SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I10P103

Volume 13 Issue 10, 36-53, October 2026
© 2026 Seventh Sense Research Group®

Original Article

Reliability-Aware Explainability for Rice Leaf Disease Classification: Weak Localization, Cross-XAI Comparison, and Cross-Source Generalization

Sujeet Kumar[1], Prajeet Sharma[2]

[1,2]*Department of Computer Science and Engineering, LNCT University, Bhopal, MP, India*.

[1]*Corresponding Author : sujeetkmr44@gmail.com;*

Received: 24 August 2026 Revised: 15 September 2026 Accepted: 18 September 2026 Published: 28 October 2026

Abstract *- This study presents a reliability-oriented framework for evaluating whether high classification performance in rice leaf disease recognition is accompanied by spatially faithful and generalizable explanations. The primary dataset comprised 4,684 images from three disease classes: bacterial blight, brown spot, and leaf smut. Perceptual-hash grouping and group-preserving partitioning were applied to reduce duplicate-related data leakage, and three ImageNet-pretrained residual networks were evaluated under a common training protocol. ResNet50 achieved the highest observed held-out performance, with an accuracy of 0.9968 and macro-F1 of 0.9967. Explanation reliability was evaluated using a 286-image Mini-GT benchmark with lesion-level bounding boxes, together with localization metrics, calibration analysis, perturbation robustness, and threshold sensitivity. Despite perfect classification on this evaluation subset, Grad-CAM++ produced limited localization, with a mean IoU of 0.1831 and pointing-game accuracy of 0.4720. A comparative evaluation of six XAI methods found no significant overall difference in paired per-image IoU (Friedman $\chi^2(5)=4.9796$, $p=0.4184$), indicating that the observed localization limitation was not specific to Grad-CAM++. Independent cross-source evaluation on a non-overlapping 120-image dataset yielded an accuracy of 0.5667 (95% CI: 0.4773–0.6519) and macro-F1 of 0.5453, revealing substantial degradation in external generalization. Overall, the findings demonstrate that high internal classification performance does not necessarily imply reliable lesion localization, stable explanations across XAI methods, or strong cross-source generalization.*

Keywords *- Rice leaf disease, Explainable Artificial Intelligence, Weak localization, XAI reliability, Lesion localization, Cross-source generalization, Precision agriculture.*

1. Introduction

Rice is one of the most important food crops in the world and remains central to food security in many regions. Because rice yield is highly sensitive to leaf diseases, timely diagnosis is critical for preventing productivity loss and reducing the cost of manual inspection. Public image repositories such as PlantVillage accelerated research on plant disease recognition by enabling reproducible model development at scale [1, 2]. Subsequent progress in residual networks and transfer learning further improved practical feasibility for leaf-image classification tasks [3]. Despite this progress, a persistent gap remains between benchmark accuracy and field trustworthiness. Deep models can predict disease classes with exceptional accuracy while still acting as opaque systems whose decision logic is difficult to validate. Explainable artificial intelligence has therefore become increasingly relevant for agricultural imaging. In particular, Grad-CAM [4] and Grad-CAM++ [5] are widely used because they offer class-specific heatmaps with minimal architectural modification. However, most plant-disease studies continue to present explanations qualitatively rather than evaluating whether the highlighted regions correspond to actual lesions. This is problematic because a visually appealing heatmap may still be spatially diffuse, biased by background cues, or only loosely aligned with disease symptoms
[6, 7].

A related concern is the reliability of both predictions and explanations. The calibration of confidence has been demonstrated to be important even with high-performing neural classifiers [8], yet the accuracy of the explanation is seldom reported when the task is agricultural image classification. Similarly, explanation robustness under perturbation is a less popular topic than classification robustness [9], though in practice, field images are blurred and have changes in lighting and compression, and other distortions. Recent surveys of rice leaf disease detection and explainability of plant diseases demonstrate that more robust validation procedures are needed to move beyond raw accuracy and anecdotal saliency maps [10-13].

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

The present study addresses these challenges through a sequential, reliability-oriented evaluation framework based on a curated rice leaf disease dataset, lesion annotations, and multiple quantitative evaluation procedures. The initial phase establishes a leakage-conscious baseline for three-class rice leaf disease classification using ResNet34, ResNet50, and ResNet152. The second stage generates Grad-CAM++ explanations for the highest-performing backbone under the defined evaluation protocol.

The third stage evaluates the explanations on a Mini-GT test set containing lesion-level bounding boxes, thereby distinguishing predictive performance from explanation faithfulness. The study also evaluates robustness to image corruption, sensitivity to heatmap thresholds, and explanation error modes in addition to overlap and pointing measures. The resulting framework provides a combined assessment of classification performance and quantitative explanation fidelity, allowing the strengths and limitations of lesion-level localization to be examined in practical agricultural settings.

The contribution of this study is methodological and evaluative rather than the introduction of a new classification loss or saliency formulation. Established classification and explanation techniques are integrated within a reliability-oriented evaluation protocol designed to distinguish predictive correctness from spatial explanation fidelity.

Specifically, the study combines perceptual-hash-based leakage control, group-preserving classification, quantitative lesion-level weak localization, localization calibration, perturbation robustness, heatmap-threshold sensitivity, statistical comparison of classification backbones, comparative evaluation of multiple XAI methods, and independent cross-source validation. This design permits the reliability of explanations to be assessed using quantitative evidence rather than visual plausibility alone.

Despite the progress in rice leaf disease classification and explainable artificial intelligence, three methodological gaps remain. First, reported classification performance may be affected by exact or near-duplicate images when dataset partitioning is performed without explicit leakage auditing. Second, XAI studies in plant disease recognition frequently rely on visual inspection of heatmaps, providing limited quantitative evidence of whether highlighted regions correspond to actual disease lesions.

Third, explanation reliability is rarely examined jointly through localization accuracy, confidence calibration, perturbation robustness, threshold sensitivity, and evaluation on an independent image source. Consequently, high classification accuracy alone does not establish that a model provides spatially faithful or transferable explanations. The present study addresses these gaps through a unified leakage-aware and reliability-oriented evaluation framework.

2. Literature Review

The list of literature that can be related to the current work can be divided into four interrelated directions: plant disease, explainable agricultural vision artificial intelligence, localization of the disease symptoms, and reliability oriented assessment of model behavior. These directions should be systematically reviewed since it is not the classification accuracy that creates the contribution of the present work. Instead, it is a combination of predictive modeling, visualization, lesion-level validation, and reliability assessment with awareness of robustness in a single experimental framework. Two interrelated approaches to research on plant disease recognition include high-accuracy visual classification and, more recently, explainable and reliable model behavior. Deep learning was able to replace hand-engineered feature pipelines by using large public datasets, particularly the PlantVillage repository [1], in a variety of crop-disease references. As Mohanty et al. [2] demonstrated at an early stage, convolutional neural networks are capable of very high-quality plant-disease classification. When trained on curated leaf images. Since that time, residual learning-based architectures [3] have become popular due to their good representational power with stable optimization.

In the case of rice, deep learning has now become the paradigm of automated diagnosis. Krishnamoorthy et al. [14] indicated good performance of rice leaf Transfer learning to predict disease, demonstrating the usefulness of pre-trained state-of-the-art in agricultural imaging. DeepRice [15] furthered this trend by incorporating deep and deep-feature-based methods to subtype rice leaf disease. Pai et al. [16] recently suggested a two-CNN model that incorporates feature fusion, providing evidence that architectural enrichment can enhance accuracy of classification on rice disease datasets. These were synthesized in a large recent review by Simhadri et al. [10], evolution and came to the conclusion that transfer learning, ensemble learning, and hybrid Rice leaf disease literature is currently centered on architectures.

However, good performance is not a solution to the interpretability issue alone. Grad-CAM was proposed by Selvaraju et al. [4] as a localization technique based on gradient to generate class-discriminative visual explanations of CNNs, and Chattopadhyay et al. [5] further expanded this concept with Grad-CAM++, which enhances localization behavior when there are multiple or diffuse sources of evidence. Explainable artificial intelligence has become widely marketed in the agricultural field as a making mechanism make decisions more transparent to practitioners. Kinger et al. [11] put an emphasis on the significance of explainability in plant-disease detectors and also argued that disease recognition systems must have visual evidence, interpretable by domain users. Wei et al. [12] also demonstrated that explainable deep learning can reveal model reliance further with lesion textures or larger cues on the appearance of leaves. Habaragamuwa et al. [13] claimed that

existing heatmap-based approaches are informative and, in many cases, not enough completely comprehending attributes that motivate plant-disease choices. Beyond qualitative weakly supervised localization, new work in saliency visualization has highlighted the necessity of evaluation procedures, which regulate the choice of the threshold, model selection, and localization scoring more attentively [17-19].

The dimension of reliability is not so mature, but it is also important. Guo et al. [8] indicated that neural networks are often overconfident, even in cases where they have a high accuracy in classification. The implications of this discovery to the agricultural vision in general are far-reaching since a reliable prediction of the disease can be seen as credible even in the case of a spatially erroneous explanation. More recent plant work has started to incorporate explainability with larger deployment issues. Shafik et al. [20] explainability on combined models with sustainable agricultural decision support, and Al-Falluji and Albahar [21] tested various deep architectures with XAI on rice leaf dataset of diseases, such as the 4,684 images three-class benchmark here. Their findings confirm a trend observed in the literature: classification accuracy can be very high on curated datasets, and explanation quality validation is often secondary.

Based on this, three gaps in methodology are not adequately tackled in the past work. To begin with, numerous studies use traditional train/validation/test splitting without auditing near-duplicate leakage explicitly. Second, explanation maps are often not quantitatively validated on a lesion level. Third, the reliability of explanation is rarely considered in terms of calibration, perturbation robustness or operating-threshold sensitivity. Table 1 summarizes these methodological differences between representative existing studies and the present work.

