

Article

X-DEP: An Explainability-Aware Development and Evaluation Pipeline for Transparent Mango Disease Detection Using Deep CNNs

Guoquan Lee ^{1, *}  and Sagar Tamang ²

¹ School of Information and Communication Engineering, Hainan University, Hainan 570228, China

² Centre for Educational Technology, Indian Institute of Technology Patna, Bihar 801106, India

* Correspondence: 25210810000012@hainanu.edu.cn (G. L.)

Abstract

The application of deep learning models in agricultural disease detection is often limited by the lack of interpretability in their predictions. Although explainable artificial intelligence (XAI) techniques such as Local Interpretable Model-Agnostic Explanations (LIME) provide visual explanations of model predictions, they are commonly applied after model training and evaluation. In this study, we propose X-DEP, an explainability-aware development and evaluation pipeline that integrates predictive performance analysis and XAI diagnostics into a common model development workflow. Rather than modifying the convolutional neural network (CNN) training objective, X-DEP uses explanation outputs as complementary diagnostic evidence for model comparison, failure analysis, and model-selection decisions. The pipeline was evaluated on the MangoDHDS dataset, comprising 1,500 images across five classes, using four pretrained CNN architectures: VGG16, ResNet50, DenseNet121, and MobileNetV2. In the comparative benchmark, VGG16 achieved the highest mean classification accuracy of 92.00%, with macro-precision, macro-recall, and macro-F1 scores of 0.925, 0.924, and 0.925, respectively. DenseNet121 and MobileNetV2 achieved mean accuracies of 87.56% and 86.67%, respectively, whereas ResNet50 achieved 63.56%. Class-wise analysis and LIME visualizations revealed differences in disease recognition and explanation patterns that were not fully captured by aggregate accuracy. In a separate ablation study, the LIME-informed feedback configuration, which incorporated HSV-based lesion-region cropping and background suppression, improved test accuracy from 91.73% to 92.89% and macro-F1 from 0.909 to 0.918. These findings highlight the potential value of integrating predictive evaluation with explanation-based diagnostics to inform agricultural vision model development.

Keywords: Mango Disease Detection; Explainable Artificial Intelligence; Deep Learning; Convolutional Neural Networks; LIME; Transfer Learning

Citation: Lee, G., Tamang, S. (2026). X-DEP: An explainability-aware development and evaluation pipeline for transparent mango disease detection using deep CNNs. *Impact in Agriculture* 2026, 2, 2. <https://doi.org/10.65500/agriculture-2026-002>

Received: 22 January 2026 | Revised: 12 March 2026 | Accepted: 21 March 2026 | Published: 26 March 2026

Copyright: © 2026 by the authors. Licensee Impaxon, Malaysia. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Mango (*Mangifera indica* L.) is one of the most economically important tropical fruit crops, widely cultivated for its nutritional value, commercial demand, and contribution to agricultural production. According to FAOSTAT-based production statistics, India is the largest

producer in the combined category of mangoes, guavas, and mangosteens, accounting for approximately 44.6% of global production in 2024, equivalent to 27.77 million tonnes [1]. Despite its agricultural importance, mango production is affected by several diseases, including anthracnose, powdery mildew, and bacterial canker, which

can reduce crop productivity, fruit quality, and market value. Early identification of disease symptoms is therefore important for effective crop management. However, conventional disease identification relies largely on visual inspection by agricultural specialists and farmers. Such practices are time-consuming, require domain expertise, and may produce inconsistent results when applied across large cultivation areas [2, 3].

Recent developments in artificial intelligence (AI), particularly deep learning (DL), have provided new opportunities for automated plant disease identification. Convolutional neural networks (CNNs) have demonstrated considerable potential for learning disease-related features directly from agricultural images, reducing the dependence on manually designed image descriptors. Sladojević et al. [14] demonstrated the applicability of deep neural networks to plant disease recognition, while Ferentinos [8] reported high classification performance across multiple plant disease categories. The effectiveness of transfer learning was further investigated by Too et al. [9], who compared pretrained deep learning architectures for plant disease identification. These developments have encouraged the application of CNN-based methods to different agricultural crops, including mango, where disease symptoms may exhibit considerable variations in color, shape, texture, and severity.

Several studies have investigated deep learning approaches for mango disease classification. Mahmud et al. [5] developed a lightweight DenseNet-based model for mango leaf disease identification, demonstrating the potential of computationally efficient CNN architectures. Hossain et al. [7] investigated Vision Transformer-based learning for mango leaf disease recognition, extending the application of deep learning beyond conventional convolutional architectures. More recently, Thaseentaj and Ilango [3] examined transfer learning with feature localization and visual explanations, while Radhakrishnan et al. [2] investigated explainable AI methods for multiclass mango leaf disease classification. These studies demonstrate the growing interest in developing automated disease identification systems with improved predictive capabilities. Nevertheless, classification accuracy alone does not provide sufficient information about the visual features responsible for individual predictions.

Explainable Artificial Intelligence (XAI) has consequently received increasing attention as a means of improving the transparency and interpretability of deep learning models. Methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) [28] and Local

Interpretable Model-agnostic Explanations (LIME) [29] provide visual evidence of image regions influencing model predictions. In agricultural disease detection, such explanations can help determine whether a classifier relies on disease-related characteristics or unrelated visual information, including background regions and image acquisition artifacts. Alhammad et al. [6] investigated the application of deep learning and explainable AI for potato leaf disease classification, while Porfirio et al. [10] explored explanation techniques for plant disease recognition. Although these approaches provide useful insights into model behavior, visual explanations do not independently establish the correctness or reliability of model predictions. Their interpretation must therefore be considered alongside predictive performance, class-wise classification behavior, and model evaluation results.

An important challenge in explainable agricultural computer vision is the limited integration of XAI findings into the broader model-development process. In many applications, explanation techniques are used to visualize predictions after model training, whereas classification performance is evaluated separately using accuracy, precision, recall, and F1-score. This separation limits the extent to which explanation outputs contribute to identifying model weaknesses, examining misclassification patterns, and informing subsequent development decisions. Moreover, a model with high classification accuracy may still produce explanations that emphasize irrelevant image regions. Therefore, a structured approach is required to combine predictive benchmarking with explanation-based diagnostics and to investigate how the resulting observations can inform model refinement. Such an approach should distinguish the diagnostic contribution of XAI from actual modifications to model training or preprocessing.

To address these challenges, this study proposes X-DEP, an explainability-aware development and evaluation pipeline for mango disease classification using pretrained CNN architectures. The proposed pipeline integrates dataset preprocessing, model training, comparative benchmarking, LIME-based diagnostic analysis, and performance validation within a common workflow. Four pretrained architectures, namely VGG16, ResNet50, DenseNet121, and MobileNetV2, are evaluated using the publicly available MangoDHDS dataset [23], comprising 1,500 images distributed equally across five disease and healthy classes. The evaluation considers overall classification performance, class-wise measures, statistical comparisons, confusion matrices, and visual explanation

patterns. In addition, a human-in-the-loop feedback stage is examined, in which observations from LIME explanations inform a subsequent preprocessing refinement involving HSV-based lesion-region cropping and background suppression. Importantly, X-DEP does not introduce a new CNN architecture or directly incorporate explanation maps into the optimization objective. Instead, its contribution lies in organizing predictive evaluation and explanation-based evidence to support model assessment and development decisions.

The main contributions of this study are summarized as follows:

- An explainability-aware development and evaluation pipeline is presented to integrate CNN training, comparative benchmarking, predictive assessment, and LIME-based diagnostic analysis within a structured workflow.
- A comparative evaluation of four pretrained CNN architectures is conducted using a consistent frozen-backbone transfer-learning protocol to examine their classification performance on the MangoDHDS dataset.
- A combined predictive and explainability assessment is performed using aggregate performance measures, class-wise analysis, statistical comparisons, confusion matrices, and LIME visualizations to identify model-specific strengths and limitations.
- A LIME-guided feedback iteration is evaluated through a controlled preprocessing modification, allowing changes in classification performance and qualitative explanation behavior to be examined without directly modifying the CNN optimization objective.

The remainder of this manuscript is organized as follows. Section 2 reviews related studies on deep learning and explainable plant disease detection. Section 3 presents the proposed X-DEP methodology and its development stages. Section 4 describes the CNN architectures selected for comparative evaluation. Section 5 presents the dataset, experimental configuration, evaluation metrics, and experimental results. Section 6 compares the findings with previous studies, while Section 7 discusses the main observations, methodological implications, and limitations. Finally, Section 8 concludes the study and outlines directions for future research.

2. Related work

In recent years, deep learning has gained considerable attention in agricultural applications, particularly for

automated plant disease detection. Mango (*Mangifera indica*) is an important tropical crop affected by various foliar diseases. Earlier studies employed image-processing techniques to assess disease severity [4]. Subsequently, convolutional neural networks (CNNs) emerged as effective approaches for automatically extracting discriminative features and improving disease classification performance.

Mahmud et al. [5] developed a lightweight DenseNet model for mango leaf disease classification, achieving 99.44% accuracy with reduced computational complexity. Alhammad et al. [6] employed transfer learning and Grad-CAM for potato leaf disease classification, achieving 98% testing accuracy while improving interpretability. Similarly, Hossain et al. [7] explored Vision Transformers (ViTs) for mango leaf disease classification, highlighting their potential for effective feature extraction.

