International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I8P113

Volume 74 Issue 8, 197-209, August 2026

Original Article

Leakage-Proof Reliability-Aware Deep Learning Framework for Rice Leaf Disease Diagnosis

Sujeet Kumar[1*], Prajeet Sharma[2]

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

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

Received: 16 March 2026 Revised: 06 July 2026 Accepted: 13 July 2026 Published: 29 August 2026

Abstract - *Timely and accurate diagnosis of rice leaf diseases is significant in order to achieve reduced loss and precision agriculture management. In this paper, a leakage-aware and reliability-oriented deep learning framework is proposed to classify rice leaf diseases based on the Mendeley rice leaf disease dataset. In order to avoid the problem of similar and duplicate images in the training and validation set, the framework incorporates perceptual hashing and stratified group K-fold cross-validation. The evaluation measures the accuracy of the in-domain classification, robustness to controlled image corruption, evaluation of the Expected Calibration Error and temperature scaling methods, selective prediction using risk-coverage analysis, external-dataset validation, and computational efficiency profiling. The in-domain evaluation resulted in an almost perfect score. The accuracy achieved by the cross-validation method applied to the folds evaluated showed high values, while the external UCI rice leaf dataset presented significantly lower values, with an accuracy of 54.17% and macro-F1 of 48.98%. This contrast suggests that high in-domain accuracy should be considered with caution if there is a domain shift. In addition, efficiency and reliability analysis indicated that mid-scale convolutional models, especially ResNet-based models, achieved a better trade-off between predictive performance and computational cost compared to larger high-capacity models. The suggested assessment scheme in the proposed framework then focuses not only on accuracy but also on deployment-oriented evaluation in the context of rice leaf disease diagnosis with the help of AI.*

Keywords - *Rice leaf disease detection, Deep learning, Reliability-first evaluation, Field-Readiness Score (FRS), Model calibration, Precision agriculture.*

1. Introduction

Rice (Oryza sativa) is one of the most widely cultivated cereal crops and remains central to food security in many rice-producing regions. Its productivity, however, is frequently affected by foliar diseases such as bacterial leaf blight, brown spot, and leaf smut. These diseases alter the visual structure of leaves, reduce photosynthetic capacity, and may cause considerable yield and quality losses when diagnosis is delayed. Conventional field diagnosis usually depends on manual inspection by trained personnel. Although expert assessment is valuable, it is time-consuming, subjective, and difficult to scale across large farming areas. Image-based automated diagnosis has therefore become an important direction for supporting early disease recognition and timely crop-management decisions.

Deep learning has made substantial progress in plant disease recognition because it can learn discriminative visual patterns directly from leaf images. Convolutional neural networks, including residual, dense, lightweight, and compound-scaled architectures, have been widely used for image-based crop disease classification. More recently, vision transformer models have also been explored because of their ability to model long-range spatial relationships. These advances have led to high accuracy in the classification of curated datasets and have reinforced the potential of AI in precision agriculture. But accuracy alone is not enough to determine if a model is appropriate for deployment in the field. In many plant disease datasets, images can look similar because they are acquired within the same plant or under the same growing conditions, from the same sample population, or through subtle variations in viewpoint, lighting, or background. When such visual related images are split across training and test partitions, the reported performance could be overly optimistic.

Several previous studies have used the standard approach of random data splitting to assess models and focus on accuracy. This protocol may not be sufficient to provide effective control of duplicate and near-duplicate samples, and may not provide the necessary generalization under realistic acquisition variability. Furthermore, many studies do not address the question of whether the confidence interval predicted matches the actual correct answer, if the model is stable to typical image corruptions, or if the lack of confidence can be detected before deployment. Although model size,

inference latency and floating-point operations are of great importance, particularly in mobile and edge-based agriculture, they are often poorly reported. Such limitations lead to a lack of confidence when applying in real field conditions, compared to high lab performance.

In order to solve this problem, this paper presents a leakage-aware and reliability-oriented deep learning diagnosis framework for rice leaf disease. It has been constructed by using a perceptual-hash-based grouping method and stratified group K-fold validation to ensure that samples in the same group have very high similarity to each other and are not placed in different folds in the training and testing sets. This design minimizes the possibility of bias in the evaluation due to visually correlated images. The evaluation protocol also goes beyond clean-set accuracy, and includes robustness testing under controlled corruptions, probability calibration via expected calibration error and temperature scaling, selective prediction via risk-coverage analysis, external dataset validation, and computational profiling. The framework also assesses the ability of a model to accurately classify the types of rice leaf diseases when trained with curated images, as well as the stability of the model output, its interpretability in terms of its level of confidence, and its usability for deployment with limited resources.

The novel aspect of the study is the combination of the leakage-controlled validation, reliability testing, external-domain assessment, and efficiency-aware comparison in a single diagnostic framework. The study does not only take the high in-domain accuracy as the only measure of model quality, but also considers other factors related to deployment. The inclusion of an external rice leaf dataset provides insight into the extent to which models trained on the Mendeley Rice Leaf Disease Dataset generalize to an independent data source. This external evaluation gives a more conservative and realistic account of generalization than does in-domain testing. The proposed Field-Readiness Score also provides an aggregate measure of the predictive performance, reliability, and computational efficiency to aid in deploying candidate models in the field.

The key findings of this study summarize as follows: First, to mitigate the impact of duplicated and visually similar samples within folds, a leakage-aware partitioning strategy is proposed, which is based on a perceptual hashing technique and stratified group K-fold validation. Second, a reliability-oriented evaluation protocol is proposed, which is achieved by integrating robustness analysis, calibration assessment, selective prediction and external-domain validation. Third, efficiency profiling is included as the number of parameters, FLOPs, model size, and inference latency to explore field applicability. Fourth, a common evaluation setting is provided for comparing convolutional and transformer-based model families. Fifth, results reveal a high level of near-perfect in domain accuracy, which should be viewed with skepticism, as there is a high domain gap based on external validation. The result confirms the necessity of reliability testing and evaluation in AI solutions for agriculture, specifically for deployment purposes.

The rest of the paper is organized as follows. Section 2 summarizes the previous work on plant disease detection, leakage awareness evaluation, model reliability and field-oriented deployment using deep learning technologies. Section 3 outlines the proposed methodology, which involves preparing the dataset, perceptual-hash grouping, training of the model, reliability metrics, and Field-Readiness score. The description of the experimental setup, implementation details, evaluation metrics and the controls for reproducibility are provided in Section 4. The experimental results presented and discussed in section 5 represent the in-domain performance, robustness, calibration, external validation, risk coverage behaviour and efficiency analysis. Section 6 summarizes this study and presents future directions on domain adaptation, interpretability and lightweight deployment.