Table 1. Methodological comparison of existing studies and the present study

Study	Classification / XAI	Quantitative localization	Reliability analysis	Cross-source
Krishnamoorthy et al. [14]	Classification	NR	NR	NR
Ritharson et al. [15]	Classification + XAI	NR	NR	NR
Pai et al. [16]	Classification	NR	NR	NR
Wei et al. [12]	Classification + XAI	NR	NR	NR
Shafik et al. [20]	Classification + XAI*	NR	NR	NR
Al-Falluji and Albahar [21]	Classification + XAI	NR	NR	Limited
Present study	**Classification + XAI**	**IoU + Pointing Game**	**Calibration + Perturbation + Threshold**	**Independent dataset**

NR = not reported in the cited study.

The comparison indicates that the present study integrates these evaluation dimensions within a single reliability-aware framework rather than treating explainability as visualization alone.

3. Dataset Description

This segment explains the datasets that were employed in the research and explains the data characteristics that have a direct impact on the classification and localization results. Since the suggested framework does not measure disease recognition only, but also faithfulness in explanations, it should be noted to define the key rice leaf dataset and the Mini-GT benchmark separately.

Special emphasis is placed on class composition, perceptual hash grouping, split integrity, and lesion-size statistics, as they are strongly affected by factors have an impact on the interpretation of the reported results.

3.1. Main Rice Leaf Dataset

The main dataset contains 4,684 leaf images belonging to three disease categories: bacterial blight, brown spot, and leaf smut. The class sizes are 1,604, 1,620, and 1,460 images, respectively.

The image statistics show that the dataset is visually standardized at 300×300 pixels, with an aspect ratio of 1.0 across the corpus. Although spatial resolution is constant, file size varies considerably, indicating noticeable diversity in compression and visual complexity. A class-distribution summary is shown in Table 2, while the corresponding visual overview is provided in Figure 1.

Table 2 and Figure 1 jointly indicate that the data partitions remain well balanced after leakage-aware grouping. This is important because later differences in model performance can therefore be attributed more credibly to the architectures rather than to strong distributional imbalance.

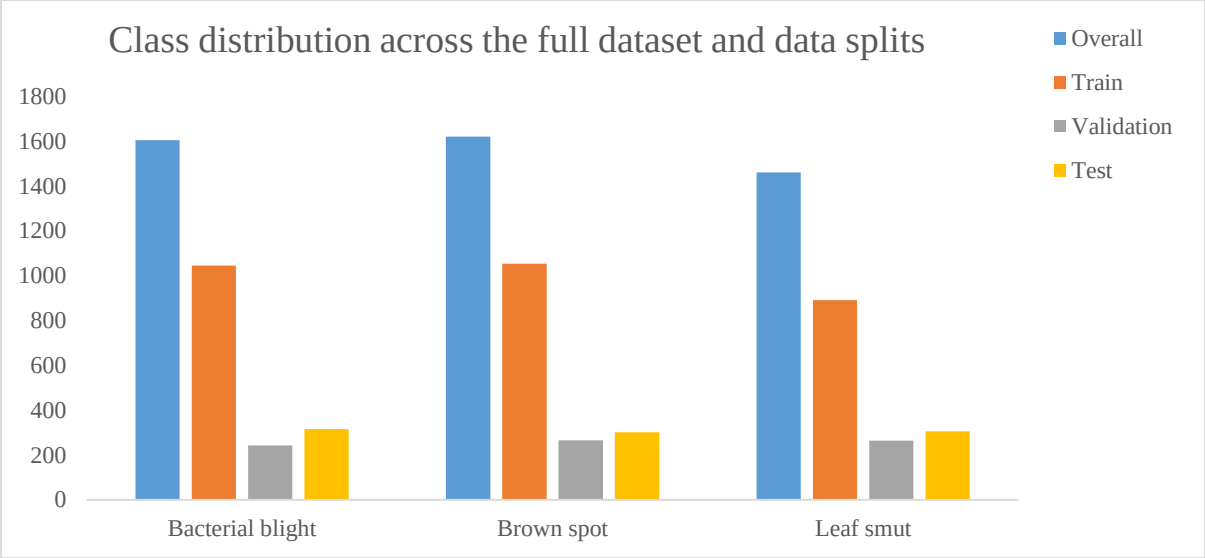


Fig. 1 Overall class distribution and class-preserving train/validation/test allocation for the rice leaf disease dataset

Table 2. Class composition of the main dataset and the group-preserving train/validation/test split

Class	Overall	Train	Validation	Test
Bacterial blight	1604	1045	243	316
Brown spot	1620	1054	265	301
Leaf smut	1460	890	264	306

3.2. Leakage-Aware Grouping Statistics

Instead of naive random splitting, the pipeline used perceptual hashing to identify exact and near-duplicate image groups. This resulted in 1,867 groups across the 4,684 images, with a median group size of 2 and a maximum group size of 12. The grouping statistics are summarized in Table 3. These numbers clarify two important aspects of the dataset. First, the large number of repeated or near-repeated visual groups justifies the use of group preserving splits. Second, the lesion-size statistics of Mini-GT foreshadow why localization will be more difficult than classification: although the mean annotated area is 15.70% of the image, the median is only 4.61%, indicating that many lesions occupy a small fraction of the frame.

Table 3. Additional dataset statistics used in the experimental protocol

Statistic	Value
Total images in main dataset	4684
Unique pHash groups	1867
Median pHash group size	2
Maximum pHash group size	12
Image resolution (mean)	300×300
Image aspect ratio (mean)	1.00
Mean file size	39.02 KB
Mini-GT images	286
Mini-GT class counts	86 / 100 / 100
Mean lesion-area ratio in Mini-GT	0.1570
Median lesion-area ratio in Mini-GT	0.0461

3.3. Mini-GT Benchmark

The classification dataset does not contain lesion-level supervision. To quantitatively evaluate explanation quality, a separate Mini-GT benchmark of 286 images with bounding-box annotations was used. The class distribution is 86 bacterial blight, 100 brown spot, and 100 leaf smut images. The lesion-area statistics vary markedly: the first quartile occupies only 2.15% of image area, the median 4.61%, and the upper quartile 17.05%. Figure 2 visualizes these characteristics. The lesion-area distribution in Figure 2 is especially important for interpreting later results. A model can be class-correct while still producing a heatmap that is only loosely aligned with the small symptomatic region. Therefore, weak localization metrics should not be expected to match the extremely high classification accuracy observed on the main test set.

3.4. External Cross-Source Dataset

An independent external dataset containing 120 rice-leaf images was used for cross-source evaluation. The dataset comprised 40 images each of bacterial leaf blight, brown spot, and leaf smut. Before evaluation, all external images were compared with the complete 4,684-image development dataset using exact SHA-256 matching and the same perceptual-hash procedure used for leakage control. No exact or perceptual-hash overlap was detected. The external dataset was used exclusively for post-development evaluation and did not contribute to model training, checkpoint selection, or hyper-parameter tuning.

4. Methodology

This section presents the complete methodological framework adopted in the study. The proposed pipeline is organized as a sequence of linked stages beginning with leakage-aware disease classification, followed by Grad-CAM++-based visual explanation, and concluding with

lesion-level validation and reliability analysis. The section first introduces the overall framework and training configuration, and then formalizes the mathematical definitions, localization procedure, evaluation metrics, and algorithmic workflow used to produce the final results.

4.1. Overall Framework

The complete framework consists of three consecutive stages: classification, explainability, and validation with reliability analysis. Figure 3 summarizes the workflow. The central methodological distinction of the framework is that classification and explanation are not treated as interchangeable evidence. Instead, a backbone is first optimized for disease recognition, after which its explanations are independently validated using lesion level annotations and reliability metrics. The framework does not introduce a new classification objective or explanation algorithm. Instead, established mathematical components are retained to provide a controlled basis for evaluating the reliability of post-hoc explanations. The methodological focus is therefore placed on the evaluation protocol and on the relationship among classification performance, spatial localization fidelity, calibration, robustness, explanation-method variability, and cross-source generalization.

4.2. Classification Backbones and Training Configuration

Three ImageNet-initialized residual networks were compared: ResNet34, ResNet50, and ResNet152. The final classification layer of each network was adapted for three-way prediction. All models were trained under the same optimization settings, summarized in Table 4. Holding the optimization protocol constant is essential for the validity of the backbone comparison. Under these settings, observed differences can be interpreted as architectural differences rather than artifacts of unequal tuning.

Table 4. Shared training configuration for all evaluated backbones

Parameter	Value
Input size	224×224
Batch size	32
Optimizer	AdamW
Learning rate	3×10^{-4}
Weight decay	1×10^{-4}
Epochs	10
Random seed	42
Augmentation	Horizontal flip, random rotation, mild color jitter
Checkpoint selection	Highest validation macro-F1

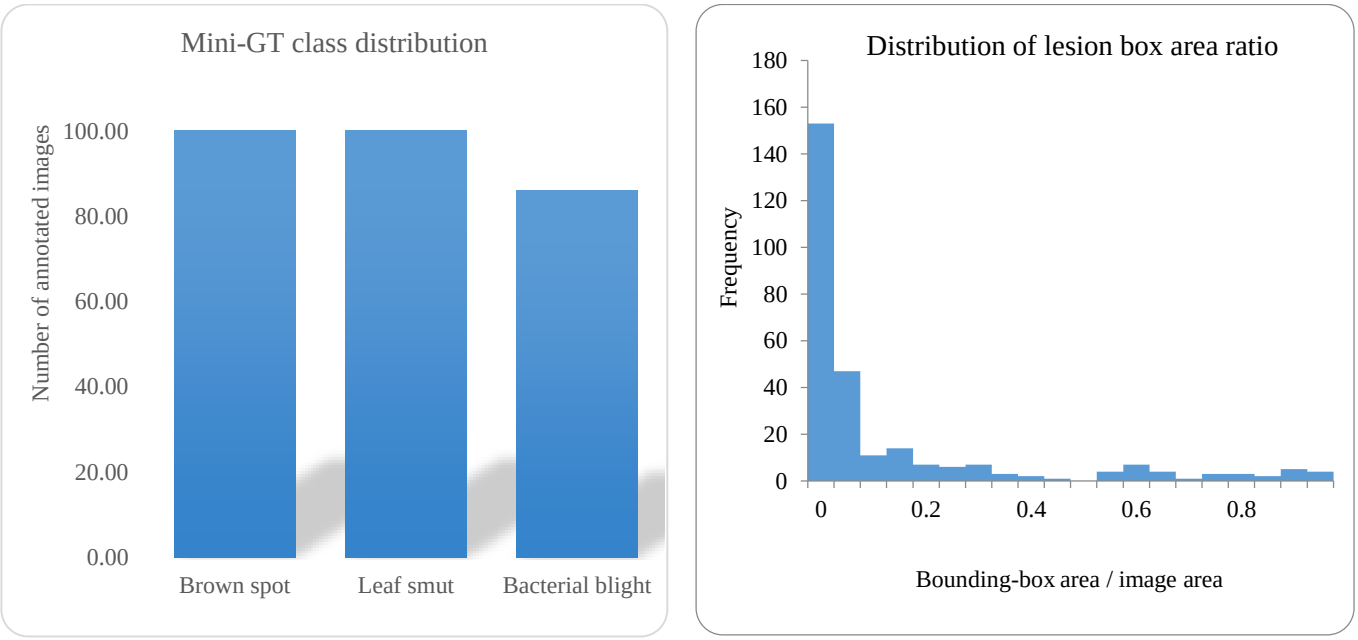


Fig. 2 Mini-GT benchmark characteristics. The class distribution is shown together with lesion-area statistics derived from the bounding-box annotations.

