

Convolutional Neural Network for Chili Plant Disease Classification: A Deep Learning Approach

Erna Dwi Astuti, Widowati, Aris Sugiharto

Doctoral Program of Information System, Postgraduate School-Universitas Diponegoro, Semarang, Indonesia

Abstract—Chili peppers are a high-value horticultural crop that is highly susceptible to foliar diseases, which can significantly reduce yield and market quality. This study proposes and evaluates a Convolutional Neural Network (CNN) model based on the MobileNet-V2 architecture for chili leaf disease classification. A combined dataset consisting of 2,690 images collected from two public repositories and one field-acquired source was used in this research. The dataset was divided into training, validation, and testing subsets using an 80:10:10 ratio and underwent preprocessing steps including image resizing, data augmentation, and normalization. The proposed model was implemented using TensorFlow 2.15 and trained on the Google Colab platform. Experimental results demonstrate strong classification performance, achieving 95.6% validation accuracy and 96.8% test accuracy with a low loss value of 0.1011. All evaluated classes, anthracnose, yellow virus, leaf spot, leaf curl, and healthy leaves achieved precision, recall, and F1-scores exceeding 0.90, accompanied by near-perfect AUC values. These findings indicate that the MobileNet-V2-based CNN exhibits effective discriminative capability and generalization across heterogeneous visual conditions, highlighting its potential applicability for AI-assisted agricultural disease monitoring systems based on image processing techniques.

Keywords—Convolutional neural network; MobileNet-V2; chili plant disease; image processing

I. INTRODUCTION

Chili peppers are a high-value horticultural commodity that plays a critical role in local consumption and regional economies [1]. Disease outbreaks in chili crops can therefore have substantial economic and social impacts, including reduced yield and quality, disrupted supply chains, price volatility, and implications for inflation and food security [1], [2]. Previous studies have reported that certain chili diseases are capable of reducing crop yields by up to 90% [3]. While fungal pathogens alone may cause losses ranging from 10% to 80%, depending on cultivar susceptibility and environmental conditions [4]. In addition, infestations by mites and other pests have been shown to degrade fruit quality and result in yield losses exceeding 60%, reaching as high as 96.4% under severe conditions [4]. These impacts represent a global challenge, constraining agricultural productivity across diverse agroecosystems rather than being confined to specific regions [4], [5], [6].

Chili plant diseases are caused by a wide range of factors, including viral, fungal, and bacterial pathogens, as well as environmental stressors such as extreme climate variability that weaken plant defense mechanisms and accelerate disease

progression [7], [8]. In practice, disease diagnosis often relies on manual inspection by trained experts, whose services may be costly, limited in availability, or inaccessible to small-scale farmers in developing regions [9]. Moreover, visual assessment performed manually is prone to inconsistency, subjectivity, and misdiagnosis [10]. Since many chili diseases manifest through observable symptoms such as leaf discoloration, lesions, curling, and fungal growth, image-based analysis offers an effective medium for capturing diagnostically relevant features [10]. Consequently, automated image-based detection systems have the potential to provide faster, more affordable, and more consistent diagnoses, helping farmers bridge expertise gaps while maintaining crop quality and meeting increasing market demands [2], [9].

Advances in image processing and deep learning have enabled significant progress in automated plant disease detection. Image-based techniques can quantify symptomatic regions, analyze color and texture variations, and identify early-stage infections more reliably than manual observation [11]. Deep learning-based image analysis, in particular, provides a scalable and fully automated framework capable of large-scale monitoring and rapid disease identification under field conditions [12]. Among deep learning approaches, Convolutional Neural Networks (CNNs) are well-suited for visual classification tasks due to their ability to hierarchically learn discriminative features directly from raw pixel data [13], [14], [15], [16]. Through convolutional operations, weight sharing, and pooling mechanisms, CNNs efficiently capture disease-related patterns ranging from low-level textures to high-level morphological indicators without requiring handcrafted feature extraction [14], [15], [16].

In agricultural settings, real-world implementation introduces constraints that are often underestimated. Farmers typically use devices with limited computational capacity, such as smartphones or low-power edge systems, and agricultural image datasets are usually small and highly diverse due to differences in illumination, background, camera specifications, and field environments. These limitations emphasize the need for lightweight CNN models that achieve a balance between predictive accuracy and computational efficiency. MobileNet-V2 fulfills this requirement by employing depthwise separable convolutions, inverted residual structures, and linear bottlenecks, which substantially reduce computational complexity and memory demands while preserving strong feature representation capabilities [10], [17], [18]. Therefore, MobileNet-V2 is well-suited for mobile or edge-based agricultural deployment scenarios [19].

Several previous studies have explored CNN-based approaches for chili leaf disease classification. Reported methods range from custom CNN architectures achieving accuracies above 95% [20] to evaluations of pre-trained models such as DenseNet, EfficientNet, and NasNet-Mobile on curated datasets [21]. Comparative studies have also shown that MobileNet-V2 can outperform conventional CNN baselines and achieve competitive performance relative to heavier architectures [10]. Despite these encouraging results, most existing studies rely on single-source datasets or images collected under controlled conditions. Such experimental setups increase the risk of overfitting and limit the evaluation of model robustness when applied to heterogeneous real-world scenarios. To date, no prior work has systematically evaluated MobileNet-V2 using a combined dataset integrating multiple public sources with field-collected images.