2. Related Work

Classifiers based on convolutional neural networks have now evolved from simple classifiers to more reliable, evaluation-based methods. As an important baseline for data-driven plant pathology, Mohanty et al. [7] showed that deep convolutional neural networks are able to detect plant diseases from leaf images. Ferantinos [10] affirmed that CNNs have the capability of operating with diverse kinds of crops and weather conditions. The results demonstrated the potential of end-to-end representation learning over traditional feature engineering pipelines. Another subsequent research was regarding transfer learning and fine-tuning. On this discovery, they found that first training on a large-scale data like ImageNet would make the model more general when it was used in agriculture and improve the convergence rate [11]. Despite these advancements, Barbedo [12, 13] indicated two key issues with the older methods: the diversity of datasets was restricted, and it had a bias on the background. These issues may lead to overly optimistic accuracy estimates when conventional random splitting is used. Due to these findings, greater efforts are now in place towards open, diverse and selectively targeted datasets such as the PlantVillage [3], and imaging environments that are more akin to those observed in the natural environments [14].

General-purpose vision systems have significantly influenced agricultural computer vision. Gradient stability and feature propagation have been improved in more profound networks [15, 16], where residual and dense designs of links are used. Meanwhile, focused on producing light models such as MobileNetV2 [5] and EfficientNet [6], it has been possible to have accurate and computationally efficient predictions that can be applied to low-power edge device research and agriculture. Other more recent attention-based approaches, such as the Vision Transformer (ViT) [17] and the Swin

Transformer [18], have utilized self-attention mechanisms to capture global spatial links. This enables one to think visually outside local convolutional filters. Recent research in rice pathology indicates that hybrid transformer-CNN models are far more effective at detecting crop diseases, with competitive recognition rates suggesting that the model can be utilized in a controlled environment [19].

The most critical aspect of farming deep learning systems till now has been accuracy, yet a recent study demonstrates the significance of systems also being reliable and trustworthy. One of the most important levels of reliability is that of calibration quality, which demonstrates the quality of expected odds in accordance with actual outcomes. Temperature scaling was initially proposed as an effective method to perform after-calibration analysis by Guo et al. [20] and has been shown to effectively work on the modern architecture by extensive testing [21]. Hendrycks and Dietterich [22] demonstrated that deep models can be vulnerable to image corruptions that are common in the wild and suggested the need for robustness testing in the presence of controlled visual perturbations. Prediction models that are selective, like the model shown in SelectiveNet, allow the models to avoid making predictions when they are not confident. It makes automated diagnostics processes safer [23].

Another critical element of accurate agricultural AI systems is out-of-distribution detection. Other common definition Base line techniques are Maximum Softmax Probability (MSP) [24] and ODIN [25], which are user-friendly and efficient in the detection of unknown or non-farm samples. Nevertheless, through careful research, they have been shown to be important for reliable implementation of the excellent uncertainty modeling and Out-of-Distribution (OOD) identification [7, 26]. The methodologies developed in these studies are used for assessing classifiers for plant diseases, in addition to the accuracy. It's an all-in-one, safe paradigm that combines calibration, robustness, selective prediction and out-of-distribution screening into a single, safe paradigm system for use at the farming edge. More recent studies, based on datasets, also suggest that plant disease recognition models might be overoptimistic if there are not controls for how the dataset is constructed, how it is acquired, duplicate samples and cross-dataset validation performed.

3. Methodology

The methodology was designed to support reproducible and deployment-oriented rice leaf disease classification. Although previous plant disease classification studies have reported high accuracy, several evaluation protocols provide limited control over data leakage, reliability analysis, and computational efficiency. To address these limitations, the proposed framework combines leakage-aware data partitioning, deep visual backbones, reliability-oriented evaluation, and efficiency profiling within a unified experimental pipeline. The graphical workflow in Figure 1 further illustrates the sequence from dataset preparation and leakage-aware splitting to model evaluation and field-readiness analysis.

In addition to accuracy, the framework proposed incorporates a collection of reliability-motivated evaluation modules, testing model behaviour against a variety of perturbations and real-world uncertainties. These are robustness testing with synthetic corruptions, probabilistic calibration, selective prediction with risk-coverage analysis, and Out-of-Distribution (OOD) detection. Indicators of efficiency, including parameter count, FLOPS, latency, and model size, are also calculated to assess the suitability of a parameter to be deployed to the field. The Field-Readiness Score (FRS) is used as a composite measure to summarize robustness, calibration, external generalization, and efficiency in a single interpretable score. The following subsections describe the major components of the proposed evaluation pipeline.

3.1. Description and Pre-Processing of the Dataset

The experiments used the Mendeley Rice Leaf Disease dataset, which contains RGB images from three disease categories: bacterial leaf blight, brown spot, and leaf smut. After dataset scanning and label verification, 4,684 valid images were retained for experimentation. The dataset contains visible variation in leaf orientation, lesion appearance, illumination, background, and image acquisition conditions. These variations make the dataset suitable for evaluating rice leaf disease classifiers under controlled in-domain conditions, while external validation is still required to assess cross-dataset generalization.

A perceptual Hashing (pHash) method was used to avoid samples that are similar in perception when experiencing leakage caused by visually similar or similar samples in data. pHash transformed each image into a small binary signature; visually similar clusters of images were grouped by Hamming distance. Images with similar perceptual-hash signatures were assigned to the same group identifier before fold generation. These group identifiers were used in stratified group K-fold validation so that visually similar images remained within a single fold and were not distributed across training and validation partitions.

This was followed by a stringent policy of data-splitting based upon Stratified Group K-Fold Cross-Validation so that images in a pHash cluster would appear only in a single fold. This ensured that visually similar images were not distributed across training and validation partitions. This resulted in five non-overlapping and class-balanced folds that are faithful to independent train and validation folds. After fold construction, train-validation group overlap was checked for each fold, and no overlapping group was observed.

Images were then downsampled to 224x224 pixels after integrity has been verified to make them compatible with convolutional architectures. ImageNet-based, normalization-based training was introduced, and middle data augmentation (random horizontal flips, small rotations and color jittering) was taken to train this model to simulate field-level variation in illumination and orientation. The image-level metadata, perceptual-hash group identifiers, and fold indices were stored and reused across all model experiments to support reproducible comparison.

3.2. Deep Learning Architectures

The implementation of deep Convolutional Neural Networks (CNNs) became the basis of the study due to their success in retrieving hierarchically structured spatial characteristics directly out of the pixel information. Several ResNet-based architectures were evaluated to examine the effect of network depth on rice leaf disease classification.