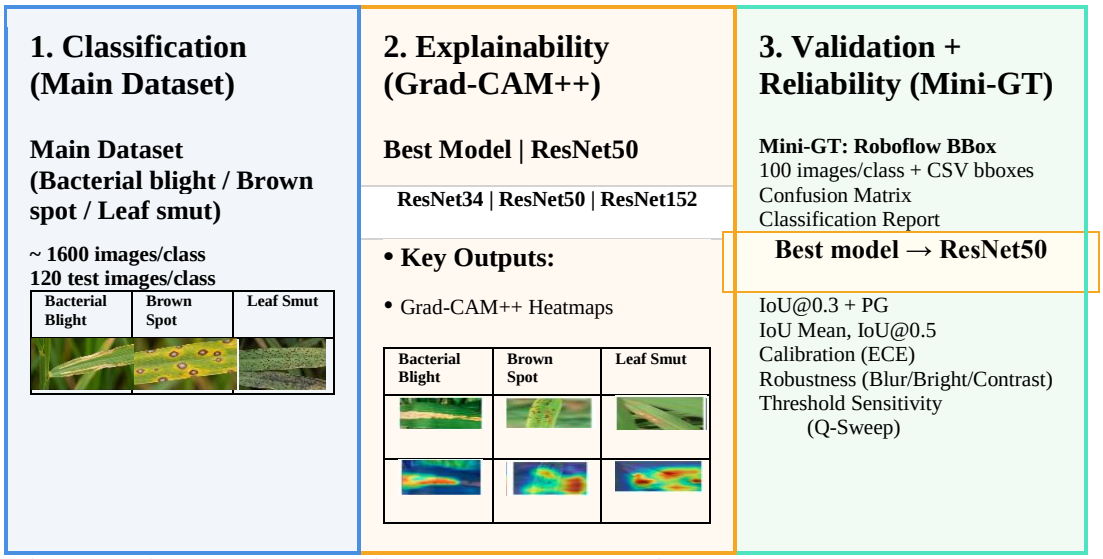


Fig. 3 Proposed framework of the study. The pipeline first trains and compares classification backbones on the main dataset, then generates Grad-CAM++ heatmaps for the best model, and finally validates explanation quality on the Mini-GT benchmark through overlap, calibration, robustness, and threshold-sensitivity analysis.

4.3. Mathematical Formulation of Classification

Let the training set be denoted by $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ where x_i is a rice leaf image and $y_i \in \{1,2,3\}$ corresponds to bacterial blight, brown spot, or leaf smut. For a network producing logits $z_i \in \mathrm{R}^3$, the class probabilities are computed through the softmax function.

The following standard classification formulation is included to define the experimental setting and notation; no new loss function or probabilistic formulation is proposed.

$$\hat{p}_{ic} = \frac{\exp(z_{ic})}{\sum_{k=1}^{3} \exp(z_{ik})}, \quad c \in \{1,2,3\} \tag{1}$$

Training minimizes the multiclass cross-entropy loss:

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{3} 1(y_i = c) \log \hat{p}_{ic} \tag{2}$$

The predicted class is $\hat{y}_i = \arg max_c \, \hat{p}_{ic}$, and test accuracy is computed as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} 1\,(\hat{y}_i = y_i) \tag{3}$$

Macro-F1 is obtained by averaging the F1-score across the three disease categories.

4.4. Grad-CAM++ Explainability

For the selected backbone, Grad-CAM++ was used to produce class-specific heatmaps. Let $A^k \in \mathbb{R}^{u \times v}$ denote the k-th activation map in the target convolutional layer, and let Y^c denote the score for class c. The Grad-CAM++ explanation map is written as:

$$L^c_{\mathrm{GC++}} = \mathrm{ReLU}\left(\sum_k w_k^c A^k\right) \tag{4}$$

where w_k^c are class-specific importance weights derived from higher-order gradients of Y^c with respect to the activation maps [5]. The resulting heatmap is upsampled to image resolution and normalized to obtain $H \in [0,1]^{W \times H}$.

In practice, activation-based explanation quality depends not only on the underlying saliency formulation but also on how the resulting map is normalized and interpreted for downstream localization analysis [22, 23].

Grad-CAM++ is used as an established post-hoc explanation method rather than as a trainable localization objective. Lesion-level annotations are not incorporated into classifier optimization.

This separation enables the subsequent Mini-GT evaluation to examine whether a classifier trained exclusively for disease recognition produces spatial evidence that independently corresponds to annotated lesion regions.

4.5. Heatmap-to-Box Conversion

To evaluate weak localization quantitatively, the normalized heatmap is transformed into a binary lesion proposal through quantile thresholding:

$$M_q(u,v) = 1\left(H(u,v) \geq Q_q(H)\right) \tag{5}$$

where $Q_q(H)$ is the q-th quantile of the heatmap values. The smallest axis-aligned bounding box enclosing the active region in M_q is then taken as the predicted lesion box B_{pred}.

This step is important because it converts a continuous explanation map into an object that can be quantitatively

compared with the ground-truth lesion box from Mini-GT. The main operating point in the study is $q = 0.8$, while the ablation study sweeps $q \in \{0.6,0.7,0.8,0.9\}$.

4.6. Localization and Reliability Metrics

If B_{gt} denotes the annotated lesion box, localization quality is measured using intersection-over-union:

$$\text{IoU}\left(B_{\text{pred}}, B_{\text{gt}}\right) = \frac{|B_{\text{pred}} \cap B_{\text{gt}}|}{|B_{\text{pred}} \cup B_{\text{gt}}|} \tag{6}$$

Localization performance is reported in terms of mean IoU, median IoU, IoU@0.3, and IoU@0.5. Additionally, the pointing game metric evaluates whether the single most active heatmap location falls inside the annotated lesion region:

$$\begin{aligned} \text{PG} &= 1((u^*, v^*) \in B_{\text{gt}}), \\ (u^*, v^*) &= \arg\max_{u,v} H(u, v). \end{aligned} \tag{7}$$

The pointing game is widely used as a localization-oriented explanation criterion because it evaluates whether the peak explanatory evidence falls inside the target region rather than requiring dense region overlap [24].

To assess whether classification confidence corresponds to localization correctness [8, 25], localization expected calibration error is computed at an IoU threshold τ:

$$\text{Loc-ECE}_\tau = \sum_{b=1}^{B} \frac{|S_b|}{N} |\text{acc}_\tau(S_b) - \text{conf}(S_b)| \tag{8}$$

where S_b is the set of samples in confidence bin b, conf(S_b) is the average confidence in that bin, and

$$\text{acc}_\tau(S_b) = \frac{1}{|S_b|} \sum_{i \in S_b} 1\,(\text{IoU}_i \geq \tau) \tag{9}$$

For perturbation analysis, heatmap stability is measured as the overlap between predicted boxes for the clean and corrupted versions of the same image:

$$\text{Stability}^{(c)} = \text{IoU}\left(B_{\text{pred}}^{(0)}, B_{\text{pred}}^{(c)}\right) \tag{10}$$

Algorithm 1: Reliability-aware rice leaf disease explainability pipeline

Require: Main image dataset D, Mini-GT annotations M, candidate backbones {ResNet34, ResNet50, ResNet152}

Ensure: Best classifier, validated Grad-CAM++ explanations, reliability metrics

1. Load all images and labels from the three disease classes
2. Compute perceptual hashes and assign group identifiers
3. Create train/validation/test splits using stratified group-preserving partitioning
4. for *each candidate backbone* do
   i) Train under the shared configuration in Table 4
   ii) Select the checkpoint with the best validation macro-F1
   iii) Evaluate the model on the held-out test split
5. Following the predefined evaluation, identify the backbone with the highest observed held-out performance for subsequent explanation analysis.
6. Generate Grad-CAM++ heatmaps for Mini-GT images.
7. Convert each heatmap into a bounding box using quantile thresholding
8. Compute IoU, IoU@0.3, IoU@0.5, and pointing game accuracy
9. Compute localization ECE at IoU thresholds of 0.3 and 0.5
10. Evaluate robustness under blur, brightness, and contrast perturbation
11. Perform threshold-sensitivity analysis for $q \in \{0.6,0.7,0.8,0.9\}$
12. Compare Grad-CAM, Grad-CAM++, Score-CAM, LayerCAM, HiResCAM, and EigenCAM under the common Mini-GT evaluation protocol.
13. Quantify uncertainty in the primary localization metrics using stratified bootstrap resampling.
14. Perform paired statistical comparisons of classification backbones and XAI methods.
15. Evaluate the selected classifier on the independent non-overlapping external dataset without retraining or fine-tuning.
16. Summarize predictive, localization, calibration, robustness, and cross-source reliability findings.