Motivated by these limitations, this study proposes an image-based chili disease classification approach using a MobileNet-V2-based CNN evaluated on a heterogeneous multi-source dataset composed of two public repositories and one field-collected dataset. This design enables the model to be assessed under diverse visual conditions, including variations in illumination, background complexity, and image quality, which more closely reflect real deployment environments. The objective of this study is to evaluate the performance of MobileNet-V2 across standard classification metrics, including accuracy, precision, recall, and F1-score [22], [23], and ROC-AUC, following preprocessing steps such as resizing, data augmentation, and normalization. The primary contribution of this work lies in evaluating a lightweight MobileNet-V2 architecture under a heterogeneous multi-source dataset, providing empirical insight into cross-domain robustness and performance consistency across different acquisition environments, which remains underexplored in existing studies. This study provides practical insight into how lightweight CNN models behave under heterogeneous real-world data conditions, particularly in terms of robustness, generalization, and suitability for resource-constrained deployment.

The remainder of this study is organized as follows. Section II describes the dataset, preprocessing procedures, and the proposed CNN architecture. Section III presents the experimental results and discussion. Finally, Section IV concludes the study and outlines directions for future research.

II. METHODS

This section presents the dataset construction, preprocessing steps, and model training strategy used in this study. A composite dataset of 2,690 chili leaf images was compiled from two public repositories and one field-observed source to capture variability in image quality and acquisition conditions. After preprocessing, the dataset was split into training, validation, and test sets using an 80:10:10 ratio. The training and validation sets were used for model learning and performance monitoring, while the test set was reserved for final evaluation. A MobileNet-V2-based Convolutional Neural Network (CNN) was employed due to its computational efficiency and effective feature representation.

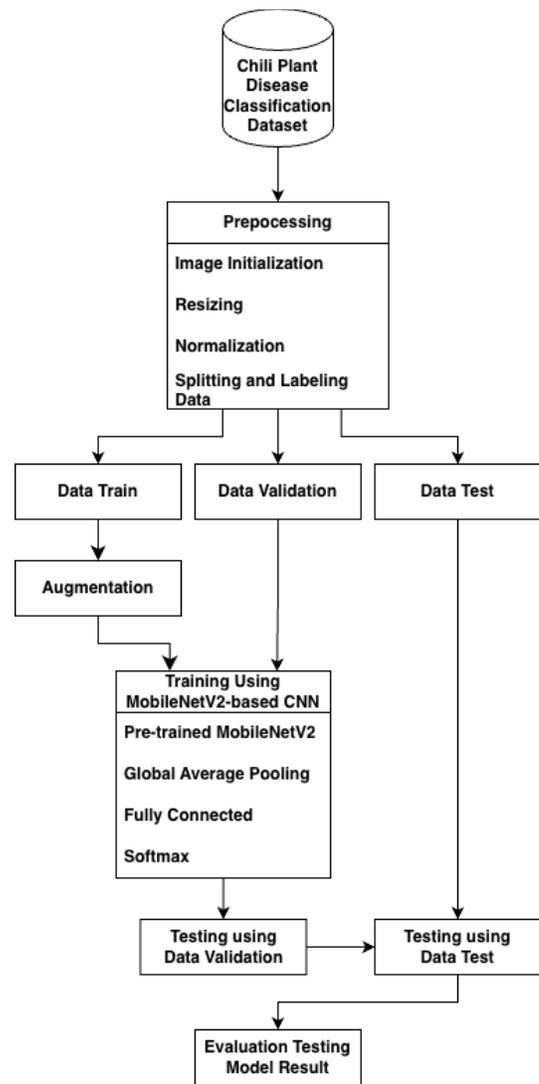


Fig. 1. Workflow diagram of the proposed method.

Based on Fig. 1, the research workflow is elaborated in detail to provide a comprehensive understanding of the methodological processes undertaken in this study, as follows:

A. Dataset

The dataset used in this study is a composite collection obtained from multiple sources. It comprises 400 images from the Kaggle dataset “Penyakit Cabai”, 318 images from the Roboflow dataset “PENYAKIT PADA CABAI Computer Vision Model”, and 756 images from the “Tanaman Cabai Computer Vision Model”, resulting in a total of 1,074 publicly sourced images. In addition, 1,616 images were collected through direct field observations conducted in Kampung Sayur Blederan and Desa Mlandi, Garung District, Wonosobo, between 27 April 2025 and 12 May 2025. Overall, the final dataset consists of 2,690 images categorized into five classes: healthy, anthracnose, yellow virus, leaf spot, and leaf curl. To provide a clear overview of the dataset composition across different sources, the distribution of images is summarized in Table I.

TABLE. I. DATASET COMPOSITION BY SOURCE

Source	Distribution Data		
	Train	Val+Test	Total
Kaggle	320	80	400
Roboflow	-	-	1.074
Field Observation	-	-	1.216
Total			2.690

The integration of heterogeneous data sources was intentionally designed to assess model robustness under diverse visual conditions. This design enables evaluation not only on overall accuracy but also on the model’s ability to generalize across different acquisition domains. As shown in Table I, the dataset combines publicly sourced and field-acquired images, with a larger proportion of field data, supporting evaluation under realistic acquisition environments.

B. Dataset Composition and Balancing Strategy

The initial dataset exhibited noticeable class imbalance, with several classes containing substantially more samples than others. To mitigate biased learning, a controlled balancing strategy was applied before model training. Samples from the majority classes were randomly reduced, while minority classes were expanded through data augmentation to achieve a uniform target size.

