ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

Deep Learning-Based Classification of Apple Leaf Diseases under Field Conditions

Supakit Mamart[1], Sumalee Sangamuang[1*], and Prompong Sugunnasil[1]

[1] Faculty of Engineering, Chiang Mai University, 50200, Thailand
* Correspondence: sumalee.sa@cmu.ac.th

Citation:
Mamart, S.; Sangamuang, S.; Sugunnasil, P. Deep learning-based classification of apple leaf diseases under field conditions. *ASEAN J. Sci. Tech. Report.* **2026**, *29*(3), e260960. https://doi.org/10.55164/ajstr.v29i3.260960.

Article history:
Received: September 14, 2025
Revised: January 16, 2026
Accepted: January 25, 2026
Available online: February 28, 2026

Abstract: Accurate identification of apple leaf diseases in field conditions is essential for sustaining crop yield and supporting precision agriculture. Variable illumination, cluttered backgrounds, and co-occurring symptoms complicate diagnosis in real orchards. This study applies a deep learning approach using a fine-tuned MobileNetV2 model to classify apple leaf diseases from a heterogeneous dataset derived from the Plant Pathology 2021 (FGVC8) benchmark. The original five labels were expanded by subdividing the "multiple disease" category into expert-defined compound subclasses, yielding 12 disease categories encompassing both single and compound infections. Data augmentation and transfer learning were employed to improve robustness, while interpretability was assessed through Grad-CAM and LIME visualizations. Results show that the model performs well on distinct single-disease categories such as rust, scab, and frogeye leaf spot, but struggles to detect overlapping or compound infections. These findings highlight both the potential and the challenges of lightweight CNN architectures for agricultural image classification. The study contributes evidence that explainable, compact deep learning models can support future efforts to build reliable tools for plant health monitoring in diverse field conditions.

Keywords: Deep learning; mobileNetV2; apple leaf disease; explainable AI; agriculture

1. Introduction

Apple cultivation is a cornerstone of temperate agriculture, contributing significantly to food security, rural livelihoods, and international trade [1]. However, apple orchards are highly susceptible to foliar diseases such as frogeye leaf spot, powdery mildew, rust, and scab. These diseases often occur simultaneously and share overlapping symptoms, complicating timely and accurate diagnosis. Early detection is essential to reduce yield losses, yet conventional monitoring relies on manual inspection by human experts, a process that is labor-intensive, time-consuming, and prone to subjective variation [2]. Recent advances in mobile imaging technologies and edge computing have therefore stimulated interest in automated, vision-based diagnostic systems powered by deep learning to support faster and more reliable decision-making in agriculture [3].

In recent years, Convolutional Neural Networks (CNNs) have emerged as a powerful tool for plant disease recognition. Architectures such as ResNet, Inception, and EfficientNet have demonstrated high diagnostic accuracy under controlled laboratory conditions [4]. However, their large memory footprint and

high computational requirements restrict their use on mobile or embedded systems, which are more practical for field-based applications. In contrast, lightweight models such as MobileNetV2 offer a promising balance between efficiency and accuracy [5], making them suitable for real-time inference on portable devices without reliance on continuous cloud connectivity. The need for offline-capable solutions is particularly acute in many parts of the Global South. In rural and mountainous regions of Southeast Asia—including northern Thailand, Laos, and Vietnam—farmers often operate in areas with unreliable internet access or intermittent electricity supply. Such constraints make cloud-dependent systems impractical in real-world scenarios. To address this gap, self-contained diagnostic models that can operate locally are required. Moreover, interpretable lightweight AI systems are essential to promote equitable access to smart agriculture technologies, especially for smallholder farmers in underserved communities [6, 7].

Despite the progress of deep learning in plant disease diagnostics, several challenges remain. First, most models are trained on curated datasets collected under ideal conditions, which do not represent the variability of field environments where images often include noise, occlusion, or uneven lighting. Second, apple leaves frequently exhibit compound infections, and standard classifiers—typically designed for single-label prediction—struggle to disambiguate overlapping lesion patterns, particularly in minority or hybrid classes. Third, the computational demands of state-of-the-art CNNs limit their use in resource-constrained settings. Without compact, interpretable models that can operate offline, the benefits of AI-driven plant disease detection remain inaccessible to the farmers who need them most. These considerations motivate the development of a lightweight, interpretable deep learning system tailored for robust performance across heterogeneous field conditions. Figure 1 outlines the conceptual framework of this study, illustrating the challenges addressed and the proposed approach. Building on this foundation, the next subsection states the problem formally and situates the research within the broader context of deep learning applications in agriculture.

The objectives of this study are to develop a lightweight MobileNetV2-based classifier for apple leaf disease recognition under real field conditions, to evaluate the robustness of the model under environmental distortions such as illumination variation, occlusion, and background clutter, and to analyze misclassification behavior using explainable AI (Grad-CAM and LIME) to understand model attention on lesion regions and compound symptoms.

1.1 Deep Learning for Plant Disease Detection

The early success of deep learning in computer vision has naturally extended to agriculture, particularly for plant disease recognition. Convolutional Neural Networks (CNNs) have proven especially effective in extracting discriminative features from leaf imagery, enabling accurate classification of disease symptoms. Mohanty et al. [3] demonstrated this capability using a CNN trained on 26 diseases across 14 crops, achieving high accuracy under controlled conditions. Ferentinos [4] further explored multiple CNN architectures—including VGG and GoogLeNet—achieving average accuracies above 93% across 58 crop-disease pairs. However, these conventional architectures are computationally demanding and thus ill-suited for real-time, in-field use. To mitigate this limitation, Too et al. [5] and Barbedo [8] investigated the use of lightweight CNNs and stressed the importance of dataset diversity for robust generalization. These developments provide the technical foundation for exploring more efficient, deployable models, as discussed next.

1.2 Lightweight CNN Architectures in Agriculture

Lightweight CNN architectures have gained attention as a practical solution for precision agriculture, where computing power is often limited.

