



IJRTSM

INTERNATIONAL JOURNAL OF RECENT TECHNOLOGY SCIENCE & MANAGEMENT

"IMAGE-BASED RICE LEAF DISEASE DETECTION: A NARRATIVE REVIEW OF MACHINE LEARNING, DEEP LEARNING, EVALUATION, AND REPRODUCIBILITY"

Vishal Harinkhere ¹, Kundan Kumar Mishra ²

¹ P. G. Scholar, IET (CSE), SAGE University Indore, Madhya Pradesh, India

² Professor, IET (CSE), SAGE University Indore, Madhya Pradesh, India

ABSTRACT

Image-based rice leaf disease recognition has developed from handcrafted feature-based methods to convolutional neural networks (CNNs), transfer learning, lightweight networks, ensemble and hybrid architectures, object detection, segmentation, and explainable artificial intelligence (XAI). This narrative review synthesizes representative literature with particular attention to dataset design, model selection, evaluation, and reproducibility. The reviewed studies frequently report high accuracy on controlled or curated image collections, but reported metrics are not directly comparable because datasets, disease taxonomies, acquisition conditions, class distributions, and train/validation/test protocols differ. The review therefore examines leakage-resistant partitioning, dataset provenance, class-wise evaluation, independent testing, repeated experiments, uncertainty, computational efficiency, and deployment conditions. It also distinguishes whole-image classification from object detection and lesion segmentation, which require different labels and evaluation metrics. A reproducibility-oriented reporting framework is presented to guide future studies in documenting data sources, preprocessing, architecture, training configuration, evaluation procedures, and external validation. The review concludes that meaningful progress in rice disease recognition depends not only on model architecture but also on transparent datasets, rigorous evaluation, and evidence of generalization beyond the conditions used for model development.

The review also considers the difference between laboratory-style classification and field-oriented decision support. A controlled image containing a single leaf on a simple background presents a substantially easier problem than a photograph containing several leaves, soil, water, shadows, overlapping plants, insects, and unrelated vegetation. Classification, object detection, and segmentation address different versions of this problem. A literature synthesis should therefore avoid treating every reported accuracy value as evidence for the same real-world capability.

For rice disease recognition, the distinction between image-level and plant-level independence is particularly important. Several photographs of the same leaf, plant, plot, or collection session may share nearly identical visual information. If related images are randomly divided across training and testing subsets, the model can encounter near-duplicates during training and evaluation. The resulting score may reflect recognition of the image source rather than robust recognition of disease symptoms. Consequently, dataset construction is treated in this review as part of model evaluation rather than as a preliminary administrative step.

Deep learning reduces some of this manual feature-engineering burden by learning hierarchical representations from images. CNNs can learn low-level edges and colour transitions in early layers and progressively more complex patterns in deeper layers. Transfer learning can further reduce the amount of task-specific data required by adapting a network pretrained on a large image corpus. However, these advantages do not remove the need for careful experimental design. A deep model can memorize background or acquisition artefacts when training images are limited, and a high test accuracy can therefore be misleading if the test set is not independent at the appropriate biological or acquisition level.

The reviewed literature includes both conventional machine-learning pipelines and deep-learning approaches. Earlier pipelines generally separated the task into image segmentation, feature extraction, and classification. Colour, texture, shape, and lesion-region descriptors were then supplied to classifiers such as K-nearest neighbours, support vector machines, or random forests. This modular design has an advantage for interpretation because each processing stage can be inspected separately. It also creates additional sources of variation: a change in segmentation threshold or feature descriptor can alter the downstream classifier even when the classifier itself is unchanged.

The practical objective of an image-based disease recognition system is not simply to assign a label to a photograph. A useful system should connect image acquisition, preprocessing, representation learning, classification, and decision support in a way that remains meaningful when images differ from the conditions used during model development. This distinction is especially important in agricultural environments, where the same disease can appear at different stages, under different illumination, and against different backgrounds. A review of the literature therefore needs to consider both algorithmic design and the evidence used to support the reported performance.

Key Words: rice leaf disease detection; rice disease diagnosis; computer vision; machine learning; convolutional neural network; transfer learning; attention; segmentation; object detection; lightweight deep learning; reproducibility; external validation.