Many studies have used transfer learning to address the lack of data in specific domains. For instance, Ferentinos [8] trained and fine-tuned CNNs on a broad dataset of 25 plant diseases, achieving over 99% accuracy, while Too et al. [9] performed a comparative study of pretrained architectures (VGG16, ResNet, InceptionV3) on leaf disease datasets. Explainable AI (XAI) has become an essential part of trustworthy artificial intelligence systems in agriculture. Techniques such as Grad-CAM and LIME allow us to visualize the attention focus of neural models, and to identify whether the predictions are grounded in legitimate disease symptoms. According to Porfirio et al. [10], it is important to integrate explanation frameworks into the training pipeline, rather than using them post-hoc. Thakur et al. [11] investigated Vision Transformer-based approaches for plant disease classification, highlighting their potential for extracting discriminative visual features. Chouhan et al. [12] explored MobileNetV2-based plant disease detection, demonstrating the applicability of lightweight deep learning architectures. Furthermore, Jackulin and Murugavalli [13] reviewed machine learning and deep learning techniques for automated plant disease identification, highlighting recent developments in agricultural diagnostics. Besides mango, early evidence for CNN effectiveness in tomato, grape and apple disease recognition came from works of Sladojević et al. [14] and Brahimi et al. [15]. Their results have sparked interest in CNN-based agricultural tools, prompting domain-specific adaptations such as the one for mango leaf disease. Despite these strides, most current systems only consider performance indicators and do not assess model trustworthiness. In this work we have emphasized the

value of explainability, continuous model evaluation and the integration of XAI methods throughout development, which together contribute to building AI tools that are not only accurate but also transparent and truly useful for real-

world agricultural applications. The comparison of existing Plant Disease classification studies is shown in Table 1.

Table 1. Critical comparison of existing plant disease classification and explainable AI studies.

Study	Dataset/Scope	Model	XAI	Main contribution	Limitation relevant to X-DEP
Ferentinos [8]	Multiple plant diseases	CNNs	Not central	Demonstrated strong deep-learning-based disease classification	Limited emphasis on explanation-based model diagnosis
Too et al. [9]	Plant disease datasets	VGG16, ResNet, Inception	Not central	Comparative transfer-learning evaluation	Explainability not incorporated into model comparison
Porfirio et al. [10]	Plant disease classification	Deep CNN	XAI	Demonstrated explainable plant disease classification	Limited evidence of explanation-driven model refinement
Thaseentaj and Ilango [3]	Mango disease images	ResNet-based	Grad-CAM	Disease classification with visual feature localization	XAI primarily used for post-hoc visualization
Radhakrishnan et al. [2]	Mango leaf diseases	Multiple deep models	XAI	Applied XAI to multiclass mango disease classification	Limited integration of explanation evidence into systematic model-development decisions
Present study	MangoDHDS	VGG16, ResNet50, DenseNet121, MobileNetV2	LIME	Unified predictive and explanation-based diagnostic workflow	Single-dataset validation; feedback iteration is human-in-the-loop and limited to a controlled preprocessing modification rather than automated explanation-guided retraining.

3. Methodology

Agricultural computer vision faces two major challenges: limited labeled training data and the lack of interpretability in deep learning predictions. To address these challenges, we propose X-DEP, an explainability-aware development and evaluation pipeline that integrates model training, benchmarking, XAI diagnostics, and performance validation. The overall workflow of the proposed X-DEP pipeline is illustrated in Figure 1, while its implementation steps are presented in Algorithm 1.

The proposed X-DEP pipeline consists of four main stages, followed by a LIME-guided feedback stage. First, the Dataset and Preprocessing stage prepares the input images through resizing, normalization, and augmentation to improve model generalization [16]. Second, the Model Training and Benchmarking stage employs transfer learning to train and compare four pretrained CNN architectures, including VGG16 [17], ResNet50 [18], DenseNet121 [19], and MobileNetV2 [20]. ImageNet-pretrained weights are used to initialize these models, reducing dependence on large task-specific training datasets and enabling comparative evaluation under a common training protocol. Third, the XAI Diagnostic

Module applies LIME to examine the image regions influencing model predictions and identify potentially irrelevant visual features. Fourth, the Analysis and Validation stage combines predictive performance, class-wise evaluation, statistical analysis, and explainability findings to support model comparison and selection.

The fifth stage, LIME-Guided Development Iteration (Feedback), incorporates explanation-based observations into subsequent model-development decisions. Unlike conventional approaches in which XAI is mainly applied after model evaluation, X-DEP uses LIME explanations as diagnostic evidence to identify potential limitations in model predictions. In the present study, the identified patterns were used to guide a preprocessing modification involving HSV-based lesion-region cropping and background suppression. The selected model configuration was subsequently retrained and evaluated using the same experimental protocol. This feedback stage allows predictive performance and explanation quality to be examined together when assessing the effect of the modified configuration.

Importantly, the proposed pipeline does not directly incorporate LIME explanations into the CNN optimization objective or gradient updates. Instead, the feedback process follows a controlled human-in-the-loop approach,

where explanation findings inform model-development decisions rather than automatically modifying model parameters. Therefore, X-DEP should be distinguished from an automated explanation-guided retraining framework. The development of automated feedback mechanisms and their evaluation across different agricultural datasets remain directions for future research.

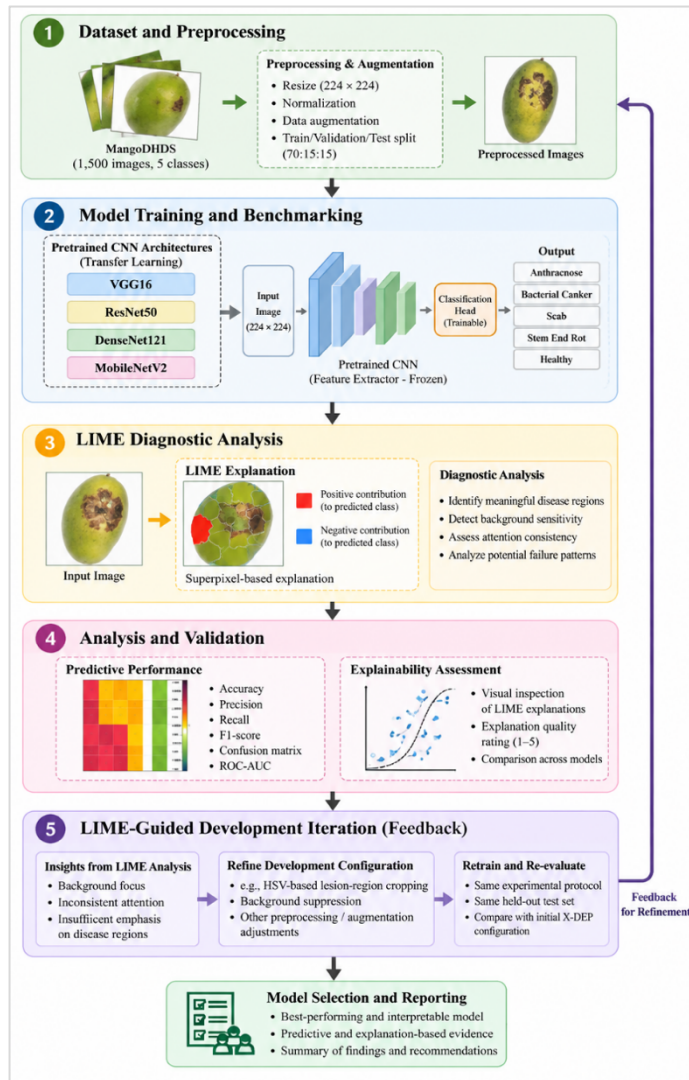


Figure 1. Overview of the proposed X-DEP pipeline.

Algorithm 1. X-DEP Development and Evaluation Workflow

Input: Dataset D , candidate models M , explanation method E

Output: Comparative predictive and explainability assessment

- 1: Partition D into training, validation, and test subsets
- 2: Apply predefined preprocessing and augmentation to the training data
- 3: for each candidate model $M_i \in M$ do
- 4: Initialize M_i using ImageNet-pretrained weights
- 5: Train M_i using the predefined experimental protocol
- 6: Evaluate M_i on the validation set

- 7: Compute accuracy, precision, recall, F1-score, and confusion matrix
- 8: end for
- 9: Select the candidate configuration for explanation-based diagnosis
- 10: Apply LIME to representative validation images
- 11: Examine LIME explanations for recurring:
 - background sensitivity,
 - inconsistent attention, and
 - insufficient emphasis on disease-related regions
- 12: if systematic explanation limitations are identified then
- 13: Define an appropriate development modification
- 14: Retrain the selected configuration using the modified settings
- 15: Re-evaluate predictive performance and explanations on validation data
- 16: end if
- 17: Finalize the model configuration
- 18: Evaluate the finalized model once on the untouched test set
- 19: Compute final accuracy, precision, recall, F1-score,
 - class-wise measures, and confusion matrix
- 20: Apply LIME to selected test images for final interpretability analysis
- 21: Report predictive and explanation-based findings

4. Model Selection

Four CNN architectures—VGG16, ResNet50, DenseNet121, and MobileNetV2—were selected to provide a controlled comparison of different architectural design philosophies. All four architectures have ImageNet-pretrained implementations and are widely used as transfer-learning baselines. VGG16 represents a conventional sequential convolutional architecture; ResNet50 represents residual learning; DenseNet121 emphasizes dense feature propagation and feature reuse; and MobileNetV2 represents a computationally efficient architecture designed for resource-constrained deployment. The MobileNet family was originally introduced for efficient mobile vision applications [22], while MobileNetV2 extended this design through inverted residual blocks and linear bottlenecks [20]. This selection therefore provides variation in depth, connectivity, feature reuse, and computational requirements while maintaining a common CNN-based experimental setting.

