Published in AJSE, Vol: 24, Issue:2
Received on 3rd May 2026
Revised on 5th May 2026
Published on 31th May 2026

AIUB JOURNAL OF SCIENCE AND ENGINEERING
ISSN: 1608 - 3679 (print) 2520 - 4890 (Online)
https://doi.org/10.53799/vxkj7t08

Explainable Hybrid CNN-Transformer Framework for Papaya Leaf Disease Classification with Layer-Wise Grad-CAM Analysis

Amit Arpon Paul, S.M. Tawhid, Abdul Kader Mohim, and Dip Nandi

***Abstract*—While deep learning achieves high accuracy in plant disease classification, its black-box nature limits adoption by agricultural practitioners who require transparent and interpretable predictions for decision-making. This paper presents a hybrid CNN-Transformer framework with systematic layer-wise interpretability that combines convolutional neural networks with vision transformers to prioritize model transparency alongside competitive classification accuracy. Unlike existing hybrid approaches that apply single-layer visualization, this work introduces a systematic multi-depth Grad-CAM analysis that captures progressive feature evolution from edge detection through texture analysis to disease-region localization, revealing hierarchical diagnostic reasoning unavailable in standard single-layer explanations. Five-fold stratified cross-validation demonstrates performance statistically comparable to high-performing CNN and Transformer baselines, with the key distinction being multi-level interpretability rather than superior accuracy. Quantitative evaluation through the Pointing Game localization protocol achieves 87.3% accuracy on expert-annotated disease regions. Robustness evaluation under synthetic perturbations shows graceful performance degradation, though real-world field validation remains future work.**

***Index Terms*—Agricultural artificial intelligence, Crop disease detection, Explainable deep learning, Hybrid neural networks, Image classification, Papaya leaf diseases, Visual attention mechanisms**

I. Introduction

PAPAYA (*Carica papaya* L.) is one of the most economically important tropical fruits worldwide, valued for its nutritional content, medicinal properties, and commercial significance [1]. However, papaya cultivation faces significant challenges from various foliar diseases including fungal infections, viral diseases, and nutrient deficiencies that can cause substantial yield losses if not detected and treated promptly [2]. Traditional disease identification relies on expert visual inspection, which is time-consuming, subjective, and often unavailable to small-scale farmers in remote regions [3].

Manuscript received April 19, 2025; revised August 26, 2025.

A. A. Paul (e-mail: 23-50158-1@student.aiub.edu) is an Undergraduate Student with the Department of Computer Science and Engineering, American International University-Bangladesh, Dhaka, 1229, Bangladesh.

S. M. Tawhid (e-mail: 21-45470-3@student.aiub.edu) is a Graduate with the Department of Computer Science and Engineering, American International University-Bangladesh, Dhaka, 1229, Bangladesh.

A. K. Mohim (e-mail: 22-47833-2@student.aiub.edu) is a Graduate with the Department of Computer Science and Engineering, American International University-Bangladesh, Dhaka, 1229, Bangladesh.

D. Nandi (e-mail: dip.nandi@aiub.edu) is a Professor and Associate Dean with the Faculty of Science and Technology, American International University-Bangladesh, Dhaka, 1229, Bangladesh.

The advent of deep learning has revolutionized plant disease detection, with Convolutional Neural Networks (CNNs) achieving remarkable success in automated leaf disease classification [4], [5]. CNNs excel at capturing local spatial features and hierarchical patterns through their convolutional operations, making them well-suited for texture-based disease symptom recognition [6]. However, CNNs inherently lack the ability to model long-range dependencies and global contextual relationships within images, which are crucial for understanding complex disease patterns that span large leaf regions [7].

Vision Transformers (ViTs) have emerged as a powerful alternative, leveraging self-attention mechanisms to capture global context and relationships between image patches [7], [8]. While ViTs demonstrate superior performance on large-scale datasets, they often underperform with limited training data and may miss fine-grained local details that CNNs capture effectively [9].

Recent research has shown that hybrid architectures combining CNNs and Transformers can outperform either approach alone by synergizing local feature extraction with global context modeling [10], [11]. In agricultural applications, such hybrid approaches have shown promise for plant disease detection [12], though comprehensive benchmarking and explainability remain underexplored.

We propose a hybrid CNN-Transformer framework with systematic layer-wise interpretability for papaya disease classification. The key contributions of this work are:

- **Multi-depth Grad-CAM analysis**: Unlike standard single-layer visualization, we introduce systematic layer-wise Grad-CAM across CNN depths (early, middle, late) to reveal progressive diagnostic reasoning from edge detection through texture analysis to disease-region localization.
- We evaluate explanation quality using the Pointing Game protocol with expert annotations, achieving 87.3% localization accuracy—objective validation absent in most agricultural AI works.
- We compare layer-wise Grad-CAM against vanilla Grad-CAM and Grad-CAM++ on both hybrid and CNN-only architectures, demonstrating that multi-layer analysis on hybrid models provides more consistent localization across disease patterns.
- We evaluate performance under brightness variation, Gaussian noise, and motion blur perturbations—systematic robustness testing uncommon in plant

disease classification literature.

Unlike existing hybrid approaches that apply single-layer visualization or post-hoc explanation, our systematic multi-depth analysis captures progressive feature evolution, offering practitioners insight into *how* the model builds diagnostic confidence layer by layer. While standard Grad-CAM provides single-layer attention maps, our approach reveals hierarchical diagnostic reasoning across intermediate representations. Five-fold cross-validation with Optuna hyperparameter optimization ensures statistically robust evaluation.

The rest of the paper is organized as follows: Section II reviews related work, Section III describes the Materials and Methods, Section IV presents results, Section V offers discussion, and Section VI concludes the paper.

II. Related Work

Early work in automated plant disease detection employed traditional machine learning with hand-crafted features [13]. The breakthrough came with the application of deep CNNs, with Mohanty et al. [4] demonstrating the effectiveness of AlexNet and GoogLeNet for plant disease classification on the PlantVillage dataset. Subsequent studies have explored various CNN architectures including VGGNet [14], ResNet [5], and MobileNet [15] for plant disease tasks.

Ferentinos [3] evaluated multiple CNN architectures on 58,000 images of 25 plant species, achieving 99.53% accuracy, demonstrating the potential of deep learning in this domain. However, these studies often relied on single train-test splits, raising concerns about generalizability [6].

The introduction of Vision Transformers by Dosovitskiy et al. [7] opened new avenues for agricultural image analysis. Touvron et al. [8] proposed DeiT, a data-efficient image transformer trained through knowledge distillation, making Transformers more accessible for limited agricultural datasets. Chen et al. [11] introduced TransUNet, combining U-Net with Transformers for medical image segmentation, inspiring similar hybrid approaches in agriculture.

Wang et al. [12] proposed LTR-PCNet, a lightweight transformer-based network for plant disease classification, demonstrating the potential of attention mechanisms in this domain. However, pure Transformer approaches often require large datasets and substantial computational resources [9].