4.7. Statistical Analysis

Statistical analyses were performed using paired observations because the compared models and explanation methods were evaluated on common samples. Accuracy uncertainty for the classification backbones was quantified using 95% Wilson confidence intervals. Differences in classification correctness across ResNet34, ResNet50, and ResNet152 were assessed using Cochran's Q test, followed by two-sided exact McNemar tests with Holm correction for pairwise comparisons. Uncertainty in the Mini-GT localization metrics was estimated using 10,000 stratified bootstrap resamples while preserving the original class distribution. Differences in per-image IoU across the six explanation methods were evaluated using the Friedman test, followed by pairwise Wilcoxon signed-rank tests with Holm correction. Statistical significance was assessed at α=0.05. The experimental pipeline is formalized in Algorithm 1, which details the proposed reliability-aware approach to rice leaf disease diagnosis.

5. Experiments

This section describes the experimental protocol used to evaluate classification performance and explanation reliability. In addition to the shared implementation settings, the evaluation considers localization quality, calibration, corruption robustness, heatmap-threshold sensitivity, comparative XAI performance, statistical uncertainty, and cross-source generalization. The Mini-GT benchmark is used specifically for quantitative weak-localization assessment.

5.1. Implementation Details

All experiments followed the common evaluation protocol described in Section 4. The classification stage evaluated ResNet34, ResNet50, and ResNet152 under identical training settings. ResNet50, which achieved the highest observed held-out classification performance, was subsequently used for the primary explanation analysis. Reliability assessment included localization calibration, corruption robustness, threshold sensitivity, comparative XAI evaluation, localization uncertainty analysis, and independent cross-source validation.

5.2. Training Dynamics

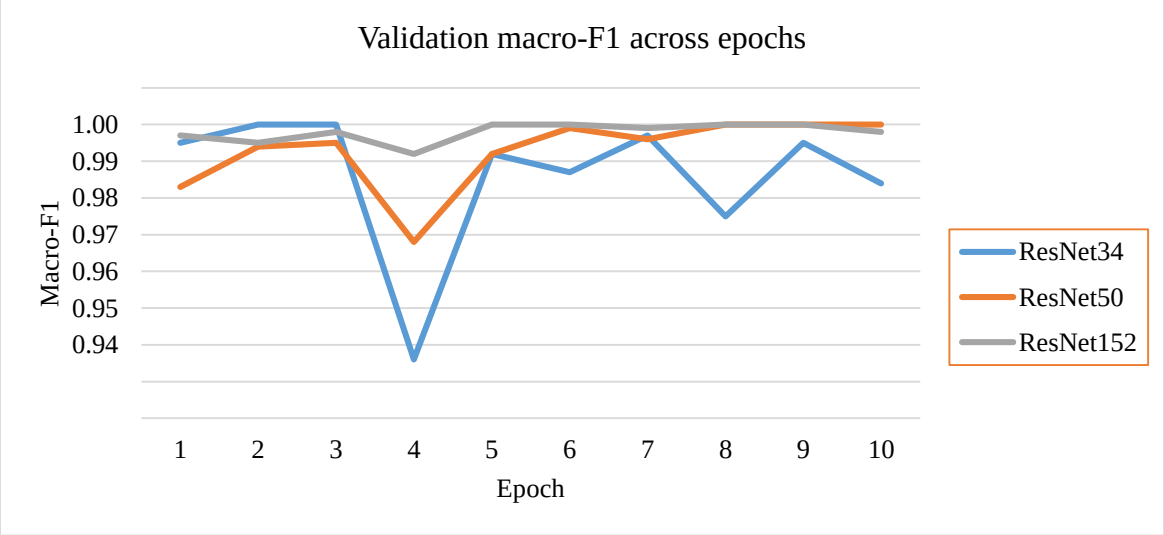
Figure 4 presents the validation trajectories of the three candidate backbones. Although all models converged rapidly, ResNet50 combined stable validation behavior with the highest observed held-out classification performance among the three evaluated backbones and was therefore used for the subsequent explanation analysis.

6. Results and Discussion

The following results evaluate the proposed framework from complementary perspectives, including predictive performance, explanation fidelity, calibration, robustness, threshold sensitivity, and error behavior. This progression is intentional, as it allows predictive accuracy and explanation reliability to be examined as related but distinct properties of the proposed framework.

6.1. Backbone Comparison and Backbone-Depth Ablation

The first ablation study concerns backbone choice. Table 5 and Figure 5 compare the three residual architectures. All three backbones performed strongly, which suggests that the dataset is highly learnable under the leakage-aware split. However, the relationship between network depth and performance was not monotonic. ResNet50 was observed to be a better model compared to the smaller ResNet34 the deeper ResNet152, which means that extra depth did not provide better generalization to this dataset. This trend is effective in that it demonstrates that it did not achieve any stronger performance by simply choosing the deepest model; ResNet50 seems to be a more appropriate compromise between representation and the visual complexity of the task.



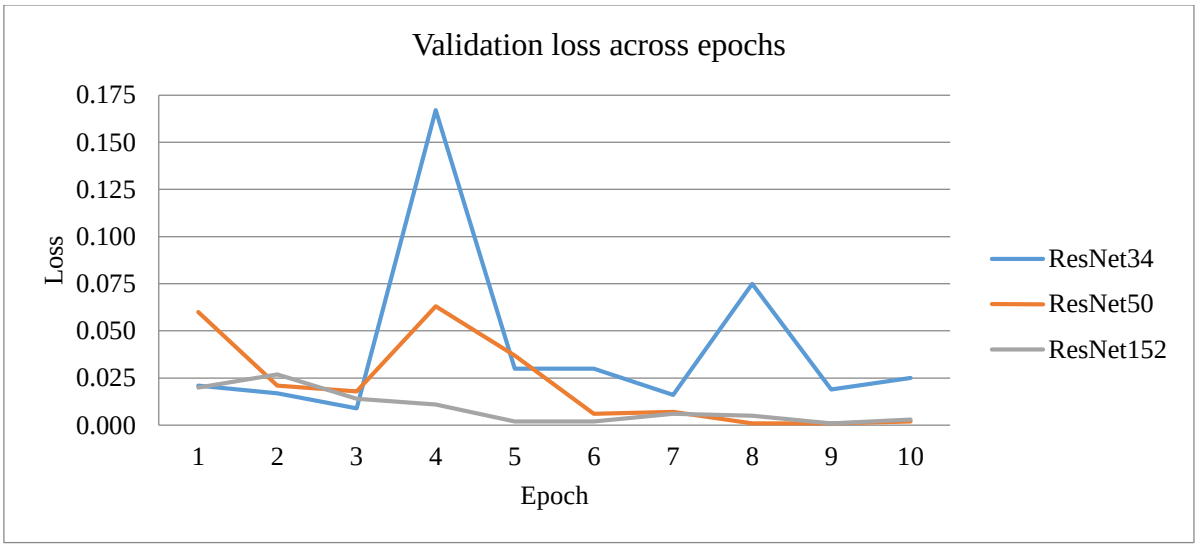


Fig. 4 Validation macro-F1 and validation loss across epochs for ResNet34, ResNet50, and ResNet152. ResNet50 shows the most favorable balance of convergence speed and stable validation performance

Table 5. Held-out test performance of the evaluated backbones

Model	Test Accuracy	Test Macro-F1
ResNet34	0.986999	0.987072
ResNet50	0.996750	0.996705
ResNet152	0.993499	0.993410

6.1.1. Statistical Comparison of Classification Backbones

To determine whether the observed performance differences among the three residual backbones were supported by paired statistical evidence, additional significance analysis was conducted on the 923 held-out test images, as shown in Table 6. Since each image was evaluated by all three classifiers and prediction correctness was represented as a paired binary outcome, Cochran's Q test was used as an omnibus test across ResNet34, ResNet50, and ResNet152. The analysis indicated an overall difference in classification correctness ($Q(2)=7.00$, $p=0.0302$). Pairwise differences were subsequently examined using two-sided exact McNemar tests with Holm correction for multiple comparisons. ResNet50 correctly classified 12 samples that were misclassified by ResNet34, whereas ResNet34 correctly classified three samples that were misclassified by ResNet50.

This comparison yielded an unadjusted $p=0.0352$, but the difference did not remain significant after Holm correction ($padj=0.1055$). For ResNet50 versus ResNet152, the corresponding discordant counts were 3 and 0 ($p=0.2500$, $padj=0.4758$). For ResNet152 versus ResNet34, the discordant counts were 12 and 6 ($p=0.2379$, $padj=0.4758$). These results indicate that although an overall difference was detectable across the three classifiers, the individual pairwise differences were not statistically significant after correction for multiple comparisons. ResNet50, therefore, represents the highest observed performer in the evaluated setting, with a test accuracy of 0.996750 and a macro-F1 of 0.996705, but the results do not support a claim of statistically established pairwise superiority.

6.2. Class-Wise Performance of the Best Model

Table 7 and Figure 6 provide a class-level view of the backbone results. The confusion matrices show that the held-out classification task is almost solved in terms of label prediction. For ResNet50, bacterial blight is recognized perfectly, while residual confusion is concentrated in the leaf smut versus brown spot boundary. For ResNet50, the residual classification errors were confined to leaf smut samples predicted as brown spot. The present analysis identifies this error pattern but does not experimentally determine the visual factors responsible for the confusion.

6.3. Weak Localization Quality on Mini-GT

The core result of the paper emerges when the best classifier is subjected to lesion-level validation. Table 8 summarizes the quantitative explanation quality for ResNet50 with Grad-CAM++. As reported in Table 8, the thresholded overlap rates remain limited, with IoU@0.3 equal to 0.2483 and IoU@0.5 equal to 0.1119. These values indicate a marked separation between classification correctness and lesion-level spatial faithfulness. Classification remains perfect on the annotated subset, yet the average spatial overlap between the explanation-derived box and the lesion annotation is modest. Therefore, a correct class label should not be conflated with a spatially faithful explanation. This is precisely the distinction that qualitative saliency inspection often obscures [7].

