IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 2, April 2026, pp. 1719~1732
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i2.pp1719-1732

Comparative deep learning study for downy mildew detection in vegetables

Supreetha Shivaraj, Manjula Sunkadakatte Haladappa
Department of Computer Science Engineering, University of Visvesvaraya College of Engineering (Bangalore University), Bengaluru, India

Article Info

Article history:

Received Nov 11, 2024
Revised Jan 22, 2026
Accepted Feb 6, 2026

Keywords:

DenseNet201
Downy mildew
Explainable artificial intelligence
Grad-CAM
Imbalanced dataset
MobileNetV2
VGG19

ABSTRACT

Several vegetable crops are affected by downy mildew, a major foliar disease resulting in notable reductions in yield. For sustainable agriculture and disease prevention, early and precise detection is crucial. To be able to detect downy mildew in five varied vegetables—bitter gourd, bottle gourd, cauliflower, cucumber (Rashid), and cucumber (Sultana)—this study evaluates three deep learning architectures: VGG19, DenseNet201, and MobileNetV2. This work focuses on imbalanced datasets collected from several sources, in opposition to prior work that depended on balanced laboratory datasets. Accuracy, precision, recall, and F1-score metrics were used to evaluate the models shortly after they were trained using transfer learning, data augmentation, and 5-fold cross-validation. Model focus regions were assessed by using gradient-weighted class activation mapping (Grad-CAM) visualizations, and statistical reliability was assessed based on paired *t*-tests and Wilcoxon signed-rank tests. By achieving mean accuracies above 98% and statistically significant results ($p < 0.05$) on cucumber datasets, DenseNet201 accomplished superior performance. Despite attaining slightly lower accuracy (89.6–100%), MobileNetV2 offered the smallest model size (12.9 MB) and minimum inference time (85 ms). The proposed approach demonstrated a transparent, generalizable, and computationally efficient deep learning pipeline for precision agriculture's real-time downy mildew detection.

Corresponding Author:

Supreetha Shivaraj
Department of Computer Science Engineering
University of Visvesvaraya College of Engineering (Bangalore University)
Bengaluru, 560001, Karnataka, India
Email: supreetha.s3191@gmail.com

1. INTRODUCTION

Plant diseases are a major threat to agricultural sustainability and global food security, and early and accurate detection is vital for yield protection. Among these, one of the most destructive foliar diseases that affects a variety of crops, including cruciferous vegetables and cucurbits, is downy mildew. In humid conditions, the infection spreads rapidly, leading to necrotic lesions, chlorosis, and significant yield loss if not addressed [1]–[3].

Disease diagnosis has traditionally relied on visual symptom recognition, laboratory assays, and manual inspection by trained agronomists. However, these methods are subjective, time-consuming, and often impractical for large-scale or resource-constrained settings. Recent developments in deep learning and computer vision [4]–[6] have significantly revolutionized plant pathology by allowing the automatic detection of plant diseases directly from images. Shallow convolutional neural networks (CNNs) were

employed in early models for disease detection [7]–[9], but these techniques suffered from overfitting and low generalization on real-world datasets. However, architectures such as VGG19, DenseNet201, and MobileNetV2 emerged later, providing robust feature extraction, high classification accuracy, and efficiency, making it suitable to be deployed on mobile and internet of things (IoT)-based systems [10]–[13]. Interpretability is another key challenge. Adoption in agriculture requires transparency (e.g., "which parts of the leaf led to the decision?"), especially when deploying models for critical crop disease monitoring. High classification accuracy alone is not sufficient. As a goal to ensure interpretability and trust, explainable artificial intelligence (XAI) techniques such as gradient-weighted class activation mapping (Grad-CAM) have been adopted increasingly to visualize significant regions influencing model predictions [14]–[16]. However, IoT of prior studies focused primarily on accuracy-based evaluations using balanced laboratory datasets, which often fail to generalize effectively under real-time, unbalanced datasets [17]. Building on the gaps identified in prior research, the present study aims to pursue four primary objectives:

- In five vegetable species (bitter gourd, bottle gourd, cauliflower, cucumber (Rashid), and cucumber (Sultana), we evaluate three prominent deep-learning architectures (VGG19, MobileNetV2, and DenseNet201) on laboratory and real-world, imbalanced datasets collected from diverse sources.
- To quantify model generalization under imbalanced conditions, we apply independent test-set evaluation and rigorous 5-fold cross-validation.
- Statistical significance testing (paired *t*-tests and [18]–[20] Wilcoxon tests) is conducted to reliably compare model performance, and we integrate explainability via Grad-CAM to visualize how models attend to disease lesions.
- Inference time and model size are assessed, which are critical for field-level or mobile use, and we give importance to both deployment readiness and predictive performance.

This study provides a reliable, transparent, and reproducible pipeline for downy mildew detection in precision agriculture by applying XAI, real-world data splits, transfer-learning, and statistical validation. For both practitioners and researchers, results contribute to choosing a model and deployment strategy. This pipeline supports practical implementation and informed decision-making in precision agriculture.

2. METHOD

To ensure accuracy and interpretability, the proposed framework for downy mildew detection across multiple vegetable species integrates XAI, cross-validation, statistical validation, and transfer learning. The complete pipeline, from data gathering to the visual interpretation of model attention via Grad-CAM, is provided in Figure 1. Each component of this pipeline was carefully designed to improve generalization on imbalanced datasets while maintaining computational efficiency for practical field deployment.

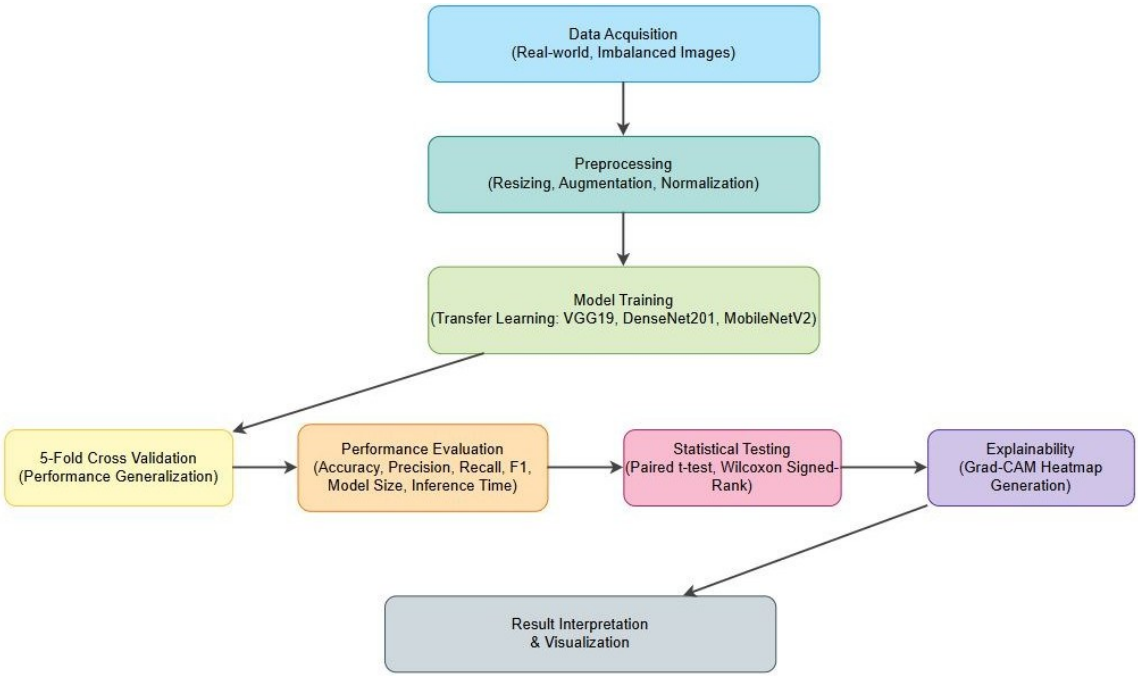


Figure 1. Architecture for downy mildew detection and interpretation using deep learning models