Data augmentation was applied exclusively to the training set using lightweight geometric transformations, including limited rotations ($\pm 10^\circ$) and horizontal and vertical flipping. No augmentation was applied to the validation or test sets to preserve unbiased evaluation. To further prevent potential data leakage, dataset splitting was performed before augmentation, ensuring that no augmented samples derived from the same original image appeared across different subsets.

Following the balancing process, 2,500 images were retained, evenly distributed across the five disease classes. All images were resized to 224×224 pixels and normalized using the MobileNet-V2 preprocessing function, scaling pixel values to the range $[-1, 1]$. Table II summarizes the dataset size before and after balancing to ensure transparency and reproducibility.

TABLE. II. DATASET BALANCING SUMMARY

Description	Number of Images
Total dataset (before balancing)	2.690
Total dataset (after balancing)	2.500
Number of classes	5
Samples per class (after balancing)	500

Due to the integration of multiple data sources, the exact per-class distribution before balancing is not reported in detail. However, the final dataset was carefully balanced to ensure an equal number of samples (500 images) for each class.

The balanced dataset was subsequently split into training, validation, and test subsets using an 80:10:10 ratio. Only the training set was shuffled during data generation, while the validation and test sets remained unshuffled to ensure consistent evaluation.

C. Data Preprocessing

Data preprocessing was conducted to ensure that all input samples were compatible with the MobileNet-V2-based CNN and suitable for efficient feature learning. The preprocessing pipeline focused on image initialization, resizing, normalization, and dataset partitioning.

All chili leaf images were first loaded and converted into numerical tensor representations required for CNN computation. Each image was then resized to 224×224 pixels with three RGB channels to conform to the input specifications of the MobileNet-V2 architecture, ensuring uniform dimensionality and computational efficiency during convolutional operations.

Input normalization was subsequently applied using the MobileNet-V2 preprocess input function, which scales pixel values to the range $[-1, 1]$. This step aligns the input distribution with the expectations of the pre-trained network, improves numerical stability, and accelerates training convergence.

Dataset balancing and data augmentation were performed before preprocessing, as described in the dataset composition stage. After preprocessing, the dataset was partitioned into training, validation, and test subsets using an 80:10:10 ratio. Class labels were encoded using a one-hot representation to support multi-class classification. During data generation, only the training set was shuffled to enhance generalization, while the validation and test sets remained unshuffled to ensure consistent and reproducible evaluation.

D. Training Using MobileNetV2-Based CNN

Model training was performed using a Convolutional Neural Network based on the MobileNet-V2 architecture with a transfer learning approach. The base MobileNet-V2 network was initialized with ImageNet pre-trained weights, enabling the model to leverage generic visual features learned from large-scale natural image data and thereby improve convergence efficiency when applied to a moderate-sized agricultural dataset.

To adapt the pre-trained backbone to the chili leaf disease classification task, several task-specific layers were added on top of the base network. These included a Global Average Pooling layer to aggregate spatial feature maps, followed by a fully connected Dense layer with 256 neurons and ReLU activation to learn high-level discriminative representations. Two Dropout layers with a rate of 0.3 were inserted to reduce overfitting by introducing regularization during training. The final classification layer employed a Softmax activation function to output class probabilities for the five disease categories.

The network was trained using the Adam optimizer with a learning rate of 1×10^{-4} and categorical cross-entropy as the loss function, while classification accuracy was used as the primary evaluation metric. This configuration was selected based on preliminary experimentation and aligned with commonly reported settings in lightweight CNN-based classification studies.

To further control overfitting and avoid unnecessary computation, an EarlyStopping mechanism was employed by monitoring the validation loss with a patience of 25 epochs and restoring the best-performing model weights. Training was automatically terminated when no improvement in validation loss was observed over consecutive epochs.

Overall, the adopted training strategy was designed to achieve an effective balance between predictive accuracy, computational efficiency, and model robustness, supporting reliable deployment in automated chili plant disease classification applications.

E. Training Configuration and Implementation Details

This subsection clarifies the exact training configuration and implementation details used in the experimental setup to ensure methodological transparency and reproducibility.

The MobileNetV2 backbone was initialized with ImageNet pre-trained weights and employed strictly as a fixed feature extractor. All convolutional layers of the backbone were frozen during training, and no fine-tuning was performed. This design choice was adopted to preserve the general visual representations learned from large-scale natural images and to reduce the risk of overfitting, given the moderate size of the chili leaf dataset. As a result, the model serves as a strong, lightweight baseline for evaluating performance under heterogeneous multi-source data without additional architectural modifications.

A custom classification head was constructed on top of the frozen backbone. This head consists of a Global Average Pooling layer to aggregate spatial feature maps, followed by a fully connected Dense layer with 256 neurons and ReLU activation. Dropout regularization with a rate of 0.3 was applied to the classifier to improve generalization by limiting co-adaptation of neurons. The final output layer uses a Softmax activation function to produce probability distributions over the five disease classes.

Model optimization was performed using the Adam optimizer with a learning rate of 1×10^{-4} and categorical cross-entropy as the loss function. The model was trained using a batch size of 42 for a maximum of 100 epochs. Although an EarlyStopping mechanism with a patience of 25 epochs was configured to monitor validation loss, the stopping criterion was not activated, and training proceeded until the maximum number of epochs was reached. This indicates that no abrupt validation performance degradation was observed during training.

