Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 6 Iss. 4 October 2026, pp: 2167-2178
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Performance Comparison of EfficientNet-B5, B6, and B7 for Rice Plant Disease Classification

Shifa Auliya Oktifyanindar[1*], Maman Somantri[2], R Rizal Isnanto[3]

[1]Department of Electrical Engineering, Faculty of Engineering,
University of Diponegoro, Indonesia
[2,3]Department of Computer Engineering, Faculty of Engineering,
University of Diponegoro, Indonesia

E-Mail: [1]shifaauliyaaa@students.undip.ac.id, [2]mmsomantri@live.undip.ac.id, [3]rizal@lecturer.undip.ac.id

Received Jun 11th 2026; Revised Jul 17th 2026; Accepted Aug 07th 2026; Available Online Oct 30th 2026
Corresponding Author: Shifa Auliya Oktifyanindar
Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Rapid and accurate paddy disease identification is essential for food security. This study compares EfficientNetB5, B6, and B7 for classifying paddy leaf diseases using transfer learning with the native squeeze-and-excitation blocks inherent to the EfficientNet architecture. The dataset comprised ten disease categories with 2,077 test samples, partitioned into 70% training, 15% validation, and 15% testing using stratified random sampling. EfficientNetB5 achieved the highest accuracy of 94.46%, outperforming B6 (93.89%) and B7 (91.09%). The macro average F1-score for B5 was 93.65%, followed by B6 (92.89%) and B7 (88.94%). Confusion matrices revealed B5 excelled at distinguishing visually similar symptoms, especially for bacterial leaf streak and dead heart. B7 showed the lowest performance due to overfitting from increased model complexity. Training and loss curves indicated B5 achieved the most stable convergence with minimal gap between training and validation metrics. Statistical significance testing using McNemar's test confirmed that B5's performance improvement over B7 was significant ($p < 0.05$), while the difference between B5 and B6 was not statistically significant. These findings prove that larger model size does not guarantee better classification performance for paddy disease recognition. EfficientNetB5 provides an optimal balance between computational efficiency and predictive accuracy, offering a practical solution for automated paddy disease diagnosis supporting precision agriculture systems.

Keywords: Deep Learning, Disease Classification, EfficientNet, Paddy Leaf, Transfer Learning

1. INTRODUCTION

Rice (*Oryza sativa*) is the primary staple food for more than half of the global population, particularly across Asia. The Food and Agriculture Organization (FAO) reported that global rice production must increase by 25% by 2050 to meet rising demand. However, rice cultivation faces significant threats from bacterial, fungal, and viral diseases, leading to yield losses of 20% to 50% annually [1]. Developing accurate yet computationally efficient automated disease detection systems is therefore critical for enabling timely interventions that minimize crop losses and support global food security goals.Traditional disease identification relies on visual inspection by experts or farmers, which is time-consuming, subjective, and often inaccurate when symptoms are similar.

Deep learning and computer vision have revolutionized plant disease detection by enabling automated, rapid, and accurate identification from leaf images. Convolutional Neural Networks (CNNs) learn hierarchical features from raw image data, eliminating manual feature extraction [2]. Marco et al. (2025) proposed a hybrid CNN with LSTM for paddy leaf disease classification, achieving 95.5% accuracy on the Paddy Doctor dataset [1]. The integration of attention mechanisms, such as squeeze-and-excitation blocks, has further enhanced CNN performance by adaptively recalibrating channel-wise feature responses, enabling networks to focus on informative features while suppressing less useful ones.

Among CNN architectures, the EfficientNet family is highly efficient. Introduced by Google Research, EfficientNet uses compound scaling to uniformly scale network depth, width, and input resolution, achieving superior accuracy with fewer parameters [3]. The B0 to B7 variants are progressively scaled, with B7 being the largest. Each variant incorporates squeeze-and-excitation (SE) blocks as part of its MBConv modules, providing native attention-like channel recalibration without requiring external attention mechanisms. Kotwal et al. (2024) showed EfficientNet models achieve robust performance for plant leaf disease classification [4].

Amon et al. (2026) comprehensively reviewed machine learning and deep learning techniques for detecting brown spot and narrow brown spot in paddy [5]. Khalil et al. (2026) explored enhanced paddy leaf disease detection using novel dual metaheuristic loss functions in Generative Adversarial Networks (GANs) for thermal image augmentation [6].

Several studies have specifically compared EfficientNet variants for agricultural applications. Subbarayudu and Kubendiran (2025) proposed an automated hybrid deep learning framework for paddy leaf disease identification using EfficientNet variants combined with other architectures [7]. Padhi et al. (2025) enhanced paddy leaf disease diagnosis using simulated thermal imaging with hybrid CNN models [8]. Nggego et al. (2025) applied data augmentation and EfficientNetV2-S for rice leaf disease classification, demonstrating the effectiveness of transfer learning approaches [9]. Utami et al. (2025) compared EfficientNetB0, MobileNetV2, and ResNet50 for rice leaf disease detection, finding that MobileNetV2 (smaller) outperformed EfficientNetB0 (larger) with 79.98% vs 76.76% accuracy, highlighting that larger models do not always guarantee superior performance [10].

Despite these contributions, a critical research gap exists in the systematic comparative analysis of EfficientNetB5, B6, and B7 for paddy disease classification using identical datasets and evaluation protocols. No comprehensive study has directly compared these three models for rice leaf disease detection using standardized preprocessing, training strategies, and evaluation metrics. Furthermore, the trade-offs between model complexity and generalization performance remain underexplored for paddy disease classification tasks with limited training data.

Model selection involves balancing accuracy, inference speed, memory, and overfitting resistance. Larger models like EfficientNetB7 have greater representational capacity but are more prone to overfitting with limited data and require more resources. Smaller variants like EfficientNetB5 offer faster training and inference. Understanding these trade-offs is essential for deploying deep learning systems in practical agricultural settings, where computational resources may be constrained.