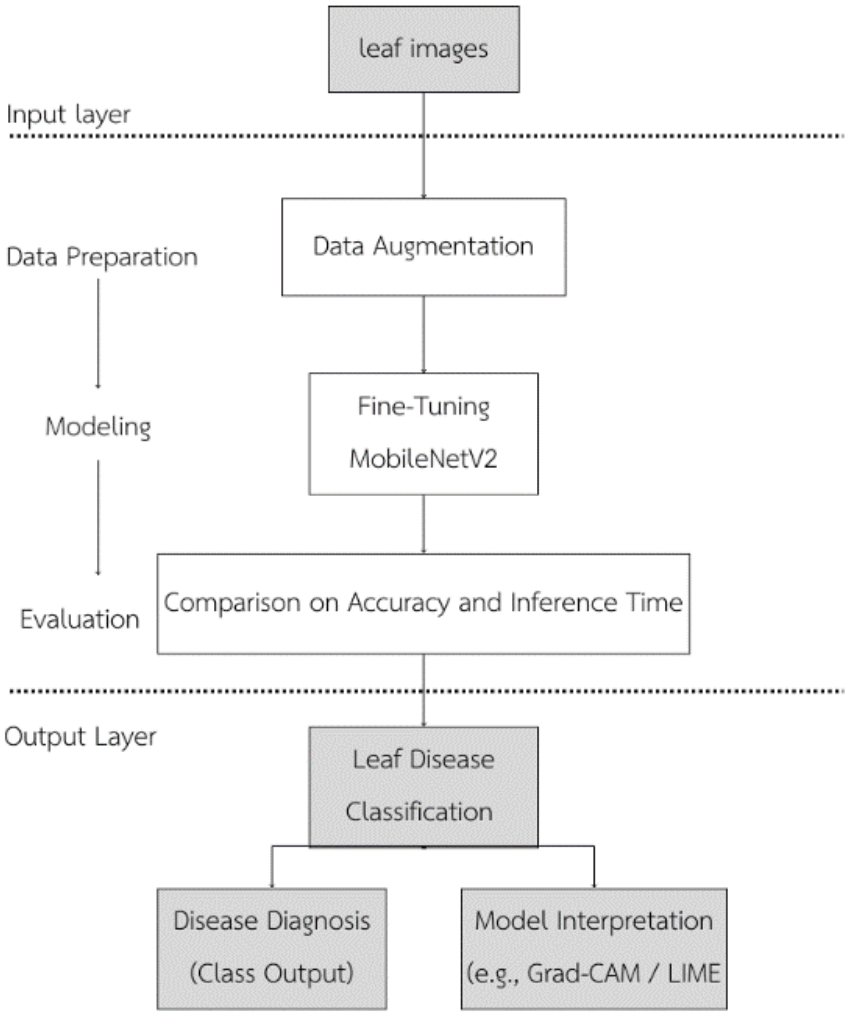

Figure 1. Conceptual framework for lightweight, interpretable, and deployable deep learning-based apple leaf disease classification. The model aims to address key challenges in real-world agricultural contexts, including compound symptoms, limited resource availability, and the need for interpretability.

Traditional networks like VGG and ResNet, though accurate, demand high memory and processing resources. In contrast, models such as MobileNet, SqueezeNet, and ShuffleNet employ parameter-efficient designs—e.g., depthwise separable convolutions—to achieve competitive accuracy while reducing complexity. Too et al. [5] found that MobileNet and SqueezeNet matched the performance of deeper models while being significantly more deployable. Fuentes et al. [9] confirmed the viability of these models for detecting tomato disease in natural environments. Kamilaris and Prenafeta-Bold'u [6] emphasized the role of lightweight CNNs in enabling real-time smart farming tools. Nevertheless, as Barbedo [8] noted, training on diverse and representative datasets remains essential for maintaining robustness. Among these options, MobileNetV2 has emerged as a particularly promising model for edge-based diagnosis, motivating further discussion in the following subsection.

1.3 MobileNetV2 in Field Applications

MobileNetV2 is increasingly adopted for field applications in agriculture due to its efficient inverted-residual architecture and linear bottlenecks. This architecture reduces model size and latency while preserving classification accuracy, making it well-suited for mobile and embedded systems. Xu et al. [10] applied MobileNetV2 to tomato leaf disease detection, achieving high performance and low inference delay under

real-world conditions. Similarly, Fuentes et al. [9] utilized MobileNet for tomato pest recognition in uncontrolled environments. Thapa et al. [11] demonstrated MobileNetV2's capacity to classify apple leaf diseases using transfer learning, while Albahli [12] reported that optimized variants of MobileNetV2 outperform deeper CNNs on mobile hardware. These applications validate MobileNetV2's suitability for edge deployment in smart agriculture. Our study builds upon these findings by customizing the model to a diverse apple leaf dataset, as further justified in the next subsection.

Table R1. Summary of related studies on plant disease classification under field conditions.

Study	Crop	Model	Dataset	Field conditions	Accuracy
Mohanty et al. (2016)	Multiple crops	AlexNet/GoogLeNet	PlantVillage	No	>99%
Ferentinos (2018)	Multiple crops	VGG/GoogLeNet	Lab images	No	>93%
Fuentes et al. (2018)	Tomato	Faster R-CNN	Field images	Yes	83–92%
Xu et al. (2022)	Tomato	MobileNetV2	Field images	Yes	94%
Thapa et al. (2020)	Apple	MobileNetV2/EfficientNet	Mixed	Limited	90%+
This study	Apple	MobileNetV2	FGVC8 + compound relabeling	Yes	65% (12-class), 71% (5-class)

This comparison highlights that most prior studies focus on single-disease labels, whereas our work explicitly analyzes compound infections and deployability constraints.

1.4 Limitations in Prior Work and Motivation

Despite the growing body of research applying deep learning to plant disease diagnosis, several key limitations hinder effective deployment in field settings. First, many existing studies rely on curated datasets captured under controlled laboratory conditions. These datasets often fail to capture the heterogeneity of real agricultural environments, including inconsistent lighting, occlusion, complex backgrounds, and compound symptoms. As Barbedo [8] notes, this lack of visual diversity limits model generalizability. Second, conventional classifiers are typically designed for single-label prediction, assuming that each leaf image contains only one disease. In practice, however, multiple infections often co-occur, leading to overlapping lesion patterns that challenge standard CNN-based methods. This limitation is particularly pronounced in apple leaf pathology, where visually ambiguous symptom combinations are common. Third, although several lightweight CNN architectures have been proposed for edge deployment, few studies provide a comprehensive evaluation that includes both classification accuracy and practical metrics such as inference latency, model size, and robustness under noisy conditions. While MobileNetV2 has shown promise in prior work [10, 12], its performance under real-world deployment constraints remains insufficiently characterized, especially for complex, multi-symptom categories. These limitations highlight the need for a field-adapted, interpretable, and computationally efficient diagnostic system that can operate reliably in resource-constrained environments. In response, this study **fine-tunes MobileNetV2** on a heterogeneous dataset of apple leaf diseases and systematically evaluates its performance under real-world deployment conditions. This includes assessing classification robustness to noise, evaluating interpretability with Grad-CAM and LIME, and benchmarking against deeper CNNs to establish a lightweight yet accurate solution for on-device agricultural diagnostics.

2. Materials and Methods

This section details the experimental methodology used to develop and evaluate the fine-tuned MobileNetV2 classifier for multi-class apple leaf disease detection. The process follows a modified CRISP-DM framework, encompassing data understanding, preprocessing, model training, and evaluation, with emphasis on field-readiness and interpretability. An overview of the workflow is illustrated in Figure 1.

2.1 Dataset and Preprocessing

The primary dataset used in this study is derived from the Plant Pathology 2021–FGVC8 competition on Kaggle [13]. It contains 18,600 high-resolution images of apple leaves, each originally annotated into one of five diagnostic categories: Apple Scab, Black Rot, Cedar Apple Rust, Healthy, and Multiple Disease, as illustrated in Figure 6. In this study, the "Multiple Disease" category was further subdivided into expert-defined compound subclasses (e.g., scab + frog eye leaf spot, rust + frog eye leaf spot, and higher-order combinations), yielding a total of 12 classes for model training and evaluation. This relabeling reflects realistic co-infection patterns observed under field conditions. Plant pathology experts manually verified all labels to ensure annotation quality and class integrity. Each disease class exhibits distinct lesion morphology and visual patterns, which are critical for both human interpretation and machine learning-based classification [8]. A brief description of the five categories is provided below:

- **Apple Scab**: Caused by Venturia inaequalis, this disease presents as olive-brown to dark circular lesions with sharply defined margins. Infected leaves often deform or drop prematurely [14].
- **Black Rot**: Induced by Botryosphaeria obtusa, black rot appears as necrotic lesions that typically originate at leaf edges and expand inward in concentric patterns [15].
- **Cedar Apple Rust**: Caused by Gymnosporangium juniperi-virginianae, this heteroecious rust produces vivid orange lesions with raised centers and chlorotic halos [16].
- **Healthy**: Leaves without visible disease symptoms are uniformly green with no deformation or lesion markings. They serve as control samples in diagnostic classification tasks [3].
- **Multiple Disease**: Co-infected leaves exhibit overlapping lesion types from two or more diseases, complicating classification due to mixed visual cues [8].