Table 6. Statistical comparison of classification performance on the held-out test set

Model/comparison	Result
ResNet34 accuracy (95% CI)	0.986999 (0.977414–0.992547)
ResNet50 accuracy (95% CI)	0.996750 (0.990488–0.998894)
ResNet152 accuracy (95% CI)	0.993499 (0.985891–0.997017)
Cochran's Q, three models	Q(2) = 7.000, p = 0.0302
ResNet50 vs ResNet34	McNemar p = 0.0352; Holm p = 0.1055
ResNet50 vs ResNet152	McNemar p = 0.2500; Holm p = 0.4758
ResNet152 vs ResNet34	McNemar p = 0.2379; Holm p = 0.4758

Table 7. Class-wise test performance of ResNet50

Class	Precision	Recall	F1-score	Support
Bacterial blight	1.0000	1.0000	1.0000	316
Brown spot	0.9901	1.0000	0.9950	301
Leaf smut	1.0000	0.9902	0.9951	306

Table 8. Localization quality of the selected model–explanation pair on the Mini-GT benchmark

Model	CAM	n	Accuracy	MacroF1	Mean IoU	Median IoU	IoU@0.3	IoU@0.5	PG accuracy
ResNet50	Grad-CAM++	286	1.0000	1.0000	0.1831	0.0855	0.2483	0.1119	0.4720

To quantify uncertainty associated with the finite Mini-GT sample, 95% confidence intervals were estimated using 10,000 stratified bootstrap resamples while preserving the original class distribution. The mean IoU of 0.1831 was associated with a 95% confidence interval of 0.1674–0.1988, whereas the median IoU interval was 0.0654–0.0991.

The corresponding intervals were 0.2168–0.2797 for IoU@0.3, 0.0804–0.1434 for IoU@0.5, and 0.4161–0.5280 for pointing-game accuracy. These intervals indicate that the conclusion of limited lesion-level spatial agreement is not dependent on a single point estimate.

6.4. Localization Calibration

The reliability analysis begins with localization calibration. Table 9 and Figure 7 summarize the results. The localization-calibration results indicate substantial disagreement between predictive confidence and spatial correctness. Loc-ECE reached 0.7517 at IoU≥0.3 and 0.8881 at IoU≥0.5, despite high classification confidence on Mini-GT. These results indicate that confidence in the predicted disease class should not be interpreted as evidence that the associated explanation is spatially aligned with the annotated lesion.

6.5. Comparative Evaluation of XAI Methods

To determine whether the observed localization behavior was specific to Grad-CAM++ or was shared by alternative explanation approaches, five additional class-activation-based methods were evaluated using the same ResNet50 classifier and the same 286-image Mini-GT benchmark. The comparison included Grad-CAM, Score-CAM, LayerCAM, HiResCAM, and EigenCAM, with Grad-CAM++ retained as the reference method. The target convolutional layer, image preprocessing, predicted-class target, heatmap normalization, and heatmap-to-box conversion procedure were held constant across methods. The main localization operating point remained q=0.80.

The results in Table 10 show modest metric-specific variation across explanation methods. Grad-CAM and HiResCAM produced the highest mean IoU (0.1846) and median IoU (0.0881), whereas EigenCAM achieved the highest IoU@0.3 rate (0.2657). Score-CAM obtained the highest pointing-game accuracy (0.5175). Grad-CAM++ produced a mean IoU of 0.1831 and a pointing-game accuracy of 0.4720, remaining close to the other evaluated methods in overlap-based localization performance. The paired per-image IoU values were further compared using the Friedman test. No statistically significant overall difference was detected among the six explanation methods ($\chi 2(5)=4.9796$, $p=0.4184$, Kendall's $W=0.00348$). Pairwise Wilcoxon signed-rank tests with Holm correction likewise identified no statistically significant difference between Grad-CAM++ and any alternative method.

These findings indicate that the weak lesion-level localization observed in the study is not attributable solely to the choice of Grad-CAM++. Instead, the evaluated methods exhibit broadly comparable spatial fidelity under the common evaluation protocol, while differing modestly in specific localization criteria.

Table 9. Localization calibration error of ResNet50 with Grad-CAM++

Model	Loc-ECE at IoU≥0.3	Loc-ECE at IoU≥0.5
ResNet50 + Grad-CAM++	0.7517	0.8881

Table 10. Comparative weak-localization and calibration performance of XAI methods on the 286-image Mini-GT benchmark

XAI method	Mean IoU	Median IoU	IoU@0.3	IoU@0.5	PG accuracy	Loc-ECE@0.3	Loc-ECE@0.5
Grad-CAM	**0.1846**	**0.0881**	0.2448	**0.1224**	0.4685	0.7552	**0.8776**
Grad-CAM++	0.1831	0.0855	0.2483	0.1119	0.4720	0.7517	0.8881
Score-CAM	0.1804	0.0808	0.2378	0.1119	**0.5175**	0.7622	0.8881
LayerCAM	0.1831	0.0846	0.2448	0.1189	0.4755	0.7552	0.8811
HiResCAM	**0.1846**	**0.0881**	0.2448	**0.1224**	0.4685	0.7552	**0.8776**
EigenCAM	0.1789	0.0771	**0.2657**	0.1014	0.4825	**0.7343**	0.8986

Note: All methods were evaluated on the same 286 Mini-GT images using the same trained ResNet50 classifier and q = 0.80 heatmap-to-box operating threshold. Lower Loc-ECE indicates better correspondence between calibrated classification confidence and localization correctness. Friedman test on paired per-image IoU: $\chi^2(5) = 4.9796$, $p = 0.4184$, Kendall's $W = 0.00348$.

6.6. External Cross-Source Validation

An external cross-source evaluation was conducted to assess whether the predictive performance observed on the held-out test set transferred to images originating from an independent dataset.

The external dataset contained 120 rice-leaf images, comprising 40 bacterial leaf blight, 40 brown spot, and 40 leaf smut samples. Prior to evaluation, the external images were compared with all 4,684 images in the development dataset using exact SHA-256 matching and the same perceptual-hash procedure employed in the leakage-control analysis. No exact or perceptual-hash overlap was detected.

The selected ResNet50 classifier was applied directly to the external images without retraining, fine-tuning, or additional model selection. The resulting accuracy was 0.5667, with a 95% Wilson confidence interval of 0.4773–0.6519. Balanced accuracy was 0.5667, while Macro-Precision, Macro-Recall, and Macro-F1 were 0.5439, 0.5667, and 0.5453, respectively. Performance varied substantially across disease categories. Brown spot achieved the highest recall at 0.8500, followed by bacterial leaf blight at 0.6250, whereas leaf smut recall decreased to 0.2250. The external cross-source performance is summarized in Table 11. The external evaluation, therefore, indicates a marked cross-source performance reduction relative to the internal held-out evaluation. The result suggests that the classifier is sensitive to differences in image acquisition, background characteristics, lesion appearance, or other dataset-specific visual factors. Consequently, the high internal classification accuracy should not be interpreted as evidence of broad generalization across independent image sources.

6.7. Reliability Analysis: Corruption Robustness

Robustness of the explanation was tested with blur, brightness and contrast perturbations. The results are shown in Table 12 and Figure 8.

Classification performance remained comparatively stable under the evaluated blur, brightness, and contrast perturbations, whereas localization quality and explanation stability showed greater variation. The largest reduction in the reported localization measures occurred under the stronger brightness perturbation. These results indicate that post-hoc spatial explanations can be more sensitive to controlled image perturbations than the corresponding class predictions.

Table 11. External cross-source performance of the selected ResNet50 classifier on the independent 120-image UCI Rice Leaf Diseases dataset

Metric	Result
Number of images	120
Correct predictions	68
Accuracy	0.5667
95% CI for accuracy	0.4773–0.6519
Balanced accuracy	0.5667
Macro-Precision	0.5439
Macro-Recall	0.5667
Macro-F1	0.5453
Mean prediction confidence	0.8271

Table 12. Robustness of ResNet50 + Grad-CAM++ under controlled corruptions

Corruption	Severity	Mean IoU	PG accuracy	Stability	Accuracy	Macro-F1
Blur	1	0.1818	0.4790	0.9396	1.0000	1.0000
Blur	2	0.1740	0.4650	0.8374	0.9895	0.9894
Brightness	1	0.1776	0.4615	0.8738	1.0000	1.0000
Brightness	2	0.1724	0.4336	0.7841	0.9965	0.9964
Contrast	1	0.1836	0.4720	0.9104	1.0000	1.0000
Contrast	2	0.1817	0.4685	0.8524	1.0000	1.0000

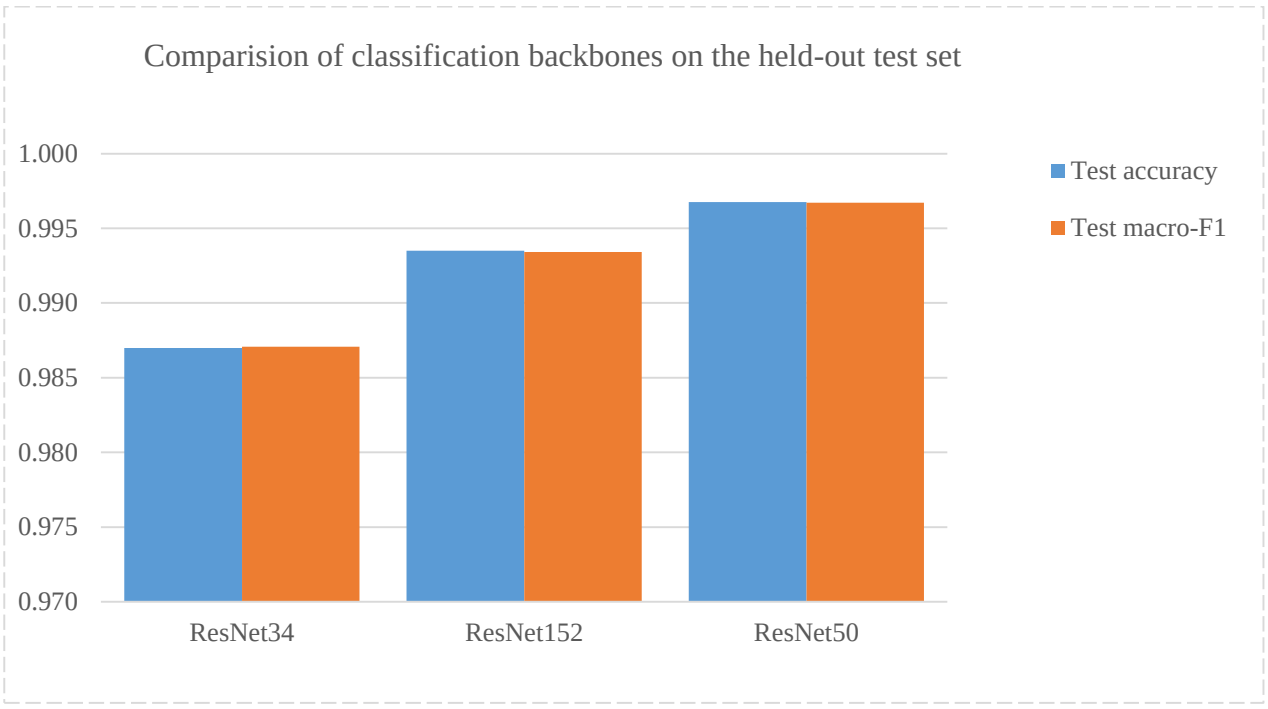


Fig. 5 Comparison of ResNet34, ResNet50, and ResNet152 on the held-out test split