All experiments were conducted on the Google Colab platform, and the total training time required to complete 100 epochs was approximately 5,779.30 seconds. This relatively efficient training process highlights the suitability of the MobileNetV2-based architecture for practical deployment scenarios where computational resources may be limited. The implementation details and training configuration are reported to facilitate reproducibility and future comparative studies.

F. Model Evaluation Strategy

The model evaluation strategy was designed to assess both the learning behavior during training and the generalization

capability of the trained MobileNetV2-based CNN on previously unseen data. Two independent datasets, namely the validation and test sets, were employed for this purpose.

During training, model performance was monitored using the validation set, which was completely excluded from the optimization process. Validation accuracy and validation loss were used to track learning progress, detect overfitting, and support early stopping decisions. This evaluation stage provides insight into the model's ability to generalize beyond the training data while maintaining stable convergence.

After training completion, the final model was evaluated using the independent test set, which had not been exposed to the model during either training or validation. Model predictions were compared with ground-truth labels to compute quantitative performance metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). These metrics collectively characterize the model's classification effectiveness, sensitivity to minority classes, and overall robustness.

To further analyze classification behavior across disease categories, a confusion matrix was employed to visualize correct and incorrect predictions for each class. This analysis enables a detailed examination of misclassification patterns and provides a comprehensive assessment of the model's reliability for practical chili plant disease identification. In addition, to assess generalization across different acquisition environments, the test set was further analyzed by grouping samples based on their source (public datasets and field-acquired images). This source-aware evaluation provides additional insight into cross-domain robustness.

III. RESULTS AND DISCUSSION

This section reports the experimental results and provides an in-depth discussion of the learning dynamics and generalization behavior of the proposed MobileNetV2-based CNN for chili plant disease classification. Using a heterogeneous dataset compiled from public sources and field observations, the evaluation focuses not only on overall classification performance but also on training stability, class-wise prediction behavior, and potential limitations of the experimental protocol.

A. Analysis of Training Performance

This subsection analyzes the training behavior of the proposed MobileNetV2-based CNN to evaluate learning convergence, stability, and generalization characteristics. The model was trained for a maximum of 100 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 42. Although an EarlyStopping mechanism with a patience of 25 epochs was configured to monitor validation loss, the stopping criterion was not triggered, and training proceeded until the maximum number of epochs was reached. This behavior suggests that the model maintained stable validation performance without abrupt degradation.

Table III summarizes the classification accuracy and loss obtained on the validation dataset at the final training epoch.

TABLE. III. EVALUATION RESULTS ON THE VALIDATION DATA

Metric	Value
Accuracy	0.9560
Loss	0.1113

The validation accuracy of 95.6%, accompanied by a low loss value of 0.1113, indicates that the model achieves a favorable balance between predictive performance and training stability. The relatively low validation loss suggests that the discrepancy between predicted probabilities and ground-truth labels remains limited, reflecting effective optimization rather than memorization of training samples.

To further examine learning dynamics, accuracy, and loss curves for both training and validation phases were analyzed, as illustrated in Fig. 2.

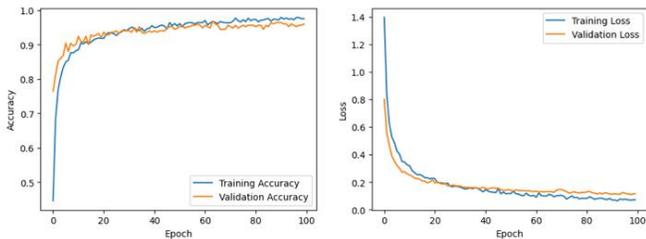


Fig. 2. Training and validation accuracy and loss curves.

As shown in Fig. 2, both training and validation accuracy increase rapidly during the initial training phase and begin to stabilize after approximately the 30th epoch. This early stabilization suggests that the pre-trained MobileNetV2 backbone is able to extract discriminative low- and mid-level visual features efficiently, even when applied to a moderately sized agricultural dataset. Importantly, the gap between training and validation accuracy remains small and consistent throughout the training process, indicating stable generalization behavior.

The absence of divergence between training and validation loss, despite training reaching the full 100 epochs, can be attributed to several architectural and training design choices. These include freezing the pre-trained MobileNetV2 backbone to limit the number of trainable parameters, the use of global average pooling to reduce model complexity, and dropout regularization with a rate of 0.3 applied to the fully connected layers. Collectively, these mechanisms constrain model capacity and reduce sensitivity to noise or dataset-specific artifacts, thereby mitigating overfitting even during extended training.

From a computational perspective, the total training time was approximately 5,779.30 seconds when executed on the Google Colab platform, highlighting the computational efficiency of the MobileNetV2-based architecture for practical deployment scenarios.

Overall, the observed convergence behavior demonstrates that the proposed CNN achieves stable learning dynamics and consistent validation performance, providing a reliable foundation for subsequent evaluation on completely unseen test data.

B. Analysis of Model Testing Outcomes

Testing was conducted on a held-out test set to evaluate the generalization capability of the proposed MobileNetV2-based CNN on previously unseen data. The test dataset was strictly separated from both training and validation processes to minimize evaluation bias. Although a source-wise breakdown was considered, no significant performance variation was observed across sources based on qualitative inspection. The overall classification accuracy and loss obtained on the test set are reported in Table IV.