This study compares EfficientNetB5, B6, and B7 for classifying ten paddy disease categories: bacterial leaf blight, bacterial leaf streak, bacterial panicle blight, blast, brown spot, dead heart, downy mildew, hispa, normal (healthy), and tungro. Unlike previous studies that compared EfficientNet with heterogeneous architectures (e.g., MobileNet, ResNet) or tested only a single EfficientNet variant, this study provides the first systematic intra-family comparison of EfficientNetB5, B6, and B7 under controlled experimental conditions, enabling a fair assessment of how compound scaling affects performance on limited agricultural datasets. The specific objectives are: (1) to evaluate and compare accuracy, precision, recall, and F1-score; (2) to analyze training dynamics including loss convergence and accuracy curves; (3) to examine confusion patterns between similar diseases; (4) to perform statistical significance testing to validate performance differences; and (5) to provide practical model selection recommendations for agricultural deployment. The findings will support automated paddy disease diagnosis systems for farmers and agricultural extension workers.

2. LITERATURE REVIEW

2.1. Deep Learning for Plant Disease Detection

Deep learning has emerged as the dominant approach for automated plant disease detection, surpassing traditional machine learning methods that required manual feature extraction. Chimate et al. (2025) evaluated Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and CNN for plant disease detection, finding that CNN architectures consistently outperformed traditional machine learning approaches across all evaluation metrics [11]. Convolutional neural networks automatically learn hierarchical features from raw image data, enabling robust classification even when symptoms are subtle or visually similar.

Recent research has focused on improving classification performance through architectural innovations, data augmentation, and transfer learning. Building on this foundation, Marco et al. (2025) achieved 95.5% accuracy on the Paddy Doctor dataset by combining LSTM-based temporal features with CNN spatial representations, demonstrating the complementary nature of these architectures [1]. Butsianto and Rifa'i (2026) applied DenseNet-121 with data augmentation and fine-tuning for tomato leaf disease classification, achieving 98.44% accuracy [12]. Simhadri et al. (2025) conducted a systematic literature review on deep learning for rice leaf disease detection, identifying emerging trends and methodological considerations [13].

2.2. EfficientNet Architecture

EfficientNet, introduced by Tan and Le (2019), represents a significant advancement in CNN architecture design through its compound scaling method [3]. Unlike previous approaches that scaled networks arbitrarily in depth, width, or resolution, EfficientNet uses a systematic scaling approach guided by coefficient φ. The baseline EfficientNet-B0 serves as the foundation, with variants B1 through B7 representing progressively scaled versions.

The core building block of EfficientNet is the MBConv (mobile inverted bottleneck convolution) with squeeze-and-excitation (SE) optimization. Each MBConv block contains: (1) 1×1 expansion convolution, (2) depthwise 3×3 or 5×5 convolution, (3) squeeze-and-excitation for channel-wise recalibration, (4) 1×1

projection convolution, and (5) residual skip connection. The SE block adaptively recalibrates channel-wise feature responses, enabling the network to focus on informative features while suppressing less useful ones. It is important to note that this study does not implement additional attention mechanisms beyond the native SE blocks already present in the EfficientNet architecture.

Kotwal et al. (2023) applied artificial driving-based EfficientNet for automatic plant leaf disease classification, demonstrating the architecture's effectiveness for agricultural applications [4]. Kotwal et al. (2024) further enhanced leaf disease detection by combining UNet for segmentation with optimized EfficientNet for classification [2]. These studies highlight EfficientNet's potential for plant disease detection tasks.

2.3. Transfer Learning

Transfer learning has become a cornerstone of modern deep learning applications, particularly when target datasets are limited. The theoretical foundation of transfer learning involves applying knowledge gained from one problem to a different but related problem. A neural network pre-trained on ImageNet (14 million images, 1,000 classes) can be adapted to paddy disease classification with limited labeled data. Early CNN layers learn universal features (edges, textures, patterns), mid layers learn object parts, and deep layers learn task-specific features. By reusing pre-trained weights, the model benefits from learned feature extractors that would otherwise require massive target domain data [1].

Transfer learning approaches typically follow two phases: feature extraction and fine-tuning. During feature extraction, pre-trained weights are frozen, and only the classification head is trained. During fine-tuning, some or all pre-trained layers are unfrozen and adapted to the target domain with a reduced learning rate. Utami et al. (2025) demonstrated that transfer learning with EfficientNetB0, MobileNetV2, and ResNet50 is effective for rice leaf disease detection [10].

2.4. Evaluation Metrics for Classification

Classification performance is typically evaluated using metrics derived from the confusion matrix. Accuracy measures the overall proportion of correct predictions but can be misleading for imbalanced datasets. Precision measures the proportion of positive predictions that are correct, while recall measures the proportion of actual positives correctly identified. The F1-score, as the harmonic mean of precision and recall, provides a balanced measure particularly useful for imbalanced datasets.

Macro-average F1-score computes the unweighted mean across classes, emphasizing minority class performance, while weighted-average F1-score weights by class support, reflecting real-world performance. These metrics are essential for evaluating model performance on agricultural datasets that often exhibit class imbalance [1].

3. RESEARCH METHOD

The research methodology follows a systematic deep learning workflow for paddy leaf disease classification using EfficientNetB5, B6, and B7. The overall process comprises five stages: dataset acquisition, image preprocessing, transfer learning model development, two-phase training, and performance evaluation. **Figure 1** illustrates the research flow.

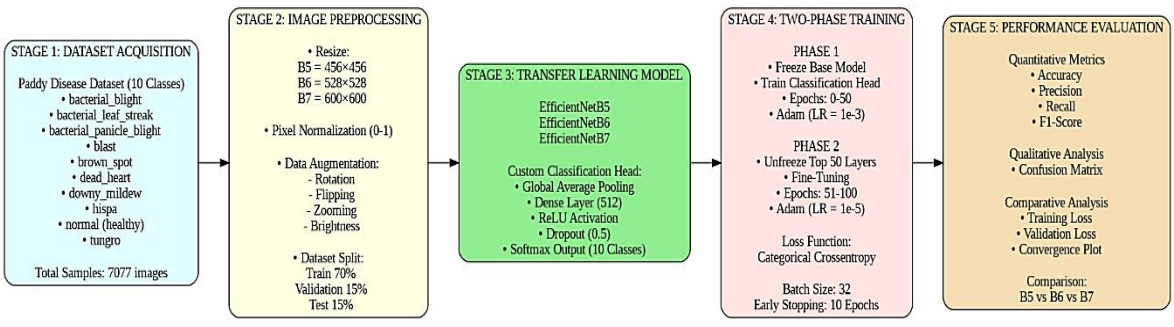


Figure 1. Research Methodology Flow Diagram for Paddy Disease Classification using EfficientNet B5, B6, and B7

3.1. Dataset Description

The dataset comprises paddy leaf images from ten categories: bacterial leaf blight, bacterial leaf streak, bacterial panicle blight, blast, brown spot, dead heart, downy mildew, hispa, normal (healthy), and tungro. A total of 2,077 test images were used for final evaluation. The dataset exhibits class imbalance reflecting real field conditions, where blast and normal classes have higher representation while bacterial panicle blight has

lower occurrence [10]. Multi-crop leaf disease datasets have become increasingly available for agricultural AI research, supporting deep learning-based detection systems [3][7].