To prepare the dataset for training, all images were resized to 224 × 224 pixels using bilinear interpolation, under the input dimensions required by MobileNetV2. Pixel values were normalized using the ImageNet dataset's mean and standard deviation to ensure compatibility with pretrained weights. A set of data augmentation techniques was implemented using the torchvision transforms module. These included random horizontal flipping, cropping, jittering brightness and contrast, and illumination shifts. Each transformation was applied probabilistically to increase robustness against visual variability encountered in real-world field conditions. An 80:20 stratified sampling strategy was used to split the dataset into training and test sets, preserving the class distribution across both splits. This split was selected because of the relatively large dataset size and the high computational cost of repeated training with deep CNNs. While k-fold cross-validation can provide more stable estimates, a stratified hold-out set is commonly adopted in large-scale vision benchmarks and was considered sufficient for comparative evaluation in this study. Cross-validation is planned as future work. These processed and well-annotated images form the input to the deep learning models discussed in the next section, where we describe the network architecture, transfer learning strategy, and baseline comparisons.

2.2 Model Architecture and Transfer Learning

Based on the preprocessed image dataset, this study employed MobileNetV2 as the backbone architecture for apple leaf disease classification. MobileNetV2 is a lightweight convolutional neural network (CNN) optimized for deployment in resource-constrained environments, such as smartphones, drones, or IoT devices commonly used in agricultural field settings. Its design incorporates inverted residual blocks and depthwise separable convolutions, effectively reducing model size and computational complexity while preserving high representational capacity. The network was initialized with pretrained ImageNet weights to leverage generalized visual feature extraction. To adapt the model to the domain-specific task, the final fully

connected classification head was replaced with a softmax layer configured for five disease classes. A transfer learning strategy was applied: the first two inverted residual stages of MobileNetV2 were frozen to retain generic low-level features, while the remaining higher-level layers were fine-tuned to specialize in distinguishing lesion patterns unique to apple leaf diseases. Regularization components—including batch normalization for training stability, dropout to prevent overfitting, and ReLU activation for non-linear transformation—were retained from the original MobileNetV2 design.

2.3 Experimental Setup and Baseline Models

An 80/20 stratified split was used to divide the dataset into training and test sets, ensuring balanced class distribution across both subsets. A consistent data augmentation pipeline—comprising resizing, horizontal flipping, color jittering, and random cropping—was applied to all models using the torchvision transforms module. To evaluate the effectiveness of our proposed fine-tuned MobileNetV2 model, we compared its performance against three widely adopted deep CNN baselines: ResNet50, InceptionV3, and EfficientNetV2L. These architectures are commonly used in plant pathology and serve as benchmarks for assessing the trade-off between diagnostic accuracy and computational cost. To enhance model transparency and support error analysis, we applied two explainable AI (XAI) techniques—Grad-CAM and LIME, which produce saliency maps and feature attribution visualizations. These methods help verify that predictions are grounded in biologically relevant lesion regions, thereby improving model interpretability.

2.4 Training and Evaluation Protocol

All models were trained using the Adam optimizer with an initial learning rate of 1×10^{-4} and categorical cross-entropy loss. A ReduceLROnPlateau scheduler was applied with a factor of 0.1 and patience of 3 epochs. Training used early stopping based on validation loss, with a patience of 5 epochs, to prevent overfitting. Model performance was evaluated using accuracy, precision, recall, and F1-score on a per-class basis. Additionally, model size (MB) and inference latency (seconds per image) were measured to assess edge deployability. A confusion matrix was constructed to analyze class-specific performance and misclassification patterns. Particular attention was given to distinguishing between visually ambiguous and compound symptom classes.

2.5 Implementation Details

All experiments were implemented in PyTorch and executed on a GPU-enabled system with 16 GB of RAM and an NVIDIA RTX 3060. The batch size was set to 32, and training continued for up to 50 epochs with an early stopping patience of 5. No explicit class balancing (e.g., weighted loss or oversampling) was used during initial training to preserve dataset realism; an ablation study on class reweighting is presented in Section 5.

3. Results and Discussion

3.1 Results

This section presents the empirical findings of our study, structured to evaluate the model from five perspectives: classification performance, benchmarking, robustness, and efficiency.

3.1.1 Classification Performance

To evaluate the per-class predictive behavior of the MobileNetV2-based classifier, a confusion matrix is presented in Figure 2. This matrix provides a detailed summary of classification outcomes across the 12 disease categories by mapping true labels against predicted ones. The results reveal strong performance for clearly defined disease categories such as healthy, frog eye leaf spot, and rust, with high diagonal values indicating precise predictions. For instance, the healthy class was correctly identified in 902 out of 1061 cases, and frog eye leaf spot achieved 484 correct predictions out of 641 samples.

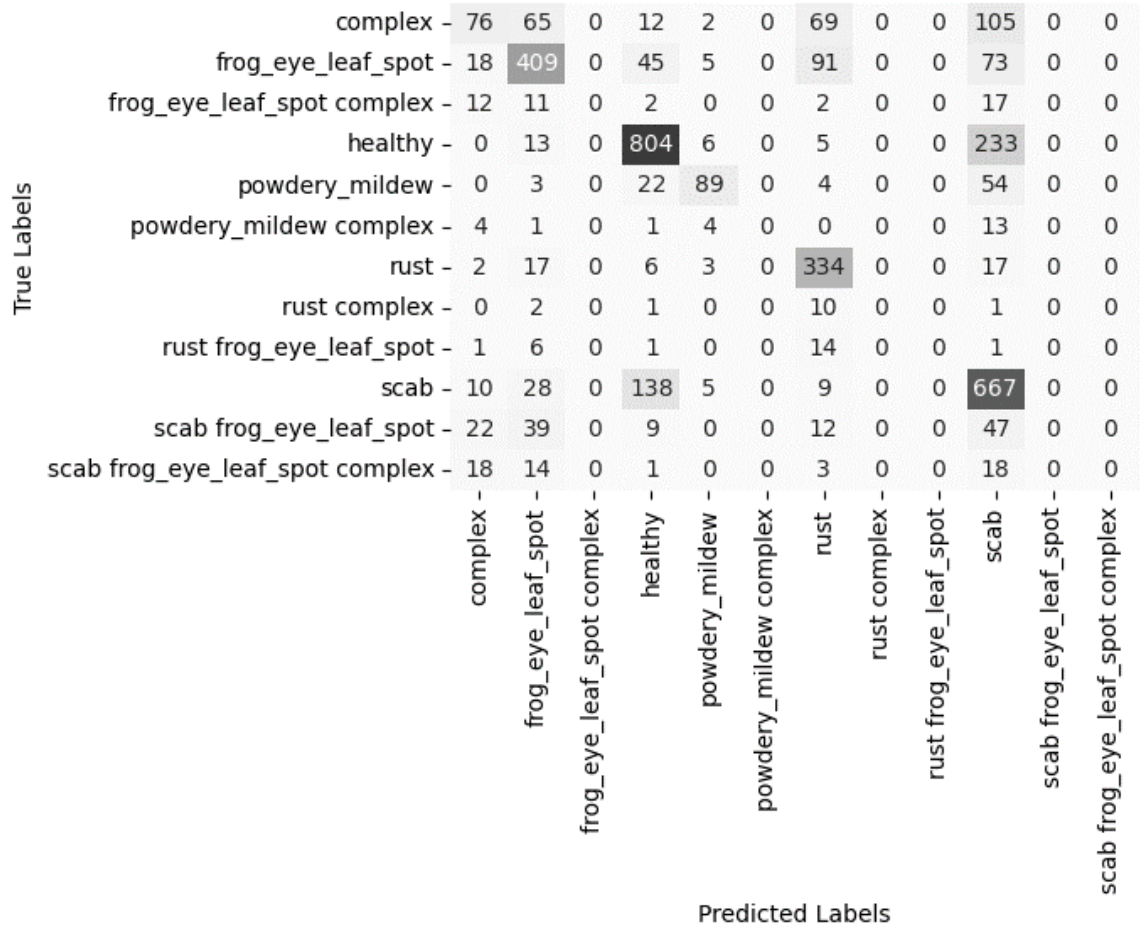

Figure 2. Confusion matrix for 12-class plant disease classification. Rows represent true labels, columns represent predicted labels.