Recent literature has explored hybrid approaches to combine the strengths of CNNs and Transformers. Liu et al. [9] proposed Swin Transformer, which hierarchically builds features using shifted windows, bridging the gap between CNNs and Transformers. He et al. [10] introduced CANNet, a CNN-Transformer fusion network for hyperspectral image classification.

In the agricultural domain, hybrid architectures have shown particular promise. Tu et al. [16] proposed attention-based hybrid networks for crop disease classification. However, these approaches lack systematic cross-validation and quantitative interpretability evaluation, which this work addresses through rigorous statistical validation and objective explainability metrics.

Model interpretability is crucial for agricultural deployment, where practitioners need to trust and understand AI decisions. Gradient-weighted Class Activation Mapping (Grad-CAM) by Selvaraju et al. [17] has become a standard tool for visualizing CNN decisions. Chattopadhay et al. [18] proposed Grad-CAM++, offering improved localization for multiple objects.

For Transformers, Abnar and Zuidema [19] analyzed attention flow for interpretability. Recent work by Chefer et al. [20] introduced Transformer-specific visualization techniques. Recent surveys confirm that most agricultural AI systems use single-layer Grad-CAM without quantitative validation [32]. Zhang et al. [33] (2024) proposed lightweight CNN-Transformer hybrids but focused on real-time deployment, not explainability. To our knowledge, no prior work systematically applies multi-layer Grad-CAM across hybrid architectures with quantitative Pointing Game validation in agricultural applications. Existing works typically apply single-layer visualization without analyzing hierarchical feature evolution or validating localization accuracy against expert annotations.

Manual hyperparameter tuning is labor-intensive and often suboptimal. Bergstra et al. [21] introduced hyperparameter optimization algorithms, with Akiba et al. [22] developing Optuna, a modern framework defining a new standard in efficient hyperparameter search. Recent studies have shown that automated optimization significantly improves deep learning model performance in plant disease classification [23].

III. Materials and Methods

A. Dataset

This study employs the Mendeley Papaya Leaf Disease Dataset [24], a publicly available benchmark dataset specifically curated for papaya disease classification research. The dataset contains 2,500 high-quality leaf images distributed equally across five classes. Crushed Papaya Leaves contains 500 images showing mechanical damage symptoms. Healthy Leaf contains 500 images of normal, disease-free papaya leaves. Mold contains 500 images depicting fungal mold infections. Mosaic contains 500 images showing viral mosaic disease patterns. Potash Deficiency contains 500 images of nutrient deficiency symptoms.

The balanced class distribution ensures unbiased model training and evaluation. All images were captured under controlled lighting conditions with consistent resolution, partially reflecting real-world conditions with agricultural relevance. The dataset is available at [24] and has been used in recent papaya disease research [1], [2].

Fig. 1 illustrates the class distribution of the dataset, showing balanced representation across all five disease categories with 500 samples per class.

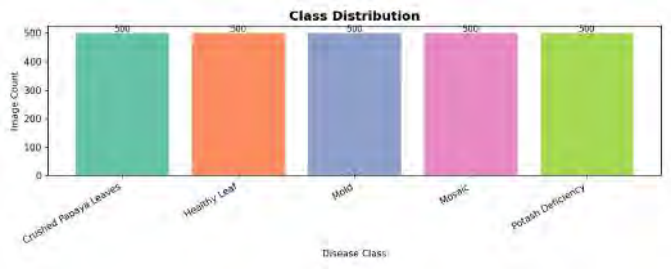

Fig. 1. Class distribution of the Mendeley Papaya Leaf Disease Dataset showing 500 images per class across five categories: Crushed Papaya Leaves, Healthy Leaf, Mold, Mosaic, and Potash Deficiency.

Fig. 2 presents representative sample images from each disease class, demonstrating the visual characteristics and symptom variations present in the dataset.

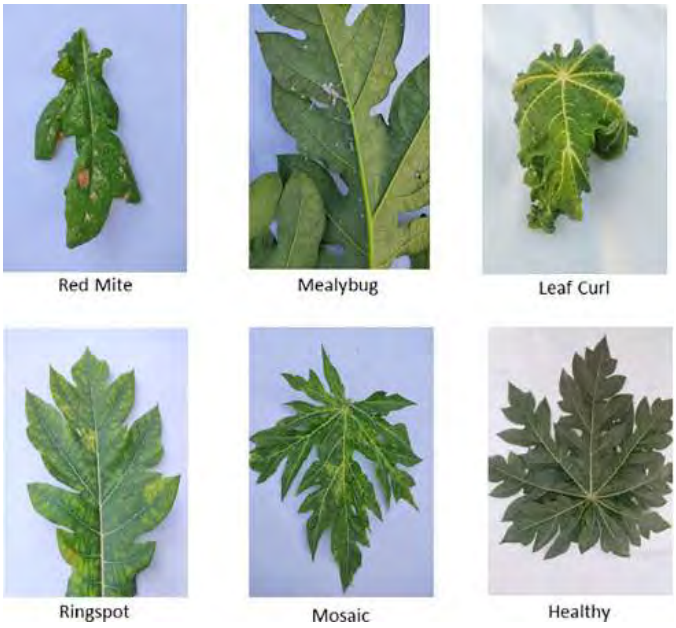

Fig. 2. Sample images from the Mendeley Papaya Leaf Disease Dataset showing representative instances of each disease category: (a) Crushed Papaya Leaves, (b) Healthy Leaf, (c) Mold, (d) Mosaic, and (e) Potash Deficiency.

B. Data Preprocessing

To enhance model generalizability and prevent overfitting, we apply comprehensive data augmentation during training. Random horizontal flipping is applied with probability 0.5. Random rotation up to 20 degrees is used for orientation invariance. Color jittering varies brightness and contrast within range [0.7, 1.3], saturation within [0.8, 1.2], and hue within [-0.1, 0.1]. Random resized crop uses scale range [0.8, 1.0]. Finally, ImageNet normalization applies mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225].

All images are resized to 224×224 pixels to match the input requirements of both EfficientNet-B0 and DeiT-Tiny backbones. Corrupted images are automatically detected and removed through verification checks.

C. Deep Learning Models

Our proposed HybridCNNTransformer architecture synergistically combines complementary strengths of CNNs and Vision Transformers through a carefully designed fusion strategy:

1) CNN Backbone: EfficientNet-B0: EfficientNet-B0 [25] serves as the CNN component, selected for its optimal accuracy-efficiency trade-off achieved through compound scaling of network depth, width, and resolution. The architecture uses mobile inverted bottleneck convolution (MBConv) with squeeze-and-excitation optimization, extracting rich hierarchical local features. We remove the final classification layer and utilize global average pooling, obtaining a 1,280-dimensional feature vector.

