Future Technology
Open Access Journal
ISSN 2832-0379

November 2026| Volume 05 | Issue 04 | Pages 246-256

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.4.20

Future Publishing LLC

Article

XLeaf-Net: an explainable deep learning framework for mango leaf disease classification and severity assessment using Grad-CAM-based visual interpretation

Shweta Amar Satao[1]*, Swati R. Maurya[2]

[1]Department of CS-IT, Somaiya School of Basic and Applied Sciences, Somaiya Vidyavihar University, Mumbai, Maharashtra, India
[2]Department of Information Technology and Computer Science, Somaiya School of Basic and Applied Sciences, Somaiya Vidyavihar University, Mumbai, Maharashtra, India

ARTICLE INFO

Article history:
Received 05 April 2026
Received in revised form
01 September 2026
Accepted 18 September 2026

Keywords:
XLeaf-Net, Mango leaf disease, Deep learning, Severity assessment, Grad-CAM, Explainable artificial intelligence

*Corresponding author
Email address:
shweta.satao@somaiya.edu

DOI: 10.55670/fpll.futech.5.4.20

ABSTRACT

Accurate mango leaf disease recognition is important for automated crop-health assessment, but reliable evaluation requires transparent data partitioning, appropriate baseline comparison, and cautious interpretation of severity and explainability outputs. This study presents XLeaf-Net, an explainable deep-learning framework for mango leaf condition classification, image-derived severity assessment, and Grad-CAM-based visual interpretation. Experiments were conducted on the balanced MangoLeafBD dataset using an image-level stratified train-validation-test protocol with file-level integrity and capture/session-group overlap checks. XLeaf-Net was trained from scratch and compared with from-scratch VGG16, ResNet50, and EfficientNetB0 models under the implemented model-specific training configurations. An auxiliary severity model used conservative, image-derived symptom-intensity proxy labels, while Grad-CAM was applied to qualitatively examine class-discriminative regions. XLeaf-Net achieved 98.67% image-level test accuracy and a 98.66% macro F1-score, outperforming the evaluated from-scratch baselines while using 2.52 million parameters. The severity model achieved 78.17% accuracy and an 80.89% macro F1-score. Grad-CAM visualizations frequently overlapped with visually symptomatic leaf regions in representative correctly classified images. These findings demonstrate strong image-level performance within the implemented protocol; however, severity labels were not expert-annotated, and the evaluation was not session-independent. XLeaf-Net should therefore be considered a compact candidate framework requiring further independent validation.

1. Introduction

CNC The production of mango is important in tropical and subtropical agriculture, and healthy leaves play an important role in photosynthesis, vegetative growth, fruit growth, orchard productivity, and quality of the orchard crop. Mango leaves can suffer from a variety of diseases and pest damage with distinctive symptoms including necrotic lesions, chlorotic spots, margins damaged by pests, powdery fungal growth, sooty surface growth, and structural damage. Appearance can shift significantly as the disease progresses, during leaf maturity, with changing light conditions, background colors, image-acquisition parameters and disease severity, and can be challenging to recognize without expert knowledge. As such, computer vision and deep learning have become a major focus of interest in automated plant-health assessment due to their ability to learn discriminative visual representations directly from leaf images. In the context of mango disease analysis, recent studies have gradually shifted from traditional categorical disease classification to architectures with interpretability, focus, computational efficiency, and decision-support-related image analysis. For example, approaches like MangoLeafNet-XAI demonstrate the larger trend of creating interpretable visualizations of mango leaf conditions that accompany predictive performance [1].

Despite these advances, automated mango leaf diagnosis remains methodologically challenging. Traditionally, diagnosis is made in the field and relies heavily on visual recognition, but its accuracy depends heavily on the observer's experience, the disease stage, environmental conditions, and symptom similarity. While automated systems can help reduce this subjectivity, they must resolve visually similar classes while tolerating relatively high within-class variation. The predictions of the model may be further influenced by background correlations, acquisition artifacts, variations in the distribution of symptoms, and variations in the visibility of the symptom intensity. While machine-learning and deep-learning approaches have been used for mango disease recognition, there is significant variability in the way these datasets were prepared, partitioned, baseline selected, interpreted, and evaluated [2]. A merely high reported classification accuracy does not provide a sound basis for a diagnostic framework, however, unless the data partition, processing method, training method, and inference range are well-documented.

CNNs have been proven to be very successful for this in the past, as they learn hierarchical image representations automatically instead of relying solely on handcrafted texture, color, or shape descriptors. Within the context of mango leaf analysis, optimized CNN structures have been shown to significantly influence recognition performance, with CNN structure and optimization technique playing significant roles [3]. However, such a comparison between the best-performing CNNs can only be useful if the experimental conditions are explicitly defined. The difference in models can be in the way they are preprocessed, augmented, optimized, and trained as well as how the dataset is split up and whether pretrained weights are used or not, among other things. Thus a model trained using the weights from the ImageNet data is not identical to the same network trained from scratch. Likewise, good performance in an image-level split cannot unconditionally be taken to be evidence of independence of acquisition sessions or original sources. Specific issues should be taken into account for situations where several images have similar acquisition properties, since arbitrary divisions of the images into subsets could leave images with similar visual features within the subsets.

The recent rise in availability of mango image datasets has provided a more solid empirical basis for experimentation with deep learning. Public image resources have helped compare architectures, explore patterns within disease images, and create frameworks for automated classification for precision-agriculture applications [4]. But researchers also have a responsibility to document data integrity and partitioning transparently, especially when datasets are available. Class distributions, corrupted files, consistent file names, duplicate paths, label mappings, and overlap between experimental subsets must be verified before describing the dataset used for model evaluation. These checks do not establish perceptual or biological independence across all images, but they do provide an auditable history of the implemented data pipeline. Therefore, to provide reliable mango leaf image analysis, the model must be effectively developed, and the inspection, partitioning, and processing of the dataset must be reported explicitly. Another challenge is handling symptom severity. Categorical classification determines the disease or leaf condition category but not directly the visible size and severity of symptoms. Previous studies have used CNN- and SVM-based methods to analyze mango diseases based on severity, showing the potential of symptom-level characterization to supplement categorical recognition [5]. The severity information has also been incorporated in mango leaf spot analysis using hybrid approaches [6]. The estimation of severity, however, needs a defensible reference label, and many public sets of diseases are distributed without severity grades or annotations of the lesions. Under these conditions, proxy labels can be derived from measurable visual properties of the image, and the method should be explicitly defined. These should be considered as heuristic symptom intensity categories and not as a form of expert agronomic ground truth. This distinction is crucial to prevent the interpretation of results from the auxiliary severity measures from being over-interpreted.