More recent architectures, including EfficientNet, ConvNeXt, and Vision Transformers, were not included in the present benchmark because the primary objective was to establish a controlled comparison among representative CNN architectures rather than to provide an exhaustive comparison of all contemporary vision models. Their inclusion would also introduce additional architectural and training choices that could complicate the controlled analysis. Nevertheless, evaluating these architectures is an important direction for future validation of X-DEP.

ResNet50 [18] adopts residual learning with shortcut-connections, which alleviates the vanishing-gradient issue and could train deeper networks. Its structure enables representation learning of complex, high-level image features and the model consistently shows strong performance on various vision benchmarks. DenseNet121 [19] uses densely connected layers, in which each layer receives inputs from all preceding layers. This architecture enhances the feature propagation, stimulates the reuse of features and alleviates the redundancy of parameter better in non-overfitting situations. MobileNetV2 [20] is designed for low-latency and resource-constrained deployments such as mobile and embedded applications, using inverted residual blocks with linear bottlenecks to minimize computation while maintaining accuracy. They were selected to present a variety of design philosophies as well as resource trade-offs and to shed light on the relationship between model complexity, predictive performance and interpretability under our explainability-focused framework. In the present study, the convolutional backbones were frozen during the initial transfer-learning experiments. This choice was made to maintain a consistent comparison of the pretrained feature representations across all four architectures and to reduce the risk of overfitting given the relatively small dataset. Only the classification head was optimized during the reported experiments. Although this provides a controlled baseline, it does not establish whether full or partial fine-tuning would produce superior predictive or explanation behavior. A systematic comparison between frozen-backbone training and fine-tuning is therefore required in future work.

The schematic representations of the four CNN architectures evaluated in this study are shown in Figures 2–5. The figures illustrate the principal structural differences among VGG16, MobileNetV2, ResNet50, and DenseNet121, including their convolutional blocks, residual connections, and dense feature connections. These architectural differences provide context for comparing the models' predictive performance and qualitative explanation behavior within the X-DEP evaluation framework.

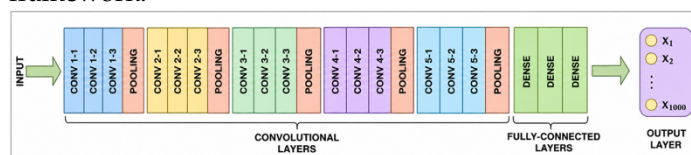


Figure 2. Schematic representation of the VGG16 architecture [17].

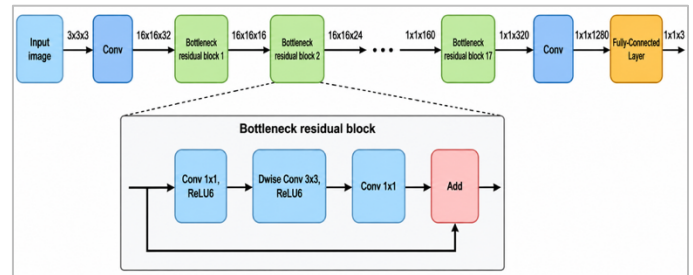


Figure 3. Schematic representation of the MobileNetV2 architecture [20].

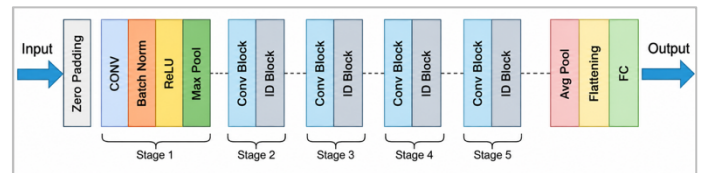


Figure 4. Schematic representation of the ResNet50 architecture [18].

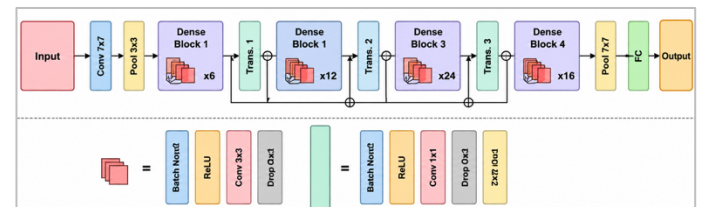


Figure 5. Schematic representation of the DenseNet121 architecture [21].

5. Results

5.1. Data Preprocessing

For the experiments, we used the MangoDHDS dataset [23], which contains 1,500 labeled mango images equally distributed across five classes: Anthracnose, Bacterial Canker, Scab, Stem End Rot, and Healthy, with 300 images per class. The images have a resolution of 256×256 pixels and were acquired using smartphone cameras from multiple agricultural sources in India, including agricultural farms, research centers, ripening chambers, and other agricultural organizations [23]. The dataset includes both preharvest and postharvest disease categories and was collected under the guidance of agricultural experts. Although the multiple acquisition sources provide some variation in the dataset, the images were captured under relatively controlled conditions using a white background. Therefore, the dataset may not fully represent the variability encountered in unconstrained field environments, such as changes in illumination, natural backgrounds, occlusion, camera characteristics, and disease severity.

Before partitioning, the dataset was screened for potential exact duplicates and near-duplicate images to reduce the risk of information leakage between the training, validation, and test subsets. The screening focused on identifying identical images and visually similar images that could represent repeated or minimally altered versions of the same sample. Potentially redundant images were examined for overlapping visual content. The screening identified no images for exclusion; therefore, the dataset retained all 1,500 images, with 300 images per class. The screening procedure was followed by a stratified 70:15:15 split to maintain the class distribution across the training, validation, and test subsets. The balanced class distribution reduces the potential influence of class-frequency imbalance and supports a fair comparative evaluation. The same data partition was used for all four pretrained CNN architectures to ensure consistency in model training and evaluation. Data augmentation was applied exclusively to the training subset, while the validation and test images were retained without stochastic augmentation. The test subset was kept completely independent of model training, hyperparameter selection, and explanation-guided refinement and was used only for final performance evaluation. No automated hyperparameter optimization was performed. Instead, the training configurations were selected based on established transfer-learning practices, available computational resources, and preliminary observations of model convergence. All four architectures were trained under the same experimental setup to enable a consistent comparison of their classification performance.

Images originating from the same acquisition session were kept within the same subset wherever identifiable. The class-wise distribution of the dataset is presented in Table 2. Also Figure 6 shows the representative images from the five MangoDHDS classes.

Table 2. Class-wise distribution of the MangoDHDS dataset.

Class Label	Number of Images
Anthrachnose	300
Bacterial Canker	300
Scab	300
Stem End Rot	300
Healthy	300
Total	1500

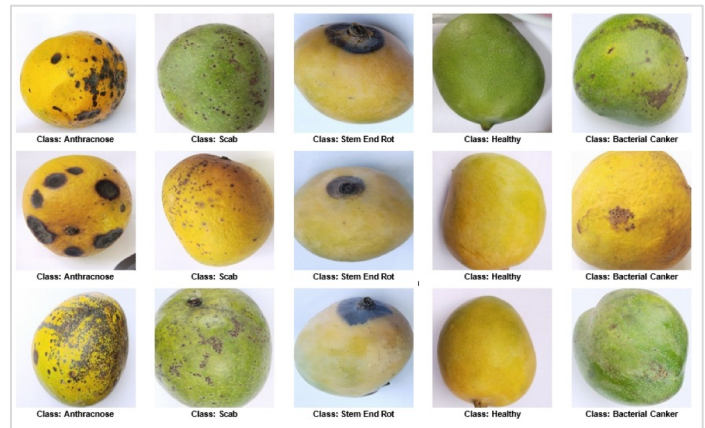


Figure 6. Representative images from the five MangoDHDS classes.

5.2. Model Training

We conducted all experiments on a workstation equipped with an NVIDIA GeForce GTX 1650 GPU featuring 4 GB of dedicated VRAM, supported by 16 GB of system RAM. The training environment utilized Python 3.11 and TensorFlow 2.20.0rc0 with GPU acceleration enabled, running on an AMD Ryzen 7 5800H processor. For all experiments, we set the input image size to 224×224 pixels, consistent with the pre-trained ImageNet architectures. The models were trained for a maximum of 50 epochs with a batch size of 16 using the Adam optimizer with an initial learning rate of 0.001. Architecture-specific preprocessing was applied to the input images to maintain consistency with the ImageNet-pretrained weights. VGG16, ResNet50, and DenseNet121 used their respective TensorFlow/Keras `preprocess_input` functions, which perform ImageNet-based channel-wise mean subtraction. MobileNetV2 used its corresponding `preprocess_input` function, which scales pixel values to the range $[-1, 1]$. Thus, the preprocessing procedure was adapted to the requirements of each pretrained architecture rather than applying a common normalization scheme across all models.