- ResNet-34: an intermediate network that finds a tradeoff between diversity of features and their cost.
- ResNet-50: a heavier architecture with bottleneck residual blocks with high capacity of representation.
- ResNet-152: an extremely deep model that is able to capture complex lesion textures and color gradients.

The same fold partitions and optimization schedule were used for the evaluated ResNet backbones to maintain a fair architectural comparison. The last connected layer was altered to give three output logits that are equivalent to the disease categories. The AdamW optimizer (learning rate 1 x 10[?]4) was used, but the training was scheduled using cosine annealing. Minimization of cross-entropy loss was used, and optional label smoothing was used to minimize prediction overconfidence. Model training was implemented with mixed-precision support where available, and the batch size was selected according to the memory requirements of each backbone.

3.2.1. Comparative Baseline Note.

To give a reasonable performance benchmark, all the ResNets versions were trained with comparable transformer-based counterparts with the same optimization parameters. This balance meant that any variation that was seen in the generalization or calibration performance was due to the architectural influences, but not a bias in the hyperparameters. The homogeneous training standard allowed subsequent comparison between CNN-based and transformer-based models using the suggested reliability-first figures of evaluation.

3.3. Transfer Learning Strategies

In order to enhance generalization even when there is limited data, we also checked ImageNet-pretrained backbones in terms of transfer learning. There were two fine-tuning strategies that were complementary:

1. Linear probing: All the convolutional layers were frozen, and only the classifier head was trained on the rice dataset, learning dataset-specific decision boundaries at a low computation cost.
2. Full fine-tuning: Once the linear probing convergence had been reached, the entire network was unfrozen and optimized together again, but with a reduced learning rate in order to put the low-level filters in line with leaf-specific textures.

The comparative evaluation included representative convolutional and attention-based visual backbones, including ResNet, DenseNet, MobileNetV2, EfficientNetB0, ConvNeXt-Tiny, ViT-Base, and Swin-T, as summarized in the results section. These models are built upon convolutional and transformer architectures, and offer the ability to evaluate compact and high-capacity transfer learning regimes. All the models were trained in the same hyperparameter schedules and conditions so as to allow equal comparison.

3.3.1. Connection with Field-Readiness Evaluation

Such pretrained transformer or hybrid backbones were added because, in addition to enhancing accuracy, trade-offs in field readiness and computational efficiency were explored. Their analysis in the FRS framework allowed them to compare the compact convolutional networks with high-capacity transformer architectures on their deployment feasibility in the agricultural setting.

3.4. Reliability-Oriented Evaluation Modules

Raw accuracy is not a sufficient measure of the trustworthiness of agricultural AI systems, and in turn, four reliability tests were incorporated into the pipeline:

- Resistance to Corruptions: Synthetic perturbations - such as reduction of brightness, contrast change and Gaussian blur were added by incrementally increasing the severity. An increment in relative resilience to visual noise in terms of Accuracy (ΔAcc) and macro-F1 (ΔF1). Here, ΔAcc and ΔF1 quantify the difference between clean-image performance and corrupted-image performance, with values closer to zero indicating stronger robustness.
- Calibration: An Expected Calibration Error (ECE) was used to measure the alignment between predicted confidence and correctness. Reliability was maintained by the use of post-hoc temperature scaling. The reliability diagrams were used to show the connection between confidence and empirical accuracy. ECE was included because a field diagnostic model should not only predict the correct class but should also provide confidence values that are consistent with observed correctness.
- Selective Prediction: In accordance with the risk-coverage paradigm, models were studied in terms of being able to abstain on uncertain samples. The Area Under the Risk-Coverage Curve (AURC) gave a scalar value of confident decision-making. Risk-coverage

analysis was used to examine whether low-confidence predictions could be deferred, which is important for reducing unreliable automated decisions in field use.

- Out-of-Distribution (OOD) screening was considered using Maximum Softmax Probability (MSP) as a confidence-based uncertainty indicator. In the present study, external rice-leaf dataset evaluation was retained as the main cross-domain generalization analysis, while detailed non-rice OOD benchmarking is treated as a future extension. MSP was used as a simple confidence-based uncertainty score, where lower maximum softmax confidence indicates a higher possibility that an input may not belong to the training distribution.

Correlation-Reliability& Efficiency: Every dimension of reliability (robustness, calibration, and selective prediction) was correlated to efficiency measures in order to derive the composite Field-Ready Score (FRS). This correlation test enabled the selection of architectures that were the most cost-effective in terms of reliability and cost. In this study, reliability denotes stable model behaviour under image perturbations, calibrated prediction confidence, uncertainty-aware selective prediction, and performance consistency under external-domain evaluation.

3.5. Field-Readiness Score and Efficiency

In actual agricultural contexts, in particular, mobile or embedded implementation, efficiency and power limitations are vital. Therefore, trained individual models were profiled:

- Total learnable parameters (in millions),
- 224 x 224 resolution floating-point operations (FLOPs).
- Mean latency on CPU and GPU of inference,
- Model size on disk (in MB).

Efficiency profiling was done consistently across all the architectures, both convolutional networks (ResNet, DenseNet, MobileNet) and transformer-based models (EfficientNet, ConvNeXt, ViT, Swin). The metrics were all normalized so that there could be a fair cross-family comparison, using the Field-Ready Score framework.

In order to combine reliability and efficiency into one interpretative measure, the Field-Readiness Score (FRS) was specified as follows:

$$FRS = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i + \varepsilon}}$$

Here, x_i represents each normalized component score, including clean accuracy, robustness score, calibration score, efficiency score, and external accuracy where applicable; ε is a small constant used to avoid division by zero. The harmonic mean punishes uneven performance, with the result that models who are strong according to one dimension but weak in another are given lower composite scores. FRS_in was computed using in-domain reliability and efficiency components, whereas FRS_ext additionally included external-dataset accuracy to reflect cross-dataset field readiness. The increased FRS, hence, implies a balanced model that may be implemented in the field in practice. For replicability, the complete workflow of dataset preprocessing, perceptual-hash grouping, fold construction, model training, reliability assessment, efficiency profiling, and FRS computation is summarized in Algorithm 1.