2.1. Data acquisition

Five vegetable datasets comprised of healthy and downy mildew-infected leaves were utilized in the study. Data were gathered from available sources, specifically vegetables like bottle gourds, cucumbers, cauliflower, and bitter gourds. The different sources from where the data is collected are listed:

- Rashid *et al.* [21]*:* cucumber leaf images from the plant leaf freshness and disease detection dataset [21], consisting of real-time data.
- Sultana *et al*. [22]: cucumber leaf images from the cucumber disease recognition dataset [22], containing images captured under various environmental conditions.
- Rashid *et al.* [21] laboratory data: bitter gourd, bottle gourd, and cauliflower leaf images collected in laboratory settings against black backgrounds. The dataset is summarized in Table 1.

Table 1. An overview of the dataset

	Original dataset	
Vegetable type	Downy mildew	Healthy
Bitter guard	570	551
Bottle guard	684	518
Cauliflower	512	526
Cucumber (Rashid)	564	527
Cucumber (Sultana)	160	160

2.2. Image preprocessing

To ensure consistency and reproducibility across experiments, a standardized pipeline was implemented to preprocess all images. To achieve the input requirements of VGG19, DenseNet201, and MobileNetV2, raw RGB images collected from heterogeneous real-field conditions were resized to 224×224 pixels. ImageNet mean and standard deviation statistics [0.485, 0.456, 0.406] [0.485, 0.456, 0.406] and [0.229,0.224, 0.225] [0.229, 0.224, 0.225] [0.229, 0.224, 0.225] increased transfer learning stability. To prevent information leakage, class proportions are maintained across training, validation, and test sets by splitting the dataset using stratified sampling prior to augmentation. The validation and test sets only comprised original, unaltered images, while augmentation was applied only on the training set using Keras generators in order to enhance robustness and address class imbalance. The integrity of performance evaluation is maintained by this strict separation that ensures that no augmented variants of training samples appear in validation or test folds. Every vegetable and cross-validation fold was subjected to the same protocol. Random rotations (up to 40°), horizontal flipping, width and height shifts (±20%), shearing, and zooming augmentation operations were carried out.

During training, augmented samples were generated dynamically in order to avoid duplicate data on disk and allow each image to appear over multiple stochastic variations across epochs. To avoid cross-crop contamination, each vegetable category—bitter gourd, bottle gourd, cauliflower, cucumber (Rashid), and cucumber (Sultana)—was processed individually. This preprocessing technique enables generalization to realfield variations, including variations in illumination, background complexity, and symptom diversity, while facilitating objective evaluation. Finally, to assess model generalization across different splits, a 5-fold cross-validation strategy was employed. Collectively, these preprocessing steps ensure that the trained models will be resilient to the challenges that exist in real-world agricultural imagery, such as noise, lighting variation, and data imbalance.

2.3. Data splitting

Prior to any data augmentation, dataset splitting was performed to ensure an unbiased evaluation and prevent data leakage. The dataset was divided into test, validation, and training sets with a 70% training ratio, 15% testing, and 15% validation. This balanced splitting ensured that each subset maintained the class distribution, thereby enabling reliable model evaluation. Table 2 provides details of the dataset after splitting.

2.4. Model training and evaluation

After preprocessing, three deep learning architectures—VGG19, DenseNet201 [23], [24] and MobileNetV2 was used to train and examine the prepared datasets for five vegetables: bitter gourd, bottle gourd, cauliflower, cucumber (Rashid), and cucumber (Sultana). Due to its deep yet stable convolutional structure, which has been widely validated in agricultural image analysis, VGG19 serves as a strong classical baseline. DenseNet201 was chosen due to its densely connected design that promotes feature reuse and efficient gradient flow, enabling the model to detect subtle and fine-grained symptoms of downy mildew in various vegetable species. In order to assess a computationally efficient alternative that can provide fast inference and low memory usage, making it suitable for real-time field applications, MobileNetV2 [25]–[27],

a lightweight architecture designed for edge deployment, was included. To leverage transfer learning for faster convergence and improved performance on limited agricultural data, all models were initialized with ImageNet-pretrained weights. A customized classification head consisting of global average pooling, dropout, and a dense SoftMax output layer with two neurons representing the classes—healthy and downy mildew—has replaced each model's final fully connected layers. Class separability was optimized using the categorical cross-entropy loss function, and the models were employed using the Adam optimizer with a learning rate of 1e-4.

To ensure reproducibility and transparency, all experimental configurations were explicitly defined. To prevent overfitting, each model was trained for 10 epochs with a batch size of 8, using early stopping with a patience of 3 epochs. To retain the best-performing weights based on validation accuracy, model checkpointing was employed. Reproducibility was ensured by fixing random seeds across NumPy, TensorFlow, and Python. All experiments were executed on Google Colab with NVIDIA GPU acceleration. In order to provide consistent and repeatable experimental outcomes, TensorFlow's deterministic behavior was additionally enabled where applicable.

Table 2. An overview of the dataset after data splitting

Dataset split summary				
Vegetable type	Class	Train	Val	Test
Bitter gourd	Healthy	385	82	84
	Downy mildew	399	85	86
Bottle gourd	Healthy	362	77	79
	Downy mildew	478	102	104
Cauliflower	Healthy	368	78	80
	Downy mildew	358	76	78
Cucumber (Rashid)	Healthy	368	79	80
	Downy mildew	394	84	86
Cucumber (Sultana)	Healthy	112	24	24
	Downy mildew	112	24	24

2.5. 5-fold cross-validation

5-fold cross-validation was implemented on each vegetable in the dataset to ensure robustness. To reduce the bias caused by a single train-test split, each model was trained and validated for 5-folds, and the results were aggregated. Accuracy, precision, recall, and F1-score were averaged over folds in order to provide a thorough evaluation of predictive performance.

2.6. Performance evaluation

To ensure reliability and adaptability across imbalanced data sets, the model's performance was assessed using both independent test-set evaluation and 5-fold cross-validation. The accuracy, precision, recall, and F1-score were obtained by comparing the network's predictions to ground-truth labels for each vegetable-model pair. Collectively, these metrics capture robustness under imbalance, sensitivity to disease detection, correctness, and class-wise reliability. Computational efficiency metrics, specifically model size (MB), number of trainable parameters, and inference time per image (ms), were reported in addition to classification performance. In order to compare the different architectures for deployment settings in mobile or edge devices, where memory and latency are important factors, these metrics are highly important. Accuracy, precision, recall, and F1-score are some of the metrics which were averaged over the 5 cross-validation folds, providing an in-depth understanding of performance stability. Further, the cross-fold variance validated each model's consistency across sampling distributions. The performance indicators employed to assess the models are as follows:

– Precision: a model's precision, or its capacity to accurately identify positive cases among all positive predictions, is represented by (1), where TP is the number of true positives, and FP is the number of false positives.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

– Recall: the model's capacity to detect each and every true positive instance is indicated by (2), which also provides the number of TP and FN.

