding architectures, limiting their suitability for resource-constrained deployment. These combined gaps motivate the present study's systematic and statistically validated comparison of decision-level voting strategies using lightweight CNNs.

Beyond these methodological limitations, tomato leaf disease datasets frequently show substantial class imbalance, with particular disease categories represented by considerably fewer samples than others. Under such conditions, accuracy alone may provide an overly optimistic assessment because it is largely influenced by majority classes, whereas minority classes, often equally important for practical disease diagnosis, may remain poorly recognized without substantially affecting the reported accuracy [13], [14]. Consequently, balanced evaluation metrics such as macro-F1 score, together with statistical significance testing, are essential for objectively comparing classification and ensemble strategies on imbalanced datasets.

To address these limitations, this study proposes a decision-level ensemble framework based on two lightweight CNN architectures, EfficientNet-B0 [15] and MobileNetV2 [16]. Three voting strategies, hard voting, soft voting, and weighted soft voting are systematically evaluated, with the voting weights optimized through grid search on the validation set. Unlike previous studies that primarily report a single ensemble configuration, this work provides a comprehensive comparison of three decision-level voting strategies under identical experimental conditions and validates the observed performance differences using McNemar's statistical significance test. Accordingly, this study aims to compare individual CNN models and three decision-level voting strategies for tomato leaf disease classification and to assess the statistical significance of their performance differences using McNemar's test.

METHODS

The overall research process was executed following a structured workflow, as illustrated in Figure 1. The workflow begins with dataset preparation, including the curation of tomato leaf disease images and a stratified train-validation-test split. Each subset then undergoes model-specific preprocessing, with data augmentation restricted solely to the training set. EfficientNet-B0 and MobileNetV2 are independently trained as feature extractors, after which their predictions are combined using hard voting, soft voting, and weighted soft voting. The voting weights for the weighted soft voting are optimized via grid search across the validation set. Finally, the two individual CNN models and the three ensemble voting strategies are evaluated using accuracy, macro precision, macro recall, macro-F1 score, confusion matrices, and McNemar's test.

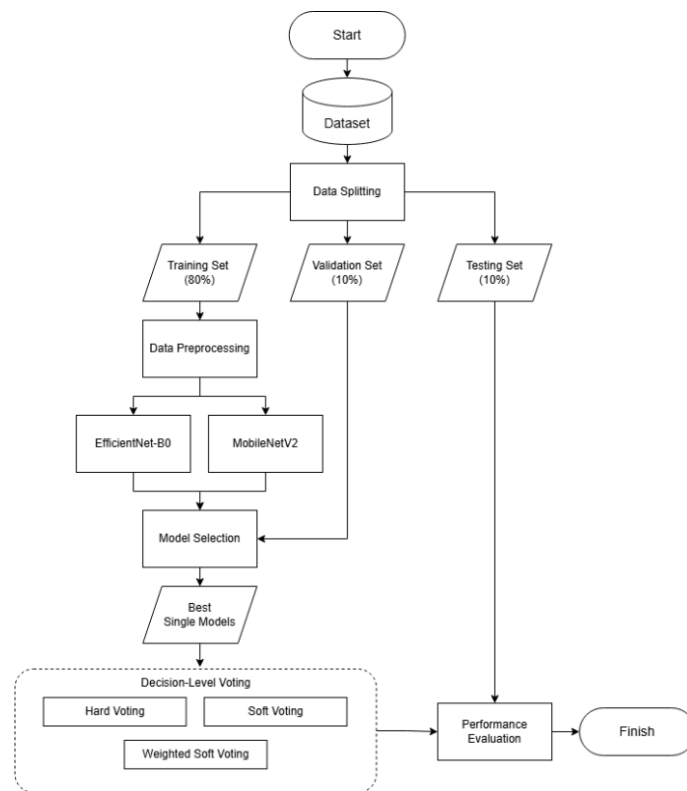


Figure 1. Flowchart of research processes

Dataset

Data for this experiment were acquired directly from the Mendeley Data Repository, “Identification of Plant Leaf Diseases Using a 9-layer Deep Convolutional Neural Network” [17]. The repository provides two versions of the dataset: with augmentation (61,486 images across 39 classes spanning multiple plant species) and without augmentation. This study used the non-augmented version and selected only the ten tomato leaf disease classes, resulting in 18,160 original images. Utilizing a stratified 80/10/10 ratio, the data was divided into three separate groups to handle training, validation, and testing, yielding 14,528, 1,816, and 1,816 images, respectively. A comprehensive breakdown of the dataset across each class can be found in Table 1.