Algorithm 1: Proposed Deep and Transfer Learning Pipeline for Rice Leaf Disease Classification

Data: Rice leaf dataset D with labels {BLB, BS, LS}

Result: Calibrated deep models and Field-Readiness Score (FRS)

1. Integrity and Preprocessing;
2. Import D (Mendeley);
3. Resize to 224 × 224, normalize to ImageNet statistics;
4. Apply augmentations (flip, rotation, color jitter);
5. Remove duplicates via perceptual Hashing (pHash);
6. Leakage-Proof Splits;
7. Generate group IDs from pHash clusters;
8. Apply Stratified Group K-Fold to create k independent folds;
9. for i ← 1 to k do
10. Split into training and validation sets;
11. Train Deep Models;
12. Initialize ResNet-34/50/152 with random weights;
13. Train using AdamW + CosineAnnealing; save best weights;
14. Transfer Learning;
15. Load pretrained EfficientNet/ConvNeXt/MobileNet/ViT;
16. Stage-1: Linear probe on frozen layers;
17. Stage-2: Fine-tune all layers at reduced learning rate;
18. Reliability Evaluation;
19. Test robustness under corruptions;
20. Compute ECE and apply temperature scaling;
21. Generate risk-coverage curves and AURC;
22. Evaluate OOD detection with MSP;
23. Efficiency Profiling;
24. Record parameter count, FLOPs, latency, and model size;
25. Compute composite Field-Readiness Score (FRS);
26. Aggregate fold-wise metrics and export summary tables;
27. Output: Calibrated, efficient models with FRS;

Overall, the proposed methodology integrates deep learning, transfer learning, leakage-aware partitioning, and reliability-oriented evaluation within a single framework. The pipeline systematically examines robustness, calibration,

selective prediction, and efficiency so that rice disease classifiers are evaluated beyond clean-set accuracy. Such a general plan helps the transition of controlled experimental research to practical field-implementable AI solutions to precision agriculture.

The later analyses in the Results and Discussion section confirm that the proposed methodology is reliable in identifying models that strike a balance between reliability and computing efficiency. The integration of transformer architecture is an example of how the framework can be extended to convolutional processes and attention-based processes, and can ensure the applicability of reliability-first evaluation to any type of neural architecture.

4. Experimental Setup

The proposed disease classification method in rice leaves was subjected to severe experimental testing to get an idea about its accuracy, reliability, and efficiency. The experiments were designed to assess model behaviour under leakage-aware data partitioning, controlled training-time augmentation, and consistent optimization settings. To support reproducibility and fair comparison, the same preprocessing rules, fold assignments, evaluation metrics, and reliability-assessment protocol were applied across the evaluated model families. This section describes the dataset preparation, augmentation strategy, training configuration, optimization settings, evaluation metrics, and reproducibility controls used in the experiments.

4.1. Dataset and Preprocessing

Experiments were conducted on the publicly available Mendeley Rice Leaf Disease dataset that comprises of three significant foliage diseases: Bacterial Leaf Blight (BLB), Brown Spot (BS), and Leaf Smut (LS), which have affected paddy crop at the most economical cost.

It contains approximately 4,620 color photos captured in all types of lighting, angles, and backgrounds, so it is a good real-life test. Following a fast visual inspection to filter out junk or low-quality shots, we downsized them to 224x224 pixels to fit in imageNet-like models without issues. ImageNet statistics ($\mu = [0.485,0.456,0.406], \sigma = [0.229,0.224,0.225]$) were used to normalize pixel intensities channel-wise.

In making sure that we had a strict separation across visually identical samples across a fold, we used perceptual Hashing (pHash) to produce a reductive binary signature of each image. The images whose Hamming distance with each other was less than a certain threshold were placed in the same cluster. This was followed by further Stratified Group K-Fold partitioning so that every member of a pHash cluster was in one fold. The approach ensures that each partition has no leakage and that there is maintenance of classes balance within the partitions.

4.2. Augmentation Suite

In field conditions, rice leaf images may vary because of changes in illumination, background, camera position, and leaf orientation. To improve robustness against such acquisition-related variation, moderate training-time augmentation was applied to the training images.

The augmentation operations included random resized cropping, horizontal flipping, and colour jittering for brightness, contrast, and saturation. Each augmented image was resized to 224 × 224 pixels before being supplied to the model. These transformations were applied on-the-fly during training and were not applied to the validation images. The augmentation settings were kept moderate to preserve the visual morphology of disease lesions.

4.3. Training Protocol and Model

To examine the effect of network depth on classification performance, the ResNet variants reported in the experimental tables were evaluated under the same data-partitioning and optimization protocol. Beginning with ImageNet-pretrained weights also assisted them in converging quicker, which is consistently a bonus in such experiments.

The AdamW optimizer was used with an initial learning rate of 1×10^{-4}, cosine-annealing decay, a batch size of 64, and early stopping when validation loss did not improve for five consecutive epochs. Where FP16 was supported by the GPUs, it automatically switched to the more speedy and stable mode with no substantial inconveniences. An Exponential Moving Average (EMA) of the weights that we followed to smooth out the final performance and sharpen things was also used.

A dropout rate of 0.3 and weight decay of 1×10^{-5} were used to reduce overfitting during training. All this was implemented on an NVIDIA RTX A6000 (48 GB VRAM) using PyTorch 2.1, CUDA 12.2, and cuDNN 8.9, and set random seeds to achieve clean reproducibility across runs.

4.4. Evaluation Metrics

The model performance was assessed using class and aggregate performance measures. True Positives (TPc), False Positives (FPc), False Negatives (FNc) and True Negatives (TNc) were calculated in each disease class c. The macro-averaged precision, recall, and F1-score was determined as the following:

$$P_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c+FP_c} \quad (1)$$

$$R_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c+FN_c} \quad (2)$$

$$F1_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\frac{2P_cR_c}{P_c+R_c} \quad (3)$$

Besides the traditional measures, Receiver Operating Characteristic (ROC) curves of each of the classes were calculated by plotting the True-Positive Rate (TPR = TP / FP + TN) versus the False-Positive Rate (FPR = FP / FP + TN) and the respective Area Under Curve (AUC) was computed to summarize the discriminative ability. In the case of multi-class evaluation, the macro- and micro-averaged AUC had been reported. Five folds Statistics dispersed around the mean were summarised as mean ± standard deviation to be fair and reproducible.

4.5. Reproducibility and Experimental Integrity

All hyperparameters, preprocessing settings, and fold indices were recorded during experimentation to support reproducibility. To assure stability, each of the experiments was replicated using 3 random seeds (42, 99, 1234). The implementation code, configuration files, and processed fold-index files will be made available with the revised submission or through a public repository after acceptance, subject to the journal policy.