In contrast, compound and visually ambiguous classes—such as powdery mildew complex, scab frog eye leaf spot, and rust complex—exhibited high misclassification rates. These categories were often confused with visually similar but simpler base classes (e.g., scab, rust, or frog eye leaf spot), indicating limitations in the model's ability to disentangle overlapping or co-occurring symptoms. Notably, the complex class was frequently misclassified as scab or frog eye leaf spot, yet this misclassification was not symmetric, suggesting potential dominance of base-class features or dataset imbalance. To gain further insight into the model's decision rationale, we incorporated explainable AI techniques. Figure 3 displays a Grad-CAM visualization for a leaf image predicted as powdery mildew with moderate confidence (31.61%). The heatmap indicates that the model focuses attention on the lesion regions, even under uncertainty, confirming the spatial relevance of the learned features.

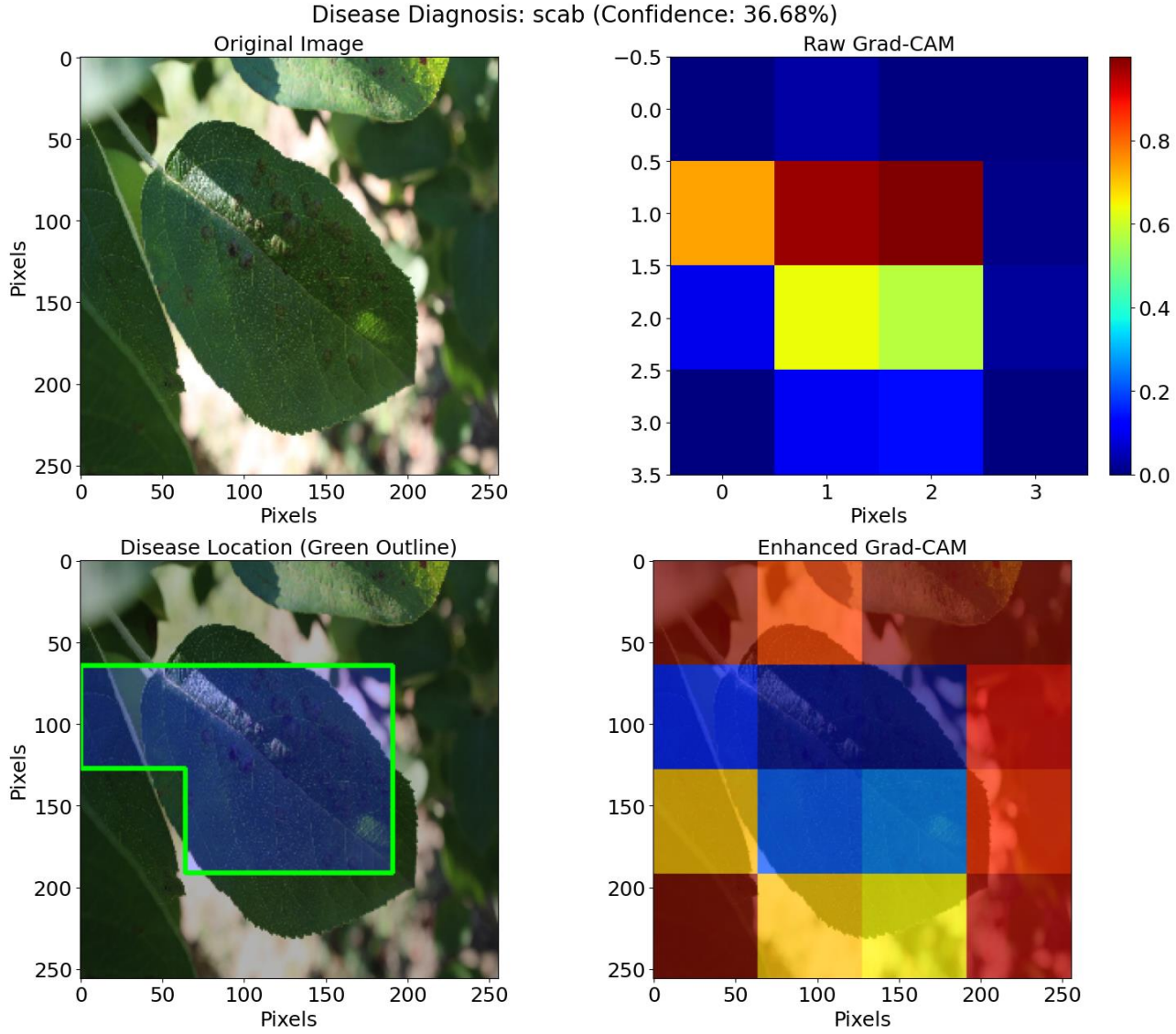

Figure 3. Grad-CAM visualization for a leaf classified as scab. The top-right panel shows raw Grad-CAM activations; the bottom-right panel shows the enhanced overlay on the original image. The green-outlined region denotes expert-annotated lesion zones. Despite moderate confidence, the model highlights relevant regions, supporting valid spatial attribution.

Figure 4 presents a LIME explanation for the same image, which the model also associated with frog eye leaf spot (35.47%). The highlighted superpixels demonstrate that visual patterns associated with necrotic lesions influenced the classifier's decision, offering a plausible explanation for inter-class confusion between visually overlapping symptoms.

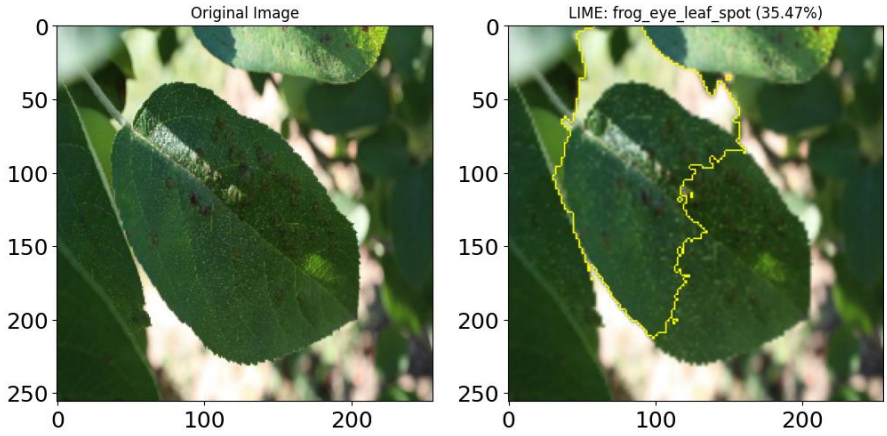

Figure 4. LIME explanation for the same apple leaf image. The right panel shows super pixels (outlined in yellow) that contributed most to the model's prediction of frog eye leaf spot. Overlapping lesion-like regions between disease classes may lead to misclassification.

Together, these interpretability tools validate that the model's attention aligns with lesion-relevant features, increasing trust in its predictions for well-defined classes. Simultaneously, they provide diagnostic insights into misclassification cases, especially among ambiguous categories, underscoring the need for future work in multi-label classification, improved data balancing, and domain-specific augmentation strategies.

