Journal of Research in Mechanics of Agricultural Machinery

Shahrekord University

Research Paper

A Lightweight Family of CNN-MLP Framework for Apple Leaf Disease Classification on Augmentation-Enhanced Heterogeneous Dataset

Forough Ehsanfar[1], Fatemeh Moayedi[2]

[1] Department of Robotics Engineering, Larestan Higher Education Center, Lar, Iran
[2] Department of Computer Engineering, Larestan Higher Education Center, Lar, Iran

Article History

Submited:2026/04/30

Revised: 2026/05/17

Accepted: 2026/06/14

Published online: 2026/06/22

Keywords:

Apple leaf disease classification
Hybrid deep learning
EfficientNet-B0
ResNet-18
Lightweight deep learning
Heterogeneous visual conditions

***Corresponding author email:**
moayedi@lar.ac.ir

ORCID:
0000-0003-0877-1566

Abstract

Accurate, computationally efficient classification of apple leaf diseases is essential to improving orchard productivity. Despite advances in deep learning, many existing models rely on intensive architectures that limit deployment in resource-constrained environments. This study proposes a lightweight CNN-MLP framework optimized for disease classification under heterogeneous visual conditions. Pre-trained ResNet-18 and EfficientNet-B0 backbones were utilized for feature extraction, with their original classification heads replaced by a task-specific, two-layer multilayer perceptron (MLP) head. This architectural adaptation, incorporating a 256-neuron hidden layer and dropout regularization, aims to enhance learning of non-linear boundaries in diverse datasets. An experimental evaluation on the 9-class AL9EE dataset, conducted across five independent runs, demonstrated that the EfficientNet-B0-MLP model achieved 98.42 ± 0.26% accuracy and a macro F1-score of 98.33 ± 0.34%, outperforming the ResNet-18-MLP variant. The lower standard deviations indicate improved training stability and reproducibility. By balancing predictive performance with reduced computational complexity, this framework provides a practical solution for edge-oriented agricultural applications, offering an effective pathway for intelligent disease monitoring in smart farming.

How to cite this paper:

Ehsanfar, F and Moayedi, F. (2026). A Lightweight Family of CNN-MLP Framework for Apple Leaf Disease Classification on Augmentation-Enhanced Heterogeneous Dataset. *Journal of Research in Mechanics of Agricultural Machinery*. 39: 99-120. https://dx.doi.org/10.22034/jrmam.2026.14992.763.

https://dx.doi.org/10.22034/jrmam.2026.14992.763

1.Introduction

Apple (*Malus domestica*) is one of the most strategic and economically significant horticultural crops, with global production estimated at approximately 97 million tons in 2023. It plays a vital role in food supply chains and international trade (FAO, 2025). Despite this economic significance, apple cultivation remains highly vulnerable to foliar diseases such as Alternaria leaf spot, powdery mildew, and apple scab (Figure 1), which can cause 20–30% yield losses globally and up to 40% in humid microclimates, resulting in major economic and ecological consequences (Agricultural Research Center of Iran, 2022; Li & Rai, 2020; Fu et al., 2022). Early and reliable detection of these diseases is therefore vital to safeguard productivity, reduce unnecessary pesticide usage, and enhance the sustainability of orchard ecosystems.

Traditional diagnostic approaches—including manual inspection and laboratory testing—are time-consuming, labor-intensive, and dependent on expert availability, making them impractical for large-scale or real-time applications. The shift toward AI-driven precision agriculture thus represents not only a technological advancement but also an economic and ecological necessity—particularly for developing nations such as Iran,

Figure 1. Example of apple scab symptoms caused by Venturia inaequalis. Typical symptoms include dark, circular to irregular necrotic lesions on fruits and leaves. Such visually variable symptoms, especially under field conditions, motivate the need for robust automated disease classification.

where climatic diversity and limited technical resources demand scalable, low-cost solutions (Agricultural Research Center of Iran, 2022).

Recent breakthroughs in deep learning and computer vision have revolutionized plant disease detection by enabling automated, high-accuracy classification systems. However, many existing models have primarily been validated on controlled or single-dataset benchmarks, and therefore their performance under broader visual variability remains insufficiently examined. Such variability may include differences in illumination, background complexity, symptom appearance, and image style. This limitation highlights the need for lightweight architectures that remain computationally efficient while maintaining robust performance across visually diverse datasets. (Fu et al., 2022; AppleLeafNet, 2024; Ali et al., 2025).

In this study, a lightweight CNN-MLP framework is developed and empirically evaluated for apple leaf disease classification on both controlled and augmentation-enhanced visually diverse datasets. The redesigned head aims to improve non-linear class separation and generalization across the visual variability in the evaluated datasets, rather than explicitly enforcing domain-invariant representations.

The remainder of this paper is organized as follows. Section 2 presents a wide-ranging literature review, summarizes previous shallow, deep, and hybrid learning approaches for apple leaf disease detection, and identifies key research gaps. Section 3 describes the methodology, detailing the baseline and proposed architectures and their optimization strategies for improved generalization within the evaluated dataset. Section 4 reports and analyzes the experimental results, including dataset descriptions, training configurations, and comprehensive comparative evaluations. Finally, Section 5 concludes the paper by summarizing the major findings, highlighting the contributions of this study, and outlining promising directions for future research.

The main contributions of this study can be summarized as follows:

1. A lightweight CNN-MLP framework for apple leaf disease classification evaluated

on controlled and augmentation-enhanced visually diverse datasets.

2. Integration of pretrained CNN (ResNet and EfficientNet) backbones with a compact two-layer MLP head to enhance non-linear class separation while maintaining computational efficiency.
3. Comprehensive evaluation on both a controlled laboratory dataset (PlantVillage) and an augmentation-enhanced visually diverse dataset (AL9EE)
4. Detailed analysis of classification performance and computational complexity for edge-oriented agricultural applications.

2. Literature Review and Previous Methods

Previous research on apple leaf disease detection can be broadly categorized into three groups: shallow learning, deep learning, and hybrid approaches, each with distinct strengths and limitations.

The first category, shallow learning methods (Table 1), primarily focuses on manual feature extraction of geometric, color, and texture descriptors from leaf images.

For example, Khan et al. (2019) utilized genetic algorithms and correlation analysis to achieve satisfactory segmentation accuracy; however, the complexity of multi-stage feature extraction and strong dependence on optimization parameters limited scalability. Similarly, Anam (2020) combined PSO and K-means, yielding Table 1. Studies based on shallow learning methods**.** better segmentation than standard K-means, but the high computational cost hindered real-time application.

Table 1. Studies based on shallow learning methods

Authors	Year	Method and Model	Strengths	Limitations / Research Gaps
Khan et al.	2019	Genetic Algorithm (GA)-based feature selection and correlation analysis	High segmentation and classification performance	Multi-stage processing; dependence on optimization algorithms; increased computational complexity
Anam	2020	PSO–K-means hybrid segmentation	Improved segmentation quality compared with conventional K-means	High computational cost; limited suitability for real-time applications
Bracino et al.	2020	Hybrid machine learning framework (KNN, ANN, GPR, and SVM)	Effective disease classification; strong performance of GPR with ARD kernel	Heavy reliance on handcrafted features; limited adaptability to unseen conditions
Chakraborty et al.	2021	Otsu thresholding, histogram equalization, and multiclass SVM	High classification accuracy (96%)	Scalability limitations of SVM; computationally intensive preprocessing
Sai and Patil	2022	Geometric and texture feature extraction with machine learning classifiers	High performance, particularly with Random Forest (98.12%)	Dependence on manual feature engineering; limited scalability for large datasets

Later studies, such as Chakraborty et al. (2021), demonstrated that integrating advanced image preprocessing with SVM classifiers could achieve high accuracy; nevertheless, computational overhead and limited generalization remained major challenges. Overall, while classical methods perform adequately on small, controlled datasets, their generalization to real-world, multi-domain conditions is limited.

The emergence of deep learning (Table 2) has helped overcome many of the limitations of shallow approaches. For instance, Ashmafee et al. (2023) proposed an EfficientNetV2S-based transfer learning framework for apple leaf disease classification using the apple leaf disease subset of the publicly available PlantVillage dataset. Their study demonstrated the effectiveness of pretrained deep feature extractors for controlled RGB image datasets. However, since the evaluation was mainly conducted on PlantVillage images, the robustness of the model under real-world field conditions, heterogeneous backgrounds, illumination variations, and diverse acquisition environments remains insufficiently explored.

Similarly, Fu et al. (2022) proposed a lightweight CNN for plant disease classification; however, the robustness of such lightweight models under broader visual variability remains an important concern.

Recent studies such as AppleLeafNet (2024) and related EfficientNet-based approaches have reported very high accuracies; however, many evaluations remain dataset-specific, and direct evidence of generalization to unseen acquisition environments is often limited. Although deep learning models have improved classification accuracy compared with traditional methods, maintaining stable performance across visual variability, such as illumination changes, background complexity, and symptom diversity, remains a major challenge.