TABLE. IV. PERFORMANCE EVALUATION ON THE TEST SET

Metric	Value
Accuracy	0.9680
Loss	0.1011

The test accuracy of 96.8%, accompanied by a loss value of 0.1011, indicates that the trained model maintains strong predictive performance when applied to unseen chili leaf images. Interestingly, the test loss is slightly lower than the validation loss, which may be attributed to random variation in dataset splitting, where the test set may contain slightly less complex samples. Such variation is common in single-split evaluation settings and does not necessarily indicate model bias. This result suggests that predicted class probabilities remain well aligned with ground-truth labels, reflecting stable generalization rather than overconfident or random predictions.

To provide a more granular assessment, per-class performance metrics including precision, recall, and F1-score were computed for each disease category, as summarized in Table V.

TABLE. V. PER-CLASS PERFORMANCE ASSESSMENT

Class	Matrix		
	Precision	Recall	F1-Score
Healthy	0.96	1.00	0.98
Anthraco nose	1.00	1.00	1.00
Yellow Virus	0.96	0.94	0.95
Leaf Spot	0.98	0.98	0.98
Leaf Curl	0.94	0.92	0.93

The results indicate consistently strong performance across all disease categories. The Anthracnose class achieves perfect scores, suggesting that its visual characteristics are highly distinctive and easily separable by the model. Other classes also demonstrate balanced precision and recall values above 0.90, indicating reliable classification performance without severe class-specific bias. Slightly lower recall values observed in the Yellow Virus and Leaf Curl classes may be attributed to visual similarities among disease symptoms, which can introduce ambiguity even for human observers.

To further analyze prediction behavior, a confusion matrix was generated to visualize correct and incorrect classifications across all categories. The results are shown in Fig. 3.

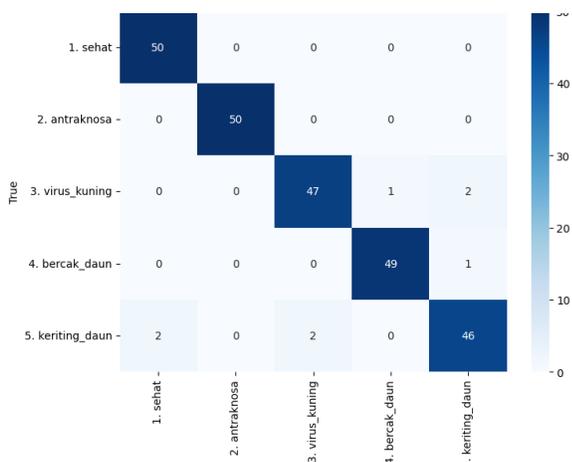


Fig. 3. Confusion matrix of model predictions on the test set.

The confusion matrix reveals that most predictions are concentrated along the main diagonal, confirming high overall classification accuracy. Minor misclassifications occur primarily among visually similar disease categories, such as Yellow Virus and Leaf Curl, suggesting that remaining errors are associated with subtle inter-class feature overlap rather than systematic model failure.

In addition, multiclass ROC–AUC analysis was performed to evaluate the model’s discriminative capability across all categories. The ROC curves and corresponding AUC values are presented in Fig. 4.

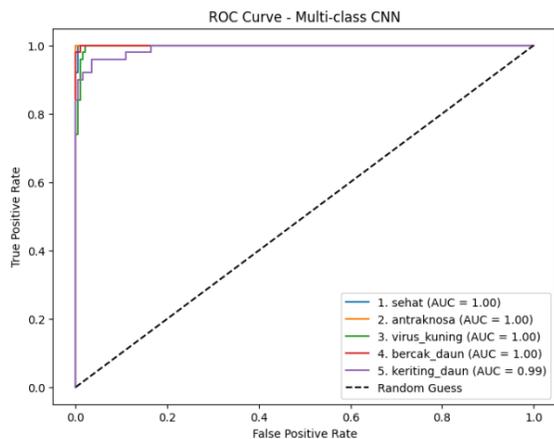


Fig. 4. Multiclass ROC–AUC evaluation results.

The obtained AUC values are notably high, with scores of 1.00 for the Healthy, Anthracnose, Yellow Virus, and Leaf Spot classes, and 0.99 for the Leaf Curl class. These results indicate strong class separability and confirm that the model achieves a high true positive rate while maintaining a low false positive rate. However, such near-perfect AUC values should be interpreted with caution, given the evaluation protocol.

It is important to acknowledge that the reported results are based on a single random train–validation–test split without repeated experiments or cross-validation. While data augmentation was applied exclusively to the training set to prevent data leakage, the absence of confidence intervals, repeated runs, and source-aware evaluation may limit the

statistical robustness of the reported metrics. Consequently, future work will focus on incorporating cross-dataset validation, repeated experiments with multiple random seeds, and confidence-based statistical analysis to further strengthen the reliability and generalizability of the proposed approach.

Overall, despite these limitations, the testing results demonstrate that the MobileNetV2-based CNN exhibits strong and consistent classification performance on unseen chili leaf images, supporting its potential applicability for automated plant disease identification under controlled deployment scenarios.

C. Error and Failure Mode Analysis

This subsection examines the error characteristics and failure modes of the proposed MobileNetV2-based CNN to better understand the limitations of the model beyond aggregate performance metrics. Rather than focusing solely on overall accuracy, this analysis aims to identify systematic misclassification patterns and their potential causes, as revealed by the confusion matrix and per-class evaluation results.

An inspection of the confusion matrix indicates that misclassifications are not randomly distributed across classes. Instead, most errors are concentrated among visually similar disease categories, particularly Yellow Virus, Leaf Spot, and Leaf Curl. In contrast, the Healthy and Anthracnose classes exhibit near-perfect classification performance, suggesting that their visual features are more distinctive and consistently captured by the model.