Another factor that needs to be taken into account is computational compactness in agricultural computer vision. Large, general-purpose CNNs can have excellent predictive accuracy, but result in very complex architectures. Hence, lightweight mango leaf disease models have been explored to minimize the model size and computational complexity [7]. However, parameter count is not a sufficient indicator of deployment efficiency as it also depends on the inference latency, memory usage, floating point operations, hardware features, and energy consumption. A small disease recognition framework should therefore be assessed on what it measures and not extrapolated to be suitable for mobile and field use. This is especially true when custom architectures are compared to popular networks like VGG16, ResNet50, and EfficientNetB0, which vary widely in performance depending on the initialization and optimization techniques used.

Multi-modal and attention-oriented architectures have also been used in recent years to expand the analysis of mango diseases. Dual-modality fusion methods that combine leaf and fruit images have been shown to be useful for the complementary visual information in mango disease classification [8]. In similar lines of work, other deep-learning studies have been expanded to encompass traditional recognition and more comprehensive disease detection and evaluation systems [9]. In the same way, Hybrid architectures with attention that incorporate convolutional feature extraction and transformer-based representations are also showing the evolution of increasingly sophisticated visual modeling strategies [10]. While these developments demonstrate the ongoing rapid development of the field, they also highlight the need for experimental methods that are transparent, reproducible, and properly interpreted.

To this end, the present study proposes XLeaf-Net, an explainable deep-learning framework for mango leaf condition classification, severity assessment from extracted images, and visual interpretation using Grad-CAM. Only the MangoLeafBD data set is used and the study is limited to eight categories including healthy leaves, disease symptoms and pest associated damage. The classification experiment was performed using an image-level stratified evaluation protocol, and the auxiliary severity component employed conservative image-derived symptom-intensity proxies, instead of expert-annotated agronomic severity. Grad-CAM is used as a qualitative post-hoc visualization method, and is not an interpretation of lesion segmentation or quantitative explanation validation. The same is also true for the baseline comparison, which is only conducted for VGG16, ResNet50 and EfficientNetB0 trained from scratch under the set-up of the experiment. Therefore, the study aims at three specific objectives: (1) creating and testing an XLeaf-Net for image-level classification of mango leaf condition into eight leaf classes; (2) comparing the performance of XLeaf-Net against the other reference models from scratch, VGG16, ResNet50, and EfficientNetB0 under the same final data partition and test set; and (3) exploring an image-derived symptom-intensity severity proxy along with qualitative visualization using Grad-CAM to add another layer of interpretation to the classification framework. These goals are explicitly stated, and the study aims to explicitly reflect transparent experimental characteristics of XLeaf-Net within the scope of the limited data set, but does not claim to be session-independent, expert-validated, or field-level generalized.

2. Literature review

Existing mango leaf disease analysis research can be broadly categorized into four methodological approaches: transfer-learning classifiers, Custom convolutional architectures, segmentation-oriented pipelines, and Augmentation- or optimization-based systems. Typical transfer-learning approaches use a pre-trained convolutional backbone to extract hierarchical visual features from leaf images. Pre-trained representations have proven useful for mango leaf disease recognition using VGG16-based models, but these models tend to focus more on predictive performance than on other aspects such as dataset-specific learning, parameter efficiency, or model interpretability [11]. Other approaches have combined localization and classification in a single pipeline. In DeepLeafNet, segmentation and recognition are treated as complementary steps, but this pipeline may require extra preprocessing and methods that are not always available across datasets [12].

Another research direction is to design a custom or conventional CNN architecture and train it for mango leaf image classification. The studies show the feasibility of end-to-end feature learning, where discriminative representations are learned directly from images, without being handcrafted. However, some implementations report only partial split integrity, class-specific errors, or visual interpretations of the learned decision process [13]. Such work has been extended to modified dense convolutional architectures that bolster feature reuse and deepen features, but more complex architectures may be more costly in terms of computational requirements for systems that must remain lightweight for image analysis [14]. Specialized models for very specific visual tasks, like the artificial-intelligence-based mango scab detection system, also illustrate their power. But the results of single-disease systems cannot be directly transferred to broader multiclass mango leaf condition recognition [15].

Another important research stream is comparative architecture selection and optimization. Studies on Bayesian optimization of pretrained CNNs show that both CNN architecture choice and hyperparameter settings can significantly affect disease-recognition results [16]. These experiments also highlight the need to separate characteristics of the target data from those inherited from the pretrained initialization. Hybrid CNN architectures have also been investigated to improve feature extraction and classification performance, but the model's behavior becomes harder to understand as the architecture grows more complex unless a separate explanation mechanism is added to the system [17]. ResNet-based approaches show that residual learning may help classify plant images; however, large residual networks may need optimization with moderately sized datasets [18]. The potential of automated image analysis has also been demonstrated with more general CNN-based crop-disease detection systems, though many focus mainly on disease-category prediction rather than more detailed symptom-intensity characterization [19].

A fourth relates to dataset diversity, model benchmarking, and synthetic augmentation. Research across different image collections, which are geographically or acquisition-specific, shows that variations in background, illumination, imaging conditions, and symptom appearance can affect reported performance across datasets [20]. While comparative analysis of various deep-learning architectures can provide a useful benchmark, it does not necessarily satisfy the multifaceted goals of compactness, methodological transparency, and visual interpretability [21]. Synthetic augmentation has also been explored as a method to boost image diversity. The evaluation in the following is conducted using CycleGAN generated images by ViT, BiT, and CNN architectures as training samples to demonstrate the potential value of artificial images, where the realism and representativeness of synthetic images are also important methodological issues [22]. The image-analysis-based mango disease identification system can also be used to assist the overall shift towards automated diagnosis, but severity-based evaluation and interpretation of model decisions is not always included in this type of system [23].

Combining all the literature, significant advances have been achieved in the classification of mango leaf diseases, however, there is some methodological differences in model initialization, data partitioning, preprocessing, complexity of architecture, model interpretation, and analysis of the severity of the disease. Therefore, conclusions of superiority cannot be made easily across studies, as they may differ in their experimental protocols. The current study is therefore not concerned with the development of XLeaf-Net as a new and different method of explainability or a substitute for the well-known pretrained architectures. Rather, it tests a model of a compact task-specific CNN in a well-documented image-level protocol, in comparison to explicit from-scratch VGG16, ResNet50 and EfficientNetB0 reference models, adds an image-derived symptom-intensity proxy in the absence of expert severity labels, and performs qualitative visual interpretation with Grad-CAM. The contribution therefore focuses on how the components could be integrated and transparently reported from the MangoLeafBD experimental setup, not on universal architectural "best" or independently validated estimation of agronomic severity.