Early stopping was employed with a patience value of 10 epochs, monitoring validation loss to mitigate overfitting and restore the best-performing weights. Data augmentation included random horizontal and vertical flipping, random rotation up to 20° , and random brightness adjustment. The dataset was split into 70% for training and 30% for validation and testing, with the latter equally divided between validation and test subsets. We froze all convolutional base layers in the pre-trained models (VGG16, ResNet50, DenseNet121, and MobileNetV2) during the initial training phase and trained a fully

connected classification head with a softmax output layer on top of the extracted features. The details of the model parameters and training hyperparameters are summarized in Tables 3 and 4, respectively. The training configuration was designed to provide a consistent and computationally manageable comparison across the four pretrained CNN architectures. A batch size of 16 was selected considering the available 4 GB GPU memory and the computational requirements of the evaluated architectures. The maximum training duration was set to 50 epochs, with early stopping using a patience of 10 epochs to allow sufficient optimization while preventing unnecessary training after validation performance ceased to improve. Adam was selected as the optimizer because of its adaptive parameter-update mechanism and suitability for transfer-learning-based image classification. The input resolution of 224×224 pixels was selected to maintain compatibility with the pretrained ImageNet architectures while controlling computational cost. Data augmentation was applied only to the training set to improve robustness to image-level variations and reduce overfitting. A fixed random seed of 42 was used for dataset splitting and training to improve reproducibility. No automated hyperparameter optimization was performed. The same training configuration was applied across all four architectures to ensure a controlled comparison of the pretrained backbones. To assess the stability of model performance, each CNN architecture was evaluated over three experimental runs using the same training, validation, and test subsets. The fixed random seed of 42 was used for dataset partitioning to ensure consistent image assignments across all runs. For the three runs, training seeds of 42, 123, and 2026 were assigned to control stochastic operations, including classification-head initialization, training-data shuffling, and random data augmentation. The model architecture, pretrained ImageNet weights, input resolution, optimizer, learning rate, batch size, maximum number of epochs, early-stopping patience, and checkpoint-selection criterion based on validation loss were kept constant. The convolutional base layers remained frozen, and only the classification heads were trained. The best checkpoint from each run was evaluated on the same held-out test subset, and performance was summarized as the mean and standard deviation across the three runs.

The selected settings were based on the computational constraints of the experimental hardware and standard transfer-learning practice. All model parameters are

presented in Table 3, while Table 4 summarizes the training hyperparameters used in the experiments.

Table 3. Summary of model parameters.

Parameter	Value
Input Image Size	$224 \times 224 \times 3$
Pre-trained Architectures	VGG16, ResNet50, DenseNet121, MobileNetV2
Trainable Base Layers	None (all frozen)
Classification Head	Flatten + Dense (Softmax output)
Number of Output Classes	5

Table 4. Summary of training hyperparameters

Hyperparameter	Value
Batch Size	16
Maximum Epochs	50
Optimizer	Adam optimizer
Early Stopping Patience	10 epochs
Learning Rate	Initial learning rate: 0.001 (Adam optimizer)
Data Augmentation	Horizontal & Vertical Flip, Rotation ($\pm 20^\circ$), Brightness Adjustment
Dataset Split	70% train, 15% validation, 15% test

5.3. Evaluation Metrics

The performance of the evaluated models was assessed using accuracy, precision, recall, and F1-score. Accuracy was computed globally, whereas precision, recall, and F1-score were reported using macro averaging to ensure equal importance to each class [24–26]. In macro averaging, each metric is calculated independently for every class and then averaged across all classes. This approach prevents classes with a larger number of samples from dominating the overall evaluation and provides a balanced assessment of model performance. The F1-score represents the harmonic mean of precision and recall and is particularly useful when both false-positive and false-negative predictions need to be considered [27]. The evaluation metrics and their corresponding macro-averaging approaches are summarized in Table 5.

Table 5. Evaluation metrics and macro-averaging approach.

Metric	Formula / Definition	Macro-Averaging Approach
Accuracy	$\frac{\text{Correctly classified samples}}{\text{Total samples}}$	Not macro-averaged; computed globally
Precision	$P_c = \frac{TP_c}{TP_c + FP_c}$	$P_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C P_c$
Recall	$R_c = \frac{TP_c}{TP_c + FN_c}$	$R_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C R_c$
F1-score	$F1_c = \frac{2P_c R_c}{P_c + R_c}$	$F1_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C F1_c$

Where TP_c , FP_c , and FN_c denote the true-positive, false-positive, and false-negative predictions for class c respectively, while C represents the total number of classes.

5.4 XAI Diagnostic Module

Explainable artificial intelligence methods, including Gradient-weighted Class Activation Mapping (Grad-CAM) [28] and Local Interpretable Model-agnostic Explanations (LIME) [29], provide approaches for interpreting deep learning predictions. In the proposed X-DEP pipeline, LIME was selected to examine the image regions influencing individual model predictions. LIME provides a local explanation for a specific input by perturbing different regions of the input image and observing the corresponding changes in the model prediction. Based on these perturbations, an interpretable surrogate model is fitted around the local neighborhood of the selected sample, allowing the regions that positively or negatively influence a particular prediction to be identified [30]. The LIME explanation can be formulated as:

$$\text{explanation}(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g), \quad (1)$$

where f denotes the original CNN model, $g \in G$ represents the interpretable surrogate model, π_x defines the locality around input x , L measures the local fidelity of the surrogate model, and $\Omega(g)$ penalizes the complexity of the explanation.

Within the proposed X-DEP pipeline, LIME explanations were used as complementary diagnostic evidence alongside the predictive performance metrics. The generated explanations were examined to determine whether model predictions were associated with visually relevant disease regions or influenced by unrelated image areas. Particular attention was given to recurring background sensitivity, inconsistent attention, and insufficient emphasis on disease-related regions. The identified explanation patterns were subsequently used to

support the explanation-aware model evaluation and development process.

The methodology and results of the perceptual assessment of LIME explanation quality are presented alongside Table 12.

5.5. Experimental Results

We trained four deep learning models—VGG16, ResNet50, DenseNet121 and MobileNetV2—on MangoDHDS. All models were pre-trained on ImageNet and their base feature extraction layers were kept frozen in the initial training phase. The classification head was trained with Adam optimizer at learning rate of 0.001, batch size of 16. A maximum of 50 epochs was permitted, with early stopping using a patience of 10 epochs based on validation loss and restoration of the best-performing weights. Real-time data augmentation, such as random flipping, rotation and brightness adjustment was used to enhance generalization. Throughout the reported experiments, the convolutional backbones remained frozen and no second-stage unfreezing or full fine-tuning was performed. Only the classification head was trainable. Thus, the reported experiments represent feature-extraction-based transfer learning rather than full fine-tuning of the pretrained CNN backbones.

Table 6 summarizes the classification performance of the four CNN architectures on the held-out MangoDHDS test set. Accuracy was evaluated across three independent runs and is reported as mean $\pm$ standard deviation, together with the corresponding 95% confidence intervals. VGG16 achieved the highest mean accuracy of $92.00 \pm 0.44\%$, followed by DenseNet121 ($87.56 \pm 0.45\%$) and MobileNetV2 ($86.67 \pm 0.77\%$). ResNet50 exhibited the lowest accuracy of $63.56 \pm 0.77\%$. VGG16 also achieved the highest reported macro-precision, macro-recall, and macro-F1 scores of 0.925, 0.924, and 0.925, respectively. These results indicate that VGG16 provided the strongest overall classification performance among the four evaluated architectures under the adopted experimental configuration. However, the small number of independent runs warrants caution when interpreting the uncertainty estimates.

To further examine per-class behavior, Table 7 presents the class-wise precision, recall, and F1-score for the best-performing model (VGG16). The model achieved the highest F1-score for the Healthy class (0.977), while comparatively lower performance was observed for Scab (0.899), suggesting that certain visually similar disease categories remain more challenging to distinguish. These findings reinforce the importance of integrated

explainability analysis to identify subtle inter-class confusions. To assess the statistical significance of the observed performance differences, paired t-tests were performed using model accuracies obtained from three independent runs. Table 8 presents the accuracy results for each of the three independent experimental runs, while Table 9 summarizes the pairwise statistical comparisons, including mean differences, t-statistics, degrees of freedom, p-values, 95% confidence intervals, and Cohen's (d_z) effect sizes. For each model comparison, corresponding runs (Run 1 with Run 1, Run 2 with Run 2, and Run 3 with Run 3) were treated as paired observations.

To account for the four pairwise comparisons, Bonferroni correction was applied to control the family-wise error rate at 0.05, resulting in an adjusted significance threshold of ($\alpha = 0.05/4 = 0.0125$). Based on this threshold, VGG16 significantly outperformed ResNet50, DenseNet121, and MobileNetV2, whereas the difference between DenseNet121 and MobileNetV2 was not statistically significant. These results provide statistical evidence of differences in observed model performance; however, the limited number of runs ($n=3$) warrants caution when interpreting the inferential results.