Table 1. Dataset distribution across each class

Class Name	Training	Validation	Testing	Total
Bacterial Spot	1,702	213	212	2,127
Early Blight	800	100	100	1000
Late Blight	1,527	191	191	1,909
Leaf Mold	761	96	95	952
Septoria Leaf Spot	1,417	177	177	1,771
Two-Spotted Spider Mites	1,341	167	168	1,676
Target Spot	1,123	140	141	1,404
Tomato Yellow Leaf Curl Virus	4,286	535	536	5,357
Tomato Mosaic Virus	298	38	37	373
Healthy	1,273	159	159	1,591

As observed in Table 1, the dataset exhibits class imbalance, with Yellow Leaf Curl Virus containing 5,357 images, while Mosaic Virus contains only 373 images. To address this imbalance, a cost-sensitive learning approach was implemented by integrating class-wise weights into the categorical cross-entropy loss function [13]. These weights were calculated inversely proportional to class frequencies, assigning higher misclassification penalties to minority classes during training to prevent majority class bias [13], [14]. To ensure fair model evaluation and avoid potential data leakage, image augmentation was restricted solely to the training subset after dataset splitting.

Data Preprocessing

All input images were rescaled to $224 \times 224 \times 3$ pixels for both architectures. Pixel normalization followed the preprocessing strategy required by each model. EfficientNet-B0 used its built-in Rescaling layer [15], whereas MobileNetV2 employed the `preprocess_input` function to map pixel values into the $[-1, 1]$ range, matching its original ImageNet pretraining [16]. No additional generic rescaling was applied to ensure compatibility with the expected preprocessing pipeline of each architecture.

Data augmentation was performed on-the-fly and applied exclusively to the training subset, including random rotations (up to 20°), width and height shifts (up to 0.1), zooming (up to 0.2), and horizontal flipping. The validation and test subsets were strictly non-augmented to preserve the integrity of the evaluation metrics.

Model Architecture and Training

Two pretrained convolutional neural network (CNN) architectures, EfficientNet-B0 and MobileNetV2, were employed as the base learners in the ensemble framework. Both models were initialized with ImageNet [18] weights and configured with `include_top=False` for transfer learning. The pretrained backbone networks remained frozen throughout training, while a custom classification head consisting of a Global Average Pooling layer, a Dropout layer (dropout rate = 0.2), and a Dense output layer with softmax activation was trained for tomato leaf disease classification.

EfficientNet-B0 and MobileNetV2 were trained using the Adam optimizer [19] with an initial learning rate of 1×10^{-3} , a batch size of 32, and a training cap of 30 epochs. EarlyStopping (patience = 6) monitored the validation loss to prevent overfitting, while ReduceLROnPlateau (factor = 0.2, patience = 3) automatically reduced the learning rate when the validation performance plateaued. ModelCheckpoint was employed to save the model achieving the lowest validation loss for subsequent evaluation. This training configuration was applied consistently to both architectures to ensure a fair comparison before ensemble construction. All models were implemented using TensorFlow 2.20.0 [20] and trained on Google Colaboratory using an NVIDIA Tesla T4 GPU.

Voting Strategies

Three decision-level ensemble voting strategies were evaluated by combining the prediction outputs of EfficientNet-B0 and MobileNetV2. Hard voting assigns each model a single vote based on its predicted class. Since only two models were combined, disagreements resulted in a tie (1–1). Such ties were resolved by selecting the prediction associated with the higher maximum softmax probability for the corresponding sample. Soft voting takes an equally weighted average of both models predicted class probabilities, and the class with the highest resulting average is chosen as the final prediction. Weighted soft voting combines the predicted class probabilities according to Formula 1.

$$P_{final}(c) = w_1 P_1(c) + w_2 P_2(c) \quad (1)$$

where $w_1 + w_2 = 1$. The optimal voting weights were determined using grid search by varying w_1 from 0 to 1 with increments of 0.05, resulting in 21 candidate weight combinations. For each combination, the macro-F1 score was computed on the validation set, and the weights achieving the highest score were selected. Macro-F1 score, rather than validation accuracy or loss, was used as the grid search objective to avoid bias toward majority classes and to align weight selection directly with the ensemble's final class decision, consistent with the macro-F1 score centric evaluation adopted throughout this study. Weight optimization drew solely on the validation set, preserving the test set strictly isolated right up to the final evaluation.