3. Methodology

3.1 Dataset description and integrity assessment

A quantitative experimental design was used to develop and evaluate XLeaf-Net for mango leaf condition classification, image-derived symptom-intensity assessment, and Grad-CAM-based visual interpretation. The experiments used the publicly available MangoLeafBD dataset, originally documented in the associated Data in Brief publication [24] and released through the Mendeley Data repository [25]. The dataset comprises 4,000 RGB images equally distributed among eight labeled categories: Anthracnose, Bacterial Canker, Cutting Weevil, Die Back, Gall Midge, Healthy, Powdery Mildew, and Sooty Mould, with 500 images in each category. The original dataset documentation reports images with a resolution of 240 × 320 pixels and indicates that the final collection includes images derived from approximately 1,800 distinct mango leaves, with additional samples generated through zooming and rotation [24,25]. Cutting Weevil was retained as the official dataset class label and represents pest-related leaf injury rather than a foliar disease.

Before model development, the dataset was audited for class-folder consistency, image counts, file formats, unreadable or corrupted images, missing metadata, duplicate image paths, duplicate filenames, and class-name/class-ID consistency. The implemented integrity checks therefore established exact file-level consistency but did not include perceptual-hash or SSIM-based near-duplicate detection. After resizing, the i-th input image was represented as:

$$X_i \in \mathbb{R}^{224\times224\times3} \quad (1)$$

where X_i denotes an RGB image supplied to the network.

3.2 Data splitting, preprocessing, and augmentation

A group-based partition was initially examined using class-specific capture-group identifiers and GroupShuffleSplit. Although this produced zero group overlap, the resulting validation and test subsets did not retain adequate representation of all eight classes; for example, the test partition contained only Anthracnose and Sooty Mould. Consequently, this split was not used for the reported eight-class experiment. The final experiment therefore used an image-level stratified 70:15:15 split, producing 2,800 training, 600 validation, and 600 test images, corresponding to 350, 75, and 75 images per class. The final stratified partition contained capture-group overlap of 16 groups between training and validation, 15 between training and testing, and 15 between validation and testing. Accordingly, the protocol is explicitly considered an image-level evaluation rather than a session-independent evaluation. Pixel intensities were normalized as:

$$\tilde{X}_i = \frac{X_i}{255} \quad (2)$$

Mapping values to $[0, 1]$. Training augmentation used rotations up to 20°, horizontal flipping, width and height shifts of 0.10, zoom of 0.15, brightness variation from 0.8 to 1.2, and nearest-neighbor filling. Validation and test images were normalized without augmentation.

3.3 XLeaf-Net disease classification framework

XLeaf-Net was implemented as a task-specific CNN trained from scratch. The network comprised four feature-extraction blocks. Each block contained two 3×3convolutions with same padding, Batch Normalization, and ReLU activation, followed by 2×2max pooling. The filter progression was 32, 64, 128, and 256. After the fourth block, a 3×3, 512-filter convolution, Batch Normalization, and ReLU activation were applied. Global Average Pooling was followed by Dense-256 and Dense-128 layers, each with ReLU activation and dropout of 0.30, before an eight-unit softmax output. The model contained 2,523,560 parameters. The layer relu4_2 denotes the activation following the second 256-filter convolution in Block 4. For class c, softmax probability was:

$$p_{ic} = \frac{exp(z_{ic})}{\sum_{j=1}^{8} exp(z_{ij})} \quad (3)$$

where $z_{ic}$is the corresponding logit. Training minimized categorical cross-entropy,

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{8} y_{ic}\, log(p_{ic}) \quad (4)$$

where $y_{ic}$is the one-hot target. The executed classification experiment used Adam with 10^{-4}learning rate, batch size 32, seed 42, and 10 training epochs. Early stopping monitored validation loss with patience 10; ReduceLROnPlateau used factor 0.3, patience 5, and minimum learning rate 10^{-7}. No explicit weight-decay term was applied.

3.4 Baseline model training and evaluation protocol

VGG16, ResNet50, and EfficientNetB0 were evaluated as from-scratch reference models using weights=None; consequently, the comparison does not represent ImageNet-pretrained transfer-learning performance. The models shared the final stratified split, 224 × 224 input dimensions, $[0, 1]$normalization, training augmentation, batch size, and test set, while model-specific optimization settings were used during the executed runs.

VGG16 was trained for 15 epochs using Adam at $10^{-4}$with categorical cross-entropy. ResNet50 used 15 epochs, Adam at 10^{-5}, and categorical cross-entropy with 0.05 label smoothing. EfficientNetB0 used 10^{-5}, 0.05 label smoothing, 0.50 dropout, and extended training to 60 epochs. Thus, results are interpreted specifically within these implemented from-scratch protocols rather than as evidence of general superiority over optimized pretrained implementations. Evaluation included accuracy, macro precision, macro recall, macro F1-score, weighted F1-score, test loss, confusion matrices, and parameter counts:

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \quad (5)$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \quad (6)$$

$$F1_c = \frac{2\, Precision_c\, Recall_c}{Precision_c + Recall_c} \quad (7)$$

3.5 Image-derived symptom-intensity assessment

MangoLeafBD does not provide expert-annotated agronomic severity grades; therefore, the implemented severity component represents an image-derived symptom-intensity proxy. Healthy images were assigned a score of zero. Diseased images were converted to HSV colour space after resizing. A leaf mask retained pixels satisfying $\big((G > 0.75R) \vee (R > 0.75G)\big)$, $V < 245$, and $S > 15$. Symptom pixels combined four implemented masks: dark regions ($V < 80$); brown regions ($5 \le H \le 25$, $S > 45$, $60 \le V < 210$); yellow regions ($25 < H \le 38$, $S > 50$, $100 < V < 235$); and pale regions ($V > 150$, $S < 60$), with all masks restricted to the estimated leaf area. The proxy score was:

$$S_i = \frac{|M_{symptom,i}|}{|M_{leaf,i}|} \quad (8)$$

For each diseased class c, the 33rd and 66th percentiles were then used to define

$$G_i = \begin{cases} \text{Mild}, & S_i \le Q_{0.33,c}, \\ \text{Moderate}, & Q_{0.33,c} < S_i \le Q_{0.66,c} \\ \text{Severe}, & S_i > Q_{0.66,c}. \end{cases} \quad (9)$$

The implemented code calculated these percentile thresholds from the complete class-wise score distributions before the generated labels were merged into the train, validation, and test dataframes. Accordingly, this task is interpreted as exploratory proxy grading rather than independently validated agronomic severity assessment.