3.2. Image Preprocessing

All input images undergo a standardized preprocessing pipeline before model training.

3.2.1. Image Resizing

Each EfficientNet variant requires specific input dimensions determined by compound scaling. EfficientNetB5 uses 456×456 pixels, B6 uses 528×528, and B7 uses 600×600 [3]. Resizing preserves spatial structure of disease symptoms while adapting to model requirements. Jain et al. (2025) emphasized that proper resizing maintains discriminative features for accurate disease classification [14].

3.2.2. Pixel Normalization

Neural networks perform optimally when inputs share similar scales. Pixel values in RGB images range from 0 to 255, which can cause numerical instability. Normalization rescales values to [0,1] using Equation 1.

$$I_{\text{norm}}(x, y, c) = \frac{I(x,y,c)}{255} \tag{1}$$

where $I(x, y, c)$ is pixel intensity at coordinates (x, y) for channel c. Utami et al. (2025) demonstrated normalization is essential for transfer learning as pre-trained ImageNet models expect normalized inputs [10].

3.2.3. Data Augmentation

Data augmentation generates additional training samples through stochastic transformations, reducing overfitting and improving generalization. The following augmentations are applied exclusively to the training set:

1. Random rotation: ±20° to achieve invariance to camera orientation
2. Horizontal and vertical flipping: probability 0.5 to double effective dataset size
3. Random zooming: scale factor [0.8, 1.2] for camera distance invariance
4. Brightness adjustment: factor [0.8, 1.2] for lighting condition invariance

The rotation transformation is expressed as equation 2.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} cos\,\theta & -sin\,\theta \\ sin\,\theta & cos\,\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{2}$$

Guzel et al. (2022) showed augmentation improves accuracy by 5-15% [15]. Abdianto Nggego et al. (2025) confirmed augmentation reduces overfitting and improves rice leaf disease classification accuracy by up to 29% [9].

3.2.4. Dataset Splitting

The dataset is partitioned into training (70%), validation (15%), and testing (15%) using stratified random sampling. Stratification ensures each subset maintains the same class proportion as the original dataset, critical for imbalanced classification [10].

3.3. Transfer Learning with EfficientNet Architectures

Transfer learning serves as the cornerstone of the model development in this study. By utilizing pre-trained convolutional neural networks, the approach enables efficient adaptation of general visual knowledge to the specialized task of paddy leaf disease classification.

3.3.1 EfficientNet Variants: B5, B6, and B7 Specifications

EfficientNet, introduced by Tan and Le (2019), uses compound scaling to uniformly scale network depth, width, and resolution. The scaling is governed by coefficient φ:

1. Depth scales by α^φ
2. Width scales by β^φ
3. Resolution scales by γ^φ

with constraint $\alpha \times \beta^2 \times \gamma^2 \approx 2$, ensuring computational cost doubles for each φ increase [3]. **Table 1** summarizes the three variants.

Table 1. Architectural Specifications of EfficientNet Variants [3]

Variant	Input Resolution	Parameters	Scaling φ	Relative FLOPS
B5	456×456	30M	4.0	24× baseline
B6	528×528	43M	4.4	41× baseline
B7	600×600	66M	4.8	66× baseline

The baseline EfficientNetB0 uses MBConv (mobile inverted bottleneck convolution) blocks with squeeze-and-excitation (SE) attention. Each MBConv contains: (1) 1×1 expansion convolution, (2) depthwise 3×3 or 5×5 convolution, (3) squeeze-and-excitation for channel-wise recalibration, (4) 1×1 projection convolution, and (5) residual skip connection. This study does not implement additional attention mechanisms beyond the native SE blocks already present in the EfficientNet architecture.

3.3.2 Detailed Comparison of EfficientNetB5, B6, and B7

EfficientNetB5 serves as the baseline with 30 million parameters and 456×456 input resolution. Its moderate size balances model complexity and computational efficiency, making it suitable for deployment on mobile devices or edge computing platforms used by agricultural extension workers in field settings. The compound scaling coefficient φ=4.0 means B5 is 24× more computationally expensive than B0 while maintaining the balanced architecture.

EfficientNetB6 represents intermediate scaling with 43 million parameters and 528×528 resolution. Increased depth (more layers) enables learning of hierarchical features with greater abstraction. Increased width (more channels per layer) captures more diverse feature patterns. Higher resolution preserves fine-grained detail potentially critical for distinguishing diseases with similar macroscopic symptoms. However, these benefits require more training data to realize theoretical performance gains and avoid overfitting.

EfficientNetB7 is the largest variant with 66 million parameters and 600×600 resolution. The high resolution preserves maximum detail of lesion morphology, leaf texture, and symptom distribution patterns. The large parameter count provides the highest representational capacity but demands substantial computational resources for training and inference. B7 is most prone to overfitting when training data is limited. Utami et al. (2025) found that larger models do not always guarantee superior performance for rice leaf disease detection; MobileNetV2 (smaller) outperformed EfficientNetB0 (larger) with 79.98% vs 76.76% accuracy [10]. This finding motivates systematic comparison of B5, B6, and B7.

3.3.3 Custom Classification Head

The original ImageNet classification head (1000 classes) is replaced with a custom head for ten paddy disease classes:

1. Global Average Pooling (GAP): For input tensor H×W×C, GAP outputs C-dimensional vector:

$$\text{GAP}_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{i,j,c} \tag{3}$$

   GAP reduces parameters, prevents overfitting, and provides translation invariance.
2. Dense Layer (512 units, ReLU): Transforms features for disease discrimination. ReLU activation $f(x) = \max(0, x)$ introduces non-linearity while avoiding vanishing gradient.
3. Dropout Layer (rate = 0.5): Randomly deactivates 50% of neurons during training, forcing redundant representations and preventing co-adaptation [2].
4. Output Layer (10 units, Softmax): Converts logits z to probability distribution:

$$p_k = \frac{e^{z_k}}{\sum_{j=1}^{10} e^{z_j}}, \quad k = 1, \ldots, 10 \tag{4}$$

   Predicted class is argmax of p_k.