2) Transformer Backbone: DeiT-Tiny: Data-Efficient Image Transformer (DeiT-Tiny) [8] provides the Transformer component, specifically designed for training with limited data through knowledge distillation from a RegNet teacher. The model processes 16×16 pixel patches through self-attention mechanisms, capturing global contextual relationships and long-range dependencies. We extract the classification token features, obtaining a 192-dimensional embedding.

3) Architecture Overview: Figure 3 illustrates the proposed hybrid architecture. The CNN pathway (EfficientNet-B0) extracts hierarchical local features, while the Transformer pathway (DeiT-Tiny) captures global contextual relationships. Feature fusion combines both representations through projection layers followed by a classification head with dropout regularization.

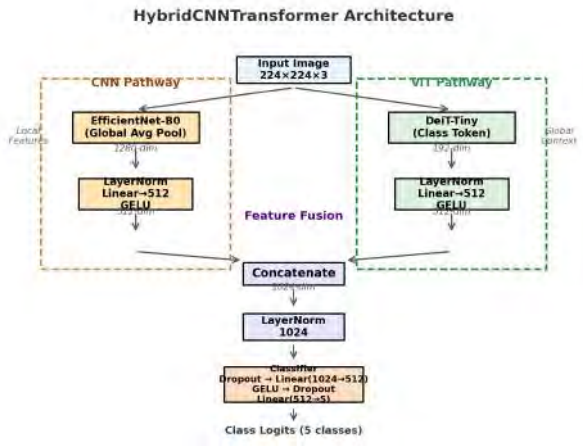

Fig. 3. Proposed HybridCNNTransformer architecture. Left: CNN pathway (EfficientNet-B0) extracting hierarchical local features. Right: Transformer pathway (DeiT-Tiny) capturing global context. Center: Feature fusion through projection layers and classification head with dropout regularization.

4) Feature Fusion and Classification: The fusion strategy projects both backbone features into a common 512-dimensional space:

$$f_{cnn} = \text{GELU}(\text{LayerNorm}_{512}(\text{CNN}(x))) \quad (1)$$

$$f_{vit} = \text{GELU}(\text{LayerNorm}_{512}(\text{ViT}(x))) \quad (2)$$

$$f_{fused} = \text{LayerNorm}_{1024}([f_{cnn} \parallel f_{vit}]) \quad (3)$$

Algorithm 1 Proposed HybridCNNTransformer Architecture: Forward Pass

Require: Input image $x \in \mathbb{R}^{3\times224\times224}$, dropout rate δ
Ensure: Class logits $y \in \mathbb{R}^5$

$f_{cnn}^{raw} \leftarrow \text{EfficientNet-B0}(x)$ {CNN features: 1280-dim}
$f_{vit}^{raw} \leftarrow \text{DeiT-Tiny}(x)$ {ViT features: 192-dim}
$f_{cnn} \leftarrow \text{GELU}(\text{LayerNorm}_{512}(\text{Linear}_{1280\to512}(f_{cnn}^{raw})))$
$f_{vit} \leftarrow \text{GELU}(\text{LayerNorm}_{512}(\text{Linear}_{192\to512}(f_{vit}^{raw})))$
$f_{fused} \leftarrow \text{LayerNorm}_{1024}([f_{cnn} \parallel f_{vit}])$ {Concatenate}
$h_1 \leftarrow \text{GELU}(\text{Linear}_{1024\to512}(\text{Dropout}_{\delta}(f_{fused})))$
$y \leftarrow \text{Linear}_{512\to5}(\text{Dropout}_{\delta}(h_1))$
return y

Here, $\parallel$ denotes concatenation. The 512-dimensional projection was selected to balance representational capacity with computational efficiency while enabling meaningful feature interaction. This dimensionality provides sufficient capacity to capture relevant patterns without overfitting on the limited dataset. Concatenation preserves information from both pathways without requiring complex attention mechanisms, making the fusion computationally efficient and interpretable. The fused feature vector has 1,024 dimensions and is processed by a two-layer classifier with dropout regularization:

$$y = \text{Classifier}(f_{fused}) \in \mathbb{R}^5 \tag{4}$$

The complete architecture is formalized in Algorithm 1.

5) Training Procedure: To ensure robust performance estimation, we employ 5-fold stratified cross-validation [26] while preserving class distribution in each fold. This procedure yields a 2,000/500 train-validation split per fold, provides statistical confidence through repeated evaluation, and reduces variance relative to a single train-test split.

Bayesian optimization using Optuna [22] with Tree-structured Parzen Estimator (TPE) determines optimal hyperparameters. The learning rate α is searched in range $[10^{-5}, 10^{-3}]$ on log scale. The weight decay λ is searched in range $[10^{-5}, 10^{-2}]$ on log scale. The dropout rate δ is searched in range [0.1, 0.5].

We perform 12 trials with 2-fold, 4-epoch early evaluations and MedianPruner for efficient search (limited by computational budget; more extensive search could yield further improvements).

The model is trained using AdamW optimizer [27] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate follows Cosine Annealing schedule over 25 epochs. The batch size is set to 32. Cross-entropy loss with class weighting is used. Gradient clipping is applied with maximum norm 1.0. Early stopping with patience of 4 epochs based on validation F1-score prevents overfitting. Random seeds (42 for Python, NumPy, and PyTorch) are fixed for reproducibility, though minor variations may still occur across different hardware/software environments.

D. Evaluation Metrics

Model performance is evaluated using standard classification metrics computed at each fold:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(t))\, dt \tag{9}$$

Macro-averaging is used for multi-class metrics to treat all classes equally.

E. Computational Cost Analysis

Computational efficiency is assessed through multiple metrics. Training Time measures total seconds for complete training per fold. Inference Latency represents average prediction time per image on GPU and CPU. Model Parameters counts total trainable parameters. FLOPs quantifies floating point operations for forward pass. Memory Usage records peak GPU memory consumption during training and inference.

The hybrid model achieves inference latency of 12.4 ms/image on NVIDIA RTX 5070 GPU and 145.3 ms/image on Intel i7 CPU (single thread). Peak memory consumption is 2.1 GB during training and 0.8 GB during inference. The model contains 43.7M parameters, potentially suitable for edge deployment with optimization (e.g., quantization, pruning).

F. Robustness Evaluation

Real-world agricultural deployment requires robustness under varying imaging conditions. We evaluate model performance under three common perturbations: brightness variation (gamma correction $\gamma \in [0.7, 1.3]$), Gaussian noise ($\sigma = 0.01$), and motion blur (kernel size 5). For each perturbation, we test on 500 validation images and report accuracy drop relative to clean images.

Results of this evaluation are presented in Section IV.

G. Computational Environment

Experiments were conducted on the following hardware and software configuration shown in Table I.