3.6 Grad-CAM-based visual interpretation

Grad-CAM was applied as a qualitative post-hoc visualization technique, not for lesion segmentation or quantitative explanation validation. Gradients of the selected class score with respect to activation maps were spatially averaged:

$$\alpha_k^c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\frac{\partial y^c}{\partial A_{ij}^k} \quad (10)$$

and the localization map was

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k A^k\right) \quad (11)$$

The implementation compared final_conv_layer, relu_final, and relu4_2 visually; relu4_2 was selected for the final figures. Representative visualizations were generated from correctly classified test samples, using median-confidence examples rather than exclusively the highest-confidence predictions. Thus, the maps were treated as coarse class-discriminative attention evidence only.

Experiments were implemented with TensorFlow/Keras, OpenCV, NumPy, Pandas, scikit-learn, Matplotlib, and Seaborn. TensorFlow 2.21.0 executed on CPU because no GPU was available. Model checkpoints, CSV logs, reports, confusion matrices, and Grad-CAM outputs were retained for traceability.

4. Results

4.1 Dataset composition and image-level split checks

The MangoLeafBD dataset used in this study contained 4,000 valid RGB images distributed equally among eight classes: Anthracnose, Bacterial Canker, Cutting Weevil, Die Back, Gall Midge, Healthy, Powdery Mildew, and Sooty Mould. Each class contained 500 images. Metadata-level inspection confirmed that all images were stored in .jpg format and that no corrupted files, missing values, duplicate image paths, duplicate filenames, or inconsistencies between class names and class identifiers were detected. These checks established file-level consistency but did not constitute a perceptual or leaf-level near-duplicate analysis. An image-level stratified partition was subsequently used to maintain equal class representation across the training, validation, and testing subsets, as presented in Table 1.

Table 1. Stratified dataset split and class distribution

Class	Training	Validation	Testing	Total
Anthracnose	350	75	75	500
Bacterial Canker	350	75	75	500
Cutting Weevil	350	75	75	500
Die Back	350	75	75	500
Gall Midge	350	75	75	500
Healthy	350	75	75	500
Powdery Mildew	350	75	75	500
Sooty Mould	350	75	75	500
Total	2,800	600	600	4,000

Table 1 shows that every class contributed 350 training images, 75 validation images, and 75 testing images. The final partition therefore comprised 2,800 training images and 600 images each for validation and testing while preserving the original balanced class distribution. Although no exact image-path or filename overlap was identified, the capture/session-group audit detected 16 overlapping groups between training and validation, 15 between training and testing, and 15 between validation and testing. Therefore, all subsequent performance measures are interpreted as outcomes of an image-level stratified evaluation protocol rather than evidence of fully session-independent generalization.

4.2 XLeaf-Net disease classification performance

XLeaf-Net achieved a test loss of 0.0454 and an image-level classification accuracy of 98.67%. Macro precision, macro recall, and macro F1-score were 98.69%, 98.67%, and 98.66%, respectively, while the weighted F1-score was 98.66%. The close correspondence between accuracy and the macro-averaged measures indicates consistently strong performance across the balanced test subset. The detailed class-wise results are presented in Table 2.

Table 2. Class-wise disease classification performance of XLeaf-Net

Class	Precision (%)	Recall (%)	F1-Score (%)	Support
Anthracnose	100.00	98.67	99.33	75
Bacterial Canker	98.68	100.00	99.34	75
Cutting Weevil	100.00	100.00	100.00	75
Die Back	100.00	100.00	100.00	75
Gall Midge	98.59	93.33	95.89	75
Healthy	98.67	98.67	98.67	75
Powdery Mildew	98.67	98.67	98.67	75
Sooty Mould	94.94	100.00	97.40	75
Macro Average	98.69	98.67	98.66	600

Table 2 shows that Cutting Weevil and Die Back achieved perfect precision, recall, and F1-score within the evaluated test subset. Gall Midge had the lowest recall and F1-score, although all class-wise F1-scores remained above 95%. The disease-classification confusion matrix in Figure 1 further illustrates the distribution of correct and incorrect predictions. Figure 1 shows that the majority of predictions are concentrated along the main diagonal, indicating a high proportion of correct classifications. Only a small number of off-diagonal predictions are visible, consistent with the high class-wise recall and F1-scores reported in Table 2. These findings represent performance on a single predefined image-level test partition. Repeated-seed experiments, cross-validation, and confidence intervals were not generated within the implemented experiment; therefore, the reported values should be interpreted as observed test-set results rather than estimates of variability across alternative data partitions.

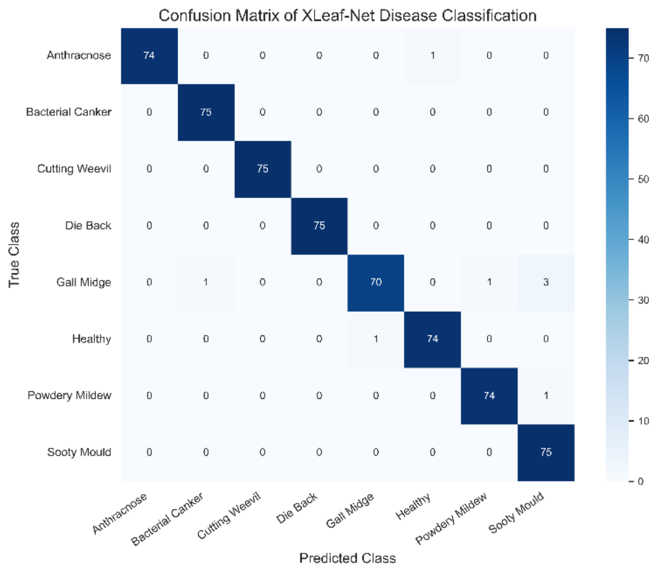

Figure 1. Confusion matrix of XLeaf-Net for mango leaf disease classification

Table 3. Comparative performance of XLeaf-Net and from-scratch baseline models

Model	Test Loss	Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro F1-Score (%)	Parameters
XLeaf-Net	0.0454	98.67	98.69	98.67	98.66	2,523,560
VGG16	0.1987	92.67	92.95	92.67	92.59	14,879,944
ResNet50	0.7868	77.33	79.62	77.33	76.02	24,146,184
EfficientNetB0	1.2263	62.83	67.63	62.83	61.20	4,411,435

4.3 From-scratch baseline model comparison

XLeaf-Net was compared with VGG16, ResNet50, and EfficientNetB0 using from-scratch initialization rather than ImageNet-pretrained weights. The models were assessed using the same final image-level data partition and test set, while model-specific optimization settings were employed where required during training. Table 3 summarizes the comparative results. Table 3 shows that XLeaf-Net achieved the highest observed accuracy and macro F1-score under the implemented from-scratch protocols. VGG16 was the closest-performing baseline, whereas ResNet50 and EfficientNetB0 yielded lower predictive performance under their corresponding configurations. Figure 2 illustrates the relative differences in accuracy and macro F1-score among the evaluated architectures. Figure 3 shows that XLeaf-Net contained approximately 2.52 million parameters, compared with 14.88 million for VGG16, 24.15 million for ResNet50, and 4.41 million for EfficientNetB0. XLeaf-Net was therefore the smallest evaluated architecture in terms of parameter count. Parameter count is interpreted only as a measure of model size. FLOPs, inference latency, memory utilization, energy consumption, and mobile-device runtime were not measured in the implemented experiments; therefore, parameter count alone is not used as evidence of deployment efficiency.