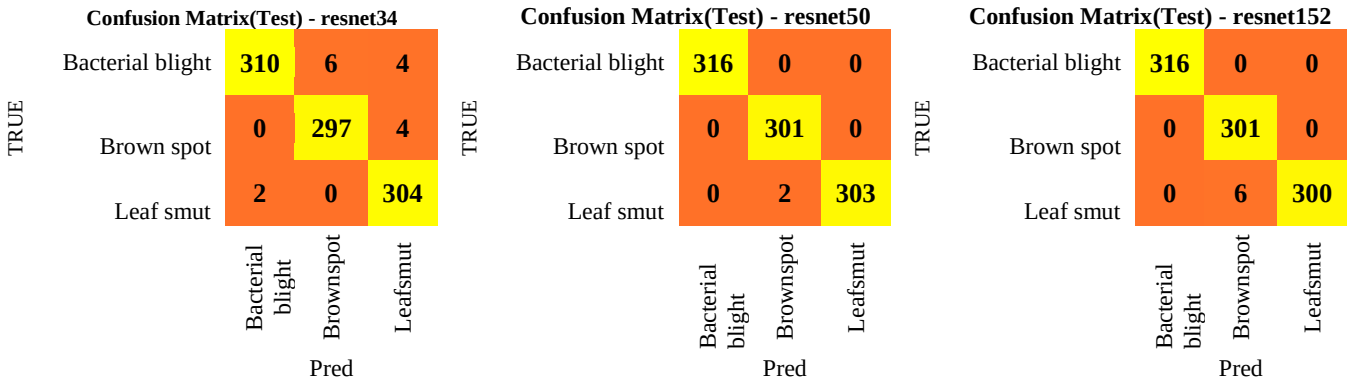


Fig. 6 Confusion matrices for ResNet34, ResNet50, and ResNet152 on the held-out test split

Table 13. Threshold-sensitivity analysis for heatmap-to-box conversion

q	Mean IoU	Median IoU	PG accuracy	IoU@0.3	IoU@0.5
0.6	0.1714	0.0594	0.4720	0.1993	0.1399
0.7	0.1788	0.0699	0.4720	0.2273	0.1364
0.8	0.1831	0.0855	0.4720	0.2483	0.1119
0.9	0.1970	0.1330	0.4720	0.2727	0.0839

6.8. Ablation Study: Heatmap Threshold Sensitivity

The second ablation study examines the effect of the quantile threshold used to convert Grad-CAM++ heatmaps into bounding boxes. Table 13 and Figure 9 summarize the results. As q increases, the heatmap becomes more concentrated, which improves the mean and median IoU. However, the stricter threshold simultaneously lowers the IoU@0.5 rate. The study, therefore, shows that heatmap evaluation is operating point dependent. Importantly, though, the central conclusion remains stable across all thresholds: localization quality remains substantially weaker than classification accuracy.

6.9. Error Analysis

Two complementary error analyses were performed: classification errors on the held-out test set and localization quality stratified by lesion size. Classification errors are summarized in Table 14 and Figure 10, whereas localization performance stratified by lesion size is reported in Table 15 and Figure 11. The test-set error pattern is narrow but informative. All three ResNet50 mistakes were leaf smut samples predicted as brown spot, and the mean confidence on these errors remained above 0.99. This suggests that the residual failure mode is not random uncertainty but a systematic visual confusion between two symptom categories.

The lesion-size analysis explains much of the gap between classification and localization. For the smallest quartile, the mean IoU falls to 0.0259 and the pointing game accuracy to 0.1806. In contrast, for the largest lesions, the mean IoU rises to 0.4953 and the pointing game accuracy to 0.9028. This indicates that weak localization is heavily constrained by lesion scale. The model can classify disease from broader contextual evidence, even when the actual symptomatic region is very small, but the explanation map becomes much more faithful when the lesion occupies a larger and more spatially distinct region.

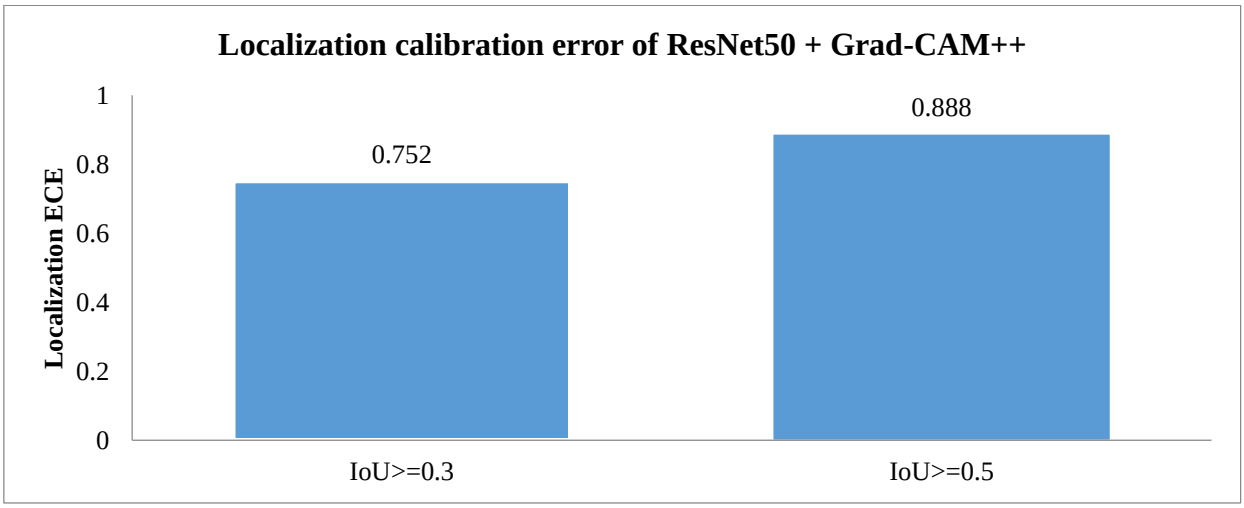
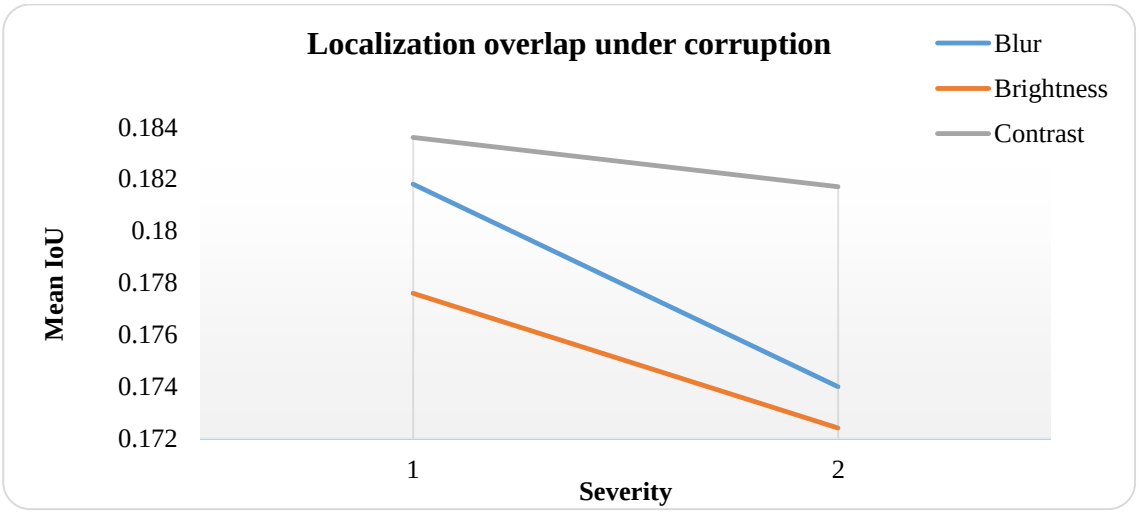


Fig. 7 Localization calibration error at two correctness thresholds. High Loc-ECE indicates that confidence is a poor proxy for spatial correctness.