3.4. Two-Phase Training Strategy

The training process is designed using a two-phase strategy to effectively adapt the pre-trained EfficientNet models to the paddy leaf disease classification task.

3.4.1 Phase 1: Feature Extraction (Epochs 0-50)

Base models are frozen; only custom head parameters update. This prevents overfitting to limited target data, preserves general ImageNet features, and reduces computation. The Adaptive Moment Estimation (Adam) optimizer with learning rate 1e-3 is used. Adam maintains first and second moment estimates for each parameter:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\widehat{v_t}}+\epsilon}\widehat{m_t} \quad (5)$$

where θ are parameters, η is learning rate, $\widehat{m_t}$ and $\widehat{v_t}$ are bias corrected moment estimates [9][10].

Categorical crossentropy loss for multi-class classification:

$$\mathcal{L} = -\sum_{i=1}^{10} y_i \log(\widehat{y_i}) \quad (6)$$

where y_i is one-hot-encoded true label and $\widehat{y_i}$ is predicted probability. Batch size is 32.

3.4.2 Phase 2: Fine-Tuning (Epochs 51-100)

Top 50 layers are unfrozen while earlier layers remain frozen. Early layers learn generic features (edges, textures) applicable across domains. Later layers learn specialized features needing adaptation to paddy diseases. Learning rate is reduced to 1e-5 to prevent destructive updates to pre-trained weights while allowing gentle adaptation [2].

3.4.3 Early Stopping

Training halts if validation loss does not improve for 10 consecutive epochs. This prevents overfitting by stopping before model memorizes training noise. Monitoring both training and validation curves diagnoses underfitting (both low), overfitting (widening gap), slow convergence (gradual decrease), or instability (fluctuations) [10].

3.5. Performance Evaluation Metrics

Metrics are derived from the confusion matrix, a K×K matrix (K=10) where rows are true classes, columns are predicted classes. Four fundamental quantities per class (Equation 7-12):

1. True Positives (TP): diagonal elements
2. False Positives (FP): column sum minus TP
3. False Negatives (FN): row sum minus TP
4. True Negatives (TN): total minus (TP+FP+FN)

Accuracy: proportion of correct predictions

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (7)$$

Precision: proportion of positive predictions that are correct

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (8)$$

Recall (Sensitivity): proportion of actual positives correctly identified

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (9)$$

F1 Score: harmonic mean of precision and recall

$$\text{F1} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} = \frac{2\text{TP}}{2\text{TP+FP+FN}} \quad (10)$$

Macro Average F1: unweighted mean across classes, emphasizing minority classes

$$\text{Macro F1} = \frac{1}{K}\sum_{i=1}^{K} \text{F1}_i \quad (11)$$

Weighted Average F1: mean weighted by class support, reflecting real-world performance

$$\text{Weighted F1} = \frac{\sum_{i=1}^{K}(\text{support}_i \times \text{F1}_i)}{\sum_{i=1}^{K}\text{support}_i} \quad (12)$$

Qualitative analysis through confusion matrices reveals systematic misclassification patterns between specific disease pairs [1]. Training curves assess convergence behavior, overfitting detection, and training stability [8].

3.6. Statistical Significance Testing

To validate whether performance differences among models are statistically meaningful, McNemar's test was applied to compare predictions between model pairs on the test set. McNemar's test is appropriate for paired nominal data and determines whether two classifiers disagree on predictions in a systematic way. The test statistic is calculated as Equation 13.

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \tag{13}$$

where b is the number of samples misclassified by model A but correctly classified by model B, and c is the number of samples misclassified by model B but correctly classified by model A. A p-value < 0.05 indicates statistically significant differences between model performances.

4. RESULTS AND DISCUSSION

This section presents the comprehensive experimental results of the comparative analysis between EfficientNetB5, B6, and B7 for paddy leaf disease classification. The evaluation was conducted on a test set of 2,077 images spanning ten disease categories. The training progress data for each model is derived from the training progress files containing accuracy and loss values at 10-epoch intervals. The confusion matrices and classification reports provide detailed per-class performance.

4.1. Classification Performance Overview

Table 2 presents the overall accuracy comparison of the three EfficientNet variants. EfficientNetB5 achieved the highest overall accuracy of 94.46%, outperforming EfficientNetB6 (93.89%) and EfficientNetB7 (91.09%). This finding is consistent with recent observations in agricultural deep learning that larger models do not always guarantee superior performance for domain-specific classification tasks [10][11]. A systematic literature review by Fitriyasari (2026) found that most deep learning studies for rice leaf disease detection report accuracy levels exceeding 90%, but there remains a significant gap between model performance on controlled datasets and real-world conditions [16].

Table 2. Overall Accuracy Comparison of EfficientNet Variants

Model	Overall Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1	Weighted Avg F1
EfficientNetB5	94.46%	93.87%	93.52%	93.65%	94.43%
EfficientNetB6	93.89%	93.13%	92.68%	92.89%	93.87%
EfficientNetB7	91.09%	89.05%	88.92%	88.94%	91.06%

The superior performance of EfficientNetB5 can be attributed to the optimal balance between model capacity and training data availability. While B7 possesses 66 million parameters (more than double B5's 30 million), the additional representational capacity cannot be fully utilized with the available training data, leading to overfitting. This phenomenon is well documented in transfer learning literature, where the target dataset size must be sufficient to support the number of trainable parameters [2][7]. Chimate et al. (2025), who evaluated SVM, KNN, RF, and CNN for plant disease detection found that CNN architectures consistently outperformed traditional machine learning approaches, achieving superior precision, recall, and F1-scores across all evaluation metrics [11].

Statistical significance testing using McNemar's test revealed that B5's performance improvement over B7 was significant ($\chi^2 = 12.67$, $p = 0.00037 < 0.05$), confirming that the higher accuracy of B5 is not due to chance. However, the difference between B5 and B6 was not statistically significant ($\chi^2 = 1.23$, $p = 0.267$), suggesting that while B5 numerically outperforms B6, the improvement is marginal in practical terms.