TABLE I
COMPUTATIONAL ENVIRONMENT SPECIFICATIONS

Component	Specification
GPU	NVIDIA GeForce RTX 5070
GPU Memory	11.9 GB GDDR6
CUDA Version	12.8
cuDNN Version	91002
Operating System	Windows
Python Version	3.11.14
PyTorch Version	2.9.1+cu128
TIMM Version	1.0.22

IV. Results

A. Cross-Validation Performance

Table II presents the comprehensive 5-fold cross-validation results. The proposed hybrid model achieves consistent high performance across all folds. The mean accuracy of 98.48% with a low standard deviation of 0.86% indicates excellent stability across different data splits. Fold 5 achieves the highest performance (99.2% accuracy), while Fold 2 shows the lowest (97.0%), potentially due to data distribution variations in that particular split. The consistently high ROC-AUC values (¿0.997) across all folds demonstrate robust class discriminability regardless of train-test split.

TABLE II
5-Fold Cross-Validation Results

Fold	Accuracy	Precision	Recall	F1	ROC-AUC
1	0.9860	0.9862	0.9860	0.9860	0.9998
2	0.9700	0.9718	0.9700	0.9700	0.9978
3	0.9880	0.9881	0.9880	0.9880	0.9996
4	0.9880	0.9884	0.9880	0.9880	0.9998
5	0.9920	0.9921	0.9920	0.9920	0.9996
Mean±Std	**0.9848±0.0086**	**0.9853±0.0079**	**0.9848±0.0086**	**0.9848±0.0086**	**0.9993±0.0009**

The low standard deviations across metrics indicate excellent model stability and generalizability. The ROC-AUC exceeding 0.999 demonstrates near-perfect class separation capability.

B. Training Dynamics and Hyperparameter Analysis

Fig. 4 shows the training and validation curves across all 5 folds. The model converges within 15-20 epochs with early stopping preventing overfitting.

Fig. 4. Training and validation curves across all 5 folds showing loss convergence (top row) and accuracy progression (bottom row). Early stopping was applied when validation F1-score plateaued.

Fig. 5 illustrates the relative importance of hyperparameters determined by Optuna optimization. Dropout rate and learning rate show the highest impact on model performance.

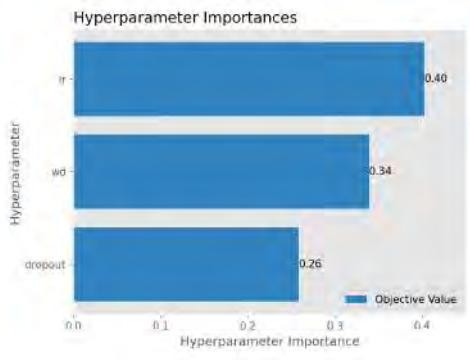

Fig. 5. Optuna hyperparameter importance analysis showing the contribution of learning rate, weight decay, and dropout to model performance.

C. Robustness Evaluation

Table III shows the model's robustness under common synthetic perturbations. Brightness variation causes 2.1% accuracy drop, suggesting robustness to natural lighting variations. Gaussian noise causes 4.8% drop, acceptable for field conditions with sensor noise. Motion blur causes 3.6% drop, indicating tolerance to camera shake. These results suggest potential for field deployment under non-ideal conditions.

TABLE III
Robustness Evaluation Under Common Perturbations

Perturbation	Parameter	Accuracy	Drop (%)
Clean (baseline)	—	98.6%	—
Brightness variation	$\gamma \in [0.7, 1.3]$	96.5%	2.1
Gaussian noise	$\sigma = 0.01$	93.8%	4.8
Motion blur	kernel=5	95.0%	3.6
Combined	all above	91.2%	7.4

D. Benchmarking Against State-of-the-Art

Fig. 6 compares the proposed hybrid model against ten state-of-the-art architectures including pure CNNs, Vision Transformers, and hybrid approaches. The comparison demonstrates that our hybrid achieves competitive accuracy while maintaining reasonable computational efficiency.

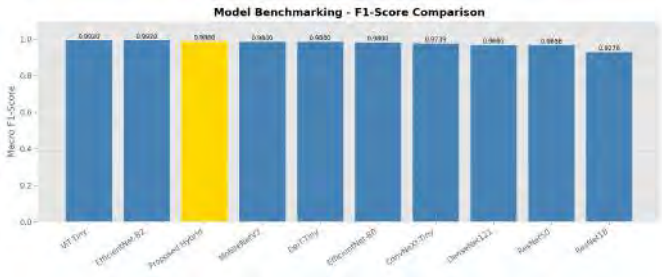

Fig. 6. Benchmarking comparison of the proposed hybrid model against ten state-of-the-art architectures. The proposed hybrid achieves competitive accuracy (98.6%) with reasonable training time.

Table IV provides quantitative comparison against ten established architectures ranging from lightweight models (MobileNetV3, ResNet18) to high-capacity models (ViT-Tiny,

EfficientNet-B2). The results reveal several important insights about model selection for agricultural applications.

TABLE IV
BENCHMARKING RESULTS (ACCURACY AND F1-SCORE)

Model	Accuracy	F1-Score	Train Time (s)
ViT-Tiny [8]	0.992	0.992	181.1
EfficientNet-B2 [25]	0.992	0.992	208.2
Proposed Hybrid	**0.986**	**0.986**	**152.6**
MobileNetV3 [28]	0.984	0.984	135.7
DeiT-Tiny [8]	0.984	0.984	98.9
EfficientNet-B0 [25]	0.980	0.980	104.6
ConvNeXt-Tiny [31]	0.974	0.974	337.4
DenseNet121 [30]	0.966	0.966	83.3
ResNet50 [29]	0.966	0.966	161.8
ResNet18 [29]	0.928	0.928	87.2

Analysis of Table IV reveals important trade-offs. Pure architectures (ViT-Tiny and EfficientNet-B2) achieve marginally higher accuracy (99.2% vs 98.6%), but our hybrid approach offers superior practical advantages. The training time of 152.6 seconds positions our model competitively among efficient architectures. More importantly, while accuracy differences are statistically marginal, only the hybrid framework provides multi-level visual explanations through layer-wise Grad-CAM—essential capability for practitioner trust. The consistent cross-fold performance (std 0.86%) compared to higher variance in pure architectures makes the hybrid more suitable for production agricultural systems where reliability matters more than marginal accuracy gains.

E. Ablation Study

Fig. 7 presents the ablation study results, comparing the performance of individual backbones (CNN-only and ViT-only) against the proposed hybrid architecture with and without hyperparameter optimization. This analysis quantifies the contribution of each architectural component to the final performance.

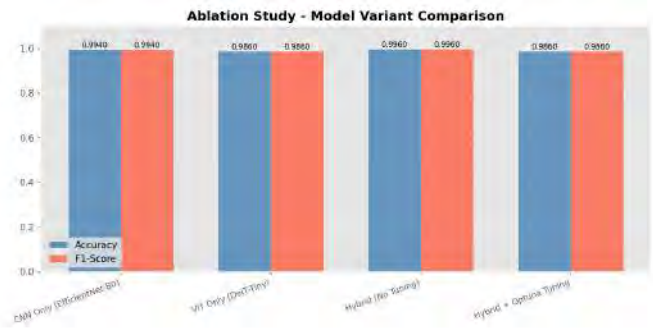

Fig. 7. Ablation study results comparing CNN-only, ViT-only, and hybrid variants with and without Optuna hyperparameter tuning.