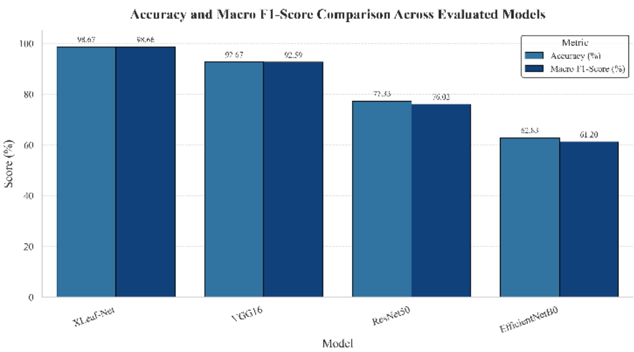

Figure 2. Accuracy and macro F1-score comparison across evaluated models

4.4 Image-derived severity assessment results

The auxiliary severity task comprised four image-derived categories: Healthy, Mild, Moderate, and Severe. These categories were generated from image-based symptom-intensity rules rather than independent expert agronomic annotations. Consequently, the reported results represent classification of an image-derived severity proxy rather than validation against plant-pathologist-assigned severity grades. The class-wise results are summarized in Table 4.

Table 4. Class-wise performance for image-derived severity assessment

Severity Class	Precision (%)	Recall (%)	F1-Score (%)	Support
Healthy	97.40	100.00	98.68	75
Mild	84.44	69.09	76.00	165
Moderate	64.29	67.24	65.73	174
Severe	79.13	87.63	83.16	186
Macro Average	81.31	80.99	80.89	600

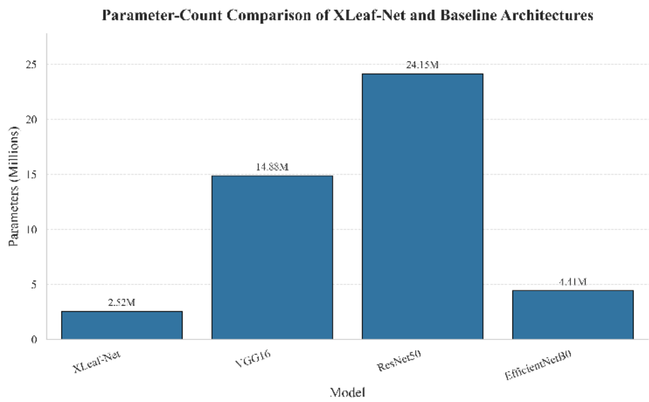

Figure 3. Parameter-count comparison of XLeaf-Net and baseline architecture

Table 4 shows that Healthy achieved the highest F1-score at 98.68%, followed by Severe at 83.16%. Moderate was the weakest category, with 64.29% precision, 67.24% recall, and 65.73% F1-score. At the overall task level, the severity model achieved a test loss of 0.5824, accuracy of 78.17%, macro precision of 81.31%, macro recall of 80.99%, macro F1-score of 80.89%, and weighted F1-score of 78.08%. Figure 4 illustrates the distribution of errors across the four image-derived categories. Figure 4 shows that all 75 Healthy images were classified correctly, while 163 of 186 Severe samples were correctly assigned to the Severe category. Most remaining errors occurred between adjacent Mild, Moderate, and Severe proxy levels. The weaker results for Mild and Moderate indicate that the intermediate image-derived categories were more difficult to distinguish than Healthy and Severe. These findings should therefore be understood as the model's ability to reproduce internally defined symptom-intensity categories rather than as evidence of expert-validated agronomic severity estimation.

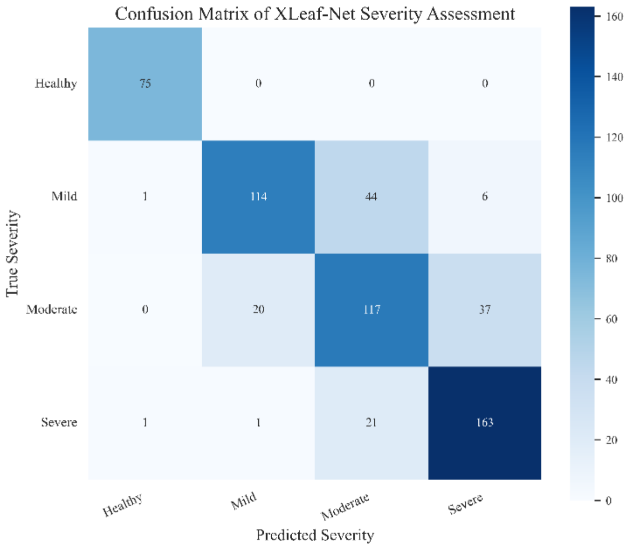

Figure 4. Confusion matrix of XLeaf-Net for image-derived severity assessment

4.5 Grad-CAM visual interpretation

Grad-CAM heatmaps were generated from the relu4_2 activation layer after candidate activation layers had been inspected visually. Representative correctly classified test images were selected across the evaluated classes to illustrate the spatial activation patterns produced by the network, as shown in Figure 5. Figure 5 shows that class-discriminative activation frequently overlapped with visually distinctive leaf regions, including necrotic areas, damaged margins, structural injury, surface discoloration, and powder-like patterns. Sooty Mould exhibited comparatively broader and more diffuse activation. These heatmaps are interpreted as qualitative post-hoc visualizations rather than lesion-segmentation masks or quantitative evidence of explanation faithfulness. Because the displayed examples were selected from correctly classified predictions, Figure 5 provides illustrative visual interpretation rather than a systematic assessment of model explanations across the complete test set.

4.6 Consolidated XLeaf-Net task summary

The two XLeaf-Net implementations were evaluated for eight-class disease classification and four-class image-derived severity assessment. Table 5 summarizes their principal test metrics and parameter counts. Table 5 shows that the disease-classification model achieved substantially higher observed accuracy and macro F1-score than the image-derived severity model.