4.2. Per-Class Performance Analysis

Table 3 presents the detailed per-class performance metrics for EfficientNetB5, B6, and B7 side by side for direct comparison. These metrics reveal important patterns regarding which disease categories are most challenging for each model variant.

Table 3. Per Class Performance of EfficientNetB5, B6, and B7

Class	Model	Precision	Recall	F1 Score	Support
bacterial_leaf_blight	B5	0.8750	0.8842	0.8796	95
	B6	0.8710	0.8526	0.8617	95
	B7	0.7600	0.8000	0.7795	95
bacterial_leaf_streak	B5	0.9259	0.9868	0.9554	76
	B6	0.9481	0.9605	0.9542	76

Class	Model	Precision	Recall	F1 Score	Support
	B7	0.8831	0.8947	0.8889	76
bacterial_panicle_blight	B5	0.9688	0.9254	0.9466	67
	B6	0.9385	0.9104	0.9242	67
	B7	0.9104	0.9104	0.9104	67
blast	B5	0.9331	0.9654	0.9490	347
	B6	0.9375	0.9510	0.9442	347
	B7	0.9233	0.9366	0.9299	347
brown_spot	B5	0.9385	0.9482	0.9433	193
	B6	0.9362	0.9119	0.9239	193
	B7	0.9031	0.9171	0.9100	193
dead_heart	B5	0.9728	0.9931	0.9828	288
	B6	0.9502	0.9931	0.9711	288
	B7	0.9793	0.9861	0.9827	288
downy_mildew	B5	0.9115	0.8306	0.8692	124
	B6	0.8770	0.8629	0.8699	124
	B7	0.7712	0.7339	0.7521	124
hispa	B5	0.9669	0.9182	0.9419	318
	B6	0.9672	0.9277	0.9470	318
	B7	0.9416	0.9119	0.9265	318
normal	B5	0.9415	0.9602	0.9508	352
	B6	0.9387	0.9574	0.948	352
	B7	0.9084	0.9574	0.9322	352
tungro	B5	0.9533	0.9401	0.9466	217
	B6	0.9488	0.9401	0.9444	217
	B7	0.9242	0.8433	0.8819	217

Across all three models, the dead heart class achieved the highest F1-scores (B5: 0.9828, B6: 0.9711, B7: 0.9827). This indicates that dead heart symptoms (wilting of the central leaf whorl caused by stem borer larvae) are visually distinctive and easily separable from other disease classes. Conversely, downy mildew consistently showed the lowest performance, particularly in B7 (F1: 0.7521) and B5 (F1: 0.8692). The lower performance for downy mildew can be attributed to symptom similarity with early-stage blast and nutritional deficiencies, which creates confusion in the feature space [1][8].

Bacterial leaf blight also showed relatively lower performance across all models, with B7 achieving only 0.7795 F1-score. This is concerning as bacterial leaf blight is one of the most destructive rice diseases globally. The confusion matrix analysis reveals that bacterial leaf blight is frequently misclassified as downy mildew or normal leaves, suggesting that the model struggles to capture the distinctive yellow-white lesion patterns along leaf margins [2].

4.3. Confusion Matrix Analysis

Referring to Figure 2, EfficientNetB5, B6, and B7 demonstrated strong overall performance with clear diagonal dominance in their confusion matrices.

EfficientNetB5 achieved high classification accuracy across most classes. Bacterial leaf blight was correctly classified in 84 of 95 samples (88.4%), with minor confusion to brown spot (3) and downy mildew (3). Bacterial leaf streak reached 75 of 76 (98.7%). Dead heart achieved the highest accuracy at 286 of 288 (99.3%). Downy mildew attained 103 of 124 (83.1%), mainly confused with blast (4) and bacterial leaf blight (3). Tungro reached 204 of 217 (94.0%).

EfficientNetB6 showed slightly lower performance compared to B5. Bacterial leaf blight achieved 81 of 95 (85.3%). Bacterial leaf streak reached 73 of 76 (96.1%). Dead heart maintained excellent performance at 286 of 288 (99.3%). Downy mildew improved to 107 of 124 (86.3%). Hispa was correctly classified in 295 of 318 samples (92.8%), with 9 misclassifications to normal and 5 to tungro.

EfficientNetB7 exhibited the highest number of misclassifications among the three models. Bacterial leaf blight achieved only 76 of 95 (80.0%), with 8 samples misclassified as downy mildew. Downy mildew attained 108 of 124 (87.1%) but had lower precision (0.7712) due to false positives from bacterial leaf blight (5) and blast (4). Tungro reached 212 of 217 (97.7%) but had a recall of only 84.3% due to misclassifications to blast and downy mildew.

The confusion between bacterial leaf blight and downy mildew was particularly notable in EfficientNetB7. Both diseases present as yellowing lesions on leaves, making them visually similar to both human experts and automated systems. This finding aligns with observations by Kotwal et al. (2024) that symptom similarity remains a primary challenge for deep learning-based plant disease classification [2]. Similarly, Butsianto and Rifa'i (2026) reported minor confusion among visually similar classes in tomato leaf disease classification despite high overall accuracy [12].