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

– F1-score: in (3) shows that the F1-score, which is calculated by taking the harmonic mean of precision and recall, is a balanced indicator of the model's accuracy.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (3)$$

– Accuracy: in (4), where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives, accuracy is the percentage of accurate predictions on the test set.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

The metrics outlined in the following sections provide the trade-off between predictive performance and deploy ability.

– Parameters (M): the network's total number of learnable weights.
– Model size (MB): the amount of memory required to save the model.
– Inference time (ms): the forward pass latency is the average time required to process one test image.

2.7. Statistical testing

Statistical significance testing was incorporated in the evaluation method to ensure that differences in performance between deep learning models were not due to random differences but merely reflected original improvements. Paired *t*-tests and the Wilcoxon signed-rank test were deployed to compare the metric values for each vegetable–model pair across the 5 cross-validation folds (accuracy, precision, recall, and F1-score). While the Wilcoxon test provides a non-parametric alternative that does not rely on distributional assumptions, making it suitable for small sample sizes like 5-fold cross-validation, the paired t-test assesses whether the mean difference between two models' performance scores is statistically significant under the assumption of normality. These tests were performed model-wise (e.g., DenseNet201 vs. MobileNetV2) on each metric and vegetable dataset. Whether one architecture consistently outperformed another beyond chance ($p < 0.05$) was determined by the resulting p-values. While combined with effect-size measures such as mean performance differences ((ΔMean)), this statistical analysis provided performance that is superior across several types of other vegetables, and model superiority was clearly showcased, most notably indicating DenseNet201's significant advantage on cucumber datasets.

2.8. Explainability analysis

Grad-CAM was employed to represent the spatial regions that contributed most to each model's predictions, required to enhance the deep-learning models' transparency and interpretability. By computing the gradient of the predicted class score with respect to the activations of the final convolutional layer, Grad-CAM highlights discriminative image regions and generates a class-specific heatmap. The best performing checkpoint of each architecture (VGG19, DenseNet201, and MobileNetV2) across all vegetable test sets was examined by Grad-CAM for this study. For each vegetable, representative images from both classes—healthy and downy mildew—were selected to ensure a consistent qualitative comparison. To see whether the models correctly focused on symptomatic areas such as chlorotic patches, angular lesions, or discoloration characteristic of downy mildew, the heatmaps were overlaid onto the original RGB images. Interpretability is essential for domain experts and field-level decision-making, the explainability results thus reinforce the reliability of the final model predictions and provide confidence for real-world agricultural deployment.

2.9. Result interpretation and visualization

To provide a complete assessment of performance across vegetables and architectures, the model outputs were examined using a combination of quantitative metrics, statistical comparisons, and qualitative visualizations. The evaluation metrics (accuracy, precision, recall, F1-score, parameter count, model size, and inference time) were first aggregated into comparative tables to clearly identify the trade-offs between computational efficiency and predictive power. In order to enhance interpretability, detailed Grad-CAM visualizations were generated for the healthy and downy mildew classes of all vegetables. Visual confirmation of whether each model appropriately focused on symptomatic leaf regions was provided by heatmaps. The incorporation of statistical significance testing further contextualized the results and verified that performance differences were not incidental. Together, the statistical outcomes, visual heatmaps, and structured tables provided an integrated interpretive framework that provided robustness and transparency in the evaluation of the proposed disease classification pipeline.

3. RESULTS AND DISCUSSION

The three deep-learning architectures' predictive performance, computational efficiency, and interpretability differ significantly in the experimental evaluation across five vegetables. Table 3 illustrates the overall model performance, which reports metrics such as model size, inference latency, F1-score, accuracy, precision, recall, and parameter count. Both DenseNet201 and VGG19 show strong feature-learning capability and robustness to intra-class variability, demonstrating near-perfect classification across all vegetables except cucumber Sultana, with accuracy and F1-scores of 98–100%. Despite having only 2.59 M parameters, MobileNetV2 produced highly competitive results (for example, 99.37% accuracy for cauliflower and 96.99% for cucumber Rashid), highlighting its suitability for deployment on embedded or mobile platforms.

Table 3. Performance comparison of VGG19, DenseNet201, and MobileNetV2 across all vegetables

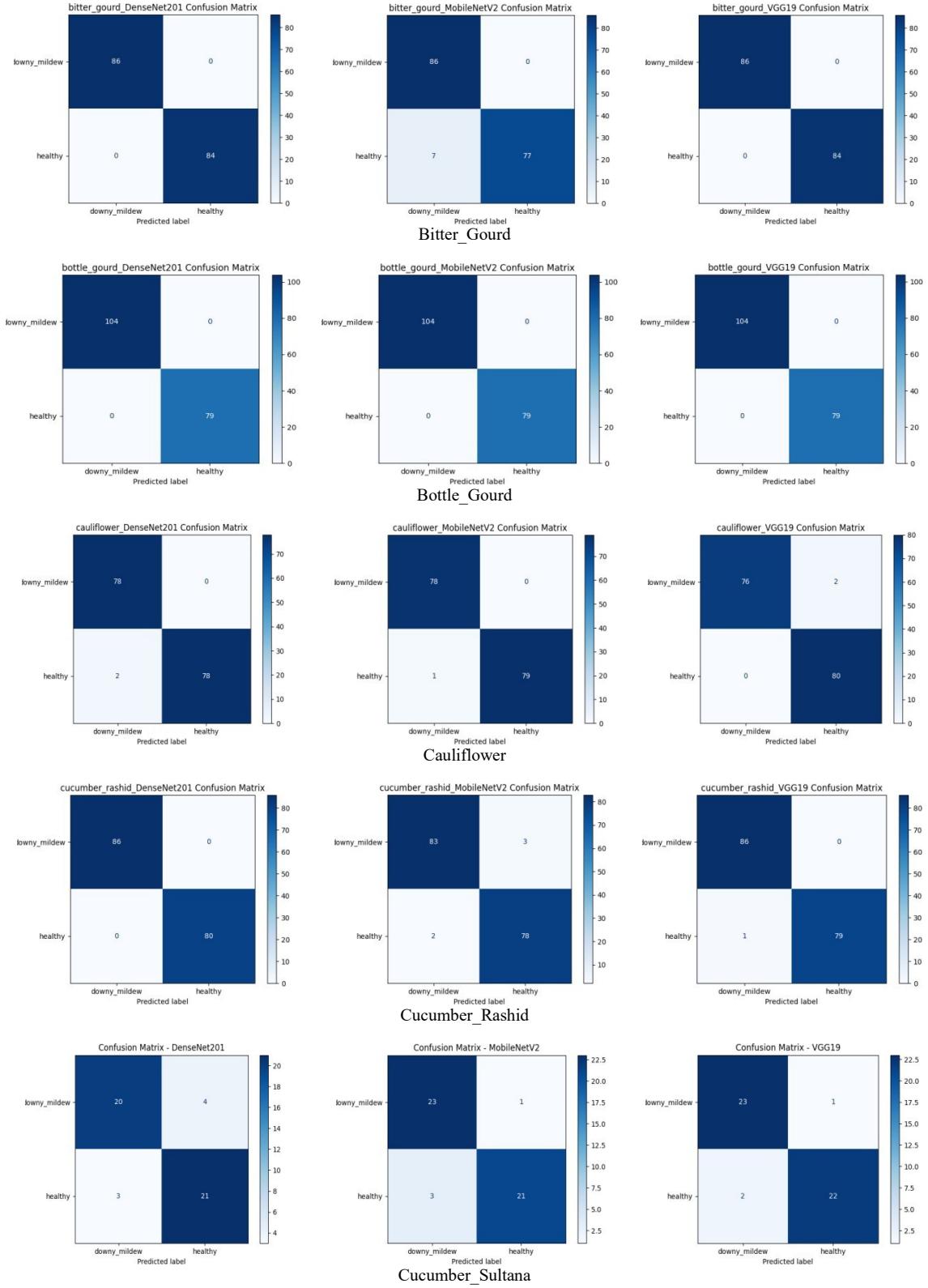
Vegetable	Model	Accuracy	Precision	Recall	F1-score	Support	Params (M)	Size (MB)	Inference (ms)
bitter_ground	DenseNet201	100	100	100	100	170	18.81	77.91	605.25
bitter_ground	MobileNetV2	95.88	96.19	95.88	95.87	170	2.59	12.93	521.63
bitter_ground	VGG19	100	100	100	100	170	20.16	77.99	979.83
bottle_gourd	DenseNet201	100	100	100	100	183	18.81	77.91	600.8
bottle_gourd	MobileNetV2	100	100	100	100	183	2.59	12.93	775.56
bottle_gourd	VGG19	100	100	100	100	183	20.16	77.99	1030.26
cauliflower	DenseNet201	98.73	98.77	98.73	98.73	158	18.81	77.91	678.87
cauliflower	MobileNetV2	99.37	99.38	99.37	99.37	158	2.59	12.93	570.84
cauliflower	VGG19	98.73	98.77	98.73	98.73	158	20.16	77.99	1018.84
cucumber_rashid	DenseNet201	100	100	100	100	166	18.81	77.91	854.98
cucumber_rashid	MobileNetV2	96.99	97	96.99	96.99	166	2.59	12.93	539.73
cucumber_rashid	VGG19	99.4	99.4	99.4	99.4	166	20.16	77.99	987.1
cucumber_sultana	DenseNet201	85.42	85.48	85.42	85.41	48	18.81	77.91	840.63
cucumber_sultana	MobileNetV2	89.58	89.65	89.58	89.58	48	2.59	12.93	575.68
cucumber_sultana	VGG19	91.67	91.96	91.67	91.65	48	20.16	77.99	1707.01