Both architectures contained approximately 2.52 million parameters, indicating closely comparable model capacities despite the difference in task complexity and output configuration. Overall, XLeaf-Net demonstrated high image-level classification performance on the predefined MangoLeafBD test subset and lower performance on the four-class image-derived severity task. The classification, baseline-comparison, severity, and Grad-CAM findings should nevertheless be interpreted within the methodological boundaries of the implemented image-level evaluation protocol rather than as evidence of session-independent, orchard-level, or expert-validated generalization.

5. Discussion

The present findings indicate that XLeaf-Net achieved a strong balance between classification performance, architectural compactness, and qualitative visual interpretability within the implemented MangoLeafBD image-level evaluation protocol. XLeaf-Net attained the highest observed test accuracy and macro F1-score among the evaluated models while using substantially fewer parameters than VGG16 and ResNet50. However, this comparison should be interpreted specifically within the from-scratch training configurations used in the study. The baseline models were not evaluated using ImageNet-pretrained weights, and model-specific optimization settings were applied where required. Therefore, the results do not establish universal superiority of XLeaf-Net over transfer-learning implementations of VGG16, ResNet50, or EfficientNetB0. Rather, they indicate that the proposed compact architecture was well matched to the characteristics of the present dataset and experimental setting. This observation is consistent with previous work emphasizing that image quality, preprocessing, and model suitability can substantially influence mango disease-recognition performance [26]. The class-wise results further demonstrate that XLeaf-Net learned highly discriminative representations for most of the evaluated categories. Cutting Weevil and Die Back achieved perfect precision, recall, and F1-score within the test subset, whereas Gall Midge showed the lowest recall and F1-score. Sooty Mould also exhibited slightly lower precision than several other categories. These residual errors may reflect the comparatively diffuse or visually overlapping appearance of symptoms in some classes. Variations in leaf morphology, pigmentation, surface condition, and symptom distribution can influence the visual boundaries available to an image classifier [27]. Nevertheless, because the test set was obtained through image-level stratification and capture/session groups overlapped across partitions, these high class-wise scores should be interpreted as performance within the defined image-level protocol rather than as evidence of independent orchard- or session-level generalization.

Table 5. Consolidated XLeaf-Net performance across disease classification and image-derived severity assessment

Task	Classes	Test Loss	Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro F1-Score (%)	Weighted F1-Score (%)	Parameters
Disease Classification	8	0.0454	98.67	98.69	98.67	98.66	98.66	2,523,560
Severity Assessment	4	0.5824	78.17	81.31	80.99	80.89	78.08	2,523,044

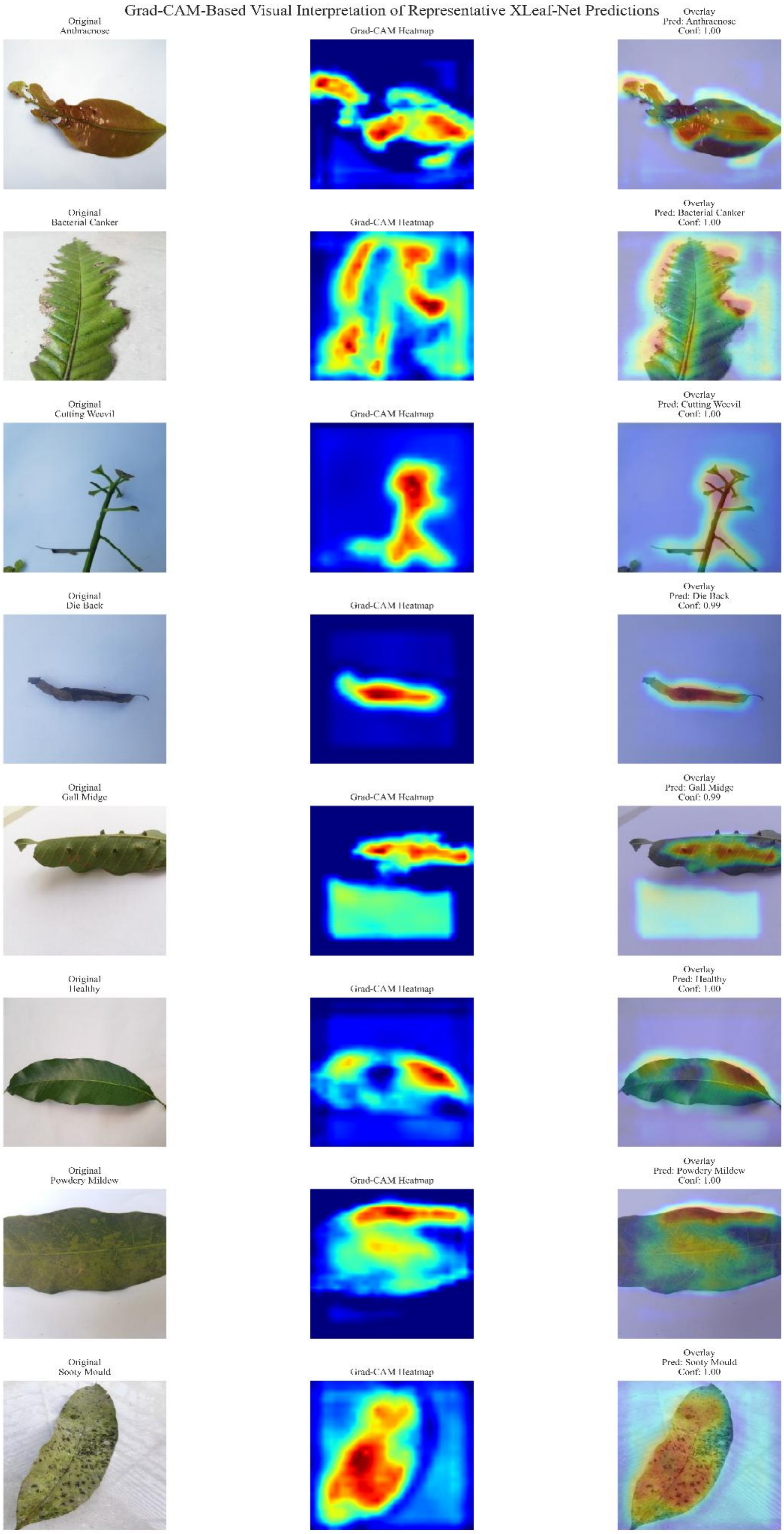

Figure 5. Grad-CAM-based visual interpretation of representative correctly classified XLeaf-Net predictions

The baseline comparison also demonstrates that parameter count alone did not determine predictive performance. XLeaf-Net contained approximately 2.52 million parameters, compared with 14.88 million for VGG16 and 24.15 million for ResNet50, yet achieved higher observed test accuracy within the implemented from-scratch experiments. EfficientNetB0 was also relatively compact but produced substantially lower performance under its evaluated configuration. These findings suggest that the interaction between architecture and dataset characteristics was important in the present experiment. Earlier mango-disease image-analysis studies have similarly emphasized the importance of appropriate model selection for automated diagnostic tasks [28]. However, compactness in this study refers strictly to parameter count. FLOPs, memory consumption, inference latency, energy use, and mobile-device performance were not measured; consequently, the current results should not be interpreted as direct evidence of field or edge deployment efficiency.