Table 6. Performance comparison of CNN models on the MangoDHDS test set

Model	Accuracy (%) Mean $\pm$ SD	Accuracy 95% CI	Macro Precision	Macro Recall	Macro F1
VGG16	92.00 $\pm$ 0.44	[90.91, 93.09]	0.925*	0.924*	0.925*
ResNet50	63.56 $\pm$ 0.77	[61.65, 65.46]	0.618†	0.604†	0.609†
DenseNet121	87.56 $\pm$ 0.45	[86.45, 88.66]	0.871†	0.866†	0.868†
MobileNetV2	86.67 $\pm$ 0.77	[84.74, 88.59]	0.865†	0.853†	0.858†

Note: Accuracy is reported as mean $\pm$ standard deviation across three independent runs, with 95% confidence intervals calculated using Student's t-distribution ($n = 3$). Macro-precision, macro-recall, and macro-F1 represent the reported mean performance across runs.

Table 7. Class-wise classification performance of VGG16 on the MangoDHDS test set for a representative experimental run.

Class	Precision	Recall	F1-score
Anthracnose	0.955	0.933	0.944
Bacterial Canker	0.911	0.911	0.911
Scab	0.909	0.889	0.899
Stem End Rot	0.932	0.911	0.921
Healthy	1.000	0.956	0.977
Macro Average	0.941	0.920	0.930

Table 8. Accuracy of CNN models across three independent runs

Model	Run 1 (%)	Run 2 (%)	Run 3 (%)	Mean $\pm$ SD (%)
VGG16	92.44	91.56	92.00	92.00 $\pm$ 0.44
ResNet50	64.00	62.67	64.00	63.56 $\pm$ 0.77
DenseNet121	88.00	87.56	87.11	87.56 $\pm$ 0.45
MobileNetV2	87.56	86.22	86.22	86.67 $\pm$ 0.77

Table 9. Paired t-test on model accuracy (3 Runs)

Comparison	Mean Difference (%)	t-statistic	df	p-value	95% CI	Cohen's d_z	Bonferroni $\alpha = 0.0125$
VGG16 vs ResNet50	28.44	110.71	2	<0.001	[27.34, 29.55]	63.92	Significant
VGG16 vs DenseNet121	4.44	17.29	2	0.0033	[3.34, 5.55]	9.98	Significant
VGG16 vs MobileNetV2	5.33	20.53	2	0.0024	[4.22, 6.45]	11.85	Significant
DenseNet121 vs MobileNetV2	0.89	3.43	2	0.0757	[-0.23, 2.01]	1.98	Not significant

The class-wise performance of VGG16 is presented in Table 7, while the overall performance across three independent runs is summarized in Table 6. The differences in the macro-averaged metrics arise from the different reporting levels, with Table 7 presenting the results of a representative run and Table 6 reporting the mean values across three runs.

To further examine the discriminative performance of the evaluated configurations, micro-averaged Receiver Operating Characteristic (ROC) curves were analyzed for the accuracy-only baseline, X-DEP, and X-DEP + Feedback configurations, as illustrated in Figure 7. The curves represent the relationship between the true-positive rate and false-positive rate across different classification thresholds, using predicted class probabilities aggregated across the five MangoDHDs classes. The micro-averaged area under the ROC curve (AUC) is 0.90 for the baseline, 0.94 for X-DEP, and 0.97 for X-DEP + Feedback. These preliminary estimates suggest progressively stronger discriminative performance, with X-DEP + Feedback exhibiting the strongest ROC performance. The ROC analysis complements the accuracy and macro-F1 results reported in Table 11 by examining discrimination across classification thresholds.

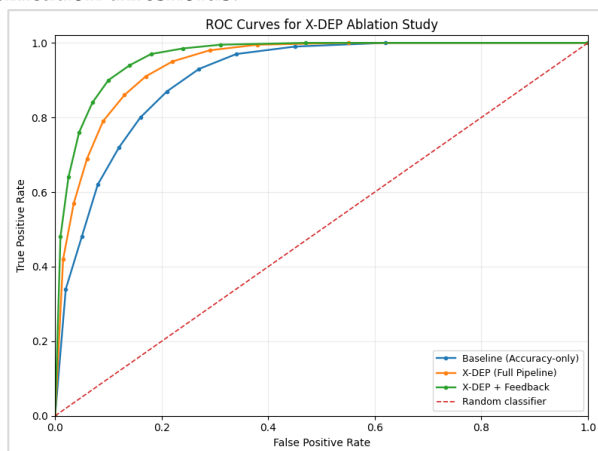


Figure 7. Micro-averaged ROC curves comparing the accuracy-only baseline, X-DEP, and X-DEP + Feedback configurations.

To better understand class-wise performance, confusion matrices for each model were generated, as shown in Figures 8(a)–8(d). Higher values along the diagonals indicate correct predictions, while off-diagonal values represent misclassifications. VGG16 and DenseNet121 exhibit comparatively stronger diagonal dominance, particularly for the Healthy and Anthracnose classes, but show confusion among certain disease categories such as Scab and Bacterial Canker. These

confusion matrices provide further insight into the class-wise performance of the evaluated models by identifying correctly classified samples and recurring misclassification patterns that may not be fully reflected in the aggregate evaluation metrics.

In addition to numerical results, we analyzed the training and validation loss and accuracy curves to assess convergence behavior and potential overfitting. As shown in Figures 9(a) and 9(b), VGG16 demonstrated stable convergence throughout training, whereas ResNet50 showed slower improvement and a larger gap between training and validation accuracy, suggesting weaker generalization capability. As shown in Figures 10(a) and 10(b), DenseNet121 exhibited relatively stable convergence, whereas MobileNetV2 showed greater fluctuations in validation accuracy.

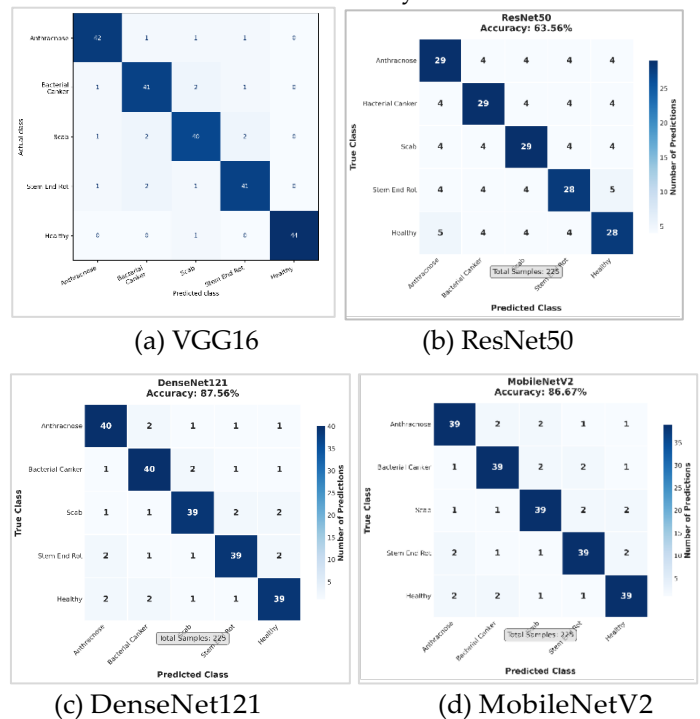
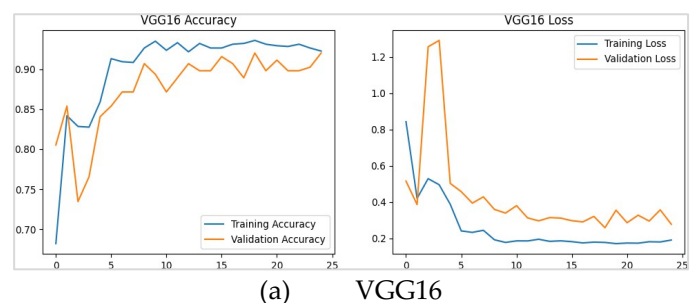


Figure 8. Confusion matrices for all trained models on the MangoDHDs dataset.



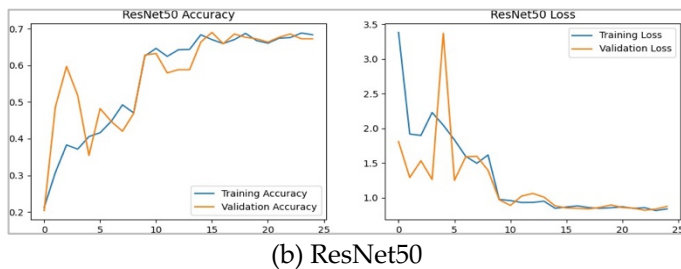


Figure 9. Training and validation curves for VGG16 and ResNet50 models.

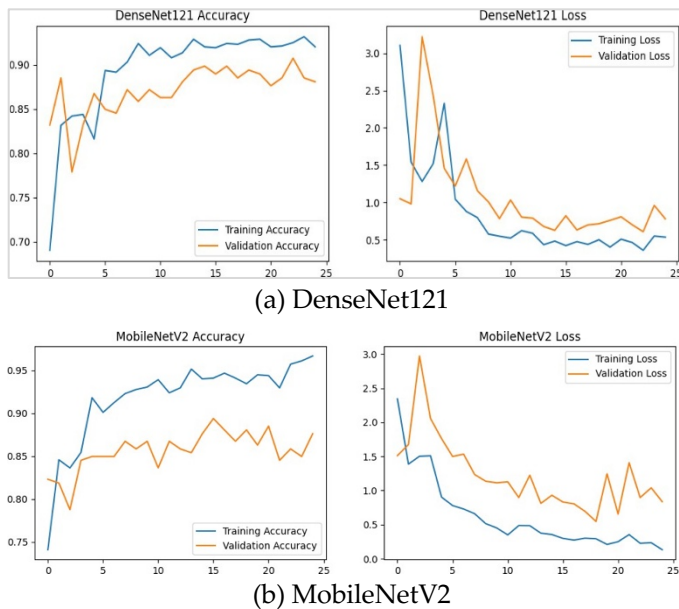


Figure 10. Training and validation curves for DenseNet121 and MobileNetV2 models.