3.1.2 Classification Accuracy and Metrics

To quantitatively assess the classifier's performance, standard metrics—including precision, recall, and F1-score—were computed for each of the 12 disease categories. The results are summarized in Table 1. While the overall accuracy reaches 65% for the 12-class (expanded-label) task, the macro-averaged F1-score drops to 0.32, indicating considerable variation in model performance across different classes, particularly among compound infection categories. Well-separated, visually distinct categories such as healthy, rust, and powdery mildew exhibit strong predictive performance, with F1-scores of 0.78, 0.77, and 0.63, respectively. The frog eye leaf spot class also demonstrates high discriminability (F1 = 0.68), benefiting from both larger support size and clear lesion morphology. In stark contrast, compound and structurally ambiguous classes—including frog eye leaf spot complex, powdery mildew complex, rust complex, and scab frog eye leaf spot complex—exhibit near-zero precision and recall. These findings are consistent with the confusion matrix analysis in Section 4.1, where misclassification was concentrated along off-diagonal cells for these hybrid labels. Particularly, the low F1-scores highlight the model's difficulty in recognizing co-occurring or overlapping symptom patterns, as further supported by the Grad-CAM and LIME visualizations presented earlier. Additionally, rare classes with limited support, such as rust frog eye leaf spot (support = 3), also suffer from negligible recall, underscoring the challenges of class imbalance and sparse representation in training data. The performance gap between simple and compound categories suggests a need for more sophisticated modeling strategies. Potential improvements include adopting multi-label training objectives, incorporating class-weighted or focal loss functions, and applying targeted data augmentation to enrich underrepresented visual patterns. Overall, while the model is effective for dominant, single-disease categories with distinct visual cues, substantial refinement is required to improve its generalization to rare, compound, or ambiguous disease instances.

3.1.3 Model Comparison with ResNet50, InceptionV3, and EfficientNetV2L

To evaluate the effectiveness of the proposed fine-tuning MobileNetV2 model, we conducted a comparative study against four widely used CNN architectures: MobileNetV3, ResNet50, InceptionV3, and EfficientNetV2L. All models were trained and tested under the same conditions to ensure a fair comparison in terms of classification performance, inference latency, and model size. As shown in Table 2, our fine-tuning MobileNetV2 model achieved the highest classification accuracy (71.04%) on the original 5-class FGVC8 label setting. At the same time, the detailed 12-class analysis is reported separately in Table 1 and Figure 2 while

maintaining a low inference time of 0.00195 seconds per image and a moderate model size of 12.79 MB. In contrast, deeper models such as ResNet50 and InceptionV3 achieved lower accuracy (26.89% and 53.78%, respectively) while having significantly larger parameter sizes (96.46 MB and 89.96 MB, respectively). Although EfficientNetV2L is known for its high capacity, it underperformed in both accuracy (28.26%) and speed, exhibiting the longest inference time (0.00773 seconds) and largest model size (455.32 MB). These findings suggest that the fine-tuned MobileNetV2 model offers a favorable trade-off among accuracy, efficiency, and deployability, making it well-suited for edge computing environments with limited computational resources.

Table 1. Classificationreportfor12categories.

Class	Precision	Recall	F1-score	Support
complex	0.42	0.18	0.25	329
frogeyeleafspot	0.58	0.82	0.68	641
frogeyeleafspotcomplex	0.00	0.00	0.00	44
healthy	0.71	0.88	0.78	1061
powderymildew	0.71	0.56	0.63	172
powderymildewcomplex	0.00	0.00	0.00	23
rust	0.84	0.70	0.77	379
rustcomplex	0.00	0.00	0.00	17
rustfrogeyeleafspot	0.00	0.00	0.00	3
scab	0.64	0.65	0.63	857
scabfrogeyeleafspot	0.25	0.03	0.06	129
scabfrogeyeleafspotcomplex	0.00	0.00	0.00	54
Accuracy			0.65	3726
Macroavg	0.34	0.32	0.32	3726
Weightedavg	0.61	0.65	0.62	3726

Table 2. Model Comparison on Accuracy and Inference Time

Model	Accuracy (%)	Inference Time (sec/image)	Model Size (MB)
Fine-TuningMobileNetV2	71.04	0.00195	12.79
MobileNetV3	26.89	0.00178	5.68
ResNet50	26.89	0.00248	96.46
InceptionV3	53.78	0.00287	89.96
EfficientNetV2L	28.26	0.00773	455.32

3.1.4 Robustness under Field-Like Conditions

To assess the robustness of the proposed model under realistic agricultural scenarios, we simulated four types of visual disturbances commonly encountered in the field: low illumination, partial occlusion, background clutter, and a composite of all three. These perturbations were applied to the test set to evaluate generalization performance beyond controlled conditions.

Table 3 summarizes the classification accuracy across noise settings. Under normal conditions, the model achieved 65.2% accuracy. Performance declined to 59.9% under low illumination, indicating moderate sensitivity to lighting. A more pronounced drop was observed under partial occlusion (48.0%) and complex backgrounds (26.6%), suggesting a reliance on unoccluded lesion features and consistent visual context. When all distortions were combined, accuracy fell to 30.2%, reflecting the compounding effect of multiple sources of noise. These results indicate that while MobileNetV2 performs reasonably well under mild degradation, its predictive stability is challenged by severe or compound noise. The findings highlight the need to integrate robust augmentation strategies and interpretability tools, such as Grad-CAM, to enhance model reliability for field deployment.

Table 3. Model Accuracy under Simulated Field Conditions

Condition	Accuracy (%)
Normal (baseline)	65.2
Low Illumination	59.9
Partial Occlusion	48.0
Complex Background	26.6
Combined Noise	30.2

3.1.5 Efficiency and Edge Readiness

The proposed MobileNetV2 model was evaluated for model size and inference latency to assess its feasibility for future deployment on resource-constrained platforms. The trained model occupies 13.6 MB of storage and achieves an average inference time of 0.00195 seconds per image on a CUDA-enabled GPU, suggesting suitability for low-latency applications. While these metrics indicate that the model is computationally efficient compared to deeper architectures such as ResNet50 and EfficientNetV2L, it is important to note that no real-world hardware deployment was conducted in this study. The reported performance metrics serve as reference indicators for potential use in edge computing environments such as mobile devices or embedded systems. Additional testing under hardware-specific constraints is required to validate real-time feasibility in practice.

3.2 Discussion

This section analyzes the empirical findings with the model's predictive behavior, interpretability, robustness, and deployability in real-world agricultural scenarios.

3.2.1 Interpretation of Findings

The results confirm that a fine-tuned MobileNetV2 can accurately classify visually distinct apple leaf diseases, especially for dominant single-disease classes such as healthy, rust, and frog eye leaf spot. These classes benefited from larger sample sizes, clearer symptom morphology, and lower intra-class variation, enabling effective feature extraction with lightweight convolutional layers. Conversely, the model consistently underperformed on compound classes with overlapping symptoms. These categories exhibited near-zero F1 Scores, highlighting the model's difficulty in disentangling co-occurring visual cues in a single-label classification setup. Explainability tools—Grad-CAM and LIME—confirmed that in such cases, attention maps often prioritized dominant lesion regions, while ignoring subtle secondary features. This reveals an attention collapse phenomenon, in which dominant disease patterns overshadow minority cues in multi-infection samples. Moreover, the asymmetry in misclassification patterns—where compound classes are misidentified as simpler ones but not vice versa—suggests that the model has implicitly learned to default to majority class prototypes. This indicates a distributional bias reinforced by class imbalance in the training data.