The image-derived severity task was more challenging than categorical disease classification. The Healthy category achieved the strongest performance, while Moderate produced the lowest precision, recall, and F1-score. Most errors occurred between adjacent Mild, Moderate, and Severe categories, which is consistent with the continuous nature of the underlying image-derived symptom score and its conversion into discrete percentile-based groups. Importantly, these categories were generated from color- and intensity-based image heuristics and were not obtained from independent plant-pathologist annotations. The resulting 78.17% accuracy therefore represents the model's ability to reproduce an internally defined symptom-intensity proxy rather than validated agronomic disease severity. The severity component nevertheless extends the image-analysis framework beyond categorical condition recognition, while remaining explicitly exploratory relative to approaches focused primarily on automated mango leaf disease classification [29].

Grad-CAM provided an additional qualitative view of the image regions contributing to model predictions. The selected relu4_2 layer produced localized activation patterns that frequently overlapped with visually distinctive regions such as necrotic areas, damaged margins, discoloration, and powder-like surface patterns. Broader activation was observed for Sooty Mould, suggesting that spatially diffuse visual characteristics were associated with less localized attention. Similar challenges have been reported in machine-vision studies dealing with mango leaf recognition under variable visual conditions [30]. These heatmaps should, however, be interpreted cautiously. Only representative correctly classified examples were displayed, the target layer was selected through qualitative inspection, and no quantitative faithfulness metric or second explanation method was evaluated. Grad-CAM therefore provides illustrative post-hoc visualization rather than validated lesion localization or causal explanation.

Several limitations constrain the interpretation of the findings. The study used a single public dataset and a single image-level train/validation/test partition. Although exact image-path and filename overlaps were not detected, capture/session-group overlap remained across the final partitions, and no perceptual near-duplicate analysis was performed. Repeated-seed evaluation, cross-validation, confidence intervals, or formal statistical significance testing were also not conducted. The baseline comparison was restricted to from-scratch models and did not include pretrained transfer-learning configurations. The severity labels were image-derived proxies rather than expert annotations, and Grad-CAM was assessed qualitatively rather than through quantitative faithfulness analysis. Accordingly, the present findings should be regarded as evidence of strong performance within the implemented image-level experimental setting. Future work should prioritize group-independent validation, expert severity annotations, pretrained baseline comparisons, repeated evaluation, quantitative explainability assessment, external datasets, and hardware-level deployment benchmarking.

6. Conclusion

This study developed and evaluated XLeaf-Net for mango leaf disease classification, image-derived severity assessment, and Grad-CAM-based visual interpretation. Within the implemented image-level MangoLeafBD protocol, XLeaf-Net achieved high classification performance while using fewer parameters than VGG16 and ResNet50. However, these comparisons apply only to the evaluated from-scratch configurations and should not be interpreted as evidence of superiority over pretrained or independently optimized baseline models. The image-derived severity module extended the framework beyond categorical recognition, but its labels were generated from visual symptom-intensity rules rather than expert agronomic annotations; therefore, its outputs should be interpreted as proxy severity categories. Grad-CAM provided qualitative insight into class-discriminative image regions, with activation often overlapping visually symptomatic leaf areas, although the maps were not validated as lesion-segmentation or quantitative faithfulness measures. The findings also remain limited by the use of a single public dataset, a single image-level split, capture/session-group overlap, and the absence of repeated-seed evaluation, external validation, or deployment benchmarking. Accordingly, XLeaf-Net should be regarded as a compact candidate framework requiring further validation rather than a field-ready diagnostic system. Future work should prioritize group-independent evaluation, expert severity annotations, pretrained baseline comparisons, quantitative explainability analysis, independent multi-location datasets, leaf segmentation or background suppression, and hardware-level testing for mobile or edge-oriented applications. Such validation would determine whether the observed image-level performance is maintained under more independent acquisition conditions and whether the framework generalizes beyond MangoLeafBD in practice.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically regarding authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with research ethics policies. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere. All survey participants provided informed consent before participating, and the anonymity and confidentiality of all respondents were strictly maintained throughout the research process.

Data availability statement

The manuscript contains all the data. However, additional data will be provided by the corresponding author upon reasonable request.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] M. A. Rahman, M. T. A. Bhuiyan, F. M. Noori, M. Z. Uddin, and A. K. M. Masum, "MangoLeafNet-XAI: An attention-enhanced deep learning architecture for accurate and interpretable mango leaf disease classification," Frontiers in Plant Science, vol. 17, Art. no. 1776537, 2026.

[2] D. Faye, I. Diop, and D. Dione, "Mango diseases classification solutions using machine learning or deep learning: A review," Journal of Computer and Communications, vol. 10, pp. 16–28, 2022, doi: 10.4236/jcc.2022.1012002.

[3] M. Prabu and B. J. Chelliah, "Mango leaf disease identification and classification using a CNN architecture optimized by crossover-based levy flight distribution algorithm," Neural Computing and Applications, vol. 34, no. 9, pp. 7311–7324, 2022.

[4] S. Jayaweera, P. Sewwandi, P. Tharaka, D. Pallewatta, T. Halloluwa, M. Wickramasinghe, et al., "MangoDB-A TJC mango dataset for deep-learning-based on classification and detection in precision agriculture," in Proc. 2024 4th Int. Conf. Advanced Research in Computing (ICARC), 2024, pp. 115–120.

[5] D. Banerjee, V. Kukreja, S. Hariharan, and V. Jain, "Enhancing mango fruit disease severity assessment with CNN and SVM-based classification," in Proc. 2023 IEEE 8th Int. Conf. Convergence in Technology (I2CT), 2023, pp. 1–6.

[6] D. Baresary, A. Saini, R. K. Sharma, D. Bordoloi, R. Sharma, and V. Kukreja, "MangoSpot: A hybrid CNN-SVM model for multi-classification of mango leaf spot disease based on seriousness levels," in Proc. 2023 3rd Int. Conf. Intelligent Technologies (CONIT), 2023, pp. 1–5.

[7] B. U. Mahmud, A. Al Mamun, M. J. Hossen, G. Y. Hong, and B. Jahan, "Light-weight deep learning model for accelerating the classification of mango-leaf disease," Emerging Science Journal, vol. 8, no. 1, pp. 28–42, 2024.