However, all models failed on cucumber Sultana, probably as a result of the test set's significantly smaller sample size (n =48), which reduces representation diversity and renders overfitting possible. Nevertheless, VGG19 achieved moderate resilience, attaining a 91.67% F1-score. VGG19 and DenseNet201 provide superior accuracy in high-capacity environments, whereas MobileNetV2 provides an accuracy–latency trade-off that is desirable for real-time or edge-based agricultural systems. MobileNetV2 remained the fastest overall, but DenseNet201 consistently performed better than VGG19 considering inference times (e.g., 605 ms vs. 979 ms for bitter gourd), proving its efficiency in reusing feature maps.

Figure 2 confusion matrices for all the models show that misclassifications were largely restricted to the cucumber Sultana subset. The models learned highly discriminative boundary features for both healthy and downy mildew categories, as demonstrated by the perfectly diagonal matrices of all other vegetables. This is further consistent with the quantitative evaluation represented in Table 3, in which three vegetables reached precision and recall values of 100%, indicating complete sensitivity and specificity for the identification of disease in those cases. The need for balanced sampling or augmentation methods is reinforced by the disparity in performance between vegetables with sparse data (e.g., cucumber Sultana) and those with abundant data (e.g., bitter gourd and bottle gourd). Additionally, the 5-fold cross-validation results showed tightly clustered performance metrics for all vegetables, confirming that the models reliably generalized across multiple sampling distributions. Despite class imbalance, DenseNet201 exhibited the least variability, reflecting stable learning. Prior to test-set evaluation, these cross-fold trends validate the trained models' robustness.

Figure 3 shows the Grad-CAM heatmaps confirm that all three architectures correctly manage symptom-rich regions associated with downy mildew infection. The attention maps produced by DenseNet201 are consistently the most exactly defined and biologically meaningful, closely capturing necrotic boundaries, vein-limited lesions, and chlorotic patches. With its lightweight architecture and reduced spatial selectivity, MobileNetV2 demonstrates a slightly broader but still accurate focus. VGG19's deeper but unregularized filters provide valid lesion localization, but with occasional background attention. Substantially, all vegetables with healthy leaves show near-uniform heatmaps with low activation, indicating low FP sensitivity. For MobileNetV2 and VGG19, harder classes like cucumber Sultana exhibit diffused and inconsistent attention that directly associates with their lower statistical performance in Table 4. Overall, the Grad-CAM results provide strong interpretability evidence that DenseNet201 is the best choice for high-accuracy deployment in agricultural decision-support systems, as it not only learns the most discriminative and symptom-faithful features but also performs best quantitatively. Using per-vegetable performance values, paired t-tests and Wilcoxon signed-rank tests were conducted across models to assess whether the observed differences were

statistically significant on larger datasets as provided in the statistical comparison of models, summarized in Table 4. While differences between VGG19 and MobileNetV2 were significant for most vegetables but not all, both tests confirmed statistically significant differences between DenseNet201 and the other two architectures ($p < 0.05$). This confirms that model discrepancies were not due to sampling variance and quantitatively validates DenseNet201's superiority.



Bitter_Gourd

Bottle_Gourd

Cauliflower

Cucumber_Rashid

Cucumber_Sultana

Figure 2. Confusion matrices of all vegetable classes examined deep learning models

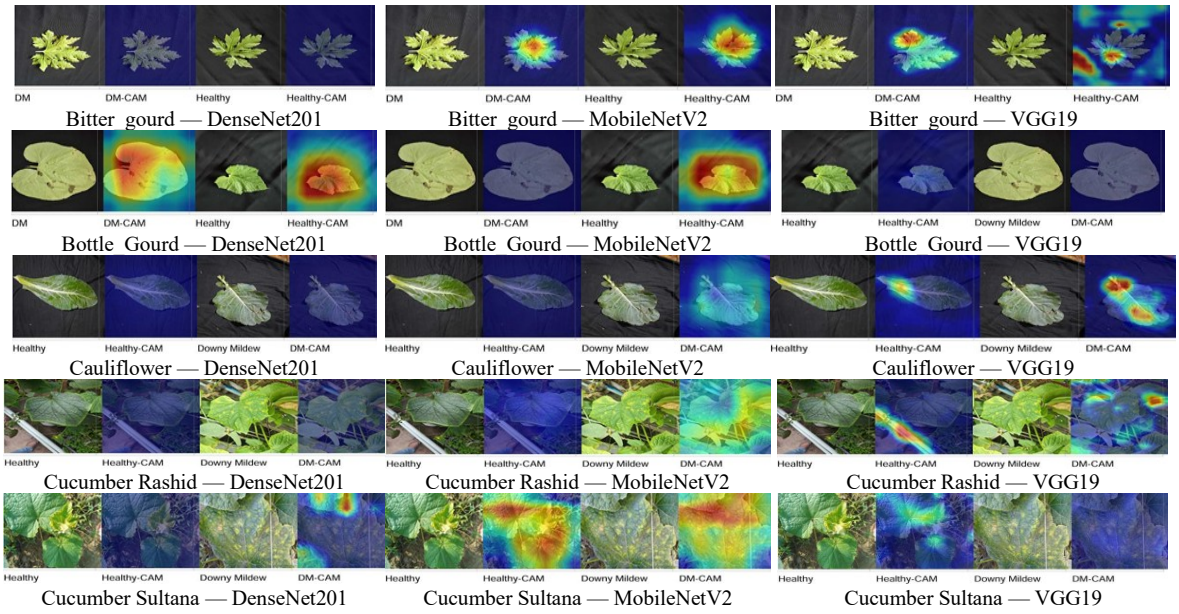


Figure 3. Grad-CAM representation of model attention for the detection of downy mildew in vegetables