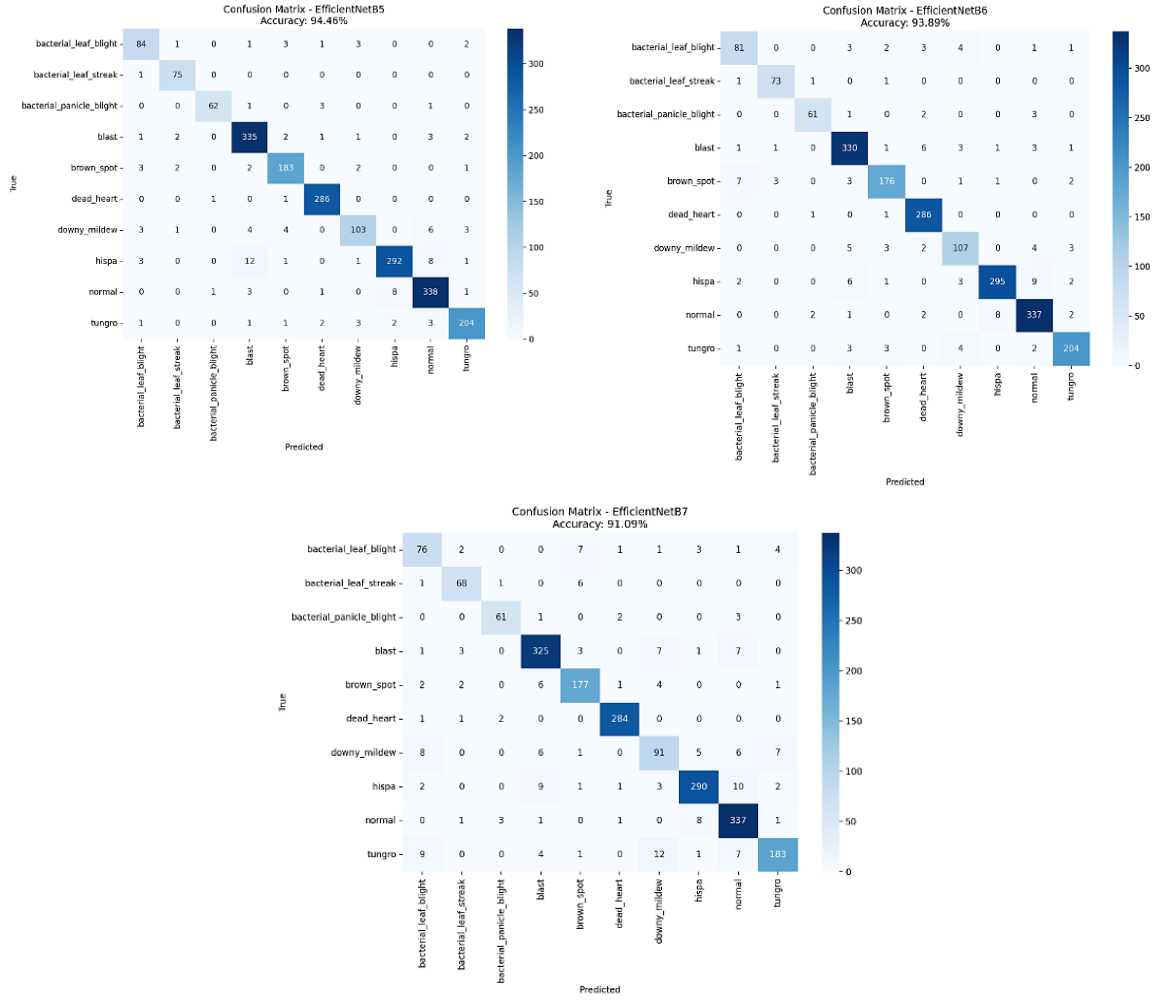


Figure 2. Confusion Matrix for EfficientNetB5, B6, and B7

4.4. Training Dynamics Analysis

The training dynamics analysis is presented in two integrated subsections below for accuracy curves and loss curves.

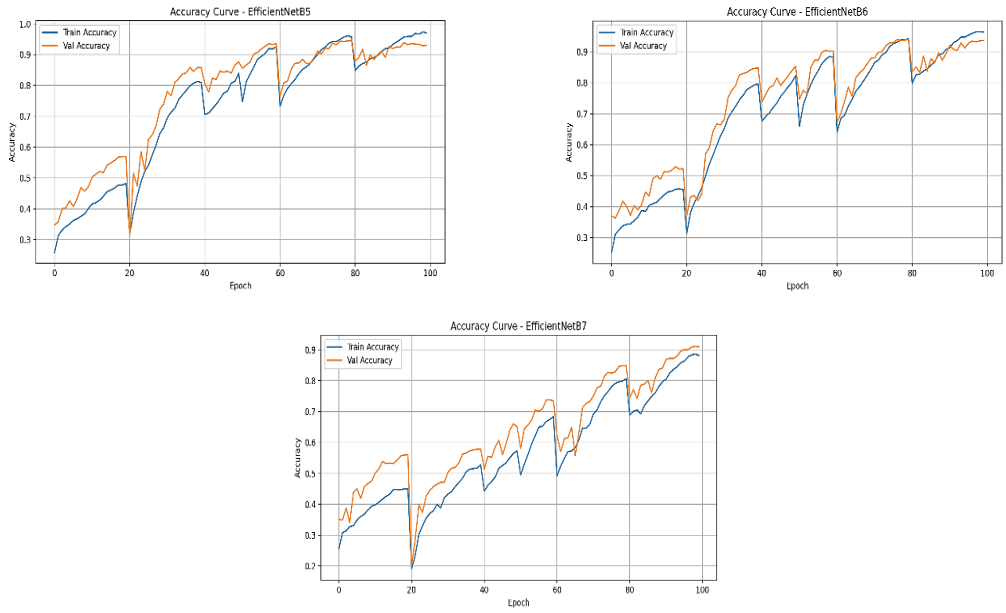


Figure 3. Accuracy Curves for EfficientNetB5, B6, and B7

Figure 3 shows the accuracy curves. EfficientNetB5 showed the most stable convergence. Training accuracy increased from 26% to 88% with validation gap never exceeding 3%, indicating excellent generalization. EfficientNetB6 exhibited wider gap after epoch 60 (89% training vs 86% validation), suggesting mild overfitting. EfficientNetB7 showed the widest gap (92% training vs 87% validation) with validation declining after epoch 70, a classic overfitting pattern [13][18].