Table 2. Studies based on deep learning methods

Authors	Year	Method and Model	Input Data Type	Strengths	Limitations / Research Gaps
Ashmafee et al.	2023	EfficientNetV2S-based transfer learning framework	RGB images from the PlantVillage apple leaf subset	Effective use of pretrained deep features; suitable for RGB-based plant disease classification	Evaluation was limited to PlantVillage images; limited evidence of robustness under field conditions, complex backgrounds, and illumination variability
Fu et al.	2022	Lightweight Convolutional Neural Network (CNN)	RGB images from AI Studio dataset	Low computational complexity; suitable for field deployment and edge devices	Reduced robustness to illumination changes and environmental variability
Ali et al.	2024	AppleLeafNet: lightweight two-stage deep learning framework	RGB images from Plant pathology dataset(FGVC8	High classification accuracy (>98%); efficient architecture with reduced computational requirements	Limited evaluation across diverse orchard conditions and heterogeneous backgrounds
Ali et al.	2025	Fine-tuned EfficientNet-B0	RGB images from PlantVillage (PV) dataset and a curated Apple PV (APV) dataset	Very high classification accuracy (>99.78%); strong feature extraction capability and computational efficiency	Experiments conducted primarily on controlled datasets (PlantVillage and Apple PlantVillage); limited field validation

In recent years, several deep learning architectures have been explored for plant disease classification. Convolutional neural networks (CNNs) remain among the most widely adopted approaches due to their strong feature extraction capability and relatively efficient computational requirements. Architectures such as ResNet and EfficientNet have demonstrated strong performance across a variety of image classification tasks, including plant disease detection. Compared with larger, more computationally intensive models, lightweight CNN architectures offer a practical balance between classification accuracy and computational efficiency. Therefore, considering the objective of developing a model suitable for efficient training and potential deployment in smart agriculture environments, this study focuses on CNN-based architectures, specifically ResNet-18 and EfficientNet-B0, as backbone networks.

To address this, some researchers have adopted hybrid approaches. For example, Krizhevsky et al. (2017) combined a pre-trained AlexNet with PSO optimization, integrating the feature-extraction power of deep learning with the optimal-feature-selection capabilities of shallow methods.

Despite improvements in feature representation, the computational cost and hardware requirements of many deep learning models still limit their practical deployment in real agricultural environments. Hybrid architectures have shown strong potential for improving classification performance; however, their efficiency and scalability under visually heterogeneous dataset conditions remain insufficiently explored.

The literature review reveals three key research gaps. First, many existing models are evaluated primarily under laboratory or single-domain conditions, which may not reflect the variability encountered in practical field environments. Second, several lightweight models proposed for practical deployment are often evaluated on limited datasets, leaving their behavior under broader visual variability insufficiently examined. Third, hybrid approaches, although often achieving high accuracy, tend to involve greater computational complexity, which limits scalability and practical deployment.

Furthermore, the lack of diverse real-world datasets, the difficulty of distinguishing visually similar diseases, and insufficient consideration of hardware constraints remain important challenges in plant disease recognition (Chakraborty et al., 2021; Sai & Patil, 2022).

To address these issues, this study proposes a lightweight deep learning framework for apple leaf disease detection. The proposed framework partially addresses environmental variability through a lightweight, regularized design and evaluation on a heterogeneous dataset; however, it lacks an explicit robustness-enhancing module. The main characteristics of the proposed study include:

• An end-to-end design that eliminates the need for manual feature extraction

• Evaluation of classification performance under dataset-level visual variability, including differences in symptom appearance, image style, and class distribution.

• Comparative analysis of two redesigned architectures based on ResNet-18 and EfficientNet-B0

• Evaluation on two datasets representing different acquisition conditions: a laboratory dataset (PlantVillage) and a heterogeneous dataset (AL9EE) (Zhu, 2023).

• Comprehensive assessment of classification performance, convergence behavior, and generalization capability, providing insights for future research.

3. Methodology

Before introducing the proposed architectures, it is essential to establish a strong baseline for comparative analysis. Therefore, this section first presents two baseline architectures—ResNet-18 and EfficientNet-B0—which are widely recognized as foundational models for plant disease detection. These baselines provide a reliable reference for evaluating the effectiveness of subsequent improvements.

Following the baseline presentation, the optimized architectures developed in this study are described in detail, emphasizing their design motivations, structural innovations, and efficiency enhancements to achieve superior performance under multi-domain conditions.

By adopting this stepwise approach, the methodology ensures fair benchmarking, transparent performance evaluation, and reproducible validation of the proposed framework.

Moreover, leveraging deep learning–based feature extraction can improve the reliability of plant disease classification when applied to datasets with heterogeneous visual characteristics. Accordingly, this study proposes two lightweight CNN-MLP pipelines for accurate and computationally efficient apple leaf disease classification.

3.1 Baseline Architectures

Two baseline architectures were examined and further developed in this study to identify and design the most efficient models for apple leaf disease detection. The following subsections describe their structures and the rationale for their selection.

3.1.1 Baseline ResNet-18 Architecture

ResNet-18 was employed as a baseline backbone network in this study due to its compact architecture, stable training behavior, and widespread use in image classification. The model is based on residual learning, in which identity-based skip connections facilitate feature propagation across layers and help alleviate the vanishing gradient problem in deep neural networks (He et al., 2016).

In the ResNet-MLP baseline configuration, ResNet-18 is used strictly as a feature-extraction backbone rather than as a newly designed architecture. The input apple leaf image is processed through the convolutional and residual blocks of ResNet-18 to extract high-level discriminative visual representations. The resulting feature maps are then passed through a global average pooling (GAP) layer to generate a compact feature vector.

Instead of using the original fully connected classification layer of ResNet-18, the classifier is replaced with a lightweight customized multilayer perceptron (MLP) head. This task-specific modification allows the model to learn non-linear decision boundaries suitable for apple leaf disease classification while maintaining a relatively low computational burden. The MLP head receives the pooled feature vector and produces the final disease class prediction.

The overall structure of the ResNet-MLP baseline framework is illustrated in Figure 2. Since the objective of this study is not to redesign the internal residual blocks of ResNet-18, detailed block-level architectural descriptions are intentionally omitted. The focus is therefore placed on the task-relevant adaptation, namely replacing the original classifier with a lightweight MLP head for apple leaf disease prediction.

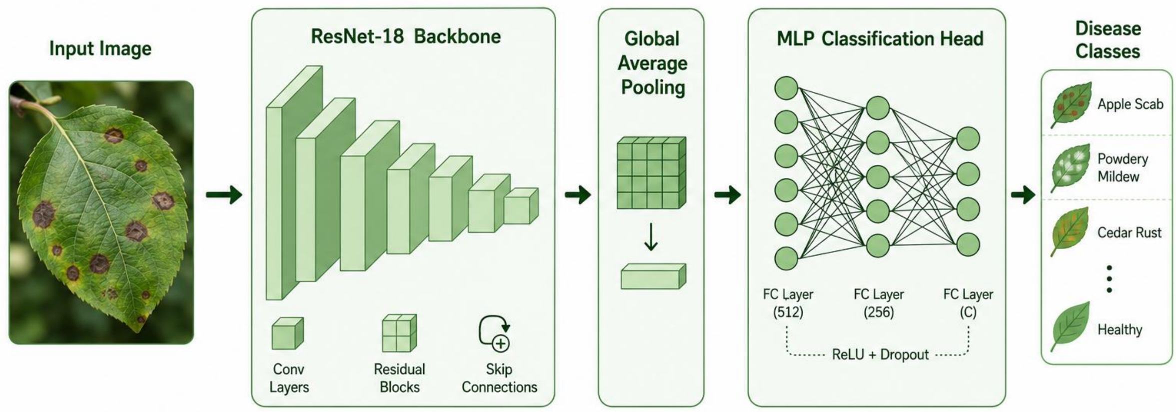

Figure 2. ResNet-MLP baseline framework for apple leaf disease classification. ResNet-18 is used as a pretrained feature-extraction backbone, followed by global average pooling and a lightweight customized MLP classification head for final disease prediction.

3.1.2 EfficientNet-B0 (Lightweight and Scalable Baseline)

EfficientNet-B0 adopts a compound scaling strategy that jointly scales network depth, width, and input resolution to improve accuracy while maintaining computational efficiency (Tan & Le, 2019). The architecture was discovered through neural architecture search and serves as the baseline model of the EfficientNet family.

Key architectural components include:

- MBConv blocks, inspired by MobileNetV2, which enable efficient feature extraction.
- Squeeze-and-Excitation (SE) modules dynamically recalibrate channel-wise feature responses by modeling inter-channel dependencies (Tan & Le, 2019).

The EfficientNet family (B0–B7) provides different trade-offs between accuracy and efficiency through compound scaling (Tan & Le, 2019). In this study, EfficientNet B0 was selected as a lightweight, scalable baseline.