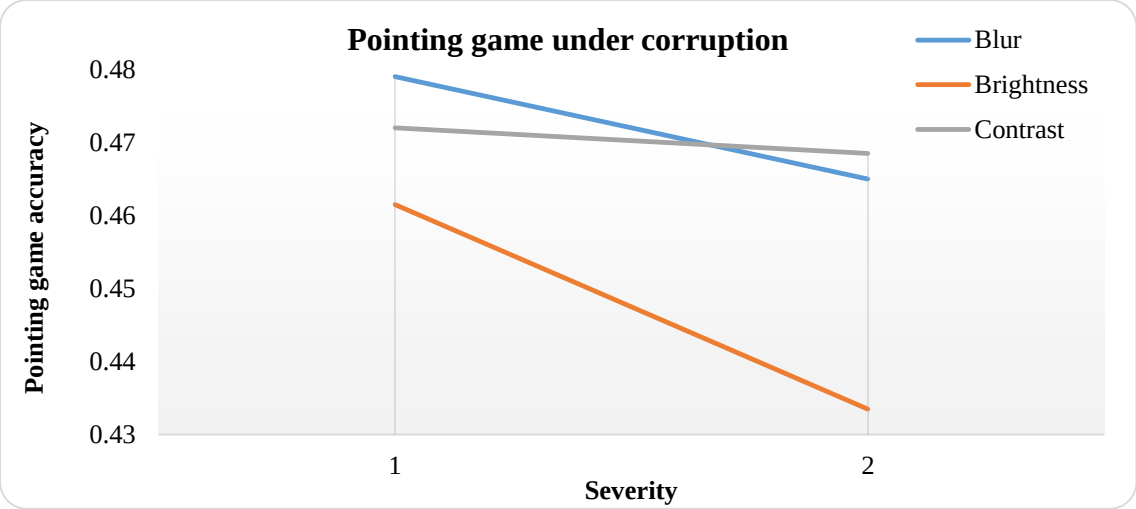
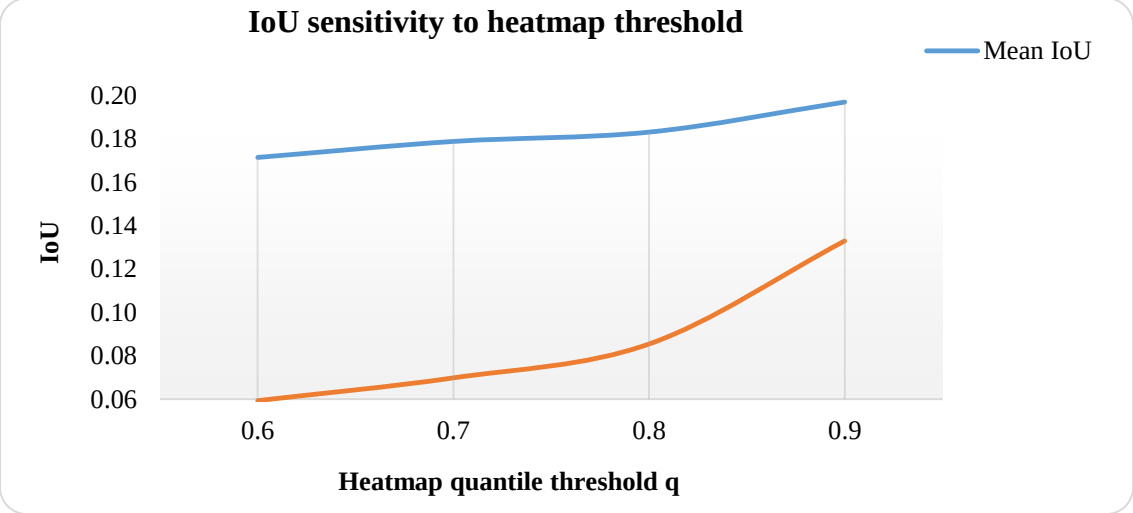


Fig. 8 Localization robustness under blur, brightness, and contrast perturbations. Classification remains highly stable, but explanation quality degrades more noticeably.



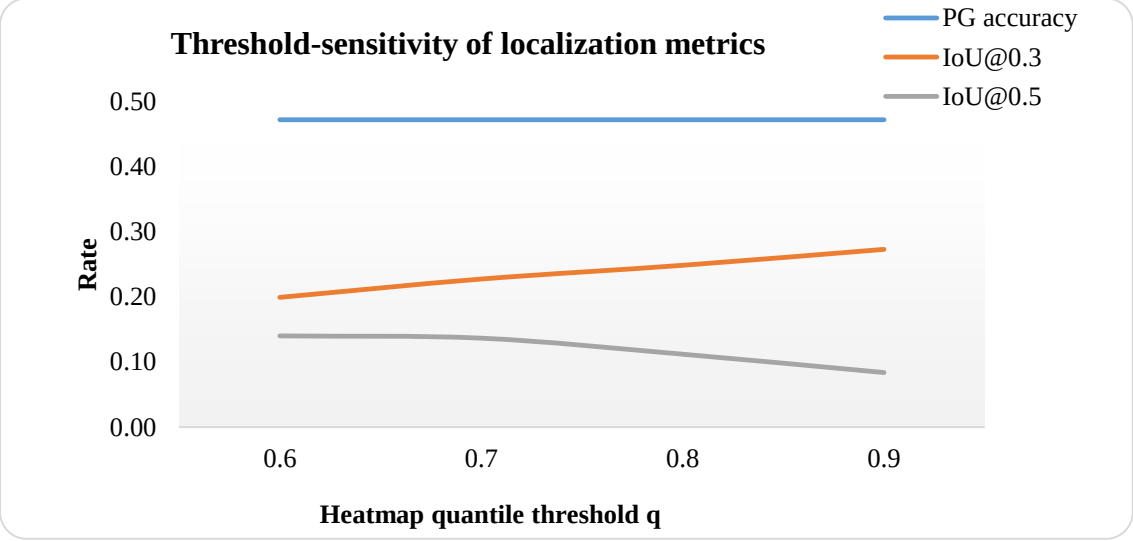


Fig. 9 Threshold-sensitivity analysis of Grad-CAM++ weak localization. Increasing the threshold concentrates the heatmap, improving coarse overlap but reducing the IoU@0.5 rate.

Table 14. Error profile of ResNet50 on the held-out test set

True class	Predicted class	Count	Mean predicted confidence
Leaf smut	Brown spot	3	0.9931

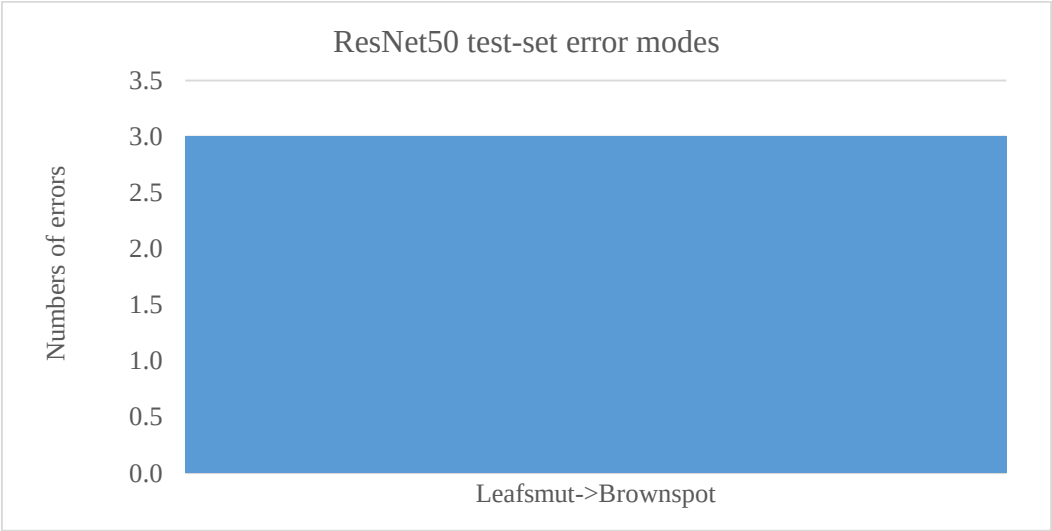


Fig. 10 Classification error modes of ResNet50, All three misclassifications correspond to leaf smut being predicted as brown spot.

Table 15. Localization quality stratified by lesion-size quartile on Mini-GT

Lesion-size quartile	*n*	Mean IoU	PG accuracy
Q1-smallest	72	0.0259	0.1806
Q2	71	0.0646	0.3662
Q3	71	0.1445	0.4366
Q4-largest	72	0.4953	0.9028

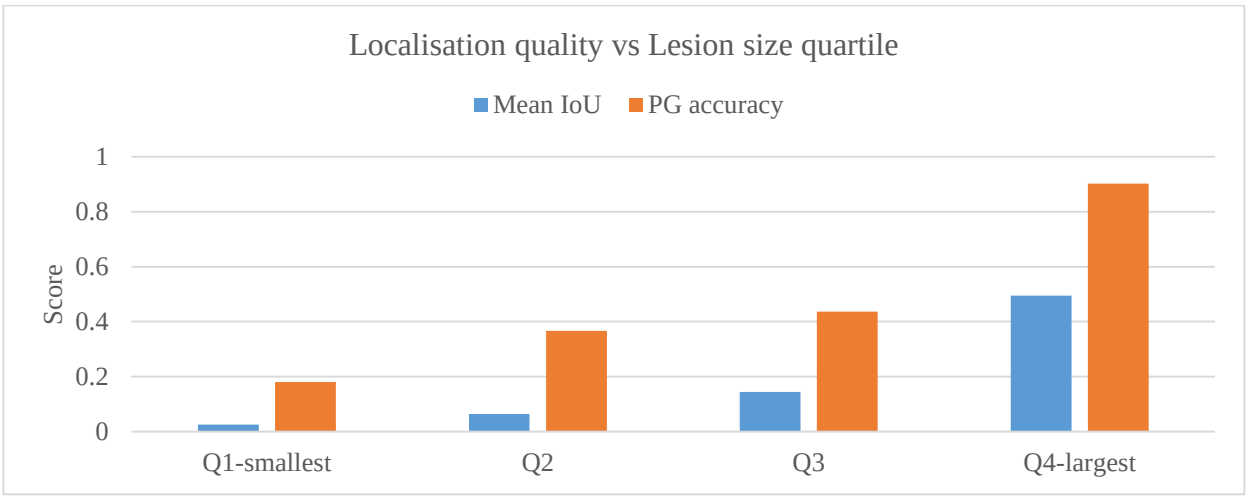


Fig. 11 Localization performance as a function of lesion-size quartile.
Grad-CAM++ localizes large lesions much more reliably than small, compact lesions.