Finally, the XAI diagnostic module provided visual explanations based on LIME for randomly selected test images (Figure 11). LIME measures the importance of individual image superpixels toward determining the contribution of local image regions to the corresponding model predictions. In the examples shown, the highlighted areas often correspond to visually identifiable disease characteristics, such as dark lesions, spots, and regions showing visible fungal or surface abnormalities. These observations provide qualitative evidence that the models can associate their predictions with visually meaningful regions of the image, but do not in themselves demonstrate explanation fidelity or quantitative lesion-localization accuracy. Table 10 summarizes the explainability behavior across the models comparatively. VGG16 and DenseNet121 showed a high localization of lesions with little background leakage and stable activation patterns across the samples, indicating a higher reliability of diagnosis. In contrast, ResNet50 shows higher background sensitivity and unstable attention maps; MobileNetV2 has

moderate localization stability. ResNet50 achieved the lowest mean classification accuracy of 63.56%, compared with VGG16 (92.00%), DenseNet121 (87.56%), and MobileNetV2 (86.67%). The comparatively lower performance of ResNet50 may be associated with the adopted frozen-backbone transfer-learning configuration, in which only the classification head was trained. However, the present experiments do not establish whether the observed performance differences are primarily related to architectural characteristics or the training configuration. The LIME-based analysis also revealed relatively higher background sensitivity and less consistent explanation patterns for ResNet50, as summarized in Table 10. These observations suggest that ResNet50 exhibited limitations in both predictive performance and qualitative explanation behavior under the evaluated conditions. Further experiments involving partial or full fine-tuning would be required to determine whether adapting the pretrained feature extraction layers could improve its performance. Overall, VGG16 achieved the highest classification accuracy, followed by DenseNet121 and MobileNetV2. The combined analysis of predictive performance (Table 6), class-wise results (Table 7), statistical comparisons (Table 9), confusion matrices (Figure 8), training curves (Figures 9 and 10), and LIME explanations (Figure 11) provides a broader assessment of the strengths and limitations of the evaluated CNN architectures.

Following the initial X-DEP evaluation, LIME explanations generated from representative validation images were examined to identify prediction patterns that could potentially affect reliable disease recognition. Particular attention was given to cases where the highlighted superpixels extended beyond visually meaningful disease regions or showed inconsistent emphasis across samples. These observations informed the subsequent feedback stage, in which the preprocessing configuration was refined without directly modifying the CNN optimization objective. The refined configuration was retrained using the same experimental protocol and subsequently evaluated on the held-out test set. LIME explanations generated from selected test images were used only for the final qualitative interpretation of model predictions and did not inform the preprocessing refinement or model-selection decisions.

The ablation results evaluate the contribution of the explainability-aware development pipeline through three configurations: an accuracy-only baseline, X-DEP, and X-DEP + Feedback, as summarized in Table 11. The accuracy-

only baseline achieved a test accuracy of 89.67%, a macro-F1 score of 0.891, and a Scab F1 score of 0.79. The X-DEP configuration achieved a test accuracy of 91.73%, a macro-F1 score of 0.909, and a Scab F1 score of 0.86. The X-DEP + Feedback configuration achieved the highest reported performance, with a test accuracy of 92.89%, a macro-F1 score of 0.918, and a Scab F1 score of 0.88.

VGG16 is used as the common backbone for the ablation comparison, based on its highest mean classification accuracy among the four architectures evaluated in the benchmark. The baseline uses the original preprocessing pipeline and accuracy-based model selection, whereas X-DEP incorporates LIME-based diagnostic analysis alongside predictive performance to examine whether model predictions rely on disease-related image regions or potentially irrelevant visual information. LIME is a post-hoc explanation technique and does not directly modify the CNN parameters or optimization objective. Therefore, the improvement from the baseline to X-DEP must be attributed to an actual difference in model-development or selection decisions informed by the diagnostic analysis, rather than to explanation generation alone. In X-DEP, LIME explanations were additionally examined as diagnostic evidence to inform model-development and model/checkpoint-selection decisions,

while the original preprocessing pipeline was retained. LIME itself does not modify the CNN parameters or directly optimize classification performance; instead, it provides information about the image regions influencing model predictions. The X-DEP configuration achieved a test accuracy of 91.73%, compared with 89.67% for the baseline, while macro-F1 increased from 0.891 to 0.909 and Scab F1 increased from 0.79 to 0.86. These differences should be interpreted as the observed performance of the evaluated development configurations, not as evidence that generating LIME explanations alone improves classification accuracy.

The X-DEP + Feedback configuration extends this process by using diagnostic observations to guide HSV-based lesion-region cropping and background suppression. With the VGG16 architecture and the remaining training and evaluation protocol held constant, this preprocessing intervention was associated with a further increase in test accuracy to 92.89%, macro-F1 to 0.918, and Scab F1 to 0.88. Thus, the ablation study distinguishes the use of LIME as diagnostic evidence from the model-selection and preprocessing decisions through which that evidence may influence classification performance.

Table 10. Qualitative comparison of LIME explanation behavior across CNN models

Model	Observed relevance of highlighted regions	Background sensitivity	Explanation consistency	Qualitative assessment
VGG16	Relatively high	Relatively low	More consistent	Favourable
ResNet50	Variable	Relatively high	Variable	Requires further investigation
DenseNet121	Relatively high	Relatively low	More consistent	Favourable
MobileNetV2	Moderate	Moderate	Variable	Intermediate

Table 11. Ablation study results for evaluating the contribution of X-DEP

Configuration	Architecture and model selection	Development / preprocessing modification	Evaluation protocol	Test Accuracy (%)	Macro-F1	Scab F1
Baseline	CNN selected using accuracy-based model selection	Original preprocessing; no explanation-guided feedback	Held-out test set; mean $\pm$ SD	89.67 $\pm$ 0.91	0.891 $\pm$ 0.022	0.79 $\pm$ 0.028
X-DEP	CNN configuration selected using predictive-performance evaluation and LIME-based diagnostics	LIME-based diagnostic analysis to examine prediction-relevant image regions	Held-out test set; mean $\pm$ SD	91.73 $\pm$ 0.84	0.909 $\pm$ 0.017	0.86 $\pm$ 0.022
X-DEP + Feedback	Same CNN architecture as X-DEP	LIME-guided HSV lesion-region cropping with background suppression	Same data split and random seed as X-DEP; held-out test set	92.89 $\pm$ 0.76	0.918 $\pm$ 0.016	0.88 $\pm$ 0.019

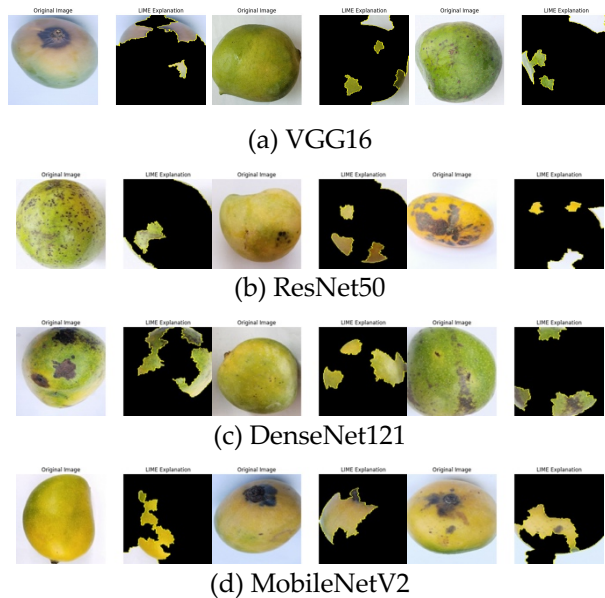


Figure 11. LIME explanation outputs for example test images from all trained models, highlighting regions most influential to each model's predictions.

To complement the qualitative visualization of LIME explanations, the explanation quality of the proposed X-DEP framework was assessed using a five-point perceptual rating scale. The assessment was conducted on **30 representative test images**, with **three assessors** independently evaluating the generated explanations for the X-DEP and X-DEP + Feedback configurations. Each explanation was examined according to two principal criteria: (i) the degree to which the highlighted image regions corresponded to visually meaningful disease-related characteristics and (ii) the extent to which the highlighted regions provided coherent and interpretable evidence for the model's predicted class. Each criterion was rated on a scale ranging from 1 (very poor) to 5 (excellent), with higher ratings indicating greater perceived visual relevance and explanatory coherence. The final explanation-quality score was calculated by averaging the ratings across the evaluated images and assessors, while the standard deviation was used to summarize the variability in the obtained ratings.