Table 4. Statistical significance analysis comparing best and next-best models across vegetables (paired *t*-test/Wilcoxon)

Vegetable	Metric	Best	Next	Mean best (%)	Mean next (%)	Δ Mean (%)	p-value	Sig.	Interpretation
bitter gourd	Accuracy	DenseNet201	MobileNetV2	100.00	99.58	0.42	0.1835	×	Models perform equivalently.
bitter gourd	F1	DenseNet201	MobileNetV2	100.00	99.54	0.46	0.1844	×	Models perform equivalently.
bitter gourd	Precision	DenseNet201	VGG19	100.00	100.00	0.00	1.0000	×	No difference in precision.
bitter gourd	Recall	DenseNet201	MobileNetV2	100.00	100.00	0.00	1.0000	×	No difference in recall.
bottle gourd	Accuracy	DenseNet201	MobileNetV2	100.00	100.00	0.00	1.0000	×	Models perform equivalently.
bottle gourd	F1	DenseNet201	MobileNetV2	100.00	100.00	0.00	1.0000	×	Models perform equivalently.
bottle gourd	Precision	DenseNet201	MobileNetV2	100.00	100.00	0.00	1.0000	×	Models perform equivalently.
bottle gourd	Recall	DenseNet201	MobileNetV2	100.00	100.00	0.00	1.0000	×	Models perform equivalently.
cauliflower	Accuracy	DenseNet201	MobileNetV2	100.00	100.00	0.00	1.0000	×	Models perform equivalently.
cauliflower	F1	DenseNet201	MobileNetV2	100.00	100.00	0.00	1.0000	×	Models perform equivalently.
cauliflower	Precision	DenseNet201	MobileNetV2	100.00	100.00	0.00	1.0000	×	Models perform equivalently.
cauliflower	Recall	DenseNet201	MobileNetV2	100.00	100.00	0.00	1.0000	×	Models perform equivalently.
cucumber rashid	Accuracy	DenseNet201	MobileNetV2	100.00	100.00	0.00	1.0000	×	Models perform equivalently.
cucumber rashid	F1	DenseNet201	MobileNetV2	100.00	100.00	0.00	1.0000	×	Models perform equivalently.
cucumber rashid	Precision	DenseNet201	MobileNetV2	100.00	100.00	0.00	1.0000	×	Models perform equivalently.
cucumber rashid	Recall	DenseNet201	MobileNetV2	100.00	100.00	0.00	1.0000	×	Models perform equivalently.
cucumber sultana	Accuracy	DenseNet201	MobileNetV2	98.20	83.01	15.19	0.0014	✓	DenseNet201 significantly better; large gap.
cucumber sultana	F1	DenseNet201	MobileNetV2	98.12	85.36	12.76	0.0017	✓	DenseNet201 significantly better on F1.
cucumber sultana	Precision	DenseNet201	MobileNetV2	97.19	74.76	22.43	0.0010	✓	DenseNet201 significantly better in precision.
cucumber sultana	Recall	MobileNetV2	DenseNet201	100.00	99.09	0.91	0.3739	×	No significant difference in recall.

3.1. Learning curve analysis

The training and validation loss and accuracy curves for the VGG19, DenseNet201, and MobileNetV2 models trained on the five-vegetable dataset are represented in Figure 4. Effective optimization is observed by all architectures' consistent convergence, monotonic loss decreases, and consistent accuracy increase. While VGG19 demonstrates gradual but steady learning, DenseNet201 and MobileNetV2 converge more rapidly and demonstrate high validation accuracy in fewer epochs. Importantly, all models' training and validation curves align closely, indicating minimal overfitting. It can be clarified by the early stopping strategy, dropout regularization, and applied data augmentation. These outcomes confirm the proposed models' robustness and ability for generalization under real-world variability.

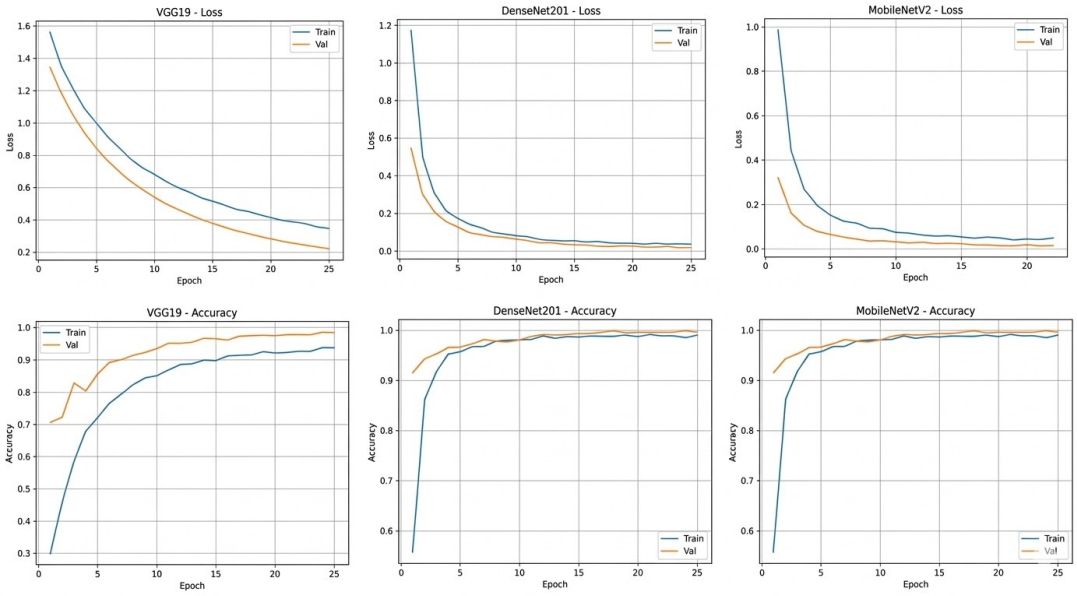


Figure 4. Learning curves showing training and validation loss and accuracy for VGG19, DenseNet201, and MobileNetV2 models

3.2. Class imbalance analysis

Multiple imbalance-handling strategies, including baseline training without balancing, class-weighted loss, and focal loss, were systematically evaluated to understand the impact of class imbalance on model performance. To ensure a fair comparison, identical data splits, network architectures, and hyperparameter settings were used in all experiments. Strong bias toward majority classes in the baseline model led to poor macro-averaged performance and almost zero recall for minority categories. Significant gains were not generated by applying class weights, and it also led to a similar prediction collapse. Conversely, by emphasizing hard and minority samples during training, focal loss significantly enhanced discrimination across all classes. With a test accuracy of 99.7%, a macro-averaged F1-score of 0.99, and balanced recall across all five classes, the DenseNet201 model trained with focal loss thus achieved near-perfect classification performance. Focal loss is an effective and reliable method to mitigate class imbalance in multi-class plant disease classification. The impact of different class imbalance handling strategies on classification performance is quantitatively provided in Table 5.