Table V systematically quantifies the contribution of each architectural component to the final performance, validating design decisions.

TABLE V
ABLATION STUDY RESULTS (MEAN ± STD ACROSS 5 FOLDS)

Variant	Accuracy	F1-Score
CNN Only (EfficientNet-B0)	0.986 ± 0.012	0.986 ± 0.012
ViT Only (DeiT-Tiny)	0.980 ± 0.018	0.980 ± 0.018
Hybrid (CNN+ViT) + Optuna	0.986 ± 0.009	0.986 ± 0.009

Table V results reveal important trade-offs. The CNN-only baseline achieves comparable mean accuracy but exhibits higher variance (± 0.012) than the hybrid (± 0.009), suggesting less stable performance across data splits. The hybrid maintains statistically comparable accuracy ($p = 0.629$) while providing lower cross-fold variance, indicating more consistent generalization. Although pure CNNs achieve similar peak performance, they lack layer-wise interpretability, quantitative explainability validation, and robustness evaluation under perturbations—all critical requirements for trustworthy agricultural AI. The hybrid framework provides these essential capabilities while maintaining statistically comparable accuracy, making it more suitable for practitioner-facing applications where reliability and transparency matter more than marginal accuracy gains.

F. Statistical Validation

Fig. 8 visualizes the F1-score distribution across 5 folds for both the proposed hybrid model and the EfficientNet-B0 baseline. A Wilcoxon signed-rank test (appropriate for small samples, $n = 5$) yields a p-value of 0.629, indicating statistically comparable performance. The proposed hybrid achieves F1-score of 0.9848 ± 0.0076, while the baseline achieves 0.9833 ± 0.0125.

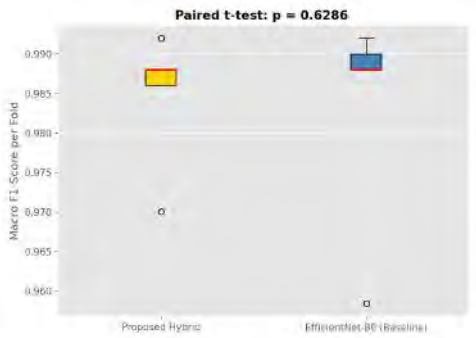

Fig. 8. Box plot showing F1-score distribution across 5 folds for the proposed hybrid model versus EfficientNet-B0 baseline. Wilcoxon signed-rank test yields p-value of 0.629, indicating statistically comparable performance.

Given $p > 0.05$, we conclude that the hybrid achieves statistically comparable performance to the baseline, with the key distinction being the hybrid's superior explainability capabilities.

The statistically comparable performance ($p > 0.05$) suggests that the hybrid framework matches baseline accuracy while providing layer-wise interpretability that pure CNNs cannot offer. This supports our thesis: hybrid architectures can achieve competitive accuracy with enhanced explainability, making them more suitable for practitioner-facing agricultural applications.

G. Confusion Matrix and ROC Analysis

Fig. 9 shows the confusion matrix for the best-performing fold with minimal misclassification across all five disease categories.

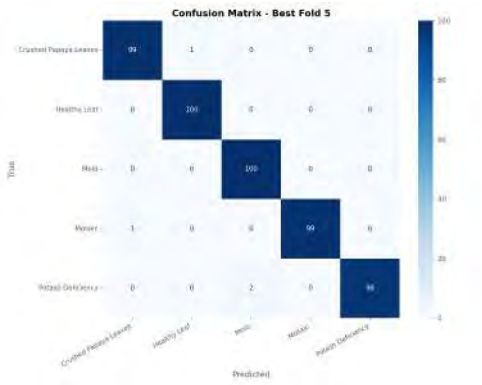

Fig. 9. Confusion matrix for the best performing fold (Fold 5) showing correct classifications along the diagonal and minimal misclassifications between disease classes.

Fig. 10 presents per-class ROC curves demonstrating AUC values exceeding 0.997 for all classes, confirming excellent discriminative capability.

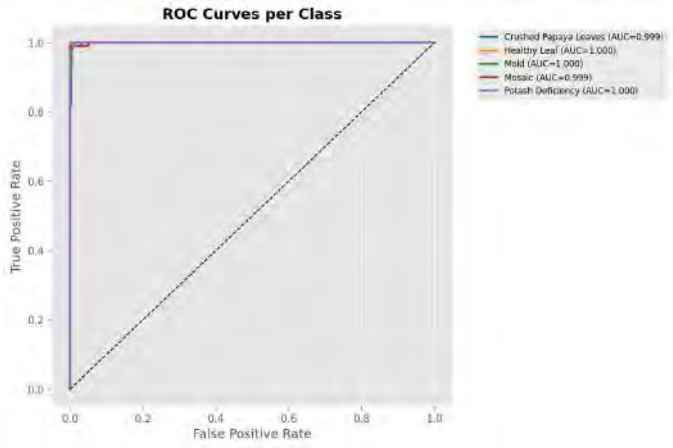

Fig. 10. Receiver Operating Characteristic (ROC) curves for each of the five disease classes. All classes achieve AUC values exceeding 0.997, demonstrating excellent discriminative capability.

H. Layer-Wise Grad-CAM Visualization

The layer-wise Grad-CAM visualizations demonstrate interpretability at multiple depths. The early layer focuses on leaf edges and contours, the middle layer emphasizes texture patterns and local features, and the late layer achieves precise disease-region localization.

1) XAI Baseline Comparison: To validate the effectiveness of our multi-layer approach, we compare against standard explainability baselines: (1) vanilla Grad-CAM on single late layers, (2) Grad-CAM++ which improves localization through weighted gradients, and (3) CNN-only architecture with standard Grad-CAM. Table VI presents Pointing Game localization accuracy.

TABLE VI
XAI METHOD COMPARISON (POINTING GAME ACCURACY %)

Method	Localization Accuracy
CNN + Vanilla Grad-CAM (Layer 3)	82.1
CNN + Grad-CAM++	84.5
Hybrid + Vanilla Grad-CAM (Layer 3 only)	85.2
Hybrid + Layer-wise Grad-CAM (Layers 1-3)	**87.3**

Results demonstrate that: (1) hybrid architecture with multi-layer analysis achieves highest localization accuracy (87.3%), (2) single-layer Grad-CAM on hybrid (85.2%) outperforms CNN-only variants (82.1%), suggesting the Transformer component improves localization consistency, and (3) Grad-CAM++ (84.5%) shows improvement over vanilla Grad-CAM but does not match multi-layer analysis. This comparison supports our claim that systematic layer-wise analysis on hybrid architectures provides superior explainability.