Performance Evaluation

The performance of the two individual CNN models and the three ensemble voting strategies was evaluated on the same reserved testing subset comprising 1,816 images. Evaluation metrics included accuracy, macro-F1 score, precision, recall, and confusion matrices. Macro-F1 score was adopted as the primary performance metric owing to the imbalanced class distribution, as it assigns equal importance to all classes regardless of their sample sizes [21], whereas accuracy and weighted-average metrics may be influenced by majority classes.

To determine if the performance gaps between classification strategies held true statistical validity, seven pairwise comparisons were performed using McNemar's test on the test set predictions. Following standard recommendations for paired nominal data, the exact McNemar test was applied when the number of discordant prediction pairs was small (< 25); otherwise, the continuity corrected chi-square approximation was used. A threshold of 0.05 was established to evaluate statistical significance ($p < 0.05$).

RESULT AND DISCUSSION

An experimental framework was developed to investigate both the optimization behavior of the individual CNN models and the comparative performance of three decision-level ensemble voting strategies. The analysis begins by examining the convergence characteristics of EfficientNet-B0 and MobileNetV2 through their training and validation curves. Subsequently, the predictive performance of the individual models and ensemble strategies is compared using accuracy, macro precision, macro recall, macro-F1 score, confusion matrices, and McNemar's statistical test. The validation accuracy and validation loss of the individual CNN models are presented in Figure 2.

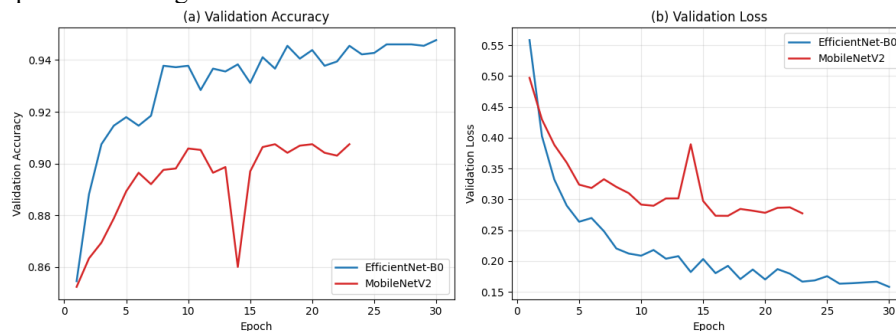


Figure 2. (a) Validation accuracy and (b) validation loss of the single models

Figure 2 presents the validation accuracy and loss trajectories of EfficientNet-B0 and MobileNetV2 throughout their optimization process. Both models exhibited progressive improvements in training accuracy together with decreasing training loss, indicating successful optimization during training. However, distinct convergence behaviors were observed between the two architectures.

EfficientNet-B0 demonstrated smooth and stable convergence across the entire training process. Both validation accuracy and validation loss improved consistently accompanied by minor fluctuations, pointing to solid generalization throughout optimization. The model completed the maximum training budget of 30 epochs without triggering early stopping. Although the learning rate remained at 1×10^{-3} for the first 29 epochs, it was reduced to 2×10^{-4} at the final epoch, where the model achieved its best validation performance, reaching a validation accuracy of 94.77% with a validation loss of 0.1583. These observations indicate that EfficientNet-B0 continued to benefit from further optimization until the final training epoch.

In contrast, MobileNetV2 converged more rapidly and triggered early stopping at epoch 23 after no further improvement in validation performance was observed. The best performing checkpoint was obtained at epoch 17, achieving a validation accuracy of 90.75% and a validation loss of 0.2734, after which the corresponding model weights were restored for all subsequent evaluations. Compared with EfficientNet-B0, MobileNetV2 converged to a higher validation loss plateau, suggesting relatively lower generalization capability for tomato leaf disease classification.