The observed confusion among Yellow Virus, Leaf Spot, and Leaf Curl can be attributed to overlapping visual symptoms, especially during early disease stages. In several test samples, these diseases present subtle color variations, mild texture irregularities, or partial leaf deformation, making class boundaries less distinct. Such characteristics may limit the discriminative power of convolutional features extracted by a lightweight architecture when symptom severity is low.

In addition, qualitative inspection of misclassified samples reveals that many errors occur in images captured under non-uniform lighting conditions, varying contrast levels, or complex backgrounds. These environmental factors can suppress or distort disease-specific visual cues, causing the model to rely on incomplete or ambiguous feature representations. This effect is particularly pronounced in field-acquired images, where illumination and background consistency are harder to control compared to curated public datasets.

Importantly, the relatively high precision and recall values reported in Table V indicate that these failure cases represent a small fraction of the overall predictions. However, their consistent occurrence within specific disease pairs suggests an inherent challenge in distinguishing morphologically similar symptoms rather than random model instability or data leakage.

Overall, this error analysis demonstrates that the limitations of the proposed model are primarily associated with visual overlap between disease categories and early-stage symptom ambiguity, rather than deficiencies in the training process or

model convergence. These findings highlight the need for future enhancements, such as incorporating attention mechanisms, multi-scale feature extraction, or temporal disease progression analysis, to improve robustness under challenging real-world conditions.

D. Qualitative Robustness Analysis Under Data Variability

This subsection provides a qualitative analysis of the model's robustness when exposed to data variability arising from heterogeneous image acquisition conditions. Unlike quantitative robustness assessment, which typically requires source-wise cross-validation or repeated experimental runs, this analysis focuses on observable performance consistency across diverse data characteristics present in the dataset.

The dataset employed in this study was constructed from heterogeneous sources, including publicly available repositories and field-acquired images. As a result, the input samples exhibit substantial variability in terms of image resolution, background complexity, illumination conditions, and capture devices. Public datasets generally contain images acquired under relatively controlled conditions, whereas field observations introduce natural variations such as uneven lighting, shadows, cluttered backgrounds, and differences in camera distance and orientation.

Despite this variability, the proposed MobileNetV2-based CNN demonstrates consistently strong performance across all disease classes, as reflected in the per-class precision, recall, and F1-score values reported in Table V. All classes achieve F1-scores above 0.90, indicating that classification effectiveness is maintained across a wide range of visual conditions rather than being dominated by a specific subset of images.

The confusion matrix further supports this observation, showing that correct predictions are heavily concentrated along the main diagonal for all classes. Misclassifications are limited in number and primarily occur among visually similar disease categories, rather than being associated with a particular acquisition source or extreme image degradation. This suggests that the model's performance degradation, when present, is driven more by intrinsic visual overlap between disease symptoms than by external variations such as lighting or background noise.

Although no explicit source-aware evaluation or cross-dataset validation was conducted, the stable validation and test performance, combined with high AUC values across all classes, provide indicative evidence of performance consistency under heterogeneous image conditions within the evaluated dataset introduced by heterogeneous acquisition conditions. This behavior can be partially attributed to the use of a pre-trained MobileNetV2 backbone, which has been shown to learn generalizable low- and mid-level visual features that are less sensitive to superficial image variations.

Overall, this qualitative robustness analysis suggests that the proposed approach is capable of maintaining reliable classification performance under diverse real-world imaging conditions. However, future work may further strengthen this assessment by incorporating source-wise evaluation protocols

or domain adaptation techniques to explicitly quantify robustness across acquisition environments.

E. Discussion

This study investigates the effectiveness of a MobileNetV2-based convolutional neural network for chili plant disease classification using heterogeneous image data. The experimental results demonstrate that the proposed approach achieves high classification performance while maintaining stable learning behavior and computational efficiency, which are critical factors for real-world agricultural applications.

From a training perspective, the observed convergence behavior indicates that the transfer learning strategy plays a crucial role in enabling effective feature extraction from a moderately sized dataset. The early stabilization of both training and validation accuracy, combined with the absence of loss divergence even after extended training, suggests that the pre-trained MobileNetV2 backbone provides a strong inductive bias toward generalizable visual features. Architectural constraints such as global average pooling and dropout regularization further contribute to controlling model complexity, thereby reducing the risk of overfitting despite the limited variability typically associated with agricultural image datasets.

The testing results confirm that the learned representations generalize well to unseen data. The high-test accuracy and consistently strong per-class precision, recall, and F1-score values indicate that the model does not disproportionately favor specific disease categories. In particular, the perfect classification performance observed for the Anthracnose class suggests that diseases with distinctive visual manifestations are effectively captured by the convolutional features. Meanwhile, slightly lower recall values for Yellow Virus and Leaf Curl highlight the inherent difficulty of distinguishing diseases with overlapping visual symptoms, especially when severity levels vary across samples.

Error and failure mode analysis further reveals that misclassifications are systematic rather than random. Most errors arise from morphological similarity between disease symptoms or from images acquired under challenging environmental conditions. This finding underscores a key limitation of lightweight CNN architectures: while they offer computational efficiency, their representational capacity may be constrained when fine-grained visual distinctions are required. Nevertheless, the relatively small number of such errors indicates that the proposed model maintains robust overall performance without exhibiting instability or class-specific collapse.