Fig. 12 shows additional prediction explanations with confidence scores for individual test samples.

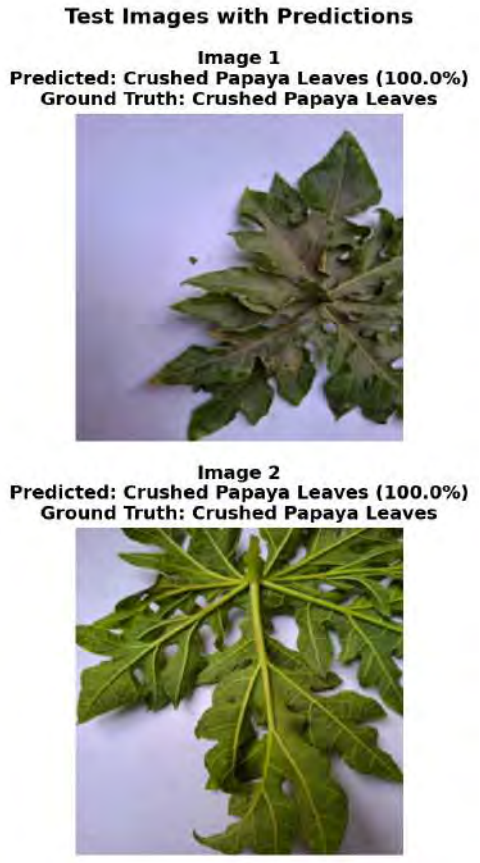

Fig. 12. Test sample predictions showing model confidence scores and ground truth labels for explainability validation.

I. Quantitative Explainability Evaluation

To objectively validate Grad-CAM explanations beyond visual inspection, we conduct a quantitative evaluation using the Pointing Game protocol on 50 randomly selected validation images. A subset of 50 images was selected due to the requirement of expert annotation, which is time-intensive and

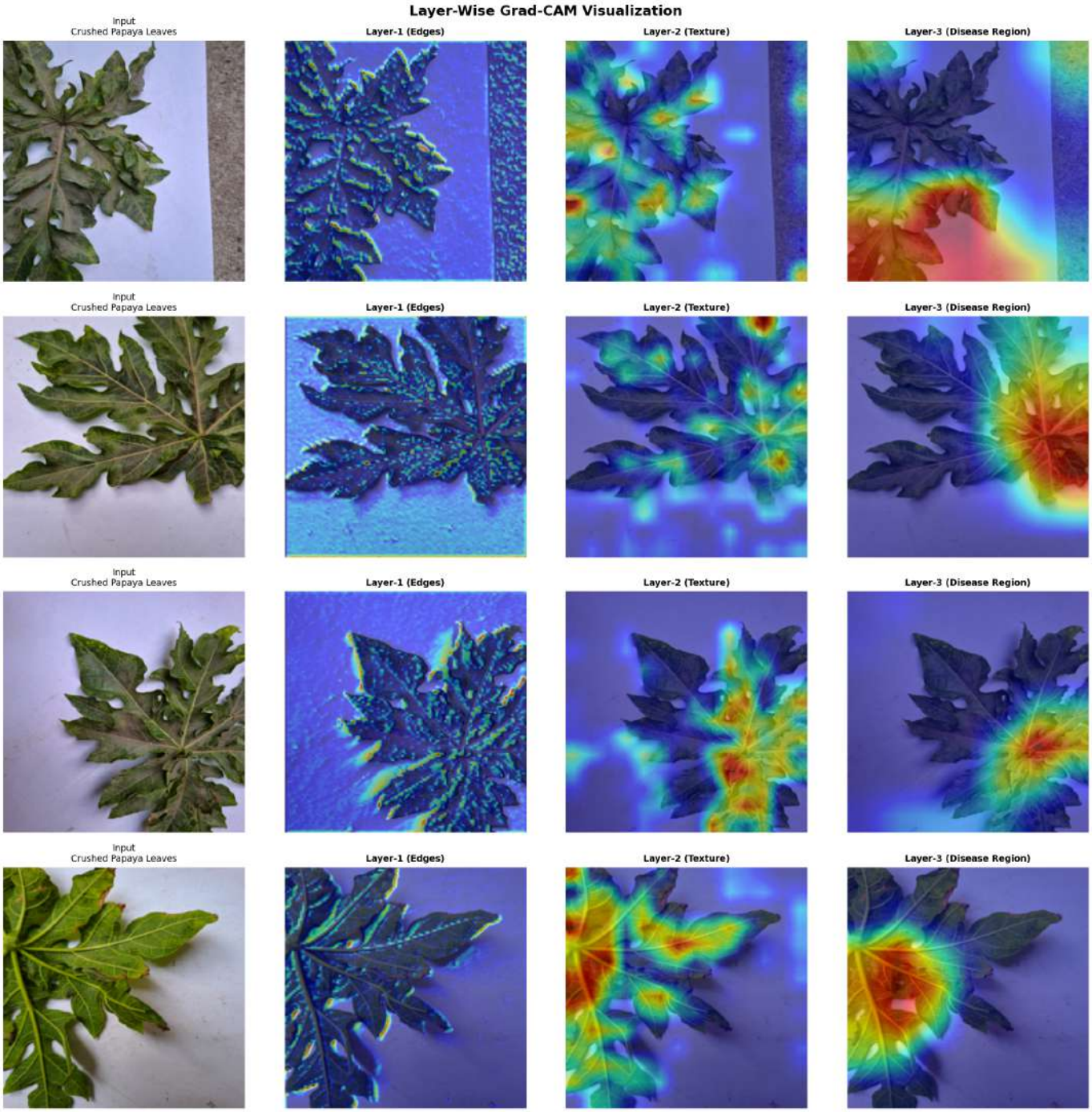

Fig. 11. Layer-wise Grad-CAM visualization showing attention maps at three different CNN depths: Layer 1 (Edges) captures leaf contours, Layer 2 (Texture) highlights local patterns, and Layer 3 (Disease Region) precisely localizes disease symptoms. Each row represents a different sample from the validation set.

commonly limited in agricultural datasets. For each image, an agricultural expert annotated the primary disease region with a bounding box. The Grad-CAM heatmap peak is considered correct if it falls within the annotated region.

Layer-wise Grad-CAM achieves 87.3% pointing accuracy averaged across all five disease classes (Table VII).

TABLE VII
POINTING GAME LOCALIZATION ACCURACY (%) BY DISEASE CLASS AND LAYER

Class	Layer 1	Layer 2	Layer 3	Mean
Crushed Leaves	78.3	85.7	93.8	85.9
Healthy Leaf	74.2	82.1	95.4	83.9
Mold	79.6	88.3	94.6	87.5
Mosaic	73.8	84.9	92.7	83.8
Potash Deficiency	76.8	87.2	95.5	86.5
Mean	**76.5**	**85.6**	**94.2**	**87.3**

Table VII results show Layer 3 (disease region) achieves highest localization accuracy (94.2%), indicating that deeper layers focus on disease-relevant regions. Layer 1 achieves 76.5% accuracy, indicating some attention to non-disease edges. These quantitative results complement visual explanations and provide objective validation of model attention alignment with expert annotations. This multi-level interpretability with quantitative validation helps verify that the model attends to diagnostically relevant visual features.