The overall structure of the EfficientNet-B0 baseline architecture employed in this study is illustrated in Figure 3.

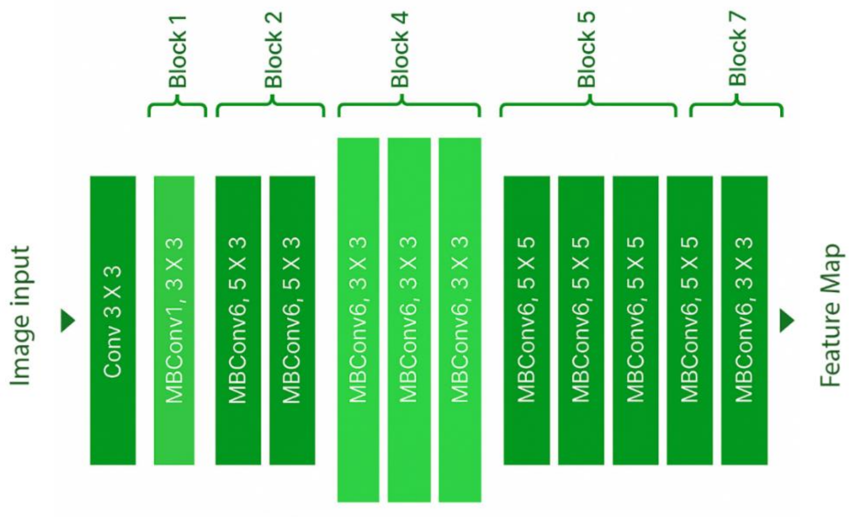

Figure 3. Baseline architecture of EfficientNet

3.2 Optimized Architectures in This Study

In this work, the two backbone architectures, ResNet-18 and EfficientNet-B0, were initialized with pretrained weights and fine-tuned (Howard & Gugger (2020)) end-to-end for apple leaf disease classification. No layer freezing was applied in any reported experiment. As illustrated in Figure 4, the original classifier layers of these backbones were replaced with a two-layer MLP head. This replacement is treated as an engineering-oriented architectural adaptation rather than a novel algorithmic contribution. The purpose of this redesign is to provide additional non-linear classification capacity on top of pretrained feature representations while maintaining a lightweight architecture suitable for edge-oriented applications.

After the forward pass through the fine-tuned backbone, the resulting feature representation is passed to the customized MLP classification head, as shown in Figure 4. This head consists of a fully connected layer with 256 neurons, followed by ReLU activation and a dropout layer for regularization. Finally, the output is mapped to a fully connected layer with the same number of neurons as the target classes. The 256-neuron hidden layer serves as an intermediate non-linear projection between the backbone feature extractor and the final classification layer. Specifically, it enables non-linear recombination of the high-level features extracted by the fine-tuned backbone and helps form more expressive decision boundaries than a direct linear classifier. The hidden dimension of 256 was selected empirically as a compact trade-off between representational capacity and computational efficiency. Smaller hidden dimensions may limit the classifier’s ability to capture discriminative inter-feature relationships, whereas larger dimensions may increase the number of trainable parameters, computational overhead, and the risk of overfitting. Therefore, this configuration provides sufficient non-linear classification capacity while keeping the classification head lightweight for practical deployment.

It should be noted that the proposed framework does not introduce explicit domain adaptation, domain generalization, adversarial alignment, feature alignment, uncertainty-aware learning, or any dedicated robustness-specific module for environmental variability. The stable performance observed under heterogeneous visual conditions mainly stems from the use of pretrained and fine-tuned backbones, the regularized MLP classification head, and the diversity of the evaluated datasets. In this study, the term heterogeneous visual conditions refers to the visual variability represented within the evaluated datasets, particularly the augmentation-enhanced AL9EE dataset, rather than to a controlled cross-domain evaluation across independent orchards, sensors, or acquisition environments.

Therefore, the observed performance should be interpreted as dataset-level generalization within the evaluated data distribution

Overall, the optimized ResNet-MLP and EffiNet-MLP pipelines leverage the representation-learning capabilities of pretrained and fine-tuned CNN backbones while incorporating a compact, regularized MLP classifier for apple leaf disease classification, as summarized in Figure 4.

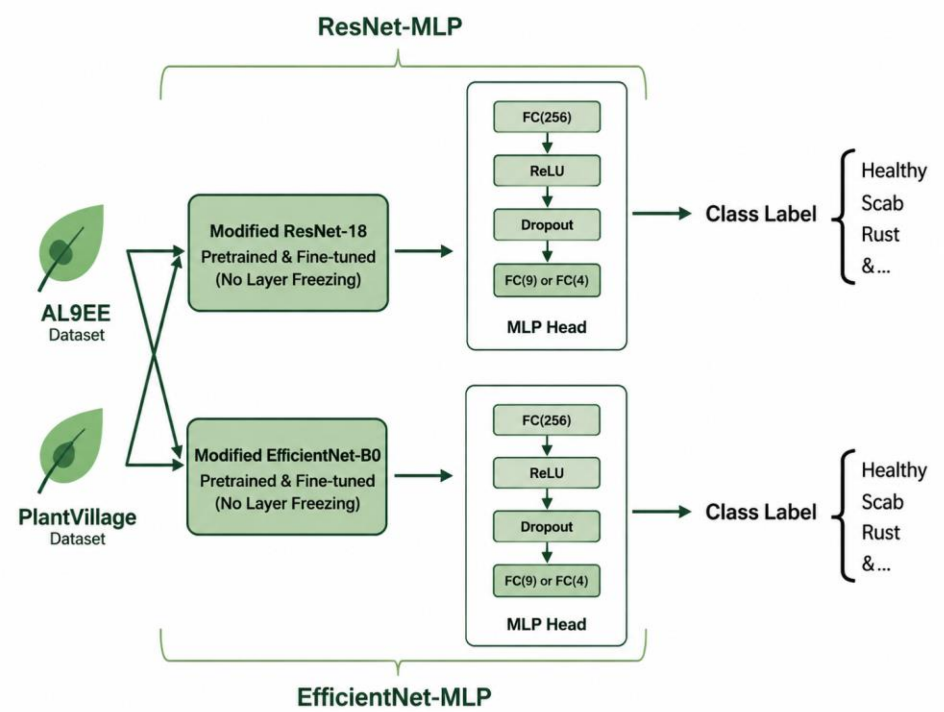

Figure 4. Architecture of the proposed lightweight CNN–MLP frameworks. The upper and lower branches illustrate the ResNet-MLP and EfficientNet-MLP pipelines, respectively. In both models, the pretrained backbone is fine-tuned end-to-end without layer freezing, and the original classifier is replaced with a customized MLP head. The MLP head consists of a 256-neuron fully connected layer, ReLU activation, dropout regularization, and a final fully connected layer matching the number of target classes.

3.2.1 ResNet-MLP (Based on ResNet-18)

This model employs ResNet-18 pretrained on ImageNet (He et al., 2016) as its backbone. The original FC layer is replaced by the proposed two-layer MLP head (see Table 3 and Figure 4).

3.2.2 EffiNet-MLP (Based on EfficientNet-B0)

Similarly, EfficientNet-B0 with pretrained ImageNet weights (Tan & Le, 2019) was adopted as the backbone. The original classifier was replaced with the proposed two-layer MLP head. The architectural modifications applied to EfficientNet-B0 are summarized in Table 4, while Figure 5 presents the overall structure of the redesigned EfficientNet-B0-MLP architecture.

Table 3. Modifications applied to ResNet-18

Component	Default (ImageNet)	Modified (ResNet-MLP)	Rationale
Output Layer	1000 neurons	9 neurons	Match the number of target classes (9 or 4 depending on the dataset)
FC Layers	512 → 1000	512 → 256 → 9	Lightweight architecture with improved non-linear representation
Dropout	None	Applied (Rate = 0.5)	Reduce overfitting during fine-tuning
Feature Extractor	Connected to original classifier	Replaced with Identity to feed the MLP head	Enable end-to-end training with a custom classifier
Efficiency	Designed for large-scale tasks	Optimized for edge/IoT scenarios	Faster inference and reduced memory usage

Table 4. Modifications applied to EfficientNet-B0

Component	Default (EfficientNet-B0)	Modified (EffiNet-MLP)	Rationale
Output Layer	1000 neurons	9 neurons	Match the number of target classes (9 or 4 depending on the dataset)
FC Layers	1280 → 1000	1280 → 256 → 9	Efficient feature transformation and non-linear separation
Dropout	None	Applied (Rate = 0.3)	Regularization suited for the lightweight backbone
Feature Extractor	MBConv feature extractor	Replaced classifier with Identity	Enable integration with the MLP head
Efficiency	Compound scaling design	Mobile-optimized classification head	Suitable for edge and IoT deployment

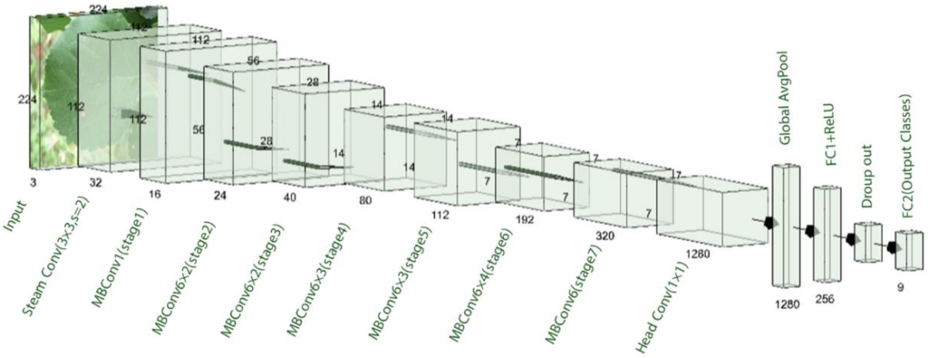

Figure 5: Illustration of overall structure of the redesigned EfficientNet-B0 model.