I. INTRODUCTION

Rice (*Oryza sativa* L.) is a major staple crop and an important component of food systems in many countries. Its productivity can be reduced by bacterial, fungal, viral, and other diseases that affect leaves and other plant tissues. Disease symptoms may appear as elongated lesions, spots, discoloration, necrosis, streaking, or changes in leaf structure. Conventional diagnosis relies heavily on visual inspection and agricultural expertise. Although expert diagnosis remains important, image-based artificial intelligence offers a complementary approach in which visible symptoms are converted into quantitative representations and disease predictions.

The attraction of image-based ML is not limited to automation. Digital images can be stored, shared, reanalysed, and used to benchmark algorithms under a defined protocol. This creates opportunities for decision-support applications on mobile phones, cameras, drones, and edge devices. Early rice-disease recognition systems generally used image segmentation followed by handcrafted colour, texture, shape, histogram, or lesion features and then applied classifiers such as support vector machines (SVMs), k-nearest neighbours, decision trees, random forests, or Bayesian methods. These pipelines remain useful as baselines because their components can be relatively compact and interpretable.

Deep learning changed the design of many plant-disease recognition systems by learning hierarchical image representations directly from data. CNNs became the dominant architecture for image classification, while transfer learning allowed pretrained networks to be adapted to agricultural datasets that are often smaller than large general-purpose image corpora. Subsequent work has explored residual networks, MobileNet, EfficientNet, Inception-family networks, ensembles, attention mechanisms, segmentation networks, and object detectors. Representative rice studies report strong performance across these approaches, but their datasets and evaluation procedures are not identical [7–20].

A central methodological issue is therefore the interpretation of accuracy. A model obtaining 99% accuracy on a small, controlled, homogeneous dataset is not necessarily solving the same problem as a model evaluated on independent images collected from different fields, cultivars, devices, seasons, or lighting conditions. Dataset leakage can also produce overly optimistic results when related images from the same plant, capture session, or source are distributed across both training and testing partitions. Reviews of plant and rice disease recognition have repeatedly identified dataset diversity and real-world generalization as important unresolved issues [6,12,13].

This article is a narrative review that synthesizes published evidence on image-based rice disease recognition and identifies methodological practices that can improve transparency and reproducibility.

1.1 Contributions of the Review

- Provides a structured narrative synthesis of classical ML, CNN, transfer-learning, lightweight, ensemble, detection,

segmentation, and XAI approaches for rice leaf disease recognition.

- Separates reported study results from cross-study interpretation so that heterogeneous accuracy values are not treated as a common benchmark.
- Highlights dataset provenance, source-level or plant-level splitting, leakage prevention, and external validation as central elements of experimental validity.
- Summarizes deployment considerations including parameters, FLOPs, memory, latency, and energy consumption.
- Proposes a reproducibility-oriented reporting checklist for future rice disease image studies.

The narrative structure also recognizes that literature availability changes over time. Recent papers can introduce new architectures or application settings, but newer publication date alone does not establish stronger evidence. A recent study with a small or narrowly sampled dataset may provide less information about generalization than an older study with a carefully separated test set. The review therefore evaluates methodological reporting alongside algorithmic novelty.

For each representative study, attention was given to the disease classes, image source, task definition, preprocessing, model family, training strategy, evaluation metrics, and any evidence of external or independent validation. When a paper provided only a final accuracy without enough information to reconstruct the evaluation protocol, the result was retained as a reported study outcome but was not treated as evidence of superiority over other studies. This approach is important because numerical comparison without protocol comparison can create a false impression of scientific precision.

The review distinguishes three levels of evidence. The first level consists of rice-specific primary studies reporting an image-recognition task and an evaluation result. The second level consists of broader plant-disease studies that provide methodological context for transfer learning, dataset diversity, or image-based diagnosis. The third level consists of foundational architecture and reporting papers used to explain the technical basis of models such as ResNet, MobileNetV2, and EfficientNet. Keeping these levels distinct prevents a general computer-vision result from being presented as if it were a direct rice-disease experiment.

The purpose of the literature-selection process was to construct a technically coherent narrative rather than to claim an exhaustive systematic review. Studies were considered representative when they contributed a clear methodological idea, a rice-disease image application, a relevant deep-learning architecture, or an evaluation practice that informs reproducibility. The resulting synthesis gives more attention to methodological transparency than to the number of papers associated with a particular architecture.