3.2.2 Deployment Feasibility and Use Cases

The fine-tuned MobileNetV2 model demonstrates favorable computational characteristics for potential use on edge devices. With a compact size of 13.6 MB and an average inference latency of less than 2 milliseconds per image on a GPU, the model demonstrates efficiency suitable for offline diagnostic scenarios in bandwidth-limited or rural agricultural contexts. Although physical deployment on platforms such as Raspberry Pi or mobile system-on-chips (SoCs) was not conducted in this study, the reported performance metrics provide preliminary evidence supporting the feasibility of such environments. Further validation through on-device testing will be required to confirm compatibility under real-world constraints, including thermal, energy, and input variability considerations. In practical use, the model may be integrated into mobile applications or drone-based crop monitoring tools to assist field workers or smallholder farmers. The inclusion of Grad-CAM and LIME facilitates visual interpretation of the model's decisions, offering an added layer of transparency and trust in high-stakes agricultural decision-making.

3.2.3 Scalability to Other Crops and Datasets

The proposed pipeline, based on transfer learning and image-based augmentation, is transferable to other crops, provided that annotated image datasets with sufficient inter-class diversity are available. For example, fine-tuning on tomato, rice, or grapevine datasets is feasible using the same backbone architecture. Explainability tools also aid in rapid domain adaptation, enabling visual validation across new phenotypes.

3.2.4 Ablation Study on Class Rebalancing

A weighted loss function was employed to assess its impact on class imbalance. The results indicate notable performance gains in previously underperforming compound classes, particularly in recall. For example, rust complex and powdery mildew complex showed a substantial increase in detection after reweighting, as illustrated in Table 4. However, this improvement often came at the cost of reduced performance in the dominant classes, underscoring the trade-off between minority-class recall and majority-class precision. To better illustrate the improvements, Figure 5 visualizes the change in F1-score across minority classes. Notably, several compound classes showed meaningful gains after applying weighted loss, reinforcing the importance of class balancing in agricultural datasets.

Table 4. Impact of Class Rebalancing on Minority Class Performance. The weighted loss improves recall and F1-score for compound and minority disease classes.

Class	Original Model			With Weighted Loss		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
complex	0.4653	0.2036	0.2833	0.2453	0.3191	0.2774
Frog eye leaf spot	0.6748	0.6443	0.6592	0.5294	0.3651	0.4321
Frog eye leaf spot complex	0.0000	0.0000	0.0000	0.2727	0.0682	0.1091
healthy	0.7767	0.7738	0.7753	0.7029	0.8096	0.7525
Powdery mildew	0.7830	0.4826	0.5971	0.5977	0.6041	0.6012
Powdery mildew complex	0.0000	0.0000	0.0000	0.1765	0.1304	0.1500
rust	0.5963	0.8575	0.7035	0.7522	0.4655	0.5681
Rust complex	0.0000	0.0000	0.0000	0.0166	0.6429	0.0324
Rust frog eye leaf spot	0.0000	0.0000	0.0000	0.0357	0.0870	0.0506
scab	0.5437	0.7993	0.6471	0.7010	0.3459	0.4629
Scab frog eye leaf spot	0.0000	0.0000	0.0000	0.1321	0.1628	0.1458
Scab frog eye leaf spot complex	0.0000	0.0000	0.0000	0.0714	0.0370	0.0488
Macro avg (all)	0.3200	0.3134	0.3055	0.3528	0.3354	0.3023

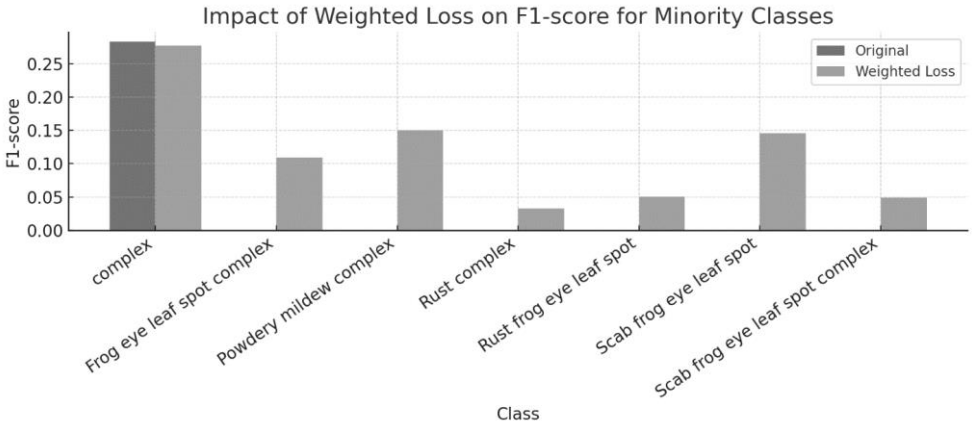

Figure 5. Per-class F1-score comparison before and after applying class weighting. Minority classes with zero F1 Scores in the original model showed notable improvements. Note that several previously underperforming compound classes (e.g., rust complex, powdery mildew complex) exhibit significant improvements in recall after class weighting.

3.2.5 Limitations, Deployment Challenges, and Ethical Considerations

Despite improved classification accuracy and explainability, several limitations remain in the proposed system. The model's performance degrades significantly under compounded visual disturbances—such as occlusion combined with low illumination—revealing sensitivity to high-entropy image conditions. In addition, the use of a single-label training framework limits the system's ability to capture real-world complexities, particularly in cases involving co-infections or overlapping symptom regions. These challenges constrain diagnostic performance in field scenarios where image quality and lesion presentation vary widely. To address these limitations, future work should explore multi-label classification frameworks that explicitly model multiple disease types within a single image. Incorporating uncertainty-aware learning objectives—such as focal loss or label smoothing—can enhance robustness in the presence of ambiguous or imbalanced class distributions. Lesion severity estimation and segmentation techniques could also support early-stage intervention and increase interpretability. Hardware benchmarking on ARM-based mobile platforms will be necessary to validate the system's practical viability for offline use in rural, low-connectivity agricultural environments. From an ethical standpoint, while automated plant disease diagnostics can offer significant value to smallholder farmers and agritech practitioners, it is important to recognize the limitations of imperfect AI systems. Misclassifications—particularly false negatives in early disease stages—may result in inaction and crop losses. Therefore, future deployments should consider human-in-the-loop frameworks that supplement AI-generated outputs with expert review or local agricultural advisory systems. Additionally, interpretability techniques such as Grad-CAM and LIME should be made accessible through user interfaces to enhance user trust and facilitate informed decision-making. These human-centered safeguards are essential to ensure responsible, reliable, and equitable deployment of AI systems in agriculture.

4. Conclusions

This study investigated a MobileNetV2-based model for apple leaf disease classification under heterogeneous field conditions. The model demonstrated strong performance in well-defined single-disease categories but struggled with compound or overlapping infections, reflecting the challenges posed by imbalanced and visually ambiguous classes. Grad-CAM and LIME confirmed that the model focused on biologically relevant lesion areas, increasing trust in predictions while revealing sources of misclassification. Key contributions of this study include: (1) the development of a lightweight MobileNetV2-based classifier tailored for heterogeneous field conditions, (2) explicit evaluation of compound infection categories derived from FGVC8 multiple-disease labels, and (3) integration of XAI methods to interpret misclassification behavior. Although physical deployment was not conducted, the results underline the value of lightweight deep learning models combined with interpretability methods for agricultural disease recognition. Future work should explore multi-label learning to capture co-infections better, employ class-balancing strategies to improve minority categories, and extend validation to larger, more diverse datasets. These steps will help advance the development of practical and trustworthy diagnostic systems to support sustainable crop management.

Appendix A: Disease Class Descriptions

This appendix provides comprehensive descriptions and pathological characteristics of the 12 apple leaf disease categories considered in this study. These categories were annotated by plant pathology experts using the PlantVillage dataset and encompass both single-disease manifestations and compound symptom co-occurrences commonly observed in real-world field conditions. Figure .6 presents representative examples of each category, illustrating the diversity in lesion morphology, co-infection patterns, and background complexity that challenge automated classification.