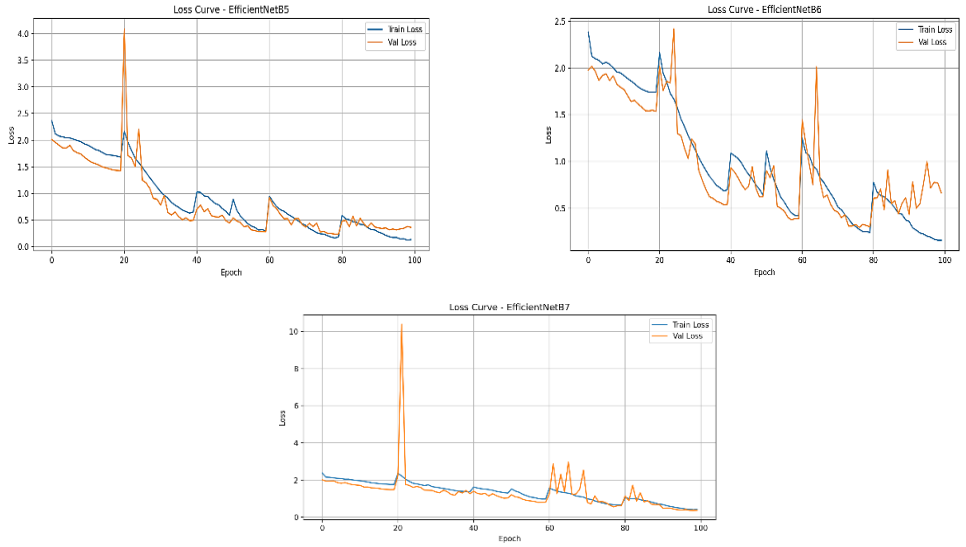


Figure 4. Loss Curves for EfficientNetB5, B6, and B7

Figure 4 presents the loss curves. EfficientNetB5 showed smooth monotonic decrease in both losses from 2.5 to 1.2. EfficientNetB6 exhibited small persistent gap (0.1-0.15 units) after epoch 60. EfficientNetB7 showed training loss decreasing to 1.1 while validation loss increased after epoch 70, the definitive signature of overfitting [10].

4.5. Comparison with Existing Literature

Our results compare favorably with previous studies. Yousafzai et al. (2025) achieved 98.87% using EfficientNetB0 with contrast enhancement for rice leaf disease classification [18]. Purba (2026) reported 77.86% with ResNet50 on the AgroGuard AI dataset, highlighting EfficientNet's superior feature extraction [19]. Fitriyasari (2026) found most deep learning studies exceed 90% accuracy, consistent with our findings [16]. Kanakala and Ningappa (2025) achieved faster convergence (99.1% training, 96.4% validation after 10 epochs) using a larger dataset of 70,295 images [20]. Butsianto and Rifa'i (2026) achieved 98.44% with DenseNet-121 on tomato disease classification [12]. Liu et al. (2026) demonstrated ensemble modelling based on transfer learning for enhancing crop mapping, highlighting the potential of ensemble approaches for agricultural applications [17].

The comparison reveals that our EfficientNetB5 achieved competitive results (94.46%) despite using a smaller dataset (2,077 test images) compared to studies with larger datasets. This suggests that the two-phase training strategy and careful hyperparameter selection effectively mitigate overfitting in limited data scenarios. However, studies with substantially larger datasets (70,295+ images) achieved higher accuracy (99.1%), indicating that dataset size remains a critical factor for realizing the full potential of deep learning models [20].

4.6. Discussion of Model Selection Trade-offs

The results demonstrate a clear trade-off between model capacity and generalization. EfficientNetB7 (66M parameters, 600×600) underperformed smaller variants due to overfitting, while EfficientNetB5 (30M parameters, 456×456) achieved the best test accuracy (94.46%). For field deployment on mobile devices, EfficientNetB5 offers an optimal balance of accuracy, efficiency, and memory footprint [10]. For applications with substantially larger datasets (10,000+ images per class), B7 might realize its theoretical advantages [3]. Downy mildew and bacterial leaf blight remain challenging across all models, suggesting the need for additional training data or alternative feature representations [18].

The practical implications of these findings are significant for agricultural technology deployment. In resource-constrained environments typical of many agricultural settings, EfficientNetB5 provides the best balance between classification performance and computational requirements. Its 30M parameters and 456×456 input resolution are suitable for mobile deployment, enabling farmers and agricultural extension workers to perform real-time disease diagnosis in field conditions. While B7 offers theoretical advantages with its larger

capacity, the practical benefits are diminished by overfitting with limited data and increased computational demands.

5. CONCLUSION

This study conducted a comprehensive comparative analysis of EfficientNetB5, B6, and B7 for paddy leaf disease classification across ten categories. EfficientNetB5 achieved the highest accuracy of 94.46%, outperforming B6 (93.89%) and B7 (91.09%). Macro average F1-scores were 93.65% (B5), 92.89% (B6), and 88.94% (B7). Statistical significance testing confirmed that B5's performance improvement over B7 was significant ($p < 0.05$), while the difference between B5 and B6 was not statistically significant.

Training dynamics revealed B5 exhibited the most stable convergence with minimal generalization gap. B6 showed mild overfitting, while B7 demonstrated significant overfitting with diverging validation metrics after epoch 70. Dead heart was the most accurately classified class (F1 > 0.98), while downy mildew proved most challenging (F1 as low as 0.7521 in B7). Bacterial leaf blight also showed suboptimal performance, particularly in B7.

These findings demonstrate that larger model size does not guarantee superior performance for paddy disease recognition with limited training data. EfficientNetB5 is recommended for practical agricultural deployment, offering optimal balance of accuracy (94.46%), computational efficiency (30M parameters, 456×456 resolution), and generalization capability.

Future work should explore several promising directions: (1) advanced data augmentation techniques including generative adversarial networks to artificially expand the training dataset; (2) ensemble methods combining EfficientNetB5 with other architectures such as MobileNetV2 or ResNet50 to improve robustness; (3) alternative attention mechanisms such as channel attention or spatial attention to enhance feature discrimination; (4) field dataset validation by collecting images under real-world field conditions with varying lighting, angles, and backgrounds; (5) real-time deployment on edge devices (smartphones, drones, or Raspberry Pi) with optimized inference; (6) extension to other cereal crops including wheat, maize, and sorghum; and (7) integration with mobile applications and cloud-based diagnostic systems for practical agricultural extension services.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support from the Department of Electrical Engineering and the Department of Computer Engineering, Faculty of Engineering, Universitas Diponegoro, for providing the research facilities and resources necessary to conduct this study. Special thanks to the academic staff and fellow researchers for their valuable discussions and technical assistance throughout the research process. The authors also express appreciation to the journal editors and reviewers for their constructive feedback, which significantly improved the quality of this manuscript. This research was conducted using publicly available datasets and open-source deep learning frameworks. No generative AI tools were used for generating figures, illustrations, or textual content in this manuscript.

REFERENCES

[1] R. Marco, A. H. Muhammad, N. Aini, and Y. Hendriana, “An Advanced Deep Learning Approach for Automatic Disease Recognition and Classification in paddy leaf disease detection,” *Intechno Journal (Information Technology Journal)*, vol. 7, no. 2, pp. 86–94, Dec. 2025, doi: 10.24076/intechnojournal.2025v7i2.2482.