4. Results, Analysis, and Evaluation

This section presents the experimental results and comprehensive analysis conducted to evaluate the performance of the proposed apple leaf disease identification method. The experiments were designed to assess the effectiveness, robustness, and computational efficiency of the developed models under both controlled and visually heterogeneous dataset conditions. The section first introduces the datasets employed in this study, followed by the training configurations and evaluation metrics. Subsequently, the obtained results are analyzed in detail and compared with existing methods to demonstrate the superiority and practical applicability of the proposed approach.

4.1. Datasets

The AL9EE dataset: The AL9EE dataset (AppleLeaf9-Enhanced Edition) is an expanded version of AppleLeaf9, developed for apple leaf disease classification. It contains a total of 26,755 apple leaf images distributed across nine categories: Alternaria leaf spot, Brown spot, Frogeye leaf spot, Grey spot, Healthy, Mosaic, Powdery mildew, Rust, and Scab. The class distribution of the AL9EE dataset is reported in Table 5.

Representative samples from the AL9EE dataset are shown in Figure 6, with three example images per disease class. Compared with controlled datasets, AL9EE shows greater visual diversity across symptom appearance, illumination conditions, leaf density, background complexity, acquisition angle, and image sharpness.

In particular, the Mosaic class samples shown in Figure 7 illustrate noticeable differences in lighting intensity, shooting angle, background characteristics, and focus quality. These variations highlight the visual heterogeneity present within the dataset. Such variability reflects realistic differences in image appearance and provides a more challenging evaluation scenario for disease classification models than strictly controlled datasets.

Since the AL9EE dataset already exhibits considerable visual variability, no additional online data augmentation techniques were applied in this study." Preprocessing was limited to resizing all images to 128 × 128 pixels and converting them into tensor format.

Table 5. Class distribution in the AL9EE dataset

Class	Alternaria leaf spot	Brown spot	Frogeye leaf spot	Grey spot	Healthy	Mosaic	Powdery mildew	Rust	Scab
NO Of samples	2743	2880	3181	2642	3096	2866	1184	2756	5410

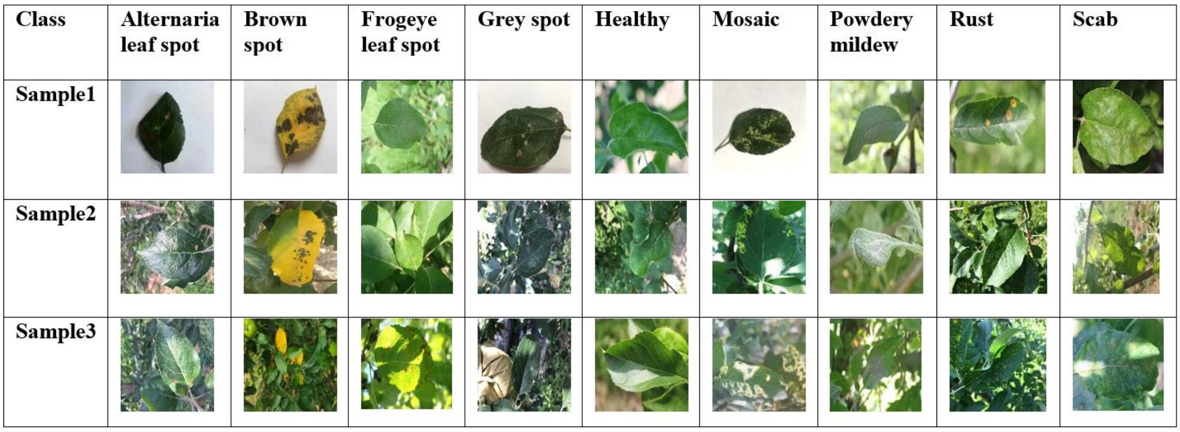

Figure 6. Representative samples from the AL9EE dataset. The samples illustrate visual diversity in symptom appearance, illumination, background complexity, acquisition angle, and image sharpness.

Figure 7. Representative samples of the Mosaic class, illustrating the visual heterogeneity in shooting angles, leaf positioning, illumination intensity, and focus quality within the AL9EE dataset.

PlantVillage Dataset: The apple subset of the PlantVillage dataset was used as a controlled benchmark. This study consisted of 3,171 images across four classes: Healthy, Rust, Scab, and Black rot. The class distribution is reported in Table 6.

Representative samples from the PlantVillage dataset are shown in Figure 8, with three images per class. In contrast to AL9EE, PlantVillage images were generally acquired under more controlled and visually uniform conditions, often with clean backgrounds and centered leaf regions. Therefore, PlantVillage provides a useful baseline for evaluating classification performance under relatively low visual variability. However, strong performance on PlantVillage should not be interpreted as direct evidence of real-world field robustness.

Table 6. Class distribution in the PlantVillage dataset

Class	Black rot	Scab	Rust	Healthy
NO of samples	621	630	275	1645

Class	Black_rot	Scab	Rust	Healthy
Sample				

Figure 8. Representative samples from the PlantVillage apple subset. Sample images are shown for each class. Compared with AL9EE, PlantVillage images generally exhibit more controlled acquisition conditions, cleaner backgrounds, and lower visual variability

4.2. Data Splitting and Training Settings

For both datasets, stratified splitting was applied to preserve class proportions across the training, validation, and test subsets. First, 20% of the images were reserved as an independent test set. The remaining 80% was further divided into training and validation subsets, with 10% of this portion used for validation.

As a result, the final split ratio was approximately 72% training, 8% validation, and 20% testing. Class labels were used as the stratification criterion in both splitting stages. To further assess statistical stability and reproducibility, additional experiments were conducted using five independent runs with different random seeds while maintaining identical data partitions, training settings, and hyperparameters. The obtained results are reported as Mean ± Standard Deviation (Mean ± SD), and the corresponding statistical summary is presented in Table 10. This protocol enables the assessment of model sensitivity to random initialization while preserving a consistent evaluation framework.

All images were resized to 128 × 128 pixels, and a batch size of 64 was used. All models were implemented and trained using the PyTorch deep learning framework. The Adam optimizer was applied with a learning rate of 0.001. For ResNet-18, a weight decay of 1×10^{-4} and an adaptive learning-rate scheduler were employed. The cross-entropy loss function was used for multi-class optimization. Dropout regularization was applied in the MLP head with rates of 0.5 for ResNet-MLP and 0.3 for EfficientNet-MLP.

The multi-seed evaluation results indicate that both models exhibit stable performance across different random initializations. In particular, EffiNet-MLP achieved lower performance variability than ResNet-MLP, demonstrating improved reproducibility and training stability.

.

4.3. Evaluation Metrics and Tools

To comprehensively assess the performance of the proposed models, standard multi-class classification metrics including Accuracy, Precision, Recall, and F1-Score were employed. These metrics provide complementary perspectives on model behavior by quantifying overall correctness, class-wise discrimination ability, and the balance between false positives and false negatives.

All evaluation metrics and confusion matrices were computed using the Scikit-learn machine learning library (Pedregosa et al., 2011), which provides reliable and widely adopted implementations for statistical performance analysis in supervised learning tasks. The mathematical definitions of the employed evaluation metrics are summarized in Table 7.

In addition to these accuracy-based metrics, computational efficiency indicators—namely Floating Point Operations (FLOPs) and the Number of Parameters—were also analyzed to evaluate the complexity and deployability of each model. The quantitative comparison of these computational complexity measures is reported in Tables 13 and 14.

These efficiency-oriented metrics complement accuracy-based evaluations by providing deeper insight into the model's computational cost, memory footprint, and inference efficiency. A smaller number of parameters indicates lower memory consumption and storage requirements, while fewer FLOPs correspond to faster inference and reduced energy demand—factors that are particularly important for edge and IoT-based agricultural monitoring systems. Therefore, combining predictive performance metrics with computational complexity analysis enables a comprehensive evaluation of both accuracy and practical deployability.