A temporary fluctuation in MobileNetV2 validation performance was observed at epoch 14, where validation accuracy decreased from 89.87% to 86.01%, accompanied by an increase in validation loss from 0.3017 to 0.3893. Despite this fluctuation, training accuracy continued to improve while training loss decreased steadily, indicating that the optimization process remained stable throughout training. Validation performance recovered over the subsequent epochs, and the best performing checkpoint was obtained at epoch 17. During this period, the ReduceLROnPlateau scheduler reduced the learning rate from 1×10^{-3} to 2×10^{-4} . Because the present study did not evaluate the learning rate scheduler independently, the observed recovery cannot be attributed solely to the reduction in learning rate. Nevertheless, the temporary fluctuation did not affect the final model selection, as the checkpoint restoration mechanism retained the

model with the highest validation performance for subsequent test-set evaluation and ensemble construction.

The best performing checkpoints of EfficientNet-B0 and MobileNetV2 obtained during training were subsequently used to construct the ensemble models. To determine the relative contribution of each base learner in the optimized weighted soft voting, a grid search was performed on the validation set by evaluating 21 candidate weight combinations. As illustrated in Figure 3, the validation macro-F1 score varies across different weight configurations. The optimal configuration was obtained at $w_1 = 0.55$ for EfficientNet-B0 and $w_2 = 0.45$ for MobileNetV2, achieving the highest validation macro-F1 score of 0.9483.

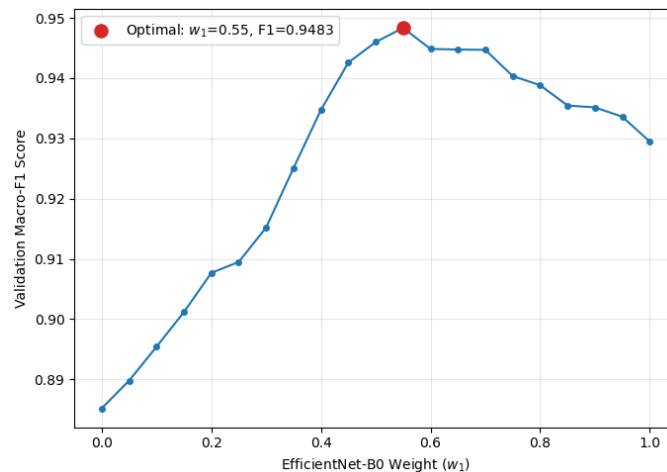


Figure 3. Grid search results for the ensemble weights on the validation set

Figure 3 illustrates the validation macro-F1 score achieved by each candidate's weight combination during the grid search process. During this evaluation, the weight allocations were subject to the constraint $w_1 + w_2 = 1$, where w_1 and w_2 denote the weights assigned to EfficientNet-B0 and MobileNetV2. The selected weights were then applied to the weighted soft voting ensemble for final evaluation on the independent test set. The evaluation results, including comparisons with the base architectures and voting strategies, are presented in Table 2.

Table 2. Performance comparison of base architecture and ensemble models

Model	Precision	Recall	F1-score	Accuracy (%)
MobileNetV2	0.8734	0.8880	0.8781	90.64
EfficientNet-B0	0.9190	0.9366	0.9266	94.05
Hard Voting	0.9319	0.9413	0.9357	94.71
Soft Voting (avg)	0.9361	0.9463	0.9404	95.04
Weighted Soft Voting	0.9359	0.9490	0.9417	95.15

Table 2 summarizes the classification performance of the two individual CNN models and the three ensemble strategies on the independent test set, measured by macro precision, macro recall, macro-F1 score, and accuracy. Across all evaluation metrics, the ensemble approaches consistently outperformed the individual CNN models across all evaluation metrics.

Among the standalone architectures, EfficientNet-B0 consistently outperformed MobileNetV2 across all evaluation metrics, achieving higher macro precision, macro recall, macro-F1 score, and accuracy. This finding is consistent with the convergence observed in Fig. 2, where EfficientNet-B0 exhibited higher validation accuracy and lower validation loss throughout training, indicating superior generalization to unseen data. In contrast, the lower performance of MobileNetV2 suggests that its learned feature representations were less discriminative for tomato leaf disease classification.