Table 5. Performance comparison of class imbalance handling strategies on the test dataset

Imbalance strategy	Accuracy	Macro precision	Macro recall	Macro F1-score
Baseline (no balancing)	0.23	0.05	0.20	0.08
Class weights	0.23	0.05	0.20	0.08
Focal loss	0.997	0.99	0.99	0.99

3.3. Error analysis

Using Grad-CAM visualizations on test samples that are misclassified, a qualitative error analysis was conducted to get better insight into the failure modes of the proposed models. Although the overall classification accuracy is over 98%, there remain a few challenging cases that are primarily characterized by background clutter, non-uniform illumination, and subtle symptom expression. The Grad-CAM maps

corresponding to the true and predicted classes are compared in Figure 5, which depicts representative failure cases for different vegetable crops. The original leaf image is provided in each example, along with attention maps for the true and predicted classes that highlight the regions that resulted in incorrect model decisions. This happens when the model's attention shifts from localized disease symptoms to non-diagnostic regions such as leaf edges, background objects, or uniformly textured areas. In addition, samples with visually ambiguous patterns or early-stage infections tend to produce diffuse activation maps, demonstrating reduced discriminative confidence. These observations highlight real challenges in field-acquired imagery and suggest that misclassifications are predominantly driven by symptom heterogeneity and environmental variability rather than model instability.

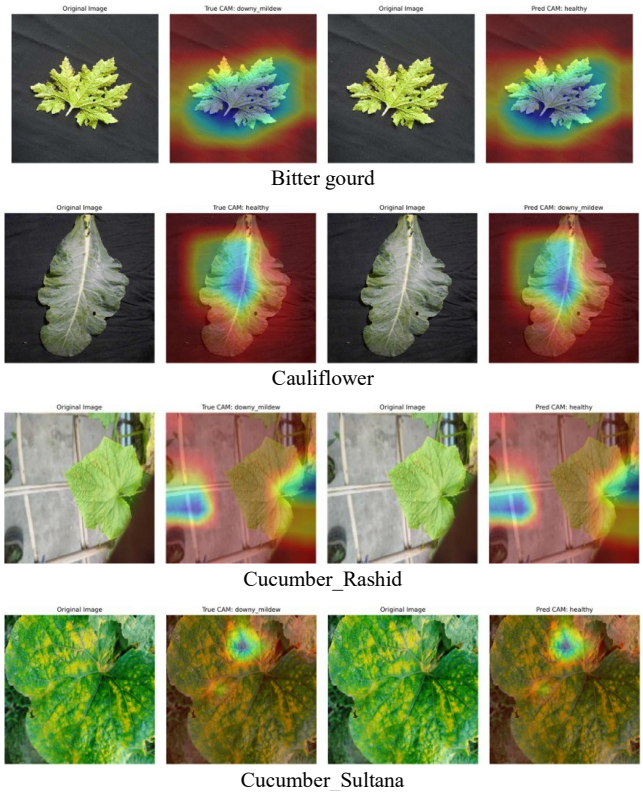


Figure 5. Grad-CAM analysis for representative misclassification

3.4. Ablation study

We conducted ablation studies with a focus on data augmentation and transfer learning in an effort to quantify the effect of key design choices. First, we evaluated DenseNet201 with and without augmentation under the identical settings to evaluate the impact of augmentation across five crop datasets. Table 6 highlights that augmentation consistently stabilized or improved performance, particularly in crops such as cucumber Sultana with few training samples. Second, by training the model with and without ImageNet pretraining, we assessed the role of transfer learning using a representative crop (bitter gourd). The model failed to achieve near-random performance (accuracy: 50.6%, F1-score: 0.00) on the removal of transfer learning as per the results in Table 7, since the network converged for predicting only the majority class. The pretrained model, in contrast, achieved strong generalization (Accuracy: 98.2%, F1-score: 98.2%). These results confirm that both transfer learning and augmentation are critical for reliable performance on limited agricultural datasets. To assess the impact of architectural complexity on classification performance, an ablation study was conducted. The lightweight MobileNetV2 model offers faster inference and reduced cost of computation, as Table 3 summarizes, but it exhibits a noticeable reduction in performance on visually complex disease patterns. The heavyweight DenseNet201, in contrast, consistently achieves superior accuracy and F1-score across all vegetables, demonstrating improved discriminative capacity and feature reuse. As an intermediate baseline, VGG19 confirms a clear trade-off between accuracy and efficiency. These results justify the selection of DenseNet201 for reliable diagnosis of plant ailments where accuracy is prioritized over minimal model size.

Table 6. Ablation study evaluating the impact of data augmentation across crop datasets

Crop	Augmentation	Accuracy	F1-score
Bitter gourd	No	0.9941	0.9941
Bitter gourd	Yes	0.9941	0.9941
Bottle gourd	No	1.0000	1.0000
Bottle gourd	Yes	1.0000	1.0000
Cauliflower	No	1.0000	1.0000
Cauliflower	Yes	1.0000	1.0000
Cucumber (Rashid)	No	0.9880	0.9873
Cucumber (Rashid)	Yes	0.9940	0.9937
Cucumber (Sultana)	No	0.9792	0.9796
Cucumber (Sultana)	Yes	0.9583	0.9600

Table 7. Ablation study analyzing the effect of transfer learning using DenseNet201 on the bitter gourd dataset

Model variant	Accuracy	F1-score
With transfer learning	0.9824	0.9825
Without transfer learning	0.5059	0.0000

3.5. Sensitivity analysis

A comprehensive sensitivity analysis was conducted with respect to dataset size, class imbalance severity, and noise/illumination variation so as to evaluate the robustness of the proposed disease classification framework under realistic constraints. Table 8 summarizes the quantitative results. Accuracy and F1-score steadily increase as the fraction of training data increases from 25% to 100% across all crops, as represented in dataset size sensitivity results. Strong data efficiency can be observed in data-rich crops such as bitter gourd and bottle gourd, which achieve near-saturation performance at 50–75% of the dataset. In contrast, at lower dataset fractions, data-scarce crops (like cucumber Sultana) exhibit substantial performance degradation, highlighting the dependence of reliable generalization on sufficient sample availability. In relation to the class imbalance sensitivity analysis, the F1-score sharply deteriorates and, in extreme cases, collapses to zero, even though overall accuracy may remain moderately high under severe imbalance. The model's failure to correctly identify minority disease classes is proof of strong majority-class bias. The findings justify the use of F1-score as the primary evaluation metric and emphasize that accuracy alone is insufficient for evaluating agricultural disease classification tasks.