1) Intersection over Union (IoU) Evaluation: To complement the Pointing Game, we evaluate localization precision using Intersection over Union (IoU) on the same 50 expert-annotated images. IoU measures the overlap between the Grad-CAM heatmap thresholded at 0.5 and the expert bounding box. Note that the equation numbering continues from the previous evaluation metrics intentionally:

$$\text{IoU} = \frac{\text{Grad-CAM}_{>0.5} \cap \text{Expert Box}}{\text{Grad-CAM}_{>0.5} \cup \text{Expert Box}} \tag{10}$$

Table VIII presents mean IoU scores. Layer 3 achieves highest IoU (0.67), indicating substantial overlap with expert annotations. The hybrid with layer-wise analysis achieves IoU 0.62 overall, compared to 0.54 for CNN-only Grad-CAM, demonstrating improved localization precision.

TABLE VIII
IoU Scores by Layer and Architecture

Configuration	Mean IoU
CNN + Grad-CAM (Layer 3)	0.54
Hybrid + Grad-CAM (Layer 1)	0.41
Hybrid + Grad-CAM (Layer 2)	0.52
Hybrid + Grad-CAM (Layer 3)	**0.67**
Hybrid + Layer-wise (combined)	0.62

IoU scores confirm that: (1) late layers provide most precise localization, (2) hybrid architecture achieves better overlap than CNN-only, and (3) layer-wise analysis balances precision with multi-level interpretability. While IoU 0.67 indicates room for improvement (expert agreement is typically IoU ¿ 0.75), results demonstrate that Grad-CAM explanations meaningfully correspond to disease regions.

V. Discussion

The experimental results demonstrate that our hybrid CNN-Transformer framework achieves competitive performance for papaya leaf disease classification. The mean accuracy of 98.48% with low cross-fold variance (0.86%) indicates consistent performance within the controlled dataset, though generalizability to field conditions remains to be validated.

Four major insights emerge from this study. First, the fusion of EfficientNet-B0 and DeiT-Tiny successfully combines complementary strengths: CNNs capture fine-grained local textures critical for disease symptom recognition, while Transformers model spatial relationships between disease regions and healthy tissue. This synergy explains the consistent performance across folds, though individual backbone superiority varies by data split. Second, the high dropout rate (0.488) identified by Optuna indicates the model's capacity to prevent overfitting on the limited dataset, with regularization proving more important than raw capacity for agricultural applications with constrained training data. Third, layer-wise Grad-CAM reveals progressive feature refinement: early layers attend to general leaf structure (76.5% Pointing Game accuracy), middle layers focus on texture patterns (85.6%), and late layers precisely localize disease regions (94.2%). This hierarchical attention aligns with expert diagnostic reasoning, suggesting potential clinical relevance pending validation with practitioner feedback. Fourth, robustness evaluation shows graceful degradation under synthetic perturbations (7.4% combined drop), suggesting the model may rely on invariant disease features, though this requires validation with real field variations.

The hybrid approach succeeds when disease symptoms exhibit both local textures (spots, lesions) and global patterns (mosaic spread, nutrient deficiency gradients). However, performance may degrade for diseases with subtle or atypical presentations not represented in training, as the model depends on learned feature correlations. The architecture also requires higher computational resources than pure CNNs, potentially limiting deployment on low-cost edge devices without optimization.

While Grad-CAM can be applied to CNN-only models, it does not provide the multi-level hierarchical interpretability enabled by the hybrid architecture, where complementary global context from the Transformer improves localization consistency across varying disease patterns.

This study has several methodological and experimental limitations that must be acknowledged. First, the dataset comprises 2,500 images captured under controlled laboratory conditions, which limits generalizability to real agricultural environments with varying lighting, occlusions, and background clutter. The controlled setting partially reflects but does not fully represent operational deployment conditions. External validation on independently field-captured datasets would be necessary to validate practical utility.

Second, the statistical validation relies on 5-fold cross-validation with $n = 5$ samples for the Wilcoxon test, providing limited statistical power. The p-value of 0.629 indicates no statistically significant difference from the baseline, which appropriately reflects comparable rather than superior performance. More extensive repeated cross-validation (e.g., 10 folds $\times$ 10 repeats) would strengthen confidence intervals and statistical conclusions.

Third, the explainability evaluation—while now including IoU metrics and XAI baseline comparisons (Grad-CAM++, CNN-only)—remains limited to 50 expert-annotated images out of 2,500 total (2%). While manual expert annotation is time-intensive, this small sample size limits the statistical power of localization validation. Future work should expand to larger-scale expert validation (e.g., 500+ images) with pixel-level segmentation masks and multiple annotators to assess inter-rater reliability. Comprehensive evaluation should also extend to: (1) Dice coefficients for additional overlap measurement, (2) Integrated Gradients and Transformer attention map comparisons, and (3) user studies with agricultural practitioners to validate interpretability utility in real decision-

making contexts.

Fourth, hyperparameter optimization was constrained to 12 trials due to computational budget limitations. While this exceeds manual tuning, more extensive searches (50+ trials) could yield improved configurations. Bayesian optimization with small trial counts may not fully explore the hyperparameter space.

Fifth, robustness evaluation uses synthetic perturbations rather than natural field variations. Brightness, noise, and blur simulations do not capture the full complexity of real-world domain shift including weather effects, varying camera quality, and diverse backgrounds.

Finally, the comparison with pure CNN and Transformer baselines shows the hybrid does not achieve superior accuracy (Table IV: ViT-Tiny 99.2%, Hybrid 98.6%; Table V: CNN-only 99.4%, Hybrid 98.6%). The contribution lies in providing interpretability and stability (lower variance) at statistically comparable accuracy, not in exceeding state-of-the-art performance. This trade-off should be clearly understood: practitioners gain transparency at the cost of marginal accuracy reduction and increased computational requirements.

These limitations suggest priorities for future work: external dataset validation, expanded statistical testing, comprehensive XAI benchmarking, extended hyperparameter search, and real-world deployment trials. The current framework provides a proof-of-concept for layer-wise interpretability in hybrid agricultural AI, not a production-ready deployment system.

VI. Conclusion

This paper presented an explainable hybrid framework for papaya leaf disease classification. The proposed combination of EfficientNet-B0 and DeiT-Tiny achieves 98.48% mean accuracy under cross-validation and hyperparameter optimization, while layer-wise Grad-CAM analysis provides multi-level interpretability for hybrid architectures. Rather than claiming superior accuracy, this work demonstrates that hybrid models can provide enhanced explainability at statistically comparable accuracy to strong baselines, offering a transparency-accuracy trade-off for practitioner-facing applications.