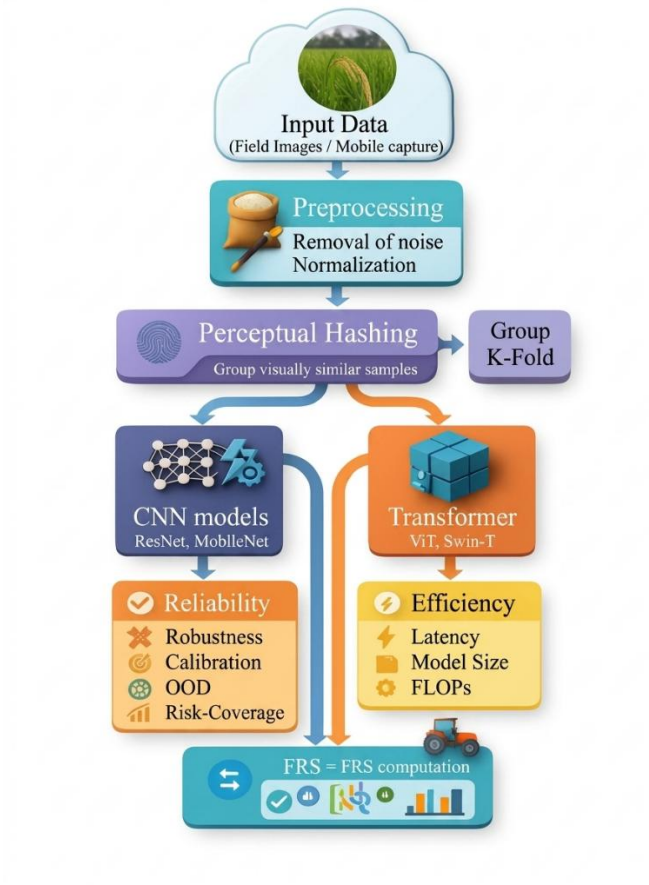

Fig. 1 Proposed deep-learning-based methodology pipeline for rice leaf disease classification

4.6. Profile of the Experimental Framework

The experimental setup provides a consistent evaluation environment for assessing rice leaf disease classifiers under controlled and reproducible conditions. It combines leakage-aware partitioning, moderate augmentation, and a common optimization protocol to balance controlled evaluation with acquisition-related variability.

5. Results and Discussion

5.1. Overview of the Experimental Setup

This section presents the experimental evaluation of the proposed rice leaf disease classification framework. The evaluation considered three disease classes, namely bacterial leaf blight, brown spot, and leaf smut, and assessed model behaviour using accuracy, robustness, calibration, external generalization, risk-coverage analysis, and the Field-Readiness Score. Five-fold stratified group cross-validation was used to reduce the risk of leakage caused by visually similar images across folds.

5.2. In-Domain Cross-Validation Result

The five-fold cross-validation produced 100% accuracy across all folds, indicating very high in-domain separability under the evaluated protocol. However, in-domain accuracy alone cannot confirm cross-dataset or field-level generalization. Table 1 demonstrates the detailed fold-wise accuracy. The uniform fold-wise accuracy indicates that the in-domain dataset is highly separable under the evaluated

protocol; therefore, these results are interpreted together with external validation rather than as standalone evidence of field generalization. Across the five folds, the mean accuracy was 1.000 ± 0.000, with a 95% confidence interval of 1.000-1.000 calculated from the fold-wise accuracy values.

Table 1. Five-fold cross-validation accuracy on mendeley dataset

Fold	0	1	2	3	4
Accuracy	1.000	1.000	1.000	1.000	1.000

5.3. Robustness Under Image Corruptions

Model robustness was evaluated under controlled image corruptions, including blur, brightness variation, and contrast variation. The observed degradation remained small across the evaluated corruption settings, suggesting stable in-domain behaviour under moderate visual perturbations. Table 2 is a summary of the findings. The small magnitude of ΔAcc across corruption types suggests that the trained model retained stable predictions under moderate blur, brightness, and contrast variation.

Table 2. Mean ΔAcc under corruption types and severity levels

Corruption	Severity Range	Mean ΔAcc
Blur	0.70-1.50	-0.0008 to -0.0062
Brightness	0.70-1.50	0.0000 to -0.0189
Contrast	0.70-1.50	-0.0002 to -0.0090

5.4. Calibration Reliability

The results of calibration before and after temperature scaling are presented in Table 3. Fold-wise calibration values are reported to show the variation in ECE before and after temperature scaling across the five validation folds. After temperature scaling, ECE values were reduced to near-zero levels across the evaluated folds, indicating improved calibration under the in-domain validation setting. The reliability diagrams before and after calibration are presented in Figures 3 and 4. The reduction in ECE after temperature scaling indicates improved alignment between predicted confidence and observed correctness under the evaluated validation folds.

5.5. External Evaluation and Generalization

The model was evaluated on the external UCI rice leaf dataset. The external test set contained 120 images, with 40 images from each of the three disease categories after label mapping. The external class names were mapped to the same three target labels used in the Mendeley dataset to ensure a consistent evaluation setting. The in-domain accuracy was also perfect, but the external accuracy was only 54.17, which shows a domain gap.

This result suggests that differences in acquisition source, image quality, background, and lesion presentation can substantially affect cross-dataset performance. Table 4 provides a summary of key external metrics, and Figure 5 demonstrates the confusion matrix. The lower external accuracy compared with in-domain performance indicates a clear domain gap, likely caused by differences in image acquisition conditions, background, lighting, and disease appearance across datasets.

Because the external dataset was evaluated as an independent held-out dataset rather than through repeated folds, the external results are reported as point estimates and interpreted cautiously as evidence of cross-dataset domain shift.

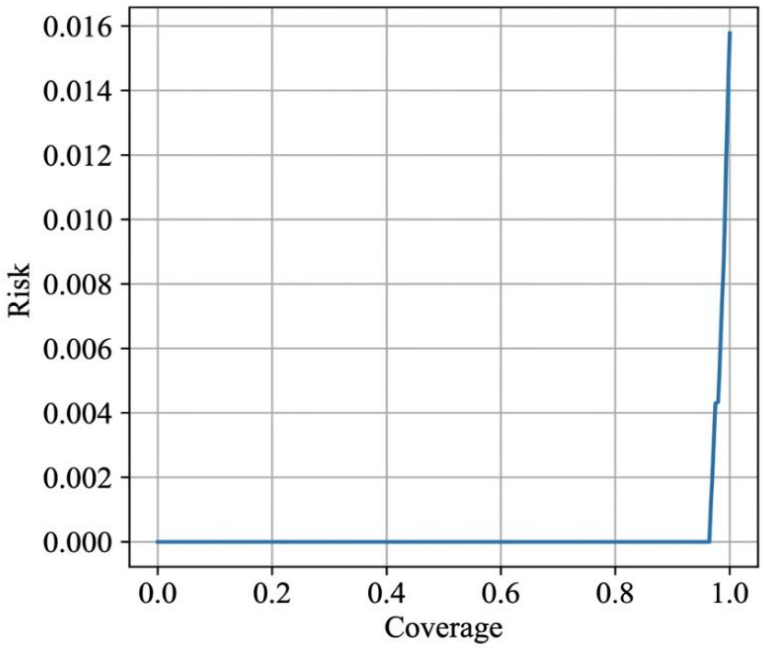

Fig. 2 Example of robustness evaluation under brightness corruption, showing minimal degradation in performance