Table 8. Consolidated sensitivity analysis of the proposed model with respect to dataset size, class imbalance, and noise/illumination variation across all crops

Crop	Condition	Setting	Accuracy	F1-score
bitter gourd	Dataset size	25%	0.8294	0.8199
	Dataset size	50%	0.9235	0.9193
	Dataset size	75%	0.9235	0.9202
	Dataset size	100%	0.9412	0.9412
	Class imbalance	Severe	0.6000	0.3200
	Noise/lighting	Perturbed	0.4882	0.6561
bottle gourd	Dataset size	25%	0.7978	0.7483
	Dataset size	50%	0.9945	0.9937
	Dataset size	75%	1.0000	1.0000
	Dataset size	100%	1.0000	1.0000
	Class imbalance	Severe	0.8033	0.7049
	Noise/lighting	Perturbed	0.4262	0.5532
cauliflower	Dataset size	25%	0.8101	0.8125
	Dataset size	50%	0.8861	0.8902
	Dataset size	75%	0.9747	0.9753
	Dataset size	100%	0.9810	0.9816
	Class imbalance	Severe	0.5570	0.2222
	Noise/lighting	Perturbed	0.5063	0.6723
cucumber rashid	Dataset size	25%	0.6446	0.5693
	Dataset size	50%	0.8012	0.7975
	Dataset size	75%	0.7470	0.7470
	Dataset size	100%	0.8916	0.8902
	Class imbalance	Severe	0.5181	0.00
	Noise/lighting	Perturbed	0.5422	0.6545
cucumber sultana	Dataset size	25%	0.3333	0.00
	Dataset size	50%	0.7917	0.7619
	Dataset size	75%	0.9167	0.9200
	Dataset size	100%	0.5833	0.5454
	Class imbalance	Severe	0.5000	0.00
	Noise/lighting	Perturbed	0.6458	0.6792

Under simulated lighting distortions, noise, and illumination sensitivity analysis shows a substantial reduction in accuracy, with accuracy values decreasing to approximately 45–55%. Nevertheless, non-zero F1-scores for most types of crops suggest that the model retains partial discriminative capability under adverse imaging conditions by using localized disease patterns. Collectively, these findings validate the robustness of the proposed model under realistic constraints and represent the importance of balanced acquisition of data and controlled imaging conditions for reliable deployment in real agricultural conditions.

3.6. Comparison with prior works

We compared the proposed framework's performance to several recent deep learning methods from the literature as mentioned in Table 9. A modified depthwise CNN integrated with squeeze-and-excitation blocks and residual skip connections was developed by Ashurov *et al*. [28]. It displayed high classification performance across multi-species leaf datasets, highlighting advanced architectural enhancements over standard CNNs. Developing transfer learning with ensemble pre-trained CNNs on the comprehensive PlantVillage dataset, Shafik *et al*. [29] achieved approximately 97.8% accuracy in multi-disease classification, emphasizing the merit of using large pre-trained representations. Khalifa *et al*. [30] showed that a custom CNN can also attain competitive performance (96.5%) on plant disease data. With systematic ablation and sensitivity analyses, along with error analysis and interpretability (Grad-CAM) to enhance model reliability and scientific insight, the proposed method delivers comparable or superior performance across multiple crops.

Table 9. Comparison of the proposed approach with recent plant disease detection studies (2023–2025)

Study	Model/architecture	Crops classes	Performance	Key limitations
Ashurov *et al.* [28]	Depthwise CNN+SE+Residual	Multiple crops/ multi-class	Accuracy ≈98%	No robustness and error analysis
Shafik *et al*. [29]	PDDNet-AE and PDDNet-LVE	Multiple crops multi-class	Accuracy 97.8%	No error analysis or deployment discussion
Khalifa *et al*. [30]	Custom CNN	Single crop/ few classes	Accuracy 96.54%	No ablation analysis
Proposed method	DenseNet201 (transfer learning)	5 crops/binary classes	Accuracy up to 89.6–100%	Higher computational cost

4. CONCLUSION

This study demonstrated that deep learning models can reliably identify downy mildew in multiple types of vegetable crops, with DenseNet201 consistently achieving good outcomes and the most informative Grad-CAM visualizations. As indicated by statistical analysis, DenseNet201 performed equivalent to VGG19 in classes with larger sample sizes, while it significantly outperformed MobileNetV2 in smaller or visually subtle datasets like cucumber Sultana. The findings indicate that feature complexity, architectural depth, and dataset scale impact model effectiveness. MobileNetV2 is better suited for real-time or edge deployment as it offers the smallest model size and fastest inference, although it displayed slightly lower predictive accuracy. Confidence in the biological validity and applicability of the proposed system was strengthened by the Grad-CAM analysis that confirmed that all models concentrated on disease-relevant regions. So as to enable reliable performance across new crop varieties, geographical regions, and unseen field environments, future work will focus on enhancing the generalization and robustness of the proposed detection framework by incorporating domain adaptation and few-shot learning strategies. To evaluate their suitability for real-time field deployment, the model benchmarking will also be extended to include recent state-of-the-art (SOTA) architectures such as vision transformers, EfficientNet-Lite, and lightweight edge-AI models. With the aim of progressing the pipeline further to a more scalable, concise, and deployment-ready AI solution for precision agriculture, future studies will further include cross-dataset validation and temporal disease progression modeling.

FUNDING INFORMATION

This research received no external funding. The study was conducted without financial support from any external agency, grant, or industrial partner. All resources, including computational facilities, software, and data collection, were provided internally by the authors and their affiliated institution. No funders had any role in the study design, data analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Supreetha Shivaraj	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Manjula Sunkadakatte Haladappa		✓				✓		✓	✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [SS], upon reasonable request.

REFERENCES

[1] A. Upadhyay *et al.*, "Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture," *Artificial Intelligence Review*, vol. 58, 2025, doi: 10.1007/s10462-024-11100-x.

[2] I. Hernández, S. Gutiérrez, and J. Tardaguila, "Image analysis with deep learning for early detection of downy mildew in grapevine," *Scientia Horticulturae*, vol. 331, 2024, doi: 10.1016/j.scienta.2024.113155.

[3] C. Yao, X. Zhang, H. Mao, H. Gao, and Q. Li, "A rapid diagnostic grading system for cucumber downy mildew based on visible light - hyperspectral imaging system," *Journal of Advances in Agriculture*, vol. 11, pp. 108–121, 2020, doi: 10.24297/jaa.v11i.8779.

[4] I. Pacal *et al.*, "A systematic review of deep learning techniques for plant diseases," *Artificial Intelligence Review*, vol. 57, no. 11, 2024, doi: 10.1007/s10462-024-10944-7.

[5] M. Shafay *et al.*, "Recent advances in plant disease detection: challenges and opportunities," *Plant Methods*, vol. 21, no. 1, 2025, doi: 10.1186/s13007-025-01450-0.

[6] R. I. Hasan, S. M. Yusuf, and L. Alzubaidi, "Review of the state of the art of deep learning for plant diseases: a broad analysis and discussion," *Plants*, vol. 9, no. 10, pp. 1–25, 2020, doi: 10.3390/plants9101302.

[7] M. H. Saleem, J. Potgieter, and K. M. Arif, "Plant disease detection and classification by deep learning," *Plants*, vol. 8, no. 11, 2019, doi: 10.3390/plants8110468.