[8] M. Mohsin, M. S. A. Hashmi, I. D. Noya, H. Garay, N. Abdel Samee, and I. Ashraf, "Dual-modality fusion for mango disease classification using dynamic attention based ensemble of leaf and fruit images," Scientific Reports, vol. 15, no. 1, Art. no. 42084, 2025.

[9] V. K. Pratap and N. S. Kumar, "Deep learning based mango leaf disease detection for classifying and evaluating mango leaf diseases," Fusion: Practice & Applications, vol. 15, no. 2, 2024.

[10] E. Ergün, "Attention-enhanced hybrid deep learning model for robust mango leaf disease classification via ConvNeXt and vision transformer fusion," Frontiers in Plant Science, vol. 16, Art. no. 1638520, 2025.

[11] G. Kaur, N. Sharma, S. Malhotra, S. Devliyal, and R. Gupta, "Mango leaf disease detection using VGG16 convolutional neural network model," in Proc. 2024 3rd Int. Conf. Innovation in Technology (INOCON), 2024, pp. 1–6.

[12] C. P. Vijay, K. Pushpalatha, D. R. Janardhana, H. R. Kumar, and K. Puttegowda, "DeepLeafNet: A novel deep learning approach for mango leaf disease segmentation and classification with optimization algorithms," SN Computer Science, vol. 6, no. 8, Art. no. 951, 2025.

[13] P. D. Rinanda, D. N. Aini, T. A. Pertiwi, S. Suryani, and A. J. Prakash, "Implementation of convolutional neural network (CNN) for image classification of leaf disease in mango plants using deep learning approach," Public Research Journal of Engineering, Data Technology and Computer Science, vol. 1, no. 2, pp. 56–61, 2024.

[14] C. Chandrashekar, K. Vijayakumar, K. Pradeep, and A. Balasundaram, "MDCN: Modified dense convolution network based disease classification in mango leaves," Computers, Materials, & Continua, vol. 78, no. 2, p. 2511, 2024.

[15] A. Kaur, V. Kukreja, D. S. Rana, A. Garg, and R. Sharma, "Deep insights into mango scab: An AI-powered approach for detection and classification," in Proc. 2024 Fourth Int. Conf. Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2024, pp. 1–5.

[16] S. Rahayu and S. F. Romdoni, "Bayesian optimized pretrained CNNs for mango leaf disease classification: A comparative study," Jurnal Teknik Informatika (Jutif), vol. 6, no. 5, pp. 3051–3078, 2025.

[17] S. B. Porna, M. F. Kabir, M. I. C. Rana, M. S. I. Sajol, T. Roy, M. A. U. Khan, et al., "Hybrid convolutional neural networks for enhanced detection of mango leaf diseases," in Proc. 2024 IEEE 6th Int. Conf. Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), 2024, pp. 547–552.

[18] N. N. Komah and D. Hermanto, "Classification of mango varieties from leaf images using ResNet-50 CNN architecture," Brilliance: Research of Artificial Intelligence, vol. 5, no. 1, pp. 559–567, 2025.

[19] S. Dinkar, K. Jayapriya, N. Pallerla, K. D. Jadhav, and A. R. Shet, "Convolutional neural networks for the intelligent and automated detection of mango leaf disease to enhance crop health management," Agricultural Science Digest, vol. 45, no. 6, 2025.

[20] S. Thaseentaj and S. Ilango, "Deep convolutional neural networks for South Indian mango leaf disease detection and classification," Computers, Materials, & Continua, vol. 77, no. 3, p. 3593, 2023.

[21] R. Varun, C. S. Chaithra, and S. Siddesha, "DeepMango: Understanding and analysis of deep models in mango leaf disease detection," in Proc. Int. Conf. Microelectronics, Electromagnetics and Telecommunication, Singapore: Springer Nature Singapore, 2024, pp. 285–295.

[22] S. T. Y. Ramadan, T. Sakib, M. A. Rahat, S. Mosharrof, F. I. Rakin, and R. Jahangir, "Enhancing mango leaf disease classification: ViT, BiT, and CNN-based models evaluated on CycleGAN-augmented data," in Proc. 2023 26th Int. Conf. Computer and Information Technology (ICCIT), 2023, pp. 1–6.

[23] S. D. Dip, S. Mahapatra, M. J. Naeen, M. U. Mojumdar, R. Sharma, and S. K. Banshal, "Automated disease identification in mango leaves using image analysis," in Proc. 2025 12th Int. Conf. Computing for Sustainable Global Development (INDIACom), 2025, pp. 1–7.

[24] S. I. Ahmed, M. Ibrahim, M. Nadim, M. M. Rahman, M. M. Shejunti, T. Jabid, and M. S. Ali, "MangoLeafBD: A comprehensive image dataset to classify diseased and healthy mango leaves," Data in Brief, vol. 47, Art. no. 108941, 2023, doi: 10.1016/j.dib.2023.108941.

[25] S. Ali, M. Ibrahim, S. I. Ahmed, M. Nadim, M. M. Rahman, M. M. Shejunti, and T. Jabid, "MangoLeafBD Dataset," Mendeley Data, Version 1, 2022, doi: 10.17632/hxsnvwty3r.1.

[26] R. Garg, A. K. Sandhu, B. Kaur, B. Goyal, and A. Dogra, "Design of filtration approach for image quality improvement in mango leaf disease detection and pharmaceutical treatment," Biomedical and Pharmacology Journal, vol. 17, no. 1, pp. 341–358, 2024.

[27] S. Patil, B. Bisen, S. K. Pandey, R. Bisen, and B. Krishna, "Assessment of variability in stem girth, leaf area, leaf thickness, chlorophyll content, and their correlation in mango hybrids," International Journal of Plant and Environment, vol. 11, no. 3, pp. 547–556, 2025.

[28] M. A. Al Rahib, N. Sultana, N. Saha, R. Mia, M. Sarker, and A. Sattar, "Image analysis and machine learning techniques for accurate detection of common mango diseases in warm climates," International Journal of Artificial Intelligence, vol. 14, no. 4, pp. 2935–2944, 2025.

[29] I. Arshad, M. S. Arshad, M. Malik, W. Raza, and Z. Maryam, "An ensemble convolutional neural network framework for automated mango leaf disease detection," Spectrum of Engineering Sciences, vol. 4, no. 3, pp. 1711–1740, 2026.

[30] P. Das, M. A. Rahman, M. T. Habib, M. A. A. Khan, and M. M. Islam, "An in-depth analysis for machine-vision-based mango leaf diseases recognition," in Proc. 2025 6th Int. Conf. Mobile Computing and Sustainable Informatics (ICMCSI), 2025, pp. 1571–1576.