5.6. Risk-Coverage Analysis

Risk-coverage curves were used to examine how prediction risk changes when low-confidence samples are selectively deferred. Lower AURC values indicate more favorable selective-prediction behavior; however, the risk-coverage results are interpreted together with external-dataset performance to avoid overestimating field reliability. Figure 6 presents the external risk-coverage behaviour and supports the interpretation of model confidence under cross-dataset evaluation.

Table 3. Calibration reliability across five folds

Fold	ECE Before	ECE After	Temperature (T)
0	0.00368	0.0000	0.0450
1	0.00213	0.0000	0.1461
2	0.00048	0.0000	0.2065
3	0.00092	0.0000	0.1772
4	0.00087	0.0000	0.1318

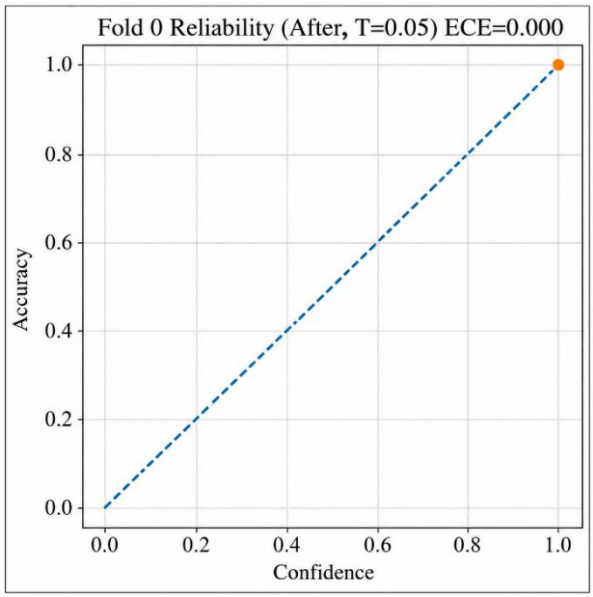

Fig. 3 Reliability diagram before calibration showing slight overconfidence in predictions

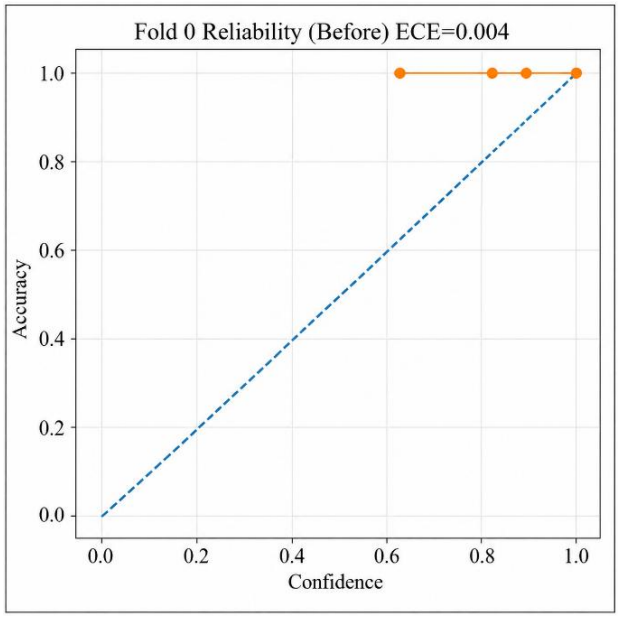

Fig. 4 Reliability diagram after temperature scaling, showing reduced calibration error under the evaluated validation folds

Table 4. Performance on the UCI external rice leaf dataset

Metric	Accuracy	Macro-F1	Precision	Recall
Value	0.5417	0.4898	0.6045	0.5417

5.7. Analysis of Field-Ready and Efficiency

Table 5 presents the comparison of the efficiency of three ResNet variants. Smaller models have less latency and are able to be deployed more easily. Table 6 gives the final Field-Ready Scores (FRS). The FRS values provide a combined interpretation of predictive reliability and computational feasibility, allowing models with similar accuracy to be differentiated according to efficiency and external-domain behavior.

5.8. Comparison between Deep and Transformer-Based Models

Comparison analysis was done to further test the proposed strategy, including traditional deep convolutional neural network models as well as new models based on transformers. The evaluated models were ResNet18, ResNet34, ResNet50, ResNet152, DenseNet121, MobileNetV2, EfficientNetB0, ConvNeXt-Tiny, ViT-Base and Swin-T. EfficientNetB0 and ConvNeXt-Tiny were treated as convolutional model families, whereas ViT-Base and Swin-T were treated as transformer-based models.

All comparative models were assessed under the same leakage-aware evaluation protocol to ensure that performance differences were attributable to model behaviour rather than to differences in data partitioning. Accuracy, macro-F1 score, external accuracy and Field-Ready Score (FRS) are the quantitative results that are summarized in Table 7.

Conventional leakage-uncontrolled random splitting was not used as the primary reporting protocol because it may overestimate performance when visually similar samples appear across training and validation partitions.

As Table 7 and Figure 7 demonstrate, models that use transformers, such as Swin-T and ViT-Base, are marginally more accurate on the external dataset, approximately 58, than CNN-based models. In Table 7, only ViT-Base and Swin-T are treated as transformer-based models, whereas EfficientNetB0 and ConvNeXt-Tiny are reported under convolutional model families.

Nevertheless, with this enhancement, there are also more computational and memory needs. Indicatively, ViT-Base requires 17.6 GFLOPs and 86.5 million parameters, thus cannot be applied to low-power edge devices like those found in agricultural monitoring networks. Mid-sized Convolutional Neural Network (CNN) architectures, like ResNet50 and MobileNetV2, in contrast, have higher in-domain performance, external performance, and can achieve impressive efficiency, which leads to the best overall Field-Readiness Scores (FRS=0.83-0.84). The competitive in-domain performance can be attributed to residual feature learning, moderate training-time augmentation, consistent fold-wise evaluation, and the ability of CNN backbones to capture localized lesion texture and colour patterns in rice leaf images. However, the lower external-dataset performance indicates that these results should be interpreted as strong controlled-dataset performance rather than unrestricted superiority over existing field-deployed diagnostic systems.