Table 7. Mathematical definitions of evaluation metrics and computational complexity parameters.

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Macro Average	$\frac{1}{N}\sum_{i=1}^{N} metric_i$
Weighted Average	$\frac{\sum_{i=1}^{N} metric_i \, w_i}{\sum_{i=1}^{N} w_i}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-Score	$2 \times \frac{Precision \times Recall}{Precision + Recall}$
Number of Parameters	$Total\ Parameters = \sum_{l=1}^{L} (w_l + b_l)$
FLOPs	$FLOPs = 2 \times \sum_{l=1}^{L} \left(C_i n^{(l)} \times K_h^{(l)} \times K_w^{(l)} \times C_{out}^{(l)} \times H_{out}^{(l)} \times W_{out}^{(l)}\right)$

4.4. Results on the AL9EE Dataset

The classification performance of the ResNet-MLP and EfficientNet-B0-MLP models on the AL9EE dataset is summarized in Tables 8 and 9, respectively. The ResNet-MLP model achieved an overall accuracy of 93%, with a macro-average F1-score of 0.92, indicating solid but slightly imbalanced performance across different classes. In contrast, the EfficientNet-B0-MLP model achieved a notably higher accuracy of 98% and a macro-average F1-score of 0.98, demonstrating more balanced and robust classification performance across the visual heterogeneity in the AL9EE dataset.

Class-level analysis, as reported in Tables 8 and 9, indicates that the ResNet-MLP model performed well for classes such as Mosaic (F1=0.98) and Brown Spot (F1=0.99), but showed weaker performance for Powdery Mildew (F1 = 0.82) and Grey Spot (F1=0.88). In contrast, the EfficientNet-B0-MLP model maintained consistently high performance across all nine classes, achieving near-perfect classification for Mosaic (F1=1.00) and Brown Spot (F1=0.99), while preserving strong precision and recall for the remaining categories.

Table 8. Performance evaluation of the ResNet-MLP model on AL9EE Dataset

Metric / Class	Alternaria Leaf Spot	Brown Spot	Frogeye Leaf Spot	Grey Spot	Healthy	Mosaic	Powdery Mildew	Rust	Scab	Accuracy / Avg
Precision	0.82	0.98	0.98	0.96	0.94	0.97	0.96	0.96	0.89	–
Recall	0.96	1.00	0.91	0.81	0.95	0.98	0.72	0.93	0.96	–
F1-Score	0.88	0.99	0.94	0.88	0.95	0.98	0.82	0.94	0.92	0.93
Macro Avg	0.94	0.91	0.92	–	–	–	–	–	–	–
Weighted Avg	0.93	0.93	0.93	–	–	–	–	–	–	–

Table 9. Performance evaluation of the EffiNet-MLP model on AL9EE Dataset

Metric / Class	Alternaria Leaf Spot	Brown Spot	Frogeye Leaf Spot	Grey Spot	Healthy	Mosaic	Powdery Mildew	Rust	Scab	Accuracy / Avg
Precision	0.94	0.99	0.98	0.99	0.99	1.00	0.94	0.96	0.97	–
Recall	0.99	0.99	0.96	0.96	0.99	0.99	0.97	0.96	0.97	–
F1-Score	0.97	0.99	0.97	0.98	0.99	1.00	0.96	0.96	0.97	0.98
Macro Avg	0.97	0.98	0.98	–	–	–	–	–	–	–
Weighted Avg	0.98	0.98	0.98	–	–	–	–	–	–	–

Table 10. Multi-seed statistical evaluation of the proposed models on the AL9EE dataset (Mean ± STD)

Model	Accuracy (%) (Mean ± SD)	F1-Score (%) (Mean ± SD)
ResNet-MLP	**93.51 ± 0.67**	**93.44 ± 0.62**
EffiNet-MLP	**98.42 ± 0.26**	**98.33 ± 0.34**

Note: Results are reported as Mean ± Standard Deviation over five independent runs using different random seeds under identical experimental settings.

To further assess the reproducibility and statistical stability of the proposed models, five independent experiments were conducted using different random seeds while maintaining identical training, validation, and testing configurations. The results, summarized in Table 10, are reported as Mean ± Standard Deviation (Mean ± SD).

The results demonstrate consistent performance across different random initializations for both architectures. ResNet-MLP achieved a mean accuracy of 93.51 ± 0.67% and a mean F1-score of 93.44 ± 0.62%, whereas EffiNet-MLP achieved a mean accuracy of 98.42 ± 0.26% and a mean F1-score of 98.33 ± 0.34%.

In addition to achieving higher average performance, EffiNet-MLP exhibited substantially lower standard deviations than ResNet-MLP, indicating greater training stability and suggesting that its reported performance is reproducible rather than attributable to a favorable random seed or a single training run.

The confusion patterns illustrated in Figures 9 and 10 provide further evidence of the class-wise performance differences between the two models. In the ResNet-MLP model, Powdery Mildew shows relatively high precision but considerably lower recall, indicating that the model is conservative in assigning this label and misses a substantial number of true Powdery Mildew samples. According to Figure 9, of the 237 test samples belonging to Powdery Mildew, only 171 were correctly classified, while 44 were incorrectly predicted as Scab and 17 as Healthy. This confirms that the lower F1-score of Powdery Mildew is mainly driven by false negatives rather than false positives. For ResNet-MLP, the number of false negatives for Powdery Mildew is calculated as:

$$FN_{PM}^{ResNet-MLP} = 237 - 171 = 66$$

Accordingly, the recall for Powdery Mildew is:

$$Recall_{PM}^{ResNet-MLP} = \frac{171}{171 + 66} = \frac{171}{237} \approx 0.7215$$

In contrast, the EfficientNet-B0-MLP model substantially reduced this confusion. As shown in Figure 10, the number of Powdery Mildew samples misclassified as Scab decreased from 44 in ResNet-MLP to only 3 in EfficientNet-B0-MLP. In addition, the number of correctly classified Powdery Mildew samples increased from 171 to 231. Therefore, the recall of Powdery Mildew in EfficientNet-B0-MLP is:

$$Recall_{PM}^{EffiNet-MLP} = \frac{231}{237} \approx 0.9746$$

This improvement from approximately 72.15% to 97.46% demonstrates that the lower performance of Powdery Mildew in

ResNet-MLP is not solely caused by inherent visual similarity among disease classes. Rather, it is more likely due to the combined influence of class imbalance and the limited feature representation capability of the ResNet-MLP model for minority-class patterns. The EfficientNet-B0 backbone, through its compound scaling strategy, MBConv blocks, and squeeze-and-excitation mechanisms, appears to extract more discriminative features and therefore better compensates for the limited number of Powdery Mildew samples. Consequently, EfficientNet-B0-MLP exhibits more balanced class-wise behavior across the visual variability in AL9EE and improves learning stability and generalization in the evaluated dataset conditions.

The training and validation trends shown in Figure 11 reveal distinct convergence behaviors. The ResNet-MLP model demonstrated slower convergence and mild fluctuations in validation loss, consistent with its lower overall accuracy (93%) and imbalance in class-level recall. By contrast, the EffiNet-MLP model achieved faster and more stable convergence, reaching an accuracy of 98% with minimal overfitting. These results emphasize that the architectural optimizations—particularly the inclusion of the two-layer MLP head and the lightweight EfficientNet backbone—improve learning stability and classification performance within the evaluated dataset conditions.