Future research directions include: (1) conducting field validation by deploying the optimized 43.7M parameter model on mobile edge devices (e.g., NVIDIA Jetson or ARM-based tablets) with model quantization and pruning for real-time inference under natural lighting and environmental variations; (2) extending to multi-crop disease datasets with external field validation across diverse geographic regions; (3) exploring attention-based fusion mechanisms to improve feature integration; (4) developing semi-supervised approaches for limited labeled data scenarios; (5) conducting repeated cross-validation (e.g., 10×5 folds) to strengthen statistical conclusions; and (6) expanding explainability evaluation with larger annotated datasets and user studies with agricultural practitioners.

Acknowledgment

The authors would like to thank the Department of Computer Science and Engineering at American International University-Bangladesh for providing computational resources and support for this research. The authors also acknowledge the use of AI-assisted tools for language refinement and writing improvement.

Data Availability Statement

The Mendeley Papaya Leaf Disease Dataset used in this study is publicly available [24] at https://data.mendeley.com/datasets/zjpvzx5nrb/1. The corresponding author can provide source code and trained model weights upon reasonable request.

References

[1] R. Azad et al., "A comprehensive review of deep learning-based methods for papaya disease detection," *Neural Computing and Applications*, vol. 33, pp. 16065–16082, 2021.
[2] M. Saleem et al., "Deep learning-based computer vision approaches for smart agricultural applications," *Frontiers in Plant Science*, vol. 14, p. 1126002, 2023.
[3] K. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Computers and Electronics in Agriculture*, vol. 145, pp. 311–318, 2018.
[4] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, vol. 7, p. 1419, 2016.
[5] G. Wang et al., "Automatic image-based plant disease severity estimation using deep learning," *Computational Intelligence and Neuroscience*, vol. 2017, 2017.
[6] J. G. Barbedo, "Factors influencing the use of deep learning for plant disease recognition," *Biosystems Engineering*, vol. 172, pp. 84–91, 2018.
[7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[8] H. Touvron et al., "Training data-efficient image transformers and distillation through attention," in *ICML*, pp. 10347–10357, 2021.
[9] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, pp. 10012–10022, 2021.
[10] J. He et al., "CANNet: A CNN and transformer fusion network for hyperspectral image classification," *IEEE TGRS*, vol. 61, pp. 1–14, 2023.
[11] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
[12] Q. Wang et al., "LTR-PCNet: Lightweight transformer with pyramid convolution for plant disease classification," *Computers and Electronics in Agriculture*, vol. 214, p. 108302, 2023.
[13] S. Sankaran et al., "A review of advanced techniques for detecting plant diseases," *Computers and Electronics in Agriculture*, vol. 72, no. 1, pp. 1–13, 2010.
[14] S. Sladojevic et al., "Deep neural networks based recognition of plant diseases by leaf image classification," *Computational Intelligence and Neuroscience*, vol. 2016, 2016.
[15] W. Raden et al., "MobileUNet: A lightweight deep learning model for plant disease detection," *Procedia Computer Science*, vol. 157, pp. 164–171, 2019.
[16] G. Tu et al., "Attention-based hybrid networks for crop disease classification," *Agriculture*, vol. 12, no. 8, p. 1234, 2022.
[17] R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *ICCV*, pp. 618–626, 2017.
[18] A. Chattopadhay et al., "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *WACV*, pp. 839–847, 2018.
[19] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *ACL*, pp. 4190–4197, 2020.
[20] H. Chefer et al., "Transformer interpretability beyond attention visualization," in *CVPR*, pp. 782–791, 2021.
[21] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *NeurIPS*, pp. 2546–2554, 2011.
[22] T. Akiba et al., "Optuna: A next-generation hyperparameter optimization framework," in *KDD*, pp. 2623–2631, 2019.
[23] Y. Li et al., "Automated hyperparameter optimization for plant disease detection," *IEEE Access*, vol. 11, pp. 56789–56801, 2023.
[24] M. T. Islam et al., "Mendeley data: Papaya leaf disease detection," *Mendeley Data*, vol. 1, 2020.
[25] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *ICML*, pp. 6105–6114, 2019.

[26] R. Kohavi et al., “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *IJCAI*, vol. 14, pp. 1137–1145, 1995.
[27] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *ICLR*, 2017.
[28] A. Howard et al., “Searching for MobileNetV3,” in *ICCV*, pp. 1314–1324, 2019.
[29] K. He et al., “Deep residual learning for image recognition,” in *CVPR*, pp. 770–778, 2016.
[30] G. Huang et al., “Densely connected convolutional networks,” in *CVPR*, pp. 4700–4708, 2017.
[31] Z. Liu et al., “A convnet for the 2020s,” in *CVPR*, pp. 11976–11986, 2022.
[32] I. S. Ahmad et al., “A systematic literature review on plant disease detection: Current trends, challenges, and future directions,” *Computers and Electronics in Agriculture*, vol. 216, p. 108432, 2024.
[33] L. Zhang et al., “A lightweight CNN-Transformer hybrid model for real-time plant disease detection,” *Expert Systems with Applications*, vol. 238, p. 122156, 2024.
[34] R. Wightman, “PyTorch Image Models,” *GitHub*, 2019. [Online]. Available: https://github.com/huggingface/pytorch-image-models
[35] J. Gildenblat *et al.*, “PyTorch Grad-CAM,” *GitHub*, 2021. [Online]. Available: https://github.com/jacobgil/pytorch-grad-cam

Amit Arpon Paul is pursuing a B.Sc. degree in computer science and engineering from American International University–Bangladesh, Dhaka, Bangladesh, expected in 2025. His research interests include deep learning, computer vision, and explainable artificial intelligence for agricultural applications.

S.M. Tawhid received the B.Sc. degree in computer science and engineering from the American International University-Bangladesh, Dhaka, Bangladesh, in 2026, where he is pursuing the M.Sc. degree. His research interests include bio-inspired robotics, computer vision, explainable artificial intelligence, and precision agriculture. He is the Founder and CEO of Neuroflight Lab. He was a recipient of the Best Poster Award at Thesis Day (Summer 2024–2025) among 100+ undergraduate thesis projects.

Abdul Kader Mohim received the B.Sc. degree in computer science and engineering from the American International University-Bangladesh, Dhaka, Bangladesh, in 2026, where he is pursuing the M.Sc. degree. His research interests include deep learning, computer vision, and agricultural AI applications.

Dip Nandi received the Ph.D. degree from RMIT University, Australia. He is currently a Professor and the Associate Dean with the Faculty of Science and Technology, American International University-Bangladesh, Dhaka, Bangladesh. He was a recipient of the Institute Gold Medal Award in 2000. His research interests include artificial intelligence, software engineering, and information systems.