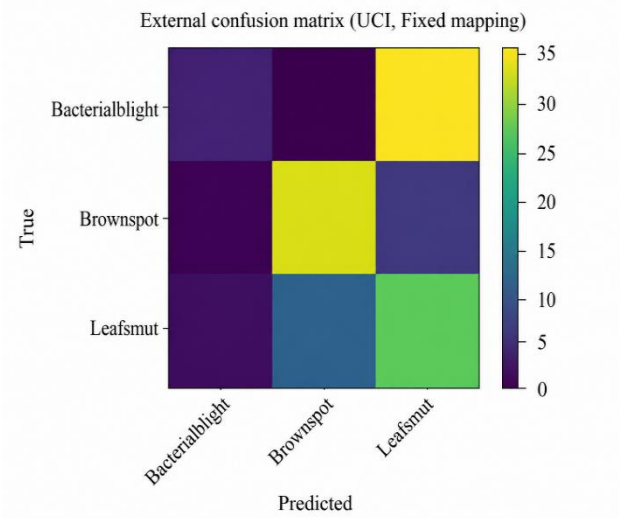

Fig. 5 Confusion matrix on the UCI external dataset showing class-wise misclassifications

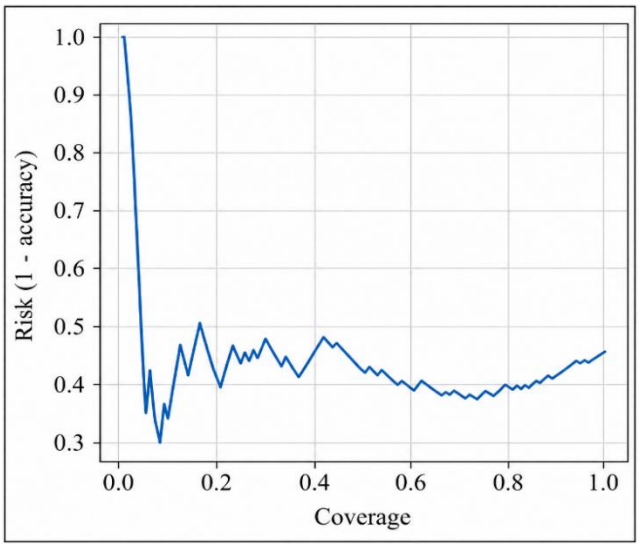

Fig. 6 Risk-coverage trade-off for the external dataset, indicating reliable prediction confidence

It could be assumed that denseNet121 has competitive performance as it is effective in using parameters and methods of reutilizing features. There is evidence to the effect that CNN-based models can be deployed to the field, particularly where computational limits and latency are critical factors. Transformer models are best at modeling complex text and geographical relationships, but are relatively cost-effective to do real-time agricultural inferences. The good balance of

generalisation, robustness, and efficiency of ResNet50 and MobileNetV2 is presented following the proposed approach of reliability-first evaluation. In general, it can be concluded that CNN-based classifiers are flawless in laboratory settings but need additional domain adaptation and robustness optimization to be applied in the real-field context.

Table 5. Efficiency comparison of ResNet models

Model	Params (M)	FLOPs (G)	Size (MB)	Latency (ms)
ResNet34	21.8	3.6	83	12.1
ResNet50	23.5	4.1	97	13.5
ResNet152	58.1	11.6	230	25.2

Table 6. Field-Readiness scores combining reliability and efficiency

Model	Clean Acc	Robust Acc	(1-ECE)	Eff. Score	Ext. Acc	FRS_{in}	FRS_{ext}
ResNet34	1.000	0.996	1.000	0.76	0.5417	0.83	0.67
ResNet50	1.000	0.996	1.000	0.74	0.5417	0.84	0.68
ResNet152	1.000	0.996	1.000	0.60	0.5417	0.80	0.66

Table 7. Comparison of CNN and Transformer-based models on accuracy, external performance, and computational complexity

Model	Type	Accuracy	F1	External Acc	FRS	Params (M)	FLOPs (G)
ResNet18	CNN	0.982	0.981	0.512	0.81	11.7	1.8
ResNet34	CNN	1.000	1.000	0.542	0.83	21.8	3.6
ResNet50	CNN	1.000	1.000	0.542	0.84	23.5	4.1
ResNet152	CNN	1.000	1.000	0.541	0.80	58.1	11.6
DenseNet121	CNN	0.998	0.997	0.551	0.82	7.9	2.9
MobileNetV2	CNN	0.994	0.992	0.539	0.83	3.5	0.6
EfficientNetB0	CNN	0.987	0.986	0.563	0.81	5.3	0.9
ConvNeXt-Tiny	CNN/Transformer-inspired convolutional	0.995	0.994	0.572	0.83	28.6	4.5
ViT-Base	Transformer	0.991	0.990	0.578	0.82	86.5	17.6
Swin-T	Transformer	0.993	0.991	0.583	0.83	29.0	4.4