The ensemble methods further improved the overall results by combining the complementary predictions of the two CNN architectures. Hard Voting outperformed both individual models by aggregating the predicted class labels from EfficientNet-B0 and MobileNetV2. Relative to EfficientNet-B0, the strongest

standalone classifier, Hard Voting increased the classification accuracy from 94.05% to 94.71% and the F1-score from 0.9266 to 0.9357. Soft Voting further enhanced the predictive performance by utilizing the predicted class probabilities, allowing the ensemble decision to exploit the confidence information contained in the probability outputs rather than relying solely on hard voting.

Weighted Soft Voting achieved the highest overall performance across all evaluated strategies, yielding the highest macro recall, macro-F1 score, and classification accuracy. Its macro precision (0.9359) was marginally lower than that of conventional soft voting (0.9361), although the difference was negligible. Unlike conventional Soft Voting, this approach assigns different voting weights to the base learners instead of treating their predictions equally. This weighting strategy produced the strongest empirical results in the present study. To provide deeper insight into these results, detailed class-level evaluations are presented in the subsequent section. The statistical significance of the observed differences between classification strategies is then assessed using McNemar's test. The class-level prediction performance of the five evaluated classification strategies, as represented by their respective confusion matrices, is presented in Figure 4.

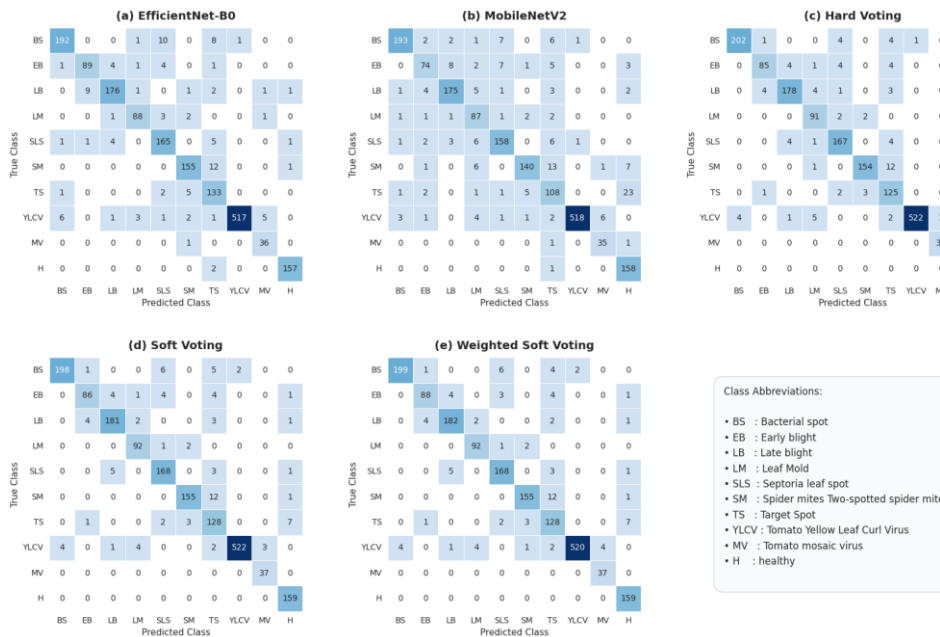


Figure 4. Confusion matrices comparing the class-level prediction performance of the five evaluated classification strategies on the tomato leaf disease test set

Figure 4 provides a visual breakdown of the confusion matrices for the five classification strategies, providing a class-level view of prediction errors that complement the aggregate metrics reported in Table 2. In general, the ensemble approaches produced a stronger concentration of predictions along the principal diagonal than the individual CNN models, indicating fewer misclassification errors and improved class discrimination. However, a detailed examination reveals that these improvements were not uniformly distributed across all disease categories and, in one notable case, were not even monotonic.