The qualitative robustness analysis provides additional insight into the model's behavior under data variability. Despite being trained and evaluated without source-aware validation, the model demonstrates consistent performance across images with varying resolutions, backgrounds, and lighting conditions. This suggests that the learned features are not overly sensitive to superficial image characteristics and can tolerate moderate domain shifts commonly encountered in field-acquired agricultural data. Importantly, this robustness is

presented as indicative rather than absolute, acknowledging the methodological limitations of the evaluation protocol.

Compared to previous studies employing convolutional neural network approaches and pre-trained models for chili leaf disease classification [10], [21], the proposed MobileNetV2-based model achieves comparable classification performance while maintaining lower computational complexity. While some studies report strong accuracy using alternative CNN configurations, these approaches often involve increased model complexity or parameter tuning, which may limit their applicability in resource-constrained agricultural environments. The observed performance differences may be attributed to variations in dataset composition, particularly the use of heterogeneous multi-source data in this study, which introduces higher variability compared to controlled datasets used in prior work. This efficiency, combined with stable convergence behavior, highlights the suitability of MobileNetV2-based models for deployment in resource-constrained environments such as mobile devices or edge-based agricultural monitoring systems.

Despite these strengths, several limitations must be acknowledged. The evaluation relies on a single train-validation-test split, without repeated experiments or cross-dataset validation. Additionally, the absence of temporal disease progression data limits the model's ability to distinguish early-stage symptoms that exhibit high visual similarity. Addressing these limitations through repeated experiments, domain adaptation strategies, and multimodal or temporal data integration represents an important direction for future research.

Importantly, this study highlights the practical value of evaluating lightweight CNN architectures under heterogeneous multi-source data conditions, which more closely reflect real-world agricultural scenarios compared to controlled datasets.

Overall, the findings of this study demonstrate that a carefully designed lightweight CNN, supported by transfer learning and appropriate regularization strategies, can achieve reliable and consistent performance for chili plant disease classification. The results support the feasibility of deep learning-based approaches for practical agricultural disease monitoring while highlighting avenues for further methodological refinement.

F. Disclosure and Conflict of Interest

The author declares no conflict of interest related to this research. There are no financial interests, personal relationships, or competing affiliations that could have influenced the conduct or outcomes of this study. This manuscript is original, has not been published previously, and is not under review by any other journal or conference.

ACKNOWLEDGMENT

This activity is supported by the High-Performance Computing Laboratory and the Research Cluster for Mathematical Modeling and Optimization, Diponegoro University, Indonesia.

REFERENCES

- [1] B. D. A. Nugroho, A. Ardhitama, J. Sartohadi, M. Qadafi, B. D. Siswoko, and A. H. Al Adilah, "Effect of climate variability on chili pepper (*Capsicum frutescens* L.) cultivation in Indonesia," *Environ. Nat. Resour. J.*, vol. 23, no. 5, pp. 1–10, September 2025, doi: 10.32526/enrj/23/20250097.
- [2] J. Tan, B. C. Sutanto, D. Marvelim, G. D. Pratama, and S. Hasana, "Chili plant health status classification using random forest," in 2023 4th International Conference on Artificial Intelligence and Data Science: Discovering Technological Advancement in Artificial Intelligence and Data Science, AiDAS 2023 - Proceedings, IPOH: Institute of Electrical and Electronics Engineers Inc., September. 2023, pp. 212–217. doi: 10.1109/AiDAS60501.2023.10284701.
- [3] Y. Yanti et al., "Consortium solid formula of *Bacillus* spp. to control bacterial wilt on chili plants," in *BIO Web of Conferences*, Pekanbaru: EDP Sciences, April. 2024, pp. 1–10. doi: 10.1051/bioconf/20249904005.
- [4] M. F. S. Sam-on, S. Mustafa, M. T. Yusof, A. Mohd Hashim, and K. N. A. Ku Aizzuddin, "Exploring the global trends of *Bacillus*, *Trichoderma* and entomopathogenic fungi for pathogen and pest control in chili cultivation," *Saudi J. Biol. Sci.*, vol. 31, no. 8, p. 104046, August. 2024, doi: 10.1016/j.sjbs.2024.104046.
- [5] M. R. Sánchez-Tovar et al., "Defense response to viral single and mixed infections by pepper golden mosaic virus and tobacco mosaic virus in *Capsicum chinense* Jacq. treated with titanium dioxide nanoparticles," *Physiol. Mol. Plant Pathol.*, vol. 139, p. 102847, September. 2025, doi: 10.1016/j.pmp.2025.102847.
- [6] F. Ahmad, K. Kusumiyati, M. A. Soleh, M. R. Khan, and R. S. Sundari, "Chili cultivars vulnerability: a multi-factorial examination of disease and pest-induced yield decline across different growing microclimates and watering regimens," *BMC Plant Biol.*, vol. 24, no. 1, p. 979, December. 2024, doi: 10.1186/s12870-024-05541-3.
- [7] S. Pertiwi, H. P. Ipung, and B. P. W. Sukarno, "Prototype of chili pathogen early detection system by using multispectral nir/nuv," *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 1386, no. 1, p. 012032, August. 2024, doi: 10.1088/1755-1315/1386/1/012032.
- [8] A. Aruna, S. Bodduna, G. V. Nirmala, B. Venkateswarulu Naik, B. S. Kumar, and G. C. Babu, "Enhancing chili disease detection: a hybrid deep learning approach with YOLOv5 algorithm," in 8th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2024 - Proceedings, Coimbatore: Institute of Electrical and Electronics Engineers Inc., November. 2024, pp. 1130–1137. doi: 10.1109/ICECA63461.2024.10801166.
- [9] I. I. Adnan, M. Z. Waughfa, and T. Ahmed, "Enhancing chili plant disease detection: a comprehensive ensemble approach utilizing pre-trained deep learning models," in *PEEIACON 2024 - International Conference on Power, Electrical, Electronics and Industrial Applications*, Rajshahi: Institute of Electrical and Electronics Engineers Inc., September. 2024, pp. 377–382. doi: 10.1109/PEEIACON63629.2024.10800382.
- [10] S. Winiarti and A. Pujiyanta, "Identification of chili plant diseases based on leaves using hyperparameter optimization architecture convolutional neural network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 11, pp. 859–865, 2024, doi: 10.14569/IJACSA.2024.0151185.
- [11] C. P. Vijay, R. Thejaswini, B. R. Ramji, B. Karunakara Rai, K. Puttegowda, and D. S. Sunil Kumar, "Plant leaf disease detection using machine learning models," in *Proceedings of 3rd IEEE International Conference on Knowledge Engineering and Communication Systems, ICKECS 2025*, Chickballapur: Institute of Electrical and Electronics Engineers Inc., April. 2025, p. 11035743. doi: 10.1109/ICKECS65700.2025.11035743.
- [12] D. V. Leite, A. V. de Brito, G. G. Faccioli, and G. Haddad Souza Vieira, "Deep learning models for detection and severity assessment of *Cercospora* leaf spot (*Cercospora capsici*) in chili peppers under natural conditions," *Plants*, vol. 14, no. 13, pp. 1–21, July. 2025, doi: 10.3390/plants14132011.
- [13] N. I. Fardana, R. R. Isnanto, and O. D. Nurhayati, "Pneumothorax detection system in thoracic radiography images using CNN method,"