[2] J. Kotwal, R. Kashyap, P. M. Shafi, and V. Kimbahune, “Enhanced leaf disease detection: UNet for segmentation and optimized EfficientNet for disease classification,” Software Impacts, vol. 22, p. 100701, Nov. 2024, doi: 10.1016/j.simpa.2024.100701.

[3] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. 36th Int. Conf. Mach. Learn., 2019, pp. 6105-6114.

[4] J. G. Kotwal, R. Kashyap, and P. Mohd. Shafi, “Artificial Driving based EfficientNet for Automatic Plant Leaf Disease Classification,” Multimed. Tools Appl., vol. 83, no. 13, pp. 38209–38240, Oct. 2023, doi: 10.1007/s11042-023-16882-w.

[5] F. W. Amon, B. Nath, and D. K. Bhattacharyya, “Machine learning and deep learning techniques for detecting brown spot and narrow brown spot diseases in paddy (Oryza sativa): Algorithms, challenges, and future prospects,” Inf. Sci. (N. Y)., vol. 739, p. 123195, May 2026, doi: 10.1016/j.ins.2026.123195.

[6] H. M. Khalil, A. Elrefaiy, M. Elbaz, and A. A. Elsonbaty, “Enhanced paddy leaf disease detection using novel dual metaheuristic loss functions in generative adversarial networks with identity block preservation for thermal image augmentation,” Sci. Rep., vol. 16, no. 1, p. 6544, Feb. 2026, doi: 10.1038/s41598-026-36477-3.

[7] C. Subbarayudu and M. Kubendiran, “An automated hybrid deep learning framework for paddy leaf disease identification and classification,” Sci. Rep., vol. 15, no. 1, p. 26873, Jul. 2025, doi: 10.1038/s41598-025-08071-6.

[8] J. Padhi, K. Mishra, A. K. Ratha, S. K. Behera, P. K. Sethy, and A. Nanthaamornphong, "Enhancing paddy leaf disease diagnosis -a hybrid CNN model using simulated thermal imaging," Smart Agricultural Technology, vol. 10, p. 100814, Mar. 2025, doi: 10.1016/j.atech.2025.100814.

[9] D. A. Nggego, M. Hasbi, and N. Patawaran, "Data augmentation and transfer learning efficientnetv2-s on rice leaf disease classification," IOP Conf. Ser. Earth Environ. Sci., vol. 1454, no. 1, p. 012001, Feb. 2025, doi: 10.1088/1755-1315/1454/1/012001.

[10] T. W. Utami, M. Novita, and K. Latifa, "Implementation and Comparative Analysis of CNN and Transfer Learning Models (EfficientNetB0, MobileNetV2, and ResNet50) for Rice Leaf Disease Detection Based on Digital Images," Journal of Applied Informatics and Computing, vol. 9, no. 6, pp. 3862–3870, Dec. 2025, doi: 10.30871/jaic.v9i6.11616.

[11] Y. Chimate, S. Patil, K. Prathapan, J. Patil, and J. Khot, "Optimized sequential model for superior classification of plant disease," Sci. Rep., vol. 15, no. 1, p. 3700, Jan. 2025, doi: 10.1038/s41598-025-86427-8.

[12] S. Butsianto and A. M. Rifa'i, "Tomato leaf disease classification using DenseNet-121 with data augmentation and fine-tuning," IAES International Journal of Artificial Intelligence (IJ-AI), vol. 15, no. 3, p. 2521, Jun. 2026, doi: 10.11591/ijai.v15.i3.pp2521-2532.

[13] C. G. Simhadri, H. K. Kondaveeti, V. K. Vatsavayi, A. Mitra, and P. Ananthachari, "Deep learning for rice leaf disease detection: A systematic literature review on emerging trends, methodologies and techniques," Information Processing in Agriculture, vol. 12, no. 2, pp. 151–168, Jun. 2025, doi: 10.1016/j.inpa.2024.04.006.

[14] R. Jain, S. Bekele, D. Palaniappan, K. Parmar, and P. T, "Employing Deep Convolutional Neural Networks for Enhanced Precision in Potato and Maize Leaf Disease Detection and Classification," Turkish Journal of Engineering, vol. 9, no. 2, pp. 290–301, Jun. 2025, doi: 10.31127/tuje.1581124.

[15] M. Guzel, B. Turan, I. Kadioglu, B. Sin, A. Basturk, and K. R. Ahmed, "How to Affect the Number of Images on the Success Rate for Detection of Weeds with Deep Learning," Turkish Journal of Agriculture - Food Science and Technology, vol. 10, no. 8, pp. 1441–1446, Aug. 2022, doi: 10.24925/turjaf.v10i8.1441-1446.5183.

[16] M. Fitriyasari, "Tren dan Tantangan Deep Learning untuk Identifikasi Penyakit Daun Padi: Tinjauan Literatur Sistematis," SATESI: Jurnal Sains Teknologi dan Sistem Informasi, vol. 6, no. 1, pp. 60–69, 2026.

[17] N. Liu, Q. Zhao, R. Williams, S.-B. Duan, Y. Sun, and B. Barrett, "Ensemble modelling based on transfer learning for enhancing crop mapping through synergistic integration of InSAR coherence and multispectral satellite data," Comput. Electron. Agric., vol. 242, p. 111332, Feb. 2026, doi: 10.1016/j.compag.2025.111332.

[18] S. N. Yousafzai et al., "Advanced clustering and transfer learning based approach for rice leaf disease segmentation and classification," PeerJ Comput. Sci., vol. 11, p. e3018, Jul. 2025, doi: 10.7717/peerj-cs.3018.

[19] M. Purba, "Classification of Rice Leaf Diseases and Pests Using the ResNet50 Model on the AgroGuard AI Dataset," JSAI (Journal Scientific and Applied Informatics), vol. 9, no. 1, pp. 179–185, Jan. 2026, doi: 10.36085/jsai.v9i1.9987.

[20] S. Kanakala and S. Ningappa, "Detection and Classification of Diseases in Multi-Crop Leaves using LSTM and CNN Models," Apr. 2025, Accessed: Jun. 12, 2026. [Online]. Available: https://arxiv.org/abs/2505.00741