Target Spot exhibited the lowest precision across all strategies (0.7347 for MobileNetV2, rising to 0.8258 for weighted soft voting). Confusion matrix inspection traces this weakness to MobileNetV2, which misclassified 23 of 141 Target Spot samples as Healthy (16.3%), an error EfficientNet-B0 never made. This produces an unexpected trade-off, although all three voting strategies improved Target Spot precision over MobileNetV2 alone, recall did not improve uniformly. Each voting strategy correctly classified only 125–128 of 141 Target Spot samples, fewer than the 133 achieved by EfficientNet-B0 alone. This indicates that incorporating MobileNetV2 predictions occasionally overturned correct EfficientNet-B0 predictions, since the voting strategies apply fixed global weights rather than class-specific reliability estimates. The finding suggests that global voting weights, despite improving aggregate metrics, can mask class-level regressions relative to the single best model.

Leaf Mold, in contrast, demonstrated a clear complementary effect. MobileNetV2 low precision for this class (0.7768) stemmed from a diffuse pattern of false positives spread across multiple other categories (Late Blight, Yellow Leaf Curl Virus, Septoria Leaf Spot, and Spider Mites all contributed misclassified samples), rather than confusion with any single class. EfficientNet-B0 exhibited a much cleaner prediction pattern for this class, and all three ensemble strategies converged toward EfficientNet-B0 cleaner pattern (reducing false positives to 6 samples under weighted soft voting). This result suggests that the ensemble successfully suppressed MobileNetV2 more scattered errors by exploiting the complementary confidence estimates from both CNN models.

Mosaic Virus, the most severely underrepresented class (only 37 test samples) achieved perfect recall (1.0000) under all three voting strategies, meaning every true Mosaic Virus sample was correctly identified. Although a small number of false positives remained from other classes (such as four instances of Yellow Leaf Curl Virus misclassified as Mosaic Virus under weighted soft voting), no Mosaic Virus samples were incorrectly classified as another disease. Despite the limited number of test samples, this result suggests that decision-level ensemble voting can preserve minority-class recognition without requiring explicit oversampling or synthetic data generation. Nevertheless, this finding should be interpreted alongside the Target Spot results, which demonstrate that the benefits of ensemble voting are not necessarily distributed uniformly across all disease categories.

Collectively, the confusion matrices demonstrate that the relative effectiveness of the evaluated decision-level voting strategies depends on how effectively they exploit the complementary prediction patterns of the constituent CNN models. Although all three ensemble strategies improved classification performance, probability-based voting consistently produced cleaner class separation than hard voting by incorporating the confidence estimates of both base learners. Out of all the evaluated strategies, Weighted Soft Voting yielded the most balanced class-level performance, reducing misclassification errors across multiple disease categories while maintaining the highest aggregate evaluation metrics. These findings complement the quantitative results presented in Table 2 and highlight the advantages of probability-based decision fusion for tomato leaf disease classification.

To determine whether the empirical outcome disparities signified genuine improvements instead of mere random sampling noise, McNemar's test was conducted across seven pairwise comparisons, as summarized in Table 3.

Table 3. Pairwise McNemar's test results for the evaluated classification strategies

Comparison	<i>p</i> – value	Significant (<i>p</i> < 0.05)
EfficientNet-B0 vs. MobileNetV2	<0.001	Yes
EfficientNet-B0 vs. Hard Voting	0.207	No
EfficientNet-B0 vs. Soft Voting	0.042	Yes
EfficientNet-B0 vs. Weighted Soft Voting	0.014	Yes
Hard Voting vs. Soft Voting	0.180	No
Hard Voting vs. Weighted Soft Voting	0.152	No
Soft Voting vs. Weighted Soft Voting	0.754	No

EfficientNet-B0 was found to significantly outperform MobileNetV2 as a standalone model ($p < 0.001$), consistent with the superior aggregate performance reported in Table 2. Both Soft Voting and Weighted Soft Voting demonstrated statistically significant improvements over EfficientNet-B0 ($p = 0.042$ and $p = 0.014$, respectively). In contrast, Hard Voting, despite achieving slightly higher accuracy and macro-F1 score than EfficientNet-B0, did not reach statistical significance ($p = 0.207$). This finding suggests that relying primarily on hard voting may not fully exploit the complementary probabilistic information produced by the two CNN models, whereas Soft Voting and Weighted Soft Voting combine the complete probability distributions during decision fusion.