II. REVIEW SCOPE AND LITERATURE SELECTION

This article uses a narrative review design with a reproducibility-focused synthesis. The search strategy was designed to identify representative rice-specific studies and recent reviews rather than to perform a systematic meta-analysis. Search concepts included combinations of “rice leaf disease,” “rice disease classification,” “rice disease detection,” “deep learning,” “CNN,” “transfer learning,” “MobileNet,” “EfficientNet,” “ResNet,” “ensemble,” “YOLO,” “segmentation,” and “explainable AI.” Literature was considered relevant when image data were a principal input and the work addressed rice disease recognition, classification, detection, segmentation, or closely related diagnostic tasks.

The review prioritizes peer-reviewed primary studies and review articles available through scholarly publisher or indexing platforms. Foundational architecture papers are included when needed to explain methods used in rice-specific studies. For each representative study, the synthesis records the study year, disease scope, image source or dataset when reported, model family, transfer-learning or ensemble strategy, reported metric, and evidence of independent or field-oriented evaluation. Because the studies use different datasets, disease classes, image-acquisition conditions, and split protocols, performance values are reported descriptively and are not statistically pooled.

To improve transparency, the review states its scope, search concepts, selection logic, extraction fields, and synthesis approach. It is a narrative review rather than a systematic review or meta-analysis.

The workflow emphasizes that literature synthesis is not simply a list of papers. Each study must be interpreted in relation to its dataset, experimental design, model, metric, and validation environment. This distinction becomes particularly important when two papers report similar accuracy values but use substantially different datasets or test protocols.

2.1 Inclusion and Exclusion Logic

- Included: rice disease image studies using classical ML, deep learning, transfer learning, ensembles, detection, segmentation, or explainability.
- Included: review papers that synthesize rice disease recognition or broader plant disease image recognition.
- Included: foundational architecture papers when they are needed to explain methods used in rice-specific studies.
- Excluded from quantitative comparison: studies for which the task, dataset, metric, or evaluation protocol is insufficiently described to interpret the reported result.
- Not pooled: accuracy values from heterogeneous datasets, because the numerical values do not represent a common experimental benchmark.

2.2 Data Extraction Fields

For each representative study, the review considers: disease classes; healthy class definition; image source; approximate dataset scale when reported; preprocessing; augmentation; model family; transfer-learning strategy; fusion or ensemble mechanism; train/validation/test design; evaluation metrics; external validation; and deployment considerations. Where a study does not provide an item, the review treats the missing information as a reporting limitation rather than inferring it.

Image quality should also be reported as part of dataset description. Resolution, compression, blur, illumination, camera type, distance, and orientation can all affect the appearance of disease symptoms. A reproducible dataset description does not require every image to be identical; instead, it should document the range of conditions represented and explain any preprocessing that changes those conditions before model training.

Background variation is another important source of distribution shift. A close-up photograph with the leaf occupying most of the frame provides a strong signal from the target object. A field photograph may include sky, soil, irrigation structures, neighbouring leaves, or other crops. If a training collection has a consistent background for one class, the background can become a shortcut feature. This is one reason that external validation and cross-environment testing are valuable even when internal accuracy is high.

The four disease categories represented in the reported literature—bacterial blight, blast, brown spot, and tungro—illustrate why visual classification is challenging. The classes can differ in lesion shape and distribution, but photographs may also contain overlapping symptom patterns. A model that is trained on a single disease stage can learn a narrow visual signature. Including images from multiple stages and acquisition conditions is therefore relevant to generalization, provided that the labels remain sufficiently reliable.

Rice leaf symptoms are expressed through combinations of colour change, lesion geometry, texture disruption, necrotic regions, chlorosis, and changes in the surrounding tissue. These characteristics are not fixed visual templates. Disease progression can modify lesion size and colour, while environmental stress can create symptoms that resemble disease. As a result, an image classifier may face both within-class variation and between-class similarity.

III. RICE LEAF DISEASES AND IMAGE CHARACTERISTICS

Rice disease recognition is biologically and visually heterogeneous. Bacterial blight may produce elongated water-soaked or yellowish lesions that extend along leaf margins. Rice blast is commonly associated with spindle- or diamond-shaped lesions with contrasting margins. Brown spot can produce circular or oval brown lesions, sometimes with darker boundaries. Tungro can cause characteristic yellowing or orange-yellow discoloration. These descriptions are useful for understanding the visual cues that image models may exploit, but actual symptom appearance changes with disease stage, cultivar, environmental stress, and imaging conditions.