- **Complex** — Hybrid or unclassified symptoms: Leaves in this category exhibit multiple overlapping lesion types without a dominant visual pattern, often arising from ambiguous or advanced-stage infections. These are the most visually heterogeneous and challenge standard classifiers due to non-distinct lesion edges and color blending.

- **frog eye leaf spot** — Cercospora spp.: Characterized by circular gray lesions with purple margins, primarily affecting mature leaves. As illustrated in Figure 6., lesions are discrete, with sharply defined boundaries that make them visually separable under standard illumination.
- **frog eye leaf spot complex** — Mixed infection: Frogeye Leaf Spot present alongside one or more additional diseases, producing lesion zones with color gradients or merged edges. Visual identification becomes more difficult due to interference from adjacent pathogen effects.
- **Healthy** — Control class: Leaves in this category are free of visible infections, exhibiting uniform green pigmentation, intact structure, and no lesion formations. They serve as negative controls in training and are clearly separable in both visual and feature space.
- **powdery mildew** — Erysiphales fungi: Exhibits as white or gray powder-like growth on the upper leaf surface. This class is visually distinct due to the uniform texture and desaturation effect, as shown in the middle-left panel of Figure 6.
- **powdery mildew complex**—Compoundinfection: PowderyMildew occurring with another disease, often resulting in visually conflicting regions—e.g., powdery overlay on necrotic or chlorotic patches. Boundary regions are often misclassified due to overlapping visual signatures.
- **rust** — Gymnosporangium spp.: Manifests as bright orange or reddish pustules, often on the underside of leaves. Figure 6. shows typical rust pustules with well-demarcated circular morphology.
- **rust complex** — Co-infection class: Rust symptoms co-occur with another pathogen, often resulting in mixed orange-brown lesions with diffuse borders or color bleeding. Co-localization complicates visual separation and feature-based segmentation.
- **rust frog eye leaf spot** — Dual infection: Co-presence of Rust and Frogeye Leaf Spot in the same leaf area. Lesions may appear as rust-colored pustules surrounded by necrotic halos, leading to spatial ambiguity in model attention maps.
- **scab** — Venturia inaequalis: Distinguished by dark, scab-like lesions with a rough surface and irregular margins. They typically appear near leaf tips or edges. The structural roughness is an important cue for both human experts and deep learning models.
- **scab frog eye leaf spot** — Compound lesion type: Visual overlap between Scab and Frogeye Leaf Spot, sometimes with concentric lesion patterns or discoloration zones that reflect both fungal pathologies. Grad-CAM heatmaps often highlight only one dominant feature, leading to misclassification.
- **Scab frog eye leaf spot complex** — Ternary or ambiguous combination: Leaves showing features of both Scab and Frogeye Leaf Spot, along with additional visual noise or secondary infection. This is one of the most challenging categories for both annotation and classification, often appearing in the lower-confidence prediction regions of LIME visualizations.

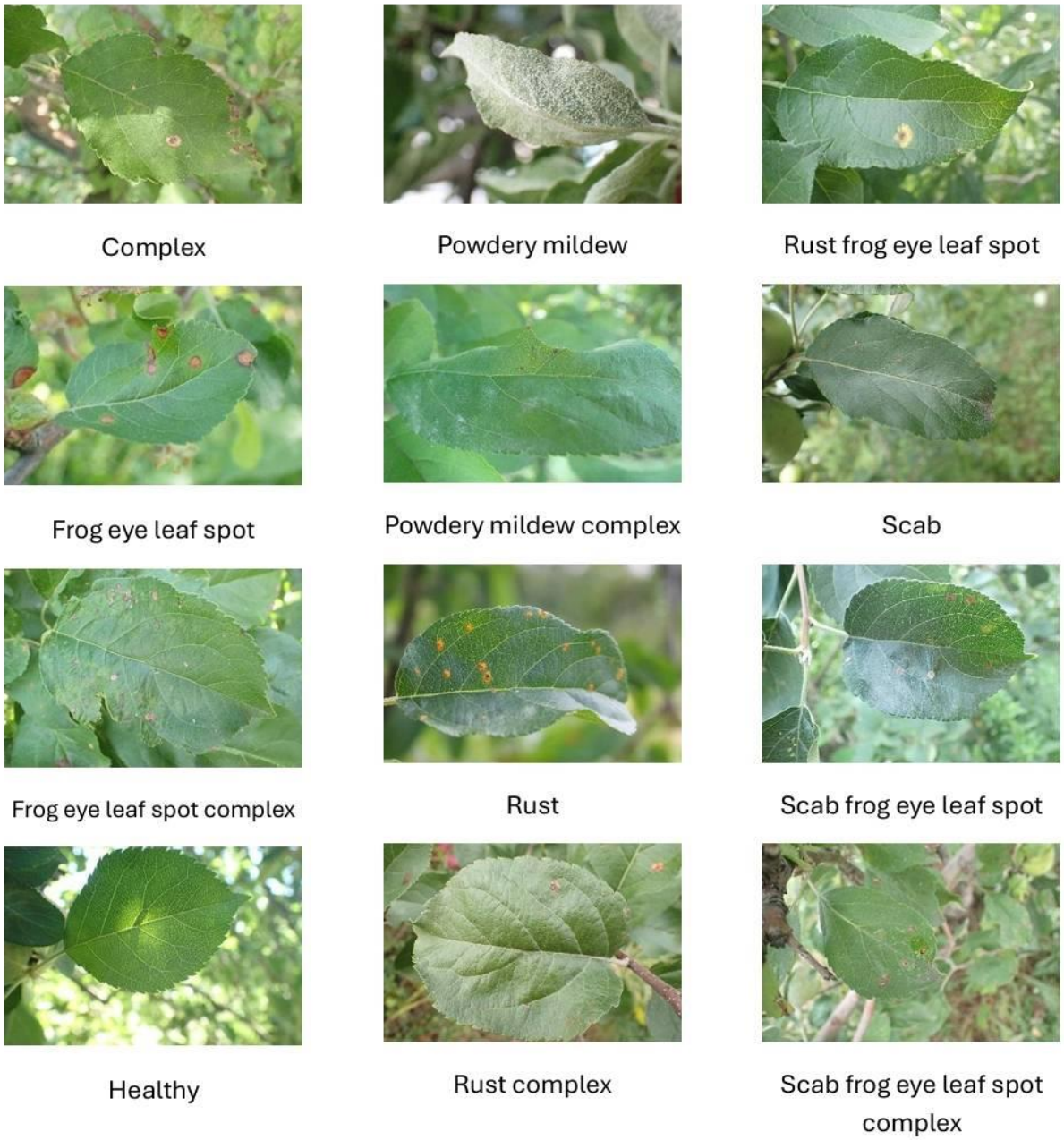

Figure 6. Representative examples of the 12 apple leaf disease categories classified by the proposed model. These include single-disease classes (e.g., scab, rust, powdery mildew, frog eye leaf spot, healthy) and compound classes (e.g., scab frog eye leaf spot, rust complex, powdery mildew complex, frog eye leaf spot complex, rust frog eye leaf spot, scab frog eye leaf spot complex, complex). The annotated images illustrate diverse lesion morphologies, co-infection interactions, and heterogeneous backgrounds, all of which significantly influence classification accuracy under real-world field conditions.