[8] V. M.-Gutiérrez *et al.,* "Comparison of convolutional neural network architectures for classification of tomato plant diseases," *Applied Sciences*, vol. 10, 2020, doi: 10.3390/app10041245.

[9] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, vol. 7, no. September, 2016, doi: 10.3389/fpls.2016.01419.

[10] A. Kumar, H. P. Monga, T. Brahma, S. Kalra, and N. Sherif, "Mobile-friendly deep learning for plant disease detection: a lightweight CNN benchmark across 101 classes of 33 crops." 2025, *arXiv:2508.10817*.

[11] M. A. A. Khandaker, Z. S. Raha, S. Islam, and T. Muhammad, "Explainable AI-enhanced deep learning for pumpkin leaf disease detection: a comparative analysis of CNN architectures," in *2024 27th International Conference on Computer and Information Technology, ICCIT 2024 - Proceedings*, 2024, pp. 2428–2433. doi: 10.1109/ICCIT64611.2024.11021957.

[12] K. Gopalan, S. Srinivasan, P. Pragya, M. Singh, S. K. Mathivanan, and U. Moorthy, "Corn leaf disease diagnosis: enhancing accuracy with ResNet152 and Grad-CAM for explainable AI," *BMC Plant Biology*, vol. 25, no. 1, 2025, doi: 10.1186/s12870-025-06386-0.

[13] H. A.-El Monsef, S. E. Smith, D. L. Rowland, and N. Abd El Rasol, "Using multispectral imagery to extract a pure spectral canopy signature for predicting peanut maturity," *Computers and Electronics in Agriculture*, vol. 162, pp. 561–572, 2019, doi: 10.1016/j.compag.2019.04.028.

[14] I. Hernández, S. Gutiérrez, I. Barrio, R. Íñiguez, and J. Tardaguila, "In-field disease symptom detection and localisation using explainable deep learning: use case for downy mildew in grapevine," *Computers and Electronics in Agriculture*, vol. 226, Nov. 2024, doi: 10.1016/j.compag.2024.109478.

[15] N. Shukla, S. A. Palwe, S. Shubham, M. Rajani, and A. Suri, "Plant disease detection and localization using GRADCAM," *International Journal of Recent Technology and Engineering*, vol. 8, no. 6, pp. 3069–3075, 2020, doi: 10.35940/ijrte.e6935.038620.

[16] J. Y. Ding, W. S. Jeon, and S. Y. Rhee, "DM-YOLOv8: cucumber disease and insect detection using detailed multi-intensity features," in *6th International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2024*, 2024, pp. 199–204. doi: 10.1109/ICAIIC60209.2024.10463255.

[17] Y. Chabalala, E. Adam, and K. A. Ali, "Exploring the effect of balanced and imbalanced multi-class distribution data and sampling techniques on fruit-tree crop classification using different machine learning classifiers," *Geomatics*, vol. 3, no. 1, pp. 70–92, 2023, doi: 10.3390/geomatics3010004.

[18] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1-30, 2006.
[19] S. Natarajan, P. Chakrabarti, and M. Margala, "Robust diagnosis and meta visualizations of plant diseases through deep neural architecture with explainable AI," *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-64601-8.
[20] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, 2020, doi: 10.1186/s12864-019-6413-7.
[21] M. R. A. Rashid, T. K. Tarin, R. Kamara, M. Y. Mou, S. F. Rabbi, and M. Hasan, "Plant leaf freshness and disease detection dataset from Bangladesh." 2024, doi: 10.17632/n67gctmjyj.3.
[22] N. Sultana, S. B. Shorif, M. Akter, and M. S. Uddin, "Cucumber disease recognition dataset," *Mendeley Data*, vol. 10. 2022, doi: 10.17632/y6d3z6f8z9.1.
[23] G. P. Kanna *et al.*, "Advanced deep learning techniques for early disease prediction in cauliflower plants," *Scientific Reports*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-45403-w.
[24] X. Gong and S. Zhang, "An analysis of plant diseases identification based on deep learning methods," *Plant Pathology Journal*, vol. 39, no. 4, pp. 319–334, 2023, doi: 10.5423/PPJ.OA.02.2023.0034.
[25] M. K. A. Mazumder, M. F. Mridha, S. Alfarhood, M. Safran, M. A. -Al-Jubair, and D. Che, "A robust and light-weight transfer learning-based architecture for accurate detection of leaf diseases across multiple plants using less amount of images," *Frontiers in Plant Science*, vol. 14, 2024, doi: 10.3389/fpls.2023.1321877.
[26] J. Lu, X. Liu, X. Ma, J. Tong, and J. Peng, "Improved MobileNetV2 crop disease identification model for intelligent agriculture," *PeerJ Computer Science*, vol. 9, 2023, doi: 10.7717/peerj-cs.1595.
[27] S. Duhan, P. Gulia, N. S. Gill, and E. Narwal, "RTR_Lite_MobileNetV2: a lightweight and efficient model for plant disease detection and classification," *Current Plant Biology*, vol. 42, 2025, doi: 10.1016/j.cpb.2025.100459.
[28] A. Y. Ashurov *et al.*, "Enhancing plant disease detection through deep learning: a depthwise CNN with squeeze and excitation integration and residual skip connections," *Frontiers in Plant Science*, vol. 15, 2024, doi: 10.3389/fpls.2024.1505857.
[29] W. Shafik, A. Tufail, C. D. S. Liyanage, and R. A. A. H. M. Apong, "Using transfer learning-based plant disease classification and detection for sustainable agriculture," *BMC Plant Biology*, vol. 24, no. 1, Feb. 2024, doi: 10.1186/s12870-024-04825-y.
[30] K. Khalifa, K. Patel, S. Parmar, D. Patel, and J. B. Upadhyay, "Plant disease detection using a deep learning approach: a custom CNN," *International Research Journal of Advanced Engineering Hub*, vol. 3, no. 10, pp. 3869–3876, 2025, doi: 10.47392/IRJAEH.2025.0562.

BIOGRAPHIES OF AUTHORS

Supreetha Shivaraj is a research scholar in the Department of Computer Science and Engineering, University of Visvesvaraya College of Engineering, Bengaluru, India. She completed her master's in Engineering in Computer Network Engineering at the Oxford College of Engineering, Bengaluru, India. She completed her bachelor's in Computer Science and Engineering from Government Sri Krishna Rajendra Silver Jubilee Technological Institute, Bengaluru, India. Currently, she is a part time research scholar pursuing the specialization "Efficient deep learning technique for plant leaf disease detection". Her subject interests include artificial intelligence, machine learning, deep learning, computer networks, and data privacy. She can be contacted at email: supreetha.s3191@gmail.com.

Manjula Sunkadakatte Haladappa is currently a professor in the Department of Computer Science and Engineering, University of Visvesvaraya College of Engineering, Bangalore University, Bengaluru, has established herself as a prominent figure in the field. Holding a B.E., M.Tech., and Ph.D. in Computer Science and Engineering. She has honed her expertise in computer networks, wireless sensor networks, data mining, cloud computing, artificial intelligence, machine learning, and federated learning. With an impressive track record, she has authored 69 journals, presented 72 conference papers, holds 5 patents, and has contributed 4 book chapters while publishing 4 books, showing her profile and diverse contributions to the academic and technological community. Additionally, she is currently a respected member of the executive council and has served as a former member of the academic senate at VTU Belagavi. She can be contacted at email: shmanjula@gmail.com.