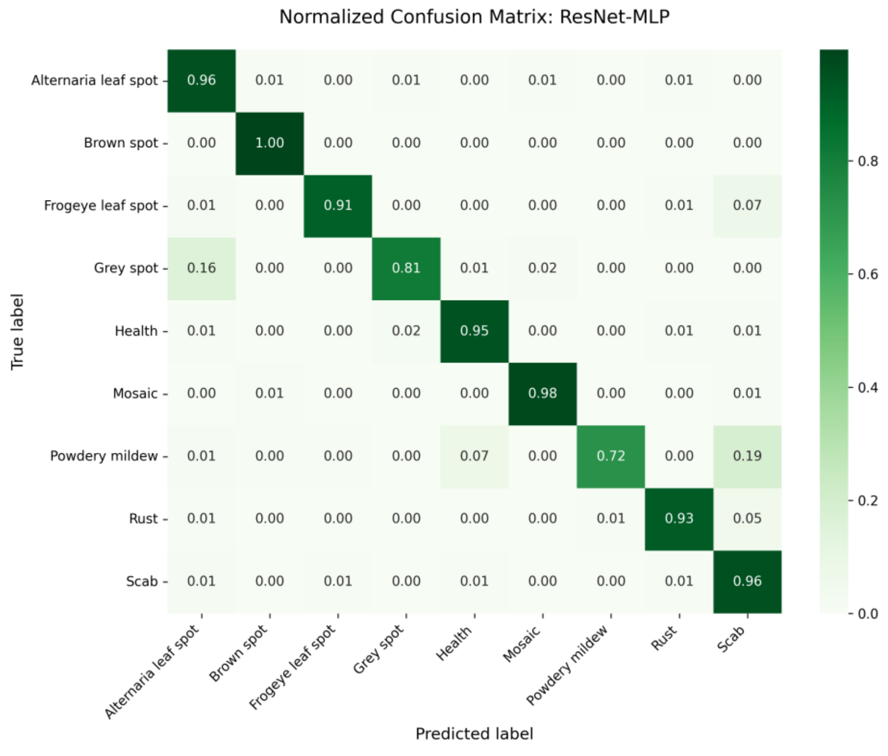

Figure 9. Confusion matrix of the Normalized ResNet-MLP model on AL9EE Dataset

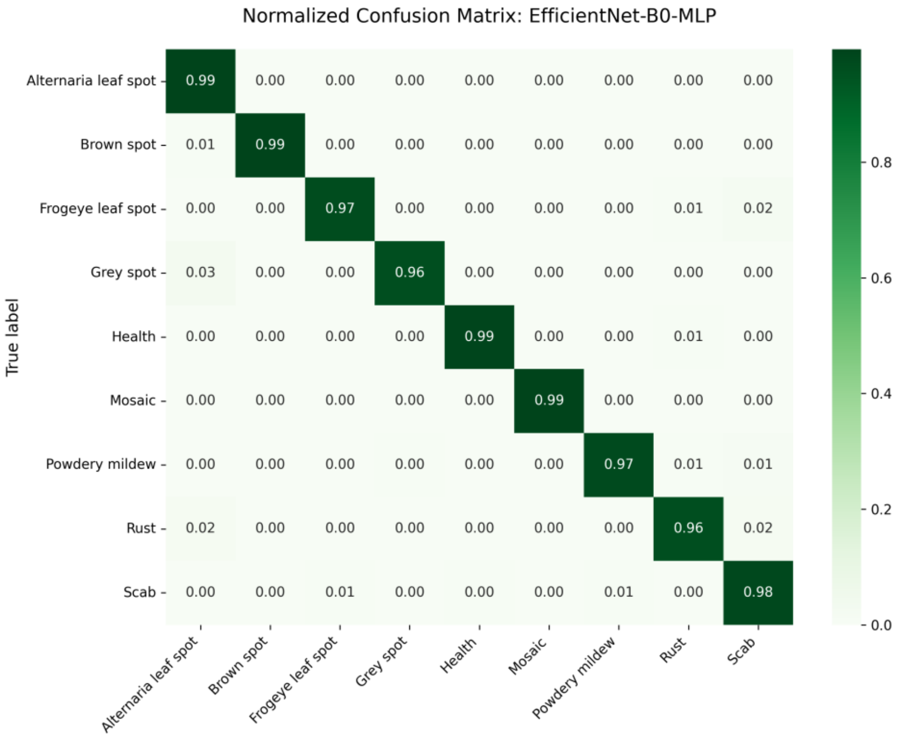

Figure 10. Confusion matrix of the Normalized EffiNet-MLP model on AL9EE Dataset

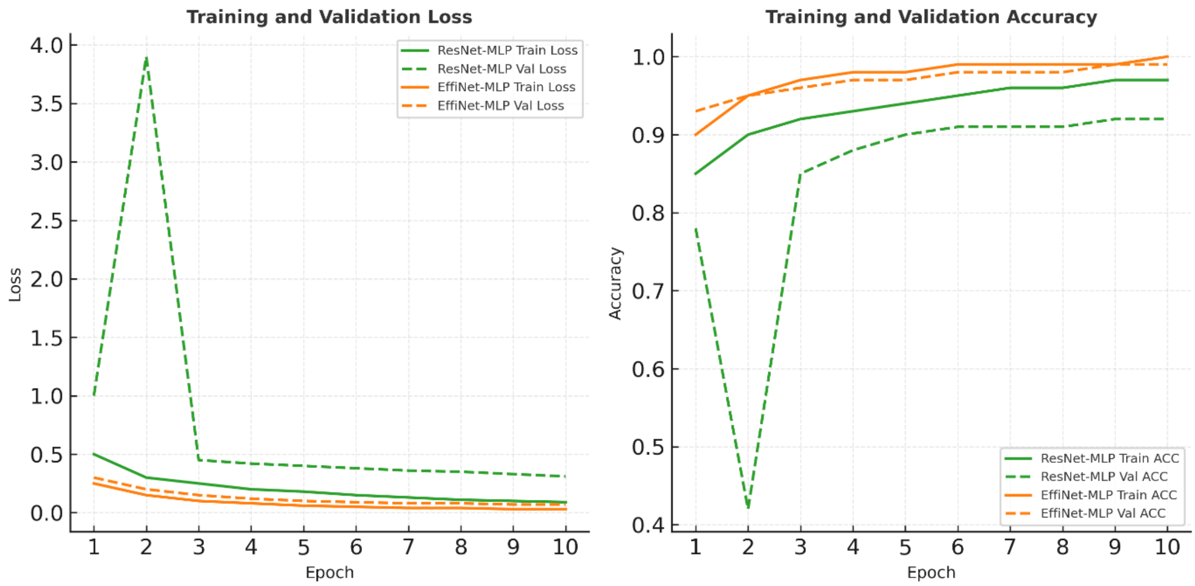

Figure 11. Training and validation performance of proposed modified ResNet-18 and proposed modified EfficientNet-B0 models on AL9EE Dataset

4.5. Results on the PlantVillage Dataset

The proposed deep learning models were initially evaluated on the PlantVillage dataset (Birant, 2024) to establish baseline performance under controlled laboratory imaging conditions. As presented in Table 11, the ResNet-MLP model achieved very high classification performance on the PlantVillage dataset (Birant, 2024), with an overall accuracy of approximately 100% and precision, recall, and F1-score values close to 1.00 across the four categories: Healthy, Rust, Scab, and Black_rot.

Table 11. Performance evaluation of the ResNet-MLP model on PlantVillage Dataset

Metric / Class	Healthy	Rust	Scab	Black_rot	Macro Avg	Weighted Avg
Precision	1.00	1.00	1.00	1.00	1.00	1.00
Recall	1.00	1.00	0.99	1.00	1.00	1.00
F1-Score	1.00	1.00	1.00	1.00	1.00	1.00

These results indicate highly effective class separation under controlled imaging conditions, although the values should not be interpreted as evidence of statistically perfect classification. Similarly, as summarized in Table 12, the EffiNet-MLP model achieved 99% accuracy, maintaining consistent results across all metrics (Macro Avg = 0.99).

While performance remained high, slightly lower recall values were observed for Scab (0.96) and Rust (0.98), suggesting limited confusion between these visually similar classes.

Table 12. Performance evaluation of the EffiNet-MLP model on PlantVillage Dataset

Metric / Class	Healthy	Rust	Scab	Black_rot	Macro Avg	Weighted Avg
Precision	0.99	0.98	0.99	0.99	0.99	0.99
Recall	1.00	0.98	0.96	1.00	0.99	0.99
F1-Score	1.00	0.98	0.98	1.00	0.99	0.99

Overall, these results show that both models exhibit strong discriminative capability and stable convergence when trained under well-controlled conditions. The high performance on PlantVillage is consistent with its controlled imaging conditions and relatively limited visual variability. In contrast, AL9EE provides a more challenging augmentation-enhanced setting with nine classes and greater visual diversity. Therefore, differences between the two datasets should be interpreted descriptively, as no formal statistical significance test or controlled cross-dataset evaluation was conducted to determine whether the observed difference is statistically meaningful.

The accuracy and loss curves shown in Figure 12 demonstrate that both models converged rapidly and stably, without signs of overfitting. Such minimal deviations indicate that the proposed EfficientNet-B0-MLP model maintains stable classification behavior under the controlled imaging conditions represented in the PlantVillage dataset.