6.10. Discussion

The results demonstrate a clear separation among internal classification performance, spatial explanation fidelity, and cross-source predictive reliability. On the leakage-aware held-out test set, all three residual networks achieved high classification performance, with ResNet50 providing the highest observed accuracy and macro-F1. However, pairwise statistical comparisons did not establish significant superiority of ResNet50 over the alternative backbones after correction for multiple comparisons.

The Mini-GT analysis revealed a substantially different pattern for explanation fidelity. Although ResNet50 classified all 286 annotated samples correctly, Grad-CAM++ achieved a mean IoU of only 0.1831 and a pointing-game accuracy of 0.4720. Bootstrap confidence intervals supported the stability of this conclusion. Localization was particularly weak for small lesions and improved markedly as lesion size increased. High Loc-ECE values further showed that classification confidence was poorly aligned with localization correctness.

The comparative XAI analysis indicated that this limitation was not specific to Grad-CAM++. Grad-CAM, Grad-CAM++, Score-CAM, LayerCAM, HiResCAM, and EigenCAM exhibited modest metric-specific differences, but the Friedman test did not identify a significant overall difference in paired per-image IoU. The evaluated methods, therefore, provided broadly comparable spatial fidelity under the common protocol.

The independent cross-source evaluation further demonstrated that high internal accuracy did not translate into equivalent performance on an external image source. Accuracy decreased to 0.5667 and macro-F1 to 0.5453, with particularly low recall for leaf smut. These results emphasize that internal benchmark performance, confidence, and visually plausible explanations should not independently be treated as sufficient evidence of reliability. Quantitative localization assessment and external validation provide complementary information that is necessary for a more cautious interpretation of agricultural vision systems.

The models in this study were optimized for disease classification rather than lesion localization. Accordingly, the Mini-GT annotations were used for post-hoc assessment and were not incorporated into classifier training. The observed localization limitations, therefore, motivate future evaluation of lesion-supervised, attention-guided, or joint classification-localization approaches rather than demonstrating the effectiveness or ineffectiveness of such methods.

The high held-out classification performance should be interpreted in the context of the characteristics of the evaluated dataset and the controlled experimental protocol. The three residual architectures were trained using the same optimization configuration and a leakage-aware group-preserving split, which provides a consistent basis for comparison. The standardized image resolution, relatively balanced class distribution, and visually constrained disease categories may also contribute to the high internal classification performance. ResNet50 produced the highest observed accuracy and macro-F1 among the evaluated backbones; however, the corrected pairwise statistical tests did not establish statistically significant superiority over the alternative backbones. Therefore, the observed result should be interpreted as the highest-performing configuration within the evaluated experimental setting rather than as evidence of universal architectural superiority. The substantially lower performance on the independent external dataset further demonstrates that high internal benchmark performance does not necessarily translate to cross-source generalization.

7. Conclusion

This study evaluated the reliability of explainable rice leaf disease classification by separating predictive accuracy from spatial explanation fidelity and cross-source generalization. A leakage-aware evaluation of ResNet34, ResNet50, and ResNet152 on a three-class dataset of 4,684 images showed high internal predictive performance, with ResNet50 achieving the highest observed held-out accuracy of 0.9968 and macro-F1 of 0.9967. However, paired statistical comparisons did not establish significant pairwise superiority after correction for multiple comparisons.

Quantitative evaluation on the 286-image Mini-GT benchmark revealed a substantial gap between disease recognition and lesion localization. Grad-CAM++ achieved a mean IoU of 0.1831, IoU@0.5 of 0.1119, and pointing-game accuracy of 0.4720 despite perfect classification on the annotated subset. Comparative evaluation of six XAI methods showed no statistically significant overall difference in paired per-image IoU, indicating that the limited localization fidelity was not specific to Grad-CAM++.

External cross-source evaluation further showed that the selected classifier achieved an accuracy of 0.5667 and macro-F1 of 0.5453 on the independent 120-image dataset. The combined findings indicate that high internal classification accuracy does not necessarily imply spatially faithful explanations or robust cross-source performance. Reliability claims for agricultural XAI systems should therefore be supported by quantitative localization analysis, uncertainty assessment, robustness evaluation, and independent external validation rather than qualitative saliency inspection alone.

Limitations and Future Work

The Mini-GT benchmark uses lesion-level bounding boxes rather than pixel-level segmentation masks. Bounding boxes provide a practical basis for weak-localization assessment, but cannot characterize fine lesion boundaries or pixel-level explanation fidelity. The present findings should therefore be interpreted as weak-localization evidence rather than segmentation accuracy. Future evaluation using independently annotated pixel-level lesion masks would permit a more spatially precise assessment of explanation fidelity.

The current study provides a solid, reliability-conscious foundation of explainable rice leaf disease analysis, although there are still multiple avenues to explore. These guidelines are based on the existing constraints of weak localization, scale of data sets, granularity of annotation, and realism of deployment. The discussion below describes the most relevant

extensions that can enhance lesion-faithfulness analysis and strengthen the feasibility of future agricultural vision systems. The current findings have a number of directions. First, bounding is not preferred; instead, lesion masks are used. Boxes would facilitate a more intensive evaluation of explanations and less ambiguity of weak localization. Second, it would be possible with several human annotators to estimate the agreement of the annotations and to decide whether there are any apparently low IoU. Cases are indicative of model failure or lesion-boundary ambiguity. Third, work in the future should test the reliability of out-of-distribution and cross-source generalization in more detail, especially in field imagery that is taken in uncontrolled light and background conditions. Future work should examine whether explicit lesion supervision can reduce the observed gap between predictive performance and explanation fidelity. Attention-supervised learning, segmentation-guided objectives, and joint classification-localization architectures provide relevant directions for determining whether spatial faithfulness can be improved when lesion information directly contributes to model optimization. Lastly, integrating localization in the training of architectures rather than post-hoc faithfulness alone may be more faithful to lesion than standard classification backbones.

References

[1] David. P. Hughes, and Marcel Salathe, "An Open Access Repository of Images on Plant Health to Enable the Development of Mobile Disease Diagnostics through Machine Learning and Crowdsourcing," *arXiv preprint*, pp. 1-13, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[2] Sharada P. Mohanty, David P. Hughes, and Marcel Salathé, "Using Deep Learning for Image-Based Plant Disease Detection," *Frontiers in Plant Science*, vol. 7, pp. 1-10, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[3] Kaiming He et al., "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ramprasaath R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 618-626, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[5] Aditya Chattopadhay et al., "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, USA, pp. 839-847, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[6] Chih-Kuan Yeh et al., "On the (in) Fidelity and Sensitivity of Explanations," *Advances in Neural Information Processing Systems*, vol. 32, 2019. [Google Scholar] [Publisher Link]

[7] Ashkan Khakzar et al., "Do Explanations Explain? Model Knows Best," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 10244-10253, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Chuan Guo et al., "On Calibration of Modern Neural Networks," *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321-1330, 2017. [Google Scholar] [Publisher Link]

[9] Amit Dhurandhar et al., "Locally Invariant Explanations: Towards Stable and Unidirectional Explanations Through Local Invariant Learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 19410-19445, 2023. [Google Scholar] [Publisher Link]

[10] Chinna Gopi Simhadri et al., "Deep Learning for Rice Leaf Disease Detection: A Systematic Literature Review on Emerging Trends, Methodologies and Techniques," *Information Processing in Agriculture*, vol. 12, no. 2, pp. 151-168, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[11] Shakti Kinger, and Vrushali Kulkarni, "Explainable AI for Deep Learning Based Disease Detection," *Proceedings of the 2021 Thirteenth International Conference on Contemporary Computing*, pp. 209-216, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Kaihua Wei et al., "Explainable Deep Learning Study for Leaf Disease Classification," *Agronomy*, vol. 12, no. 5, pp. 1-15, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Harshana Habaragamuwa, Yu Oishi, and Kenichi Tanaka, "Achieving Explainability for Plant Disease Classification with Disentangled Variational Autoencoders," *Engineering Applications of Artificial Intelligence*, vol. 133, pp. 1-24, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] N. Krishnamoorthy et al., "Rice Leaf Diseases Prediction using Deep Neural Networks with Transfer Learning," *Environmental Research*, vol. 198, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] P. Isaac Ritharson et al., "DeepRice: A Deep Learning and Deep Feature Based Classification of Rice Leaf Disease Subtypes," *Artificial Intelligence in Agriculture*, vol. 11, pp. 34-49, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Prameetha Pai et al., "A Twin CNN-Based Framework for Optimized Rice Leaf Disease Classification with Feature Fusion," *Journal of Big Data*, vol. 12, no. 1, pp. 1-27, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[17] Junsuk Choe et al., "Evaluating Weakly Supervised Object Localization Methods Right," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 3133-3142, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Jeesoo Kim et al., "Normalization Matters in Weakly Supervised Object Localization," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 3407-3416, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Shakeeb Murtaza et al., "A Realistic Protocol for Evaluation of Weakly Supervised Object Localization," *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Tucson, AZ, USA, pp. 5367-5376, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[20] Wasswa Shafik et al., "Deep Learning Technique for Plant Disease Classification and Pest Detection and Model Explainability Elevating Agricultural Sustainability," *BMC Plant Biology*, vol. 25, no. 1, pp. 1-31, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[21] Ruaa A. Al-Falluji, and Marwan Ali Albahar, "Enhancing Green AI Through Explainable Deep Learning-Based Multi-Model for Automated Rice Leaf Disease Classification," *Discover Computing*, vol. 28, no. 1, pp. 1-26, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[22] Hyungsik Jung, and Youngrock Oh, "Towards Better Explanations of Class Activation Mapping," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 1336-1344, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[23] Sylvestre-Alvise Rebuffi et al., "There and Back Again: Revisiting Backpropagation Saliency Methods," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 8839-8848, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Jianming Zhang et al., "Top-Down Neural Attention by Excitation Backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084-1102, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[25] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain, "Trainable Calibration Measures for Neural Networks from Kernel Mean Embeddings," *Proceedings of the 35th International Conference on Machine Learning*, pp. 2805-2814, 2018. [Google Scholar] [Publisher Link]