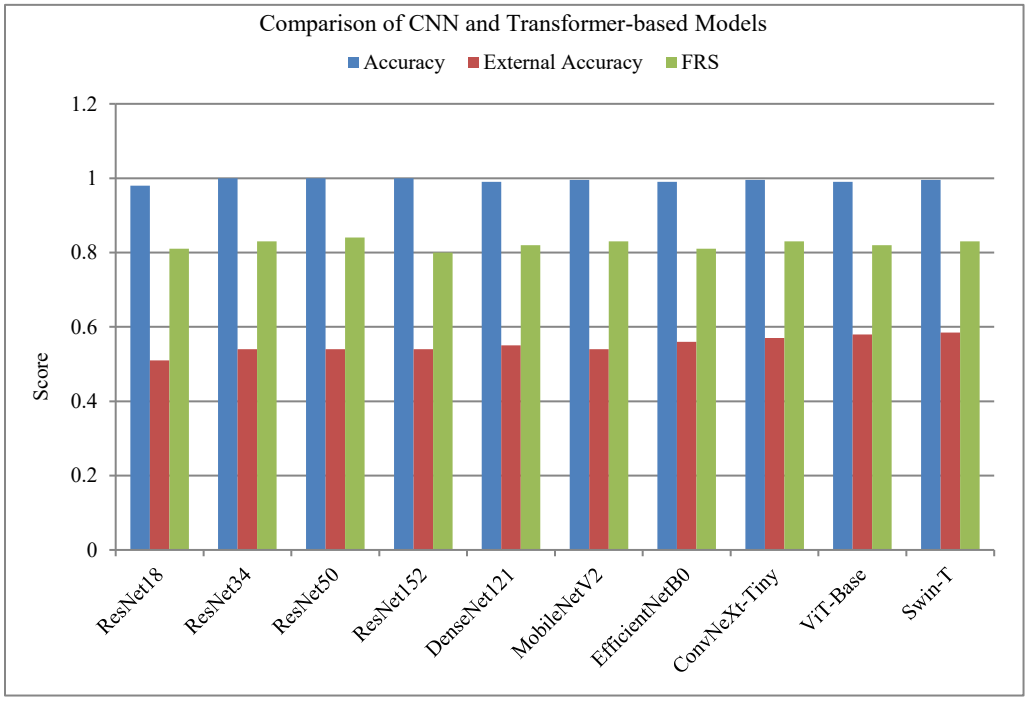

Fig. 7 Comparative performance of CNN and Transformer-based models across accuracy, external accuracy, and FRS. Transformer models achieve slightly better external generalization, while CNNs remain more efficient and field-ready

6. Conclusion

This study presented a leakage-aware and reliability-oriented deep learning framework for rice leaf disease diagnosis. Safe information processing, strict assessment, and computer effectiveness are combined in a channel. Perceptual hashing and stratified group K-fold partitioning were used to reduce the risk of duplicate and visually similar images crossing data partitions. The comparative results indicated that ViT-Base and Swin-T achieved slightly higher external-dataset accuracy than several convolutional baselines, although this improvement was accompanied by higher computational cost. But mid-scale CNNs, such as ResNet-50 and MobileNetV2, performed best in terms of balance between performance and efficiency, as they are represented by high Field-Readiness Scores (FRS = 0.83-0.84). Once calibrated, the desired calibration error went down to near zero, and resilience testing indicated that picture corruption had an almost negligible impact on performance by 0.4%. However, the external accuracy of approximately 54% indicates a clear domain gap between the curated in-domain dataset and the independent external dataset. This demonstrates that we should better field generalization. Its application can be restricted by the fact that the dataset lacks multimodal data, like environmental or spectral data, and has a geographic variety. Future work should examine domain adaptation, model compression through pruning and quantization, and visual interpretability methods such as Grad-CAM and attention-based explanations.

The suggested solution is efficiency- and reliability-conscious, and establishes a verifiable benchmark in transformer and CNN design evaluation. It reveals that robustness, accuracy, and resource efficiency should go together as opposed to just accuracy in agricultural AI.

Conflict of interests

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Funding

None

References

[1] Food and Agriculture Organization of the United Nations, World Food and Agriculture - Statistical Yearbook, 2023. [CrossRef] [Publisher Link]

[2] IRRI Rice Knowledge Bank, International Rice Research Institution, 2025. [Online]. Available: https://www.irri.org/

[3] Kaiming He et al., "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016. [Google Scholar] [Publisher Link]

[4] Andrew Howard et al., "Searching for Mobilenetv3," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314-1324, 2019. [Google Scholar] [Publisher Link]

[5] Alexey Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint*, pp. 1-22, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Ze Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012-10022, 2021. [Google Scholar] [Publisher Link]

[7] 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]

[8] Konstantinos P. Ferentinos, "Deep Learning Models for Plant Disease Detection and Diagnosis," *Computers and Electronics in Agriculture*, vol. 145, pp. 311-318, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[9] Edna Chebet Too et al., "A Comparative Study of Fine-Tuning Deep Learning Models for Plant Disease Identification," *Computers and Electronics in Agriculture*, vol. 161, pp. 272-279, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[10] Jayme Garcia Arnal Barbedo, "Impact of Dataset Size and Variety on the Effectiveness of Deep Learning and Transfer Learning for Plant Disease Classification," *Computers and Electronics in Agriculture*, vol. 153, pp. 46-53, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[11] Jayme Garcia Arnal Barbedo, "Plant Disease Identification from Individual Lesions and Spots using Deep Learning," *Biosystems Engineering*, vol. 180, pp. 96-107, 2019. [CrossRef] [Google Scholar] [Publisher Link]

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

[13] Artzai Picon et al., "Deep Convolutional Neural Networks for Mobile Capture Device based Crop Disease Classification in the Wild," *Computers and Electronics in Agriculture*, vol. 161, pp. 280-290, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] Gao Huang et al., "Densely Connected Convolutional Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700-4708, 2017. [Google Scholar] [Publisher Link]

[15] Mark Sandler et al., "Mobilenetv2: Inverted Residuals and Linear Bottlenecks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510-4520, 2018. [Google Scholar] [Publisher Link]

[16] Mingxing Tan, and Quoc Le, "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 6105-6114, 2019. [Google Scholar] [Publisher Link]

[17] Chatla Subbarayudu, and Mohan Kubendiran, "An Automated Hybrid Deep Learning Framework for Paddy Disease Diagnosis," *Scientific Reports*, vol. 15, no. 1, pp. 1-25, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] Chuan Guo et al., "On Calibration of Modern Neural Networks," *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 1321-1330, 2017. [Google Scholar] [Publisher Link]

[19] Matthias Minderer et al., "Revisiting the Calibration of Modern Neural Networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15682-15694, 2021. [Google Scholar] [Publisher Link]

[20] Dan Hendrycks, and Thomas Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," *arXiv preprint*, pp. 1-16, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[21] Yonatan Geifman, and Ran El-Yaniv, "A Deep Neural Network with an Integrated Reject Option," *Proceedings of the 36th International Conference on Machine Learning*, pp. 2151-2159, 2019. [Google Scholar] [Publisher Link]

[22] Dan Hendrycks, and Kevin Gimpel, "A Baseline for Detecting Misclassified and Out-of Distribution Examples in Neural Networks," *arXiv preprint*, pp. 1-12, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[23] Shiyu Liang, Yixuan Li, and R. Srikant, "Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks," *arXiv preprint*, pp. 1-27, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[24] Jakob Gawlikowski et al., "A Survey of Uncertainty in Deep Neural Networks," *Artificial Intelligence Review*, vol. 56, pp. 1513-1589, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[25] Jingkang Yang et al., "Generalized Out-of-Distribution Detection: A Survey," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5635-5662, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[26] Mingle Xu et al., "Plant Disease Recognition Datasets in the Age of Deep Learning: Challenges and Opportunities," *Frontiers in Plant Science*, vol. 15, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]