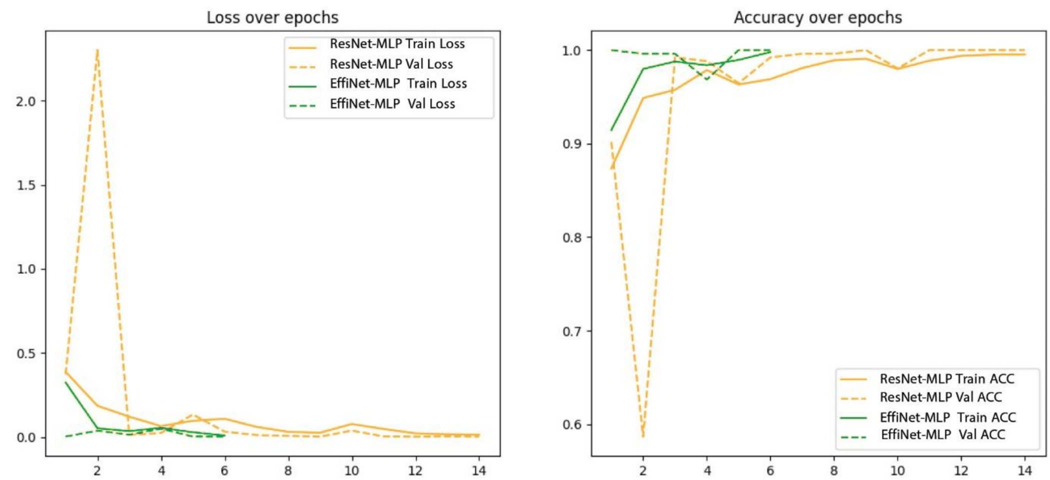

Figure 12. Accuracy and loss curves of PlantVillage: ResNet-MLP and PlantVillage: EffiNet-MLP in classifying apple leaf diseases

The ResNet-MLP curve exhibits a smooth rise in accuracy reaching 100%, while the EffiNet-MLP curve converges slightly earlier, indicating faster optimization and superior computational efficiency. The confusion matrices illustrated in Figure 13(a) and Figure 13(b) further validate these observations. For ResNet-MLP, all samples were correctly classified across all classes, confirming perfect agreement between predictions and ground truth.

For EffiNet-MLP, only a negligible number of samples from the *Scab* and *Rust* classes were misclassified, consistent with the slightly reduced recall scores reported in Table 12.

a) b)

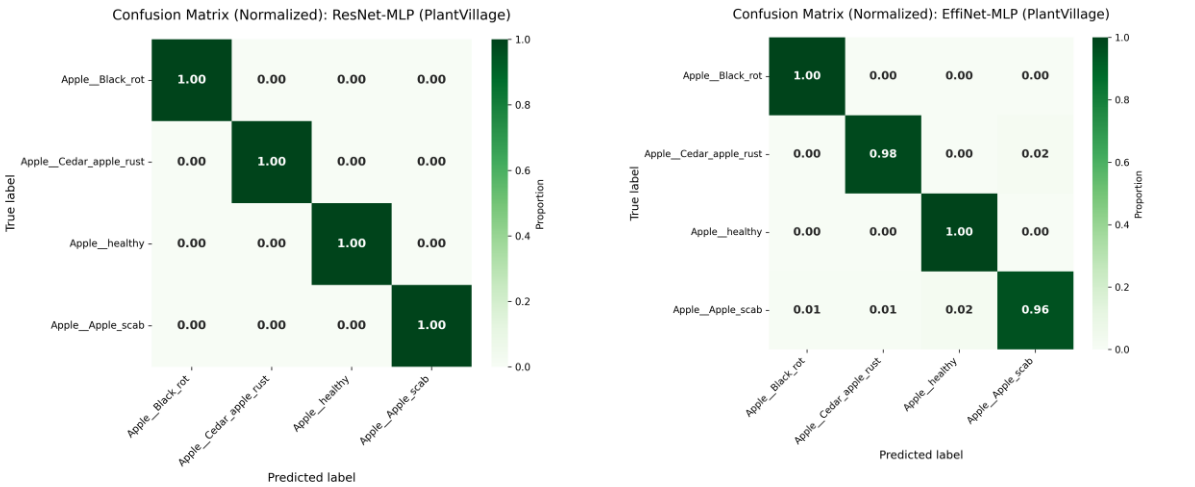

Figure 13.Normalized Confusion matrices for apple leaf disease classification: (a) PlantVillage: ResNet-MLP model; (b) PlantVillage: EffiNet-MLP model

4.6. Comparative Analysis of Results with Previous Studies

Table 13 presents a descriptive comparison between the proposed models and selected previous studies in terms of classification accuracy, parameter size, computational complexity, and reported evaluation conditions. It is important to emphasize that this comparison does not represent a strictly controlled head-to-head benchmark, since the compared studies differ in terms of datasets, class definitions, image acquisition conditions, preprocessing strategies, training protocols, and evaluation procedures.

Most previous works in plant disease recognition have been evaluated on the PlantVillage dataset, which is widely used under relatively controlled imaging conditions (Hughes & Salathé, 2015). In contrast, the present study evaluates the proposed models not only on PlantVillage but also on the augmentation-enhanced AL9EE dataset, which contains nine apple leaf disease/health classes and provides greater dataset-level visual variability in disease categories, symptom appearances, and generated image styles. (AL9EE, 2024).

Although some recent studies have reported results on AppleLeaf9, they have mainly focused on predictive performance metrics such as accuracy, precision, recall, and F1-score. To the best of our knowledge, prior AL9EE-based studies have not consistently reported computational complexity metrics, such as trainable parameter count, GFLOPs, model size, or inference latency. Therefore, a fully controlled, directly comparable efficiency-oriented benchmark for AL9EE is currently unavailable in the literature.

A fair head-to-head comparison would require re-implementing all baseline models under identical training, validation, and testing protocols on AL9EE and computing their parameter counts and FLOPs under consistent input-size and hardware settings. Such an extensive re-evaluation was beyond the scope of this work. Consequently, the comparisons presented in Table 14 should be interpreted as literature-supported descriptive comparisons rather than direct experimental evidence of favorable performance.

Earlier studies, such as the ResNet-based transfer learning approach by Li and Rai (2020) and the lightweight CNN model proposed by Fu et al. (2022), reported accuracies of approximately 97% on PlantVillage. More recent approaches, including AppleLeafNet (Ali et al., 2024) and a fine-tuned EfficientNet-B0 model (Ali et al., 2025), reported accuracies exceeding 99% in their respective experimental settings. However, these results were obtained using datasets with different class definitions, imaging conditions, and evaluation protocols. Therefore, they should not be interpreted as directly comparable to the AL9EE results reported in the present study.

On the AL9EE dataset, the proposed ResNet-MLP and EfficientNet-B0-MLP models achieved approximately 93% and 98% accuracy, respectively. The performance difference between PlantVillage and AL9EE reflects the increased visual heterogeneity and greater number of classes in AL9EE. Therefore, the slightly higher accuracy observed on PlantVillage should be considered a descriptive observation rather than evidence of statistical significance, particularly because no formal statistical testing was conducted.

Table 13. Descriptive literature-based comparison of selected apple leaf disease classification studies and the proposed models

Study / Model	Year	Dataset / Classes	Method	Accuracy	Model Size	FLOPs	Evaluation condition
Li & Rai (ResNet-18)	2020	PlantVillage (~4 classes)	Transfer Learning (ResNet-18)	97.0%	~11.7 M	~1.8G	No real-world test
Fu et al. (Lightweight CNN)	2022	PlantVillage (~4 classes)	Lightweight CNN	97.36%	~1.3M	2.55G	No real-world test
AppleLeafNet (Two-stage)	2024	PlantVillage (5 classes)	Two-stage CNN	98.6%	~1.3M	~0.31G	No real-world test
EfficientNet-B0 (fine-tuned)	2025	PlantVillage (~4 classes)	Transfer Learning	99.78%	~5.3M	0.39G	No real-world test
PlantVillage:*ResNet-MLP* (Proposed)	2026	PlantVillage (4 classes)	Modified ResNet	100%	~11.3 M	~1.82G	Controlled dataset evaluation
PlantVillage:*EffiNet-MLP* (Proposed)	2026	PlantVillage (4 classes)	Modified EfficientNet	99%	~4.3M	0.38G	Controlled dataset evaluation
AL9EE:ResNet-MLP (Proposed)	2026	AL9EE (9 classes)	Modified ResNet	93%	~11.7 M	~1.8G	Evaluated on augmentation-enhanced AL9EE
AL9EE:EffiNet-MLP (Proposed)	2026	AL9EE (9 classes)	Modified EfficientNet	98%	~5.3M	~0.39G	Accurate and stable on augmentation-enhanced AL9EE

Note: The results are not directly comparable because the listed studies differ in datasets, class definitions, preprocessing, training protocols, and evaluation settings. The AL9EE results reported for the proposed models were obtained in this study, while the other values are literature-reported.

Table 14. Experimental AL9EE results of the proposed models and literature-supported architectural complexity comparison of non-implemented candidate backbones.