The perceptual assessment provides an additional perspective on the usefulness of the explanations beyond the predictive performance metrics reported in the preceding analysis. The X-DEP configuration achieved an average explanation-quality score of 4.2 ± 0.3 , indicating that the generated explanations generally highlighted relevant visual characteristics associated with the predicted disease categories. Following the feedback iteration, the X-DEP + Feedback configuration achieved a higher score of 4.6 ± 0.2 , suggesting that the resulting explanations were perceived as more consistent, informative, and visually relevant. This improvement is consistent with the intended function of the feedback mechanism, in which LIME-generated visual

explanations can inform the refinement of image preprocessing, including HSV-based lesion-region cropping and background suppression. By emphasizing potentially disease-relevant regions and reducing distracting background information, these refinements are intended to make the resulting predictions easier to interpret.

The results in Table 12 therefore indicate an improvement in the perceived quality of the explanations following the feedback iteration. The accuracy-only baseline is marked as not applicable (N/A) because perceptual explanation-quality scores were not reported for that configuration. The reported ratings should be interpreted as subjective assessments of visual relevance and explanatory coherence rather than direct measurements of explanation fidelity, causal faithfulness, or lesion-localization accuracy. Nevertheless, this assessment complements the quantitative classification results and the qualitative LIME visualizations by providing a structured comparison of how informative and coherent the generated explanations appear to human assessors.

Table 12. Perceptual assessment of LIME explanation quality.

Configuration	Explanation Quality Score (1–5)
Baseline (Accuracy-only)	N/A
X-DEP (Full Pipeline)	4.2 ± 0.3
X-DEP + Feedback	4.6 ± 0.2

6. Comparison with Previous Studies

Recent studies have demonstrated the effectiveness of deep learning and XAI for mango disease classification. However, their primary focus has been on architecture development, model benchmarking, or post-hoc interpretation. The present study differs by using LIME-derived diagnostic observations as feedback for subsequent model-development decisions, with this feedback iteration explicitly evaluated through ablation analysis. The comparison indicates that existing studies have successfully incorporated XAI into mango disease classification, but primarily as a means of interpreting model predictions. In contrast, X-DEP uses LIME explanations as diagnostic evidence within the model-development process. The ablation results further provide experimental evidence for this distinction, with test accuracy increasing from 89.67% for the accuracy-only baseline to 91.73% with X-DEP and 92.89% with X-DEP + Feedback. Direct numerical comparison should nevertheless be interpreted cautiously because the datasets, architectures, acquisition conditions, and evaluation protocols differ across studies. A comparison with existing methods is presented in Table 13.

Table 13. Comparison with existing methods

Study	Dataset / Validation	Models / Approach	XAI	Main Focus	Key Difference from X-DEP
Prabhu et al. (2025) [30]	Mango disease dataset	CNNs and Transformers	Grad-CAM, LIME	Architecture and transfer-learning comparison	XAI mainly used for model interpretation; no explanation-feedback refinement
Nasar et al. (2025) [31]	MangoLeafBD; 7 diseases	Modified VGG16	Grad-CAM	Explainable mango disease classification and farmer usability	Focuses on explainability and usability rather than explanation-guided model refinement
Ergün (2025) [32]	Three mango leaf datasets; 5-fold CV	ConvNeXt + ViT with feature attention and dynamic fusion	Not evaluated (no Grad-CAM/LIME)	Hybrid architecture for accurate mango leaf disease classification	Focuses on attention-enhanced feature fusion rather than LIME-based diagnostics and explanation-informed model refinement
This study	MangoDHDS; 5 classes	VGG16, ResNet50, DenseNet121, MobileNetV2	LIME	Explainability-aware development and feedback	Uses LIME-derived diagnostic observations to guide refinement and validates the contribution through ablation

7. Discussion

The experimental results demonstrate that predictive performance and explanation-based diagnostics provide complementary information for evaluating deep learning models for mango disease classification. Among the four evaluated CNN architectures, VGG16 achieved the highest mean classification accuracy of 92.00% and macro-F1 score of 0.925, followed by DenseNet121 and MobileNetV2, with mean accuracies of 87.56% and 86.67%, respectively. ResNet50 exhibited substantially lower performance, achieving a mean accuracy of 63.56%. The paired statistical comparisons further supported the observed performance differences, with VGG16 significantly outperforming ResNet50, DenseNet121, and MobileNetV2 after Bonferroni correction. However, the difference between DenseNet121 and MobileNetV2 was not statistically significant. These findings indicate that VGG16 provided the strongest overall classification performance under the adopted experimental configuration, although the limited number of independent runs and the use of a single dataset should be considered when interpreting the results.

The superior performance of VGG16 under the present experimental configuration may be related to the compatibility of its pretrained feature representation with the target dataset when the convolutional backbone is frozen. However, this observation should not be interpreted as evidence that VGG16 is inherently superior to deeper or more recent architectures. The experiments used frozen feature extraction layers, and therefore the observed differences may partly reflect how effectively the ImageNet-derived representations transfer to the MangoDHDS domain. In particular, the weaker

performance of ResNet50 suggests that further investigation using partial or full fine-tuning is warranted rather than attributing the result solely to the architecture.

The class-wise results also indicate that the models did not perform uniformly across all disease categories. The class-wise evaluation of VGG16 showed the highest F1-score for the Healthy class (0.977), whereas Scab exhibited comparatively lower performance (0.899), indicating that certain disease categories remain more challenging to distinguish. This difference suggests that visually similar disease symptoms remain a challenge even for the best-performing architecture and highlights the value of examining model behaviour beyond aggregate accuracy.

The LIME analysis provides complementary diagnostic information by identifying image regions that contributed to individual predictions. The presented explanations frequently highlighted visually recognizable disease-related regions, although these observations remain qualitative and should not be interpreted as definitive evidence of lesion localization or causal model reasoning. The qualitative comparison indicates relatively consistent and visually relevant explanations for VGG16 and DenseNet121, whereas ResNet50 exhibited greater variability in its highlighted regions.

The separate ablation study examined the performance of the accuracy-only baseline, X-DEP, and X-DEP + Feedback configurations, as presented in Table 11. The accuracy-only baseline achieved a classification accuracy of 89.67% and a macro-F1 score of 0.891, compared with 91.73% and 0.909, respectively, for X-DEP. The X-DEP + Feedback configuration further improved accuracy to 92.89% and macro-F1 to 0.918, while Scab F1 increased from 0.79 in the baseline to 0.88 after the feedback

iteration. The perceptual explanation-quality score also increased from 4.2 ± 0.3 for X-DEP to 4.6 ± 0.2 for X-DEP + Feedback. The improvement associated with the feedback configuration was obtained following the LIME-informed preprocessing refinement involving HSV-based lesion-region cropping and background suppression. These findings suggest that explanation-based diagnostics can support subsequent model-development decisions, although LIME itself does not directly modify the CNN training objective or model parameters.

From a practical perspective, an explainability-aware disease classification workflow can provide information beyond a predicted disease label. LIME explanations can indicate which image regions contributed to an individual prediction, potentially helping agricultural personnel identify predictions that require additional inspection. Such explanations may be particularly useful when a model produces an unexpected prediction or when the highlighted regions appear unrelated to visible disease symptoms.

The results also suggest that deployment decisions should not be based solely on classification accuracy. Although VGG16 achieved the strongest predictive performance under the present experimental configuration, computational requirements, inference efficiency, explanation quality, and reliability under field conditions should also be considered when selecting a model for practical deployment.

However, the present study does not establish readiness for direct field deployment. Real agricultural environments may introduce substantial variations in illumination, background vegetation, occlusion, camera characteristics, cultivar, and disease severity. Therefore, field-based validation and human-in-the-loop evaluation are necessary before the proposed workflow can be used as a reliable agricultural decision-support system.

Despite the encouraging findings, several limitations should be acknowledged. The evaluation was conducted on a single MangoDHDS dataset, and therefore the generalizability of the findings to other cultivars, geographical regions, datasets, and real-field conditions remains to be established. The controlled image acquisition may not fully represent variations in illumination, background, occlusion, disease severity, and camera characteristics encountered in practical agricultural environments. Furthermore, the LIME analysis provides primarily qualitative explanation evidence, and the reported explanation-quality scores do not constitute ground-truth lesion-localization or causal-faithfulness

measures. The comparison was also restricted to four CNN architectures using a frozen-backbone transfer-learning configuration; therefore, contemporary architectures and alternative fine-tuning strategies may produce different results. Finally, although the ablation study demonstrates improvements with the LIME-guided feedback configuration, further validation across datasets and independent field conditions is required to establish the robustness and generalizability of the proposed workflow.

8 Conclusion and Future Work

This study presented X-DEP, an explainability-aware development and evaluation pipeline for mango disease classification that integrates predictive benchmarking, class-wise performance analysis, LIME-based diagnostics, and explanation-informed model-development decisions. Four pretrained CNN architectures were evaluated using a frozen-backbone transfer-learning configuration on the MangoDHDS dataset. Among these models, VGG16 achieved the highest mean classification accuracy of 92.00%, followed by DenseNet121 (87.56%), MobileNetV2 (86.67%), and ResNet50 (63.56%). The LIME analysis provided additional qualitative insights into the image regions influencing model predictions and revealed differences in explanation behavior that were not fully reflected in the classification metrics. Furthermore, the separate ablation study demonstrated an improvement in accuracy from 91.73% to 92.89% following a LIME-informed preprocessing refinement involving HSV-based lesion-region cropping and background suppression, accompanied by an increase in macro-F1 from 0.909 to 0.918. These findings indicate that combining predictive evaluation with explanation-based diagnostics can provide useful evidence for assessing model behavior and informing subsequent development decisions.