Notably, no statistically significant differences were observed among the three ensemble voting strategies. Pairwise comparisons between Hard Voting and Soft Voting ($p = 0.180$), Hard Voting and Weighted Soft Voting ($p = 0.152$), and Soft Voting and Weighted Soft Voting ($p = 0.754$) all failed to reject the null hypothesis. Collectively, these findings indicate that although decision-level ensemble learning consistently improved performance relative to the strongest individual CNN model, the choice among the evaluated voting strategies resulted in only modest differences in prediction behavior under the present experimental

setting. Nevertheless, Weighted Soft Voting consistently achieved the highest aggregate evaluation metrics while maintaining a statistically significant improvement over EfficientNet-B0.

From a practical perspective, these findings suggest that preserving the probability outputs of the constituent CNN models provides measurable benefits over hard voting when constructing decision-level ensembles. For applications employing a small number of lightweight CNN models, probability-based fusion (Soft Voting or Weighted Soft Voting) offers a simple and effective approach for improving classification performance with minimal additional computational overhead.

Comparison with Previous Research

Table 4 compares the best performing voting strategies in this study against the most directly comparable prior research, Ulutaş and Aslantaş [8], which used the identical 18,160-image, 10-class tomato leaf disease dataset.

Table 4. Comparison with previous research using the same dataset

Study	Model Combination	Voting Strategies	Accuracy
Ulutaş and Aslantaş [8]	MobileNetV3Small + EfficientNetV2L + InceptionV3 + CNN1 + CNN2 (5-EnsCNNs)	Unweighted Average (Hard Voting)	99.60%
This Study	EfficientNet-B0 + MobileNetV2 (2 lightweight CNNs)	Weighted Soft Voting	95.15%

Although Ulutaş and Aslantaş [8] reported a higher absolute accuracy, their result was obtained by fine tuning multiple heterogeneous architectures including larger models such as EfficientNetV2L combined in three to five model ensembles with hyperparameter optimization via particle swarm optimization. In contrast, this study deliberately restricted the ensemble to exactly two lightweight, frozen (non-fine-tuned) models, prioritizing a controlled, statistically validated comparison of voting strategies and lower training and deployment cost over maximizing absolute accuracy. Notably, Ulutaş and Aslantaş [8] found that unweighted averaging outperformed their weighted (maximum voting) strategies, the opposite of the trend observed in this study, where weighted soft voting achieved the highest and most consistently significant performance. This divergence suggests that the relative benefit of weighting may depend on the performance gap and architectural diversity among the constituent models. With only two comparably capable lightweight models, as in this study, weighting appears to offer a more consistent advantage than with larger, more heterogeneous ensembles.

Further contextualizing the behavior of ensemble weighting, Javidan *et al.* [10], [11] consistently reported that weighted voting outperformed simple unweighted averaging across both classical machine learning and deep learning base classifiers. Although their research utilized a different dataset and is therefore excluded from the direct empirical benchmark presented in Table 4, their methodological findings provide critical external validation. Their conclusions are directionally consistent with the present study, where weighted soft voting achieved the highest overall numerical performance. While the performance gain of the weighted strategies over unweighted soft voting in this study did not reach statistical significance according to McNemar's test, the consistent numerical superiority aligns with the broader literature, reinforcing the premise that intelligently weighting constituent models generally offers a robust, albeit sometimes marginal, advantage over simple aggregation.

CONCLUSION

This study systematically compared three decision-level voting strategies, hard voting with confidence-based tie breaking, soft voting, and weighted soft voting for combining EfficientNet-B0 and MobileNetV2 in tomato leaf disease classification. All ensemble strategies outperformed the best single model (EfficientNet-B0), with weighted soft voting achieving the highest overall performance, reaching 95.15% accuracy and a macro-F1 score of 0.9417 compared with 94.05% accuracy and 0.9266 macro-F1 score for the best single model. McNemar's test further showed that only soft voting and weighted soft voting yielded statistically significant improvements over the best single model, whereas no statistically significant differences were observed among the three ensemble voting strategies. These findings demonstrate that probability-based decision-level ensemble learning is an effective approach for improving the predictive performance of lightweight convolutional neural networks while maintaining a simple ensemble framework. Comparative results provide empirical evidence for selecting appropriate decision-level voting strategies in lightweight CNN ensembles for plant disease classification and aid in engineering more dependable machine learning solutions for precision agriculture.