Architecture	Status in This Study	No. of Parameters	Approx. FLOPs	AL9EE Accuracy /Evaluation Status	Advantages	Limitations	Reference
ResNet-MLP	Implemented and evaluated	11.7M	~1.8 GFLOPs	~93% AL9EE Acc.	Lightweight, simple architecture, fast training, easy implementation	Lower accuracy than EffiNet-MLP; less stable validation and slower convergence	This study (2026)
EffiNet-MLP (B0)	Implemented and evaluated	5.3M	~0.39–0.4 GFLOPs	~98% AL9EE Acc.	Best observed accuracy–complexity trade-off; lightweight design; promising for future Edge/IoT deployment	Lower representational capacity than larger EfficientNet variants	This study (2026)
EfficientNet-B4	Not evaluated on AL9EE in this study; literature-based comparison	~19M	~4.2 GFLOPs	Not evaluated on AL9EE	High accuracy–efficiency trade-off; significantly fewer parameters and FLOPs than comparable CNNs; strong transfer learning performance	Higher computational cost than smaller EfficientNet variants; requires GPU/accelerator for efficient training and inference	Tan and Le (2019)
EfficientNet-B5	Not evaluated on AL9EE in this study; literature-based comparison	~30M	~9.9 GFLOPs	Not evaluated on AL9EE	Higher representational capacity and accuracy than smaller EfficientNet models; strong performance on ImageNet and transfer tasks	Increased computational and memory requirements; less suitable for low-resource or real-time edge deployment	Tan and Le (2019)
Inception-ResNet-v2	Not evaluated on AL9EE in this study; literature-based comparison	~55–56M	~13 – 14GFLOPs	Not evaluated on AL9EE	Combines multi-scale Inception modules with residual connections; faster convergence; high ImageNet accuracy (Top-5 ≈ 4.1%)	High computational and memory cost; complex architecture; unsuitable for lightweight deployment	Szegedy et al. (2017)

Table 14 further provides the architectural rationale for selecting ResNet-18 and EfficientNet-B0 as backbone networks. ResNet-18 is based on residual learning (He et al., 2016), whereas EfficientNet-B0 follows the compound scaling strategy introduced by Tan and Le (2019). These architectures were selected because they offer different trade-offs between computational complexity and representational capacity.

Among the experimentally evaluated models, EfficientNet-B0-MLP achieved the most favorable balance between accuracy and efficiency. It obtained approximately 98% accuracy on AL9EE with only 5.3 million parameters and approximately 0.39–0.40 GFLOPs, making it a promising candidate for future Edge/IoT deployment in smart agriculture scenarios.

In comparison, ResNet-MLP required approximately 11.7 million parameters and 1.8 GFLOPs, achieving approximately 93% accuracy on AL9EE. Although this model is relatively simple, lightweight, and straightforward to implement, it demonstrated lower classification performance and less stable convergence compared with EfficientNet-B0-MLP.

Larger EfficientNet variants, such as EfficientNet-B4 and EfficientNet-B5, as well as Inception-ResNet-v2, provide stronger representational capacity according to their original architectural studies. However, their substantially higher parameter counts and computational demands reduce their practicality for lightweight real-time agricultural deployment. Since these architectures were not implemented or evaluated on AL9EE in the present study, they are included in Table 14 solely as literature-supported architectural references, not as experimental baselines.

Overall, the comparative analysis supports selecting EfficientNet-B0-MLP as the most appropriate architecture for this study's objectives. Within the AL9EE evaluation setting, it achieved a favorable trade-off among high observed classification accuracy, compact model size, low computational complexity, and practical suitability for Edge/IoT-based smart agriculture deployment. While larger architectures may offer higher theoretical representational capacity, their increased computational burden limits their applicability in resource-constrained agricultural environments. Consequently, the proposed EfficientNet-B0-MLP model provides a balanced and practically deployable solution for apple leaf disease classification under heterogeneous visual conditions. It should be noted that only ResNet-MLP and EffiNet-MLP were implemented and experimentally evaluated on the AL9EE dataset in this study. The remaining architectures in Table 14 are included only as literature-supported architectural references to contextualize the computational complexity of the proposed models.

5. Conclusion and Future Work

In this study, a lightweight CNN-MLP framework was developed and empirically evaluated for apple leaf disease classification on controlled and augmentation-enhanced visually diverse datasets. Two modified deep learning pipelines based on ResNet-18 and EfficientNet-B0 backbones were investigated. The experimental results demonstrated that the proposed EffiNet-MLP model achieved the best overall trade-off between classification performance and computational efficiency, achieving 98% accuracy with approximately 5.3 million parameters and 0.39 GFLOPs. Compared with ResNet-MLP, which achieved 93% accuracy with a simpler architectural configuration, EffiNet-MLP provided superior classification performance while maintaining a lightweight computational profile.

Due to its relatively low parameter count and computational cost, the proposed EfficientNet-based model appears promising for edge-oriented deployment. However, it should be emphasized that real-device deployment experiments were not conducted in this study. Therefore, practical deployment suitability in terms of inference latency, memory consumption, model size, throughput, and energy efficiency remains to be validated on actual edge platforms. Accordingly, the current results should be interpreted as evidence of computational efficiency at the model-design level, rather than as a confirmed demonstration of real-time performance on low-power hardware.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this work. This statement is approved by all authors.

Data Availability

The datasets used in this study are publicly available. The PlantVillage dataset is available on Figshare (Birant, 2024): https://doi.org/10.6084/m9.figshare.26096800.v1. The AL9EE dataset is available on Figshare (Zhu, 2023): https://doi.org/10.6084/m9.figshare.23606010.v1. All experimental results, model configurations, and evaluation metrics are fully reported within the article. No additional data were created.

Author Contributions

The contributions of the authors to this research are as follows:

- First Author (Forough Ehsanfar): Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft.
- Second Author (Fatemeh Moayedi: Corresponding Author): Supervision, Validation, Writing - review & editing, Project administration.

Ethical Approval

The authors confirm that all ethical principles were observed in conducting and publishing this research, and this has been approved by all authors.

References:

Agricultural Research Center of Iran. (2022). *Survey on plant diseases in apple orchards*. (In Persian).

Ali, H., Shifa, N., Benlamri, R., Farooque, A. A., & Yaqub, R. (2025). A fine tuned EfficientNet-B0 convolutional neural network for accurate and efficient classification of apple leaf diseases. *Scientific Reports, 15*(1), 25732. https://doi.org/10.1038/s41598-025-04479-2

Ali, M. U., Khalid, M., Farrash, M., Lahza, H. F. M., Zafar, A., & Kim, S. H. (2024). AppleLeafNet: A lightweight and efficient deep learning framework for diagnosing apple leaf diseases. *Frontiers in Plant Science, 15*, 1502314. https://doi.org/10.3389/fpls.2024.1502314

Anam, S. (2020). Segmentation of leaf spot disease in apple plants using particle swarm optimization and K-means algorithm. *Journal of Physics: Conference Series, 1562*, 012011. https://doi.org/10.1088/1742-6596/1562/1/012011

Ashmafee, M. H., Ahmed, T., Ahmed, S., Hasan, M. B., Jahan, M. N., & Rahman, A. A. (2023, February). An efficient transfer learning-based approach for apple leaf disease classification. In *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1-6). IEEE. https://doi.org/10.1109/ECCE57851.2023.10101542

Birant, D. (2024). *PlantVillage dataset*. Figshare. https://doi.org/10.6084/m9.figshare.26096800.v1

Bracino, A. A., Concepcion, R. S., Bedruz, R. A. R., Dadios, E. P., & Vicerra, R. R. P. (2020, December). Development of a hybrid machine learning model for apple (Malus domestica) health detection and disease classification. In *2020 IEEE 12th international conference on humanoid, nanotechnology, information technology, communication and control, environment, and management (HNICEM)* (pp. 1-6). IEEE. https://doi.org/10.1109/HNICEM51456.2020.9400139

Chakraborty, S., Paul, S., & Rahat-uz-Zaman, M. (2021). Prediction of apple leaf diseases using multiclass support vector machine. In *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (pp. 1–6). https://doi.org/10.1109/ICREST51555.2021.9331132

Food and Agriculture Organization of the United Nations. (2025). *FAOSTAT: Crops and livestock products*. Retrieved August 30, 2025, from https://www.fao.org/faostat/en/#data

Fu, J., Li, S., Sun, Y., Mu, Y., Hu, T., & Gong, H. (2022). *Lightweight CNN for apple leaf disease recognition*. Frontiers in Plant Science, 13, 831219. https://doi.org/10.3389/fpls.2022.831219

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90

Howard, J., & Gugger, S. (2020). *Deep learning for coders with fastai & PyTorch*. O'Reilly Media.

Hughes, D., & Salathé, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*.

Khan, M. A., Lali, M. I. U., Sharif, M., Javed, K., Aurangzeb, K., Haider, S. I., ... & Akram, T. (2019). An optimized method for segmentation and classification of apple diseases based on strong correlation and genetic algorithm based feature selection. *IEEE Access*, *7*, 46261-46277.https://doi.org/10.1109/ACCESS.2019.2908040

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*, 84–90. https://doi.org/10.1145/3065386

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.Sai, A. M., & Patil, N. (2022). Comparative analysis of machine learning algorithms for disease detection in apple leaves. In *2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)* (pp. 239–244). https://doi.org/10.1109/DISCOVER55800.2022.9974840

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)* (pp. 6105–6114).

Zhu, F. (2023). *AppleLeaf9-Enhanced Edition (AL9EE) dataset*. Figshare. https://doi.org/10.6084/m9.figshare.23606010.v1