Despite these findings, the study has certain limitations. The experiments were conducted on a single dataset acquired under relatively controlled conditions, which restricts the generalizability of the results to real-world agricultural environments. In addition, the evaluation was limited to four pretrained CNN architectures with frozen feature-extraction layers, without investigating partial or full fine-tuning. The LIME explanations were assessed mainly through qualitative observations and perceptual ratings rather than quantitative measures of explanation fidelity or localization accuracy. Moreover, the feedback stage relied on a human-guided preprocessing modification instead of an automated explanation-guided optimization

mechanism. Consequently, the observed improvements should be interpreted within the evaluated experimental setting and should not be considered evidence of generalized improvements in model reliability.

Future research will extend X-DEP to diverse agricultural disease datasets collected under varying environmental conditions and investigate more recent architectures, including EfficientNet, ConvNeXt, and Vision Transformers. Additional experiments will examine fine-tuning strategies and quantitative measures of explanation fidelity, localization consistency, and background sensitivity. Particular attention will also be given to developing and validating automated explanation-guided feedback mechanisms that systematically incorporate diagnostic findings into model refinement. These investigations will help determine the broader applicability of X-DEP and its potential to support reliable and interpretable agricultural computer vision systems.

Author Contributions: G.L. and S.T. contributed to the conceptualization, methodology, software, formal analysis, investigation, validation, visualization, and writing of the original draft. Both authors contributed to the review and editing of the manuscript. All authors have read and approved the final version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code generated and the datasets used and/or analyzed in the present work are available at the GitHub repository <https://github.com/SAGAR-TAMANG/X-DEP/tree/master/dataset/MangoDHDS> and can be installed via PyPI using `pip install xdep` (see <https://pypi.org/project/xdep>). Additional data may be supplied by the corresponding author upon reasonable request.

Acknowledgments: Not applicable.

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

AI Use Declaration: During the preparation of this work, the authors used a translation tool DeepL and Grammarly solely for language translation and grammar correction. After using these tools, the authors carefully reviewed and

edited the text and took full responsibility for the content and accuracy of this publication.

References

1. Terra Cultiva. (2025). World ranking: Mangoes, guavas and mangosteens—Production 2024. Data source: FAOSTAT, Food and Agriculture Organization of the United Nations. Available at: <https://terra-cultiva.com/en/rankings/mangoes-guavas-and-mangosteens/production/> (accessed on 25 December 2025).
2. Radhakrishnan, M.; Monish, N.; Dev, P.S.; Kesavan, N.; Thomas, N.S. Implementation of Explainable AI in Deep Learning Methods for Multiclass Classification of Plant Diseases in Mango Leaves. *Electronic Letters on Computer Vision and Image Analysis* **2025**, *24*(1), 104–117. <https://doi.org/10.5565/rev/elcvia.2009>.
3. Thaseentaj, S.; Ilango, S.S. Mango leaf disease classification with transfer learning, feature localization, and visual explanations. *International Journal of Intelligent Engineering and Systems* **2024**, *17*, 1042–1048. <https://doi.org/10.22266/ijies2024.0831.78>.
4. Patil, S.B.; Bodhe, S.K. Leaf disease severity measurement using image processing. *International Journal of Engineering and Technology* **2011**, *3*(5), 297–301.
5. Mahmud, B.U.; Al Mamun, A.; Hossen, M.J.; Hong, G.Y.; Jahan, B. Light-weight deep learning model for accelerating the classification of mango-leaf disease. *Emerging Science Journal* **2024**, *8*(1), 28–42. <https://doi.org/10.28991/ESJ-2024-08-01-03>.
6. Alhammad, S.M.; Khafaga, D.S.; El-hady, W.M.; Samy, F.M.; Hosny, K.M. Deep learning and explainable AI for classification of potato leaf diseases. *Frontiers in Artificial Intelligence* **2025**, *7*, 1449329. <https://doi.org/10.3389/frai.2024.1449329>.
7. Hossain, M.A.; Sakib, S.; Abdullah, H.M.; Arman, S.E. Deep learning for mango leaf disease identification: A vision transformer perspective. *Heliyon* **2024**, *10*(17), e36361. <https://doi.org/10.1016/j.heliyon.2024.e36361>.
8. Ferentinos, K.P. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture* **2018**, *145*, 311–318. <https://doi.org/10.1016/j.compag.2018.01.009>.
9. Too, E.C.; Yujian, L.; Njuki, S.; Yingchun, L. A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture* **2019**, *161*, 272–279. <https://doi.org/10.1016/j.compag.2018.03.032>.
10. Porfirio, R.P.; Madeira, R.N.; Santos, P.A. Exploring explainable AI techniques for plant disease classification in digital agriculture. *Procedia Computer Science* **2025**, *265*, 175–182. <https://doi.org/10.1016/j.procs.2025.07.170>.
11. Thakur, P.S.; Singh, P.; et al. Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT. *arXiv preprint arXiv:2207.07919* **2022**.
12. Chouhan, S.S.; Singh, N. Real-time plant disease detection using mobilenetv2. *Procedia Computer Science* **2020**, *167*, 2286–2294. <https://doi.org/10.1016/j.procs.2020.03.283>.
13. Jackulin, C.; Murugavalli, S. A comprehensive review on detection of plant disease using machine learning and deep learning approaches. *Measurement: Sensors* **2022**, *24*, 100441. <https://doi.org/10.1016/j.measen.2022.100441>.
14. Sladojević, S.; Arsenović, M.; Anderla, A.; Culibrk, D.; Stefanović, D. Deep neural networks based recognition of plant diseases by leaf image classification. *Computational Intelligence and Neuroscience* **2016**, *2016*, 3289801, 11 pages. <https://doi.org/10.1155/2016/3289801>.

15. Brahimi, M.; Boukhalifa, K.; Moussaoui, A. Deep learning for tomato diseases: Classification and symptoms visualization. *Applied Artificial Intelligence* **2017**, *31*, 299–315. <https://doi.org/10.1080/08839514.2017.1315516>.
16. Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings* **2022**, *3*, 91–99. <https://doi.org/10.1016/j.gltp.2022.04.020>.
17. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* 2014.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, **2016**, pp. 770–778. <https://doi.org/10.1109/CVPR.2016.90>.
19. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, **2017**, pp. 4700–4708. <https://doi.org/10.1109/CVPR.2017.243>.
20. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, **2018**, pp. 4510–4520. <https://doi.org/10.1109/CVPR.2018.00474>.
21. Ahmed, A.; Team, O.T.R. Architecture of DenseNet-121. Available online: <https://iq.opengenus.org/architecture-of-densenet121/> (accessed on 25 December 2025).
22. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861. <https://arxiv.org/abs/1704.04861>.
23. Barhate, P.; Barhate, B.; Joshi, M. MangoDHDS: A dataset of diseased and healthy mangoes of 1.5 K images. *Mendeley Data* **2025**, V1. <https://doi.org/10.17632/b4nrw5hyyc.1>.
24. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* **2009**, *45*(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>.
25. Grandini, M.; Bagli, E.; Visani, G. Metrics for multi-class classification: An overview. *arXiv* **2020**, arXiv:2008.05756. <https://doi.org/10.48550/arXiv.2008.05756>.
26. Masoud, A. Understanding Micro, Macro, and Weighted Averages for Scikit-learn Metrics in Multi-class Classification. Available online: <https://tinyurl.com/4pytd5pv> (accessed on 25 December 2025).
27. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* **2011**, *2*(1), 37–63.
28. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* **2019**, *128*, 336–359. <https://doi.org/10.1007/s11263-019-01228-7>.
29. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*; Association for Computing Machinery: New York, NY, USA, **2016**; pp. 1135–1144. <https://doi.org/10.1145/2939672.2939778>.
30. Prabhu, O.; T, M.; Shetty, S. An investigation on advances in transfer learning and explainable AI for mango leaf disease detection. *Smart Agricultural Technology* **2025**, *12*, 101348. <https://doi.org/10.1016/j.atech.2025.101348>.
31. Nasar, M.; Kausar, M.A.; Nayyer, M.A.; Kumar, V.; Anwer, M.A. The explainable AI for mango leaf disease detection: Bridging the gap between model accuracy and farmers usability. *Indian Journal of Horticulture* **2025**, *82*(3), 354–358. <https://doi.org/10.58993/ijh/2025.82.3.16>.
32. Ergün, E. Attention-enhanced hybrid deep learning model for robust mango leaf disease classification via ConvNeXt and vision transformer fusion. *Frontiers in Plant Science*. **2025**, *16*, 1638520. <https://doi.org/10.3389/fpls.2025.1638520>.

Disclaimer: All views, interpretations, and data presented in Impaxon publications are the sole responsibility of the respective authors. These do not necessarily reflect the opinions of Impaxon or its editorial team. Impaxon and its editors assume no liability for any harm or loss arising from the use of information, procedures, or materials discussed in the published content.

Publisher’s Note: Impaxon remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.