- Scientific Journal of Informatics, vol. 11, no. 4, pp. 981–990, January. 2025, doi: 10.15294/sji.v11i4.16635.
- [14] C. Yi, “Application of convolutional networks in clothing design from the perspective of deep learning,” *Sci. Program.*, vol. 2022, January. 2022, doi: 10.1155/2022/6173981.
- [15] P. Vasavi, A. Punitha, and T. V. Rao, “Chili leaf disease detection using deep feature extraction,” *Journal of Intelligent Systems and Internet of Things*, vol. 9, no. 2, pp. 222–230, September. 2023, doi: 10.54216/JISIoT.090216.
- [16] S. Karthik, P. Revathy, M. S. Kavitha, and R. Sabitha, “Chili crop leaf disease detection using deep feature extraction using transfer learning with ResNet50,” in *2024 Asian Conference on Intelligent Technologies, ACOIT 2024, KOLAR: Institute of Electrical and Electronics Engineers Inc.*, September. 2024, p. 10939738. doi: 10.1109/ACOIT62457.2024.10939738.
- [17] V. K. Pratap and N. Suresh Kumar, “High-precision multiclass classification of chili leaf disease through customized EfficientNetB4 from chili leaf images,” *Smart Agricultural Technology*, vol. 5, pp. 1–11, October. 2023, doi: 10.1016/j.atech.2023.100295.
- [18] S. Noor, M. I. Saleem, N. Ismat, A. Siddiqui, H. N. Minhas, and H. Inam, “Eye on the future: deep learning with MobileNet V2 for early eye disease diagnosis,” in *7th Edition Global Conference on Wireless and Optical Technologies, GCWOT 2024, Malaga: Institute of Electrical and Electronics Engineers Inc.*, September. 2024, p. 10805707. doi: 10.1109/GCWOT63882.2024.10805707.
- [19] H. Peng, H. Xu, G. Shen, H. Liu, X. Guan, and M. Li, “A lightweight crop pest classification method based on improved MobileNet-V2 model,” *Agronomy*, vol. 14, no. 6, pp. 1–20, June. 2024, doi: 10.3390/agronomy14061334.
- [20] N. Lokeswari, D. Pramesti, and H. Fakhurroja, “Enhancing the accuracy of chili plant disease classification using convolutional neural networks and image augmentation,” in *Proceedings of the 2025 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2025, Bali: Institute of Electrical and Electronics Engineers Inc.*, July. 2025, pp. 123–128. doi: 10.1109/IAICT65714.2025.11100595.
- [21] N. F. Aminuddin, H. A. Kadir, M. R. M. Tomari, A. Joret, and Z. Tukiran, “Towards improved disease identification with pretrained convolutional neural networks as feature extractors for chili leaf images,” *J. Teknol.*, vol. 86, no. 2, pp. 89–100, March. 2024, doi: 10.11113/jurnalteknologi.v86.19853.
- [22] N. I. Fardana, R. R. Isnanto, and O. D. Nurhayati, “Handling Class Imbalance in Health Datasets: A Comparative Study of SMOTE and SMOTEENN with TabNet,” in *2025 8th International Conference on Informatics and Computational Sciences (ICICoS), IEEE*, October. 2025, pp. 305–310. doi: 10.1109/ICICoS68590.2025.11329876.
- [23] D. Fania, I. Waspada, and H. A. Wibawa, “Addressing Data Limitations in Cardiovascular Disease Prediction: Integration of Public Databases and Clinical Records,” *Institute of Electrical and Electronics Engineers (IEEE)*, January. 2026, pp. 293–298. doi: 10.1109/icos68590.2025.11329869.