REFERENCES

- [1] I. N. Yulita, N. A. Amri, and A. Hidayat, "Mobile Application for Tomato Plant Leaf Disease Detection Using a Dense Convolutional Network Architecture," *Computation*, vol. 11, no. 2, Feb. 2023, doi: 10.3390/computation11020020.
- [2] D. Peng, W. Li, H. Zhao, G. Zhou, and C. Cai, "Recognition of Tomato Leaf Diseases Based on DIMPCNET," *Agronomy*, vol. 13, no. 7, Jul. 2023, doi: 10.3390/agronomy13071812.
- [3] M. Bouni, B. Hssina, K. Douzi, and S. Douzi, "Synergistic use of handcrafted and deep learning features for tomato leaf disease classification," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-71225-5.
- [4] M. S. A. M. Al-gaashani, F. Shang, M. S. A. Muthanna, M. Khayyat, and A. A. Abd El-Latif, "Tomato leaf disease classification by exploiting transfer learning and feature concatenation," *IET Image Process.*, vol. 16, no. 3, pp. 913–925, Feb. 2022, doi: 10.1049/ipr2.12397.
- [5] 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, Dec. 2024, doi: 10.1186/s12870-024-04825-y.
- [6] J. Sharma *et al.*, "Deep learning based ensemble model for accurate tomato leaf disease classification by leveraging ResNet50 and MobileNetV2 architectures," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-98015-x.
- [7] N. Ullah *et al.*, "A Lightweight Deep Learning-Based Model for Tomato Leaf Disease Classification," *Computers, Materials and Continua*, vol. 77, no. 3, pp. 3969–3992, 2023, doi: 10.32604/cmc.2023.041819.
- [8] H. Ulutaş and V. Aslantaş, "Design of Efficient Methods for the Detection of Tomato Leaf Disease Utilizing Proposed Ensemble CNN Model," *Electronics (Switzerland)*, vol. 12, no. 4, Feb. 2023, doi: 10.3390/electronics12040827.
- [9] A. H. Ali, A. Youssef, M. Abdelal, and M. A. Raja, "An ensemble of deep learning architectures for accurate plant disease classification," *Ecol. Inform.*, vol. 81, Jul. 2024, doi: 10.1016/j.ecoinf.2024.102618.
- [10] S. M. Javidan, A. Banakar, K. A. Vakilian, and Y. Ampatzidis, "Tomato leaf diseases classification using image processing and weighted ensemble learning," *Agron. J.*, vol. 116, no. 3, pp. 1029–1049, May 2024, doi: 10.1002/agj2.21293.
- [11] S. M. Javidan, Y. Ampatzidis, A. Banakar, K. A. Vakilian, and K. Rahnama, "An Intelligent Group Learning Framework for Detecting Common Tomato Diseases Using Simple and Weighted Majority Voting with Deep Learning Models," *AgriEngineering*, vol. 7, no. 2, Feb. 2025, doi: 10.3390/agriengineering7020031.
- [12] H. Raval and J. Chaki, "Ensemble transfer learning meets explainable AI: A deep learning approach for leaf disease detection," *Ecol. Inform.*, vol. 84, Dec. 2024, doi: 10.1016/j.ecoinf.2024.102925.
- [13] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0192-5.
- [14] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, Oct. 2018, doi: 10.1016/j.neunet.2018.07.011.
- [15] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning*, 2019, pp. 6105–6114. Accessed: Jul. 11, 2026. [Online]. Available: <https://proceedings.mlr.press/v97/tan19a.html>
- [16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.
- [17] G. Geetharamani and A. P. J., "Identification of plant leaf diseases using a nine-layer deep convolutional neural network," *Computers and Electrical Engineering*, vol. 76, pp. 323–338, Jun. 2019, doi: 10.1016/j.compeleceng.2019.04.011.
- [18] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
- [19] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 2017, [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [20] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," May 2016, [Online]. Available: <http://arxiv.org/abs/1605.08695>
- [21] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.