5. Acknowledgements

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Mahlein, A.-K. Plant Disease Detection by Imaging Sensors–Parallels and Specific Demands for Precision Agriculture and Plant Phenotyping. *Plant Dis.* **2016**, *100*(2), 241–251. https://doi.org/10.1094/PDIS-03-15-0340-FE

[2] Yamamoto, K.; Togami, T.; Yamaguchi, N. Super-resolution of Plant Disease Images for the Acceleration of Image-based Phenotyping and Vigor Diagnosis in Agriculture. *Sensors* **2017**, *17*(11), 2557. https://doi.org/10.3390/s17112557

[3] Mohanty, S. P.; Hughes, D. P.; Salathé, M. Using Deep Learning for Image Based Plant Disease Detection. *Front. Plant Sci.* **2016**, *7*, 1419. https://doi.org/10.3389/fpls.2016.01419

[4] Ferentinos, K. P. Deep Learning Models for Plant Disease Detection and Diagnosis. *Comput. Electron. Agric.* **2018**, *145*, 311–318. https://doi.org/10.1016/j.compag.2018.01.009

[5] Too, E. C.; Yujian, L.; Njuki, S.; Yingchun, L. A Comparative Study of Fine-tuning Deep Learning Models for Plant Disease Identification. *Comput. Electron. Agric.* **2019**, *161*, 272–279. https://doi.org/10.1016/j.compag.2018.03.032

[6] Kamilaris, A.; Prenafeta-Boldú, F. X. Deep Learning in Agriculture: A Survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

[7] Xu, J.; Gu, B.; Tian, G. Review of Agricultural IoT Technology. *Artif. Intell. Agric.* **2022**, *6*, 10–22. https://doi.org/10.1016/j.aiia.2022.01.001

[8] Barbedo, J. G. A. Impact of Dataset Size and Variety on the Effectiveness of Deep Learning and Transfer Learning for Plant Disease Classification. *Comput. Electron. Agric.* **2018**, *153*, 46–53. https://doi.org/10.1016/j.compag.2018.08.013

[9] Udiani, O.; Mason, S.; Smith, G.; Riviere, J. E.; Baynes, R. E. Automation and Applications of the Tolerance Limit Method in Estimating Meat Withdrawal Periods for Veterinary Drugs. *Comput. Electron. Agric.* **2018**, *146*, 125–135. https://doi.org/10.1016/j.compag.2018.02.005

[10] Paul, K.; Chatterjee, S. S.; Pai, P.; Bhadra, B.; Dasgupta, S. Viable Smart Sensors and Their Application in Data Driven Agriculture. *Comput. Electron. Agric.* **2022**, *198*, 107096. https://doi.org/10.1016/j.compag.2022.107096

[11] Agu, C. M.; Menkiti, M. C.; Ekwe, E. B.; Agulanna, A. C. Modeling and Optimization of *Terminalia catappa* L. Kernel Oil Extraction Using Response Surface Methodology and Artificial Neural Network. *Artif. Intell. Agric.* **2020**, *4*, 1–11. https://doi.org/10.1016/j.aiia.2020.01.001

[12] Lu, M.; Chen, C.-L. Detection and Classification of Bearing Surface Defects Based on Machine Vision. *Appl. Sci.* **2021**, *11*(4), 1825. https://doi.org/10.3390/app11041825

[13] Thapa, R.; Zhang, K.; Snavely, N.; Belongie, S.; Khan, A. *Plant Pathology 2021 - FGVC8*. Kaggle Competition. https://kaggle.com/competitions/plant-pathology-2021-fgvc8 (accessed 2026-02-24).

[14] MacHardy, W. E. *Apple Scab: Biology, Epidemiology, and Management*; APS Press: St. Paul, MN, **1996**.

[15] Kim, Y. K.; Kwak, J. H.; Aguilar, C. G.; Xiao, C. L. First Report of Black Rot on Apple Fruit Caused by *Diplodia seriata* in Washington State. *Plant Dis.* **2016**, *100*(7), 1499. https://doi.org/10.1094/PDIS-12-15-1463-PDN

[16] Glawe, D. A. The Powdery Mildews: A Review of the World's Most Familiar (Yet Poorly Known) Plant Pathogens. *Annu. Rev. Phytopathol.* **2008**, *46*, 27–51. https://doi.org/10.1146/annurev.phyto.46.081407.104740

[17] Zhang, S.; Wu, X.; You, Z.; Zhang, L. Leaf Image Based Cucumber Disease Recognition Using Sparse Representation Classification. *Comput. Electron. Agric.* **2017**, *134*, 135–141. https://doi.org/10.1016/j.compag.2017.01.014

[18] Wang, S.; Xu, D.; Liang, H.; Su, C.; Wei, W. Advances in Deep Learning Applications for Plant Disease and Pest Detection: A Review. *Remote Sens.* **2025**, *17*(4), 698. https://doi.org/10.3390/rs17040698

[19] Pacal, I.; Kunduracioglu, I.; Alma, M. H.; Slany, V.; Martinek, R. A Systematic Review of Deep Learning Techniques for Plant Diseases. *Artif. Intell. Rev.* **2024**, *57*, 304. https://doi.org/10.1007/s10462-024-10944-7

[20] Das, P. K.; Rupa, S. S.; Pumrin, S.; Das, U. C.; Hossen, M. K. Deep Learning for Plant Disease Detection and Classification: A Systematic Analysis and Review. *Curr. Agric. Sci. Technol.* **2024**, *24*, 259016. https://doi.org/10.55003/cast.2024.259016

[21] Yang, B.; Li, M.; Li, F.; Li, C.; Wang, J. A Novel Plant Type, Leaf Disease and Severity Identification Framework Using CNN and Transformer with Multi-label Method. *Sci. Rep.* **2024**, *14*, 11664. https://doi.org/10.1038/s41598-024-62452-x

[22] Nnamdi, U. V.; Abolghasemi, V. Optimised MobileNet for Very Lightweight and Accurate Plant Leaf Disease Detection. *Sci. Rep.* **2025**, *15*, 43690. https://doi.org/10.1038/s41598-025-27393-z

[23] Islam, M.; Azad, A. K. M.; Arman, S. E.; Alyami, S. A.; Hasan, M. M. PlantCareNet: An Advanced System to Recognize Plant Diseases with Dual-mode Recommendations for Prevention. *Plant Methods* **2025**, *21*, 52. https://doi.org/10.1186/s13007-025-01366-9

[24] Prince, R. H.; Mamun, A. A.; Peyal, H. I.; Khandakar, A.; Ayari, M. A. CSXAI: A Lightweight 2D CNN-SVM Model for Detection and Classification of Various Crop Diseases with Explainable AI Visualization. *Front. Plant Sci.* **2024**, *15*, 1412988. https://doi.org/10.3389/fpls.2024.1412988

[25] Toda, Y.; Okura, F. How Convolutional Neural Networks Diagnose Plant Disease. *Plant Phenomics* **2019**, *2019*, 9237136. https://doi.org/10.34133/2019/9237136

[26] Wang, G.; Sun, Y.; Wang, J. Automatic Image-based Plant Disease Severity Estimation Using Deep Learning. *Comput. Intell. Neurosci.* **2017**, *2017*, 2917536. https://doi.org/10.1155/2017/2917536

[27] Parashar, P.; Johri, P. Enhancing Apple Leaf Disease Detection: A CNN-based Model Integrated with Image Segmentation Techniques for Precision Agriculture. *Int. J. Math. Eng. Manage. Sci.* **2024**, *9*(4), 943–964. https://doi.org/10.33889/IJMEMS.2024.9.4.050

[28] Bhatti, M.; Zeeshan, Z.; Ms, S.; Asif, M.; Afzal, T. Advanced Plant Disease Segmentation in Precision Agriculture Using Optimal Dimensionality Reduction With Fuzzy C-Means Clustering and Deep Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 18264–18277. https://doi.org/10.1109/JSTARS.2024.3437469