A classification system trained on clean, close-up images may learn texture and lesion shape effectively, yet the same

model can encounter difficulty when leaves are partially occluded, photographed at oblique angles, affected by shadows, or surrounded by soil and other vegetation. Disease symptoms can also overlap. For example, discoloration caused by nutrient stress or environmental damage may resemble some disease patterns. This creates a distinction between visual pattern classification and agronomic diagnosis: an image model can classify the visual distribution on which it was trained without necessarily establishing the biological cause of that appearance.

Image acquisition therefore forms part of the problem definition. Important variables include camera sensor, focal length, distance from the leaf, illumination, white balance, background, field versus laboratory environment, and whether the image contains one leaf or multiple leaves. If multiple photographs are taken sequentially from the same plant, they may be highly correlated. Treating such images as independent observations can make the test set appear larger than its effective information content.

The disease categories themselves also require careful definition. A study may use four classes, such as healthy, bacterial blight, brown spot, and blast, while another study may use a larger taxonomy containing several disease stages or additional diseases. The number of classes affects task difficulty, and a model trained for four visually distinct classes should not be described as equivalent to a model trained for ten or more categories simply because both report accuracy.

3.1 Dataset Provenance

A reproducible rice-disease study should identify the exact dataset or image repository, version or access date, source institution or collection process, disease labels, image counts, acquisition conditions, and licence or permission status. If images are collected by the authors, the study should describe the collection location at an appropriate level, camera or device, approximate environmental conditions, sampling strategy, and annotation process. If a public dataset is used, the dataset citation should be included independently of the model paper citation.

3.2 Annotation Quality

Disease labels can be produced by farmers, researchers, plant pathologists, crowd-sourced annotation, or inherited dataset labels. These sources do not have identical reliability. When expert labels are available, the annotation protocol should identify the expertise of the annotators and explain how disagreements were handled. For object detection or segmentation, bounding-box or mask quality should also be evaluated. Annotation noise is particularly important when disease symptoms are subtle or when healthy and diseased classes overlap visually.

Classical classifiers remain useful as baselines because they provide a reference against which more complex models can be evaluated. SVMs, random forests, and K-nearest neighbours can be trained on compact feature representations and may require less computational resources than large CNNs. Their scientific value in a modern study is not limited to deployment: a strong baseline can reveal whether the additional complexity of a deep architecture produces a meaningful improvement under the same data partition and evaluation protocol.

Segmentation attempts to isolate the relevant plant region before feature extraction. The reviewed literature describes segmentation as part of the workflow used before classification. In controlled images, simple thresholding or clustering may be sufficient; in complex field scenes, segmentation becomes substantially harder because foreground and background colours can overlap. A segmentation error can propagate into every downstream feature and can therefore affect the classifier even when the classifier is otherwise well tuned.

Texture features capture spatial relationships that are not represented by individual pixel values. Measures derived from local binary patterns, grey-level co-occurrence matrices, gradients, or related descriptors can represent the roughness and regularity of diseased tissue. Texture can be useful when lesions create repeated local patterns, but it may also be affected by image compression, focus, leaf veins, and background material. The same texture descriptor can therefore behave differently across acquisition environments.

Colour descriptors can encode global or local distributions of pixel values. RGB is directly available from ordinary digital cameras, while HSV and related spaces can separate colour-related information from brightness more explicitly. However, colour is sensitive to illumination and camera settings. A robust study should therefore explain whether colour normalization, white balancing, histogram adjustment, or other transformations were used. Without this information, reproducing a reported feature vector can be difficult.

A conventional rice-leaf recognition pipeline can be understood as a sequence of transformations: acquisition,

background or leaf segmentation, feature extraction, feature selection or dimensionality reduction, and classification. This architecture is useful for methodological analysis because the source of a prediction can often be traced to a specific feature family. It also makes clear why preprocessing decisions can have as much practical importance as the classifier selected at the final stage.

IV. CLASSICAL MACHINE LEARNING FOR RICE DISEASE RECOGNITION

Before deep CNNs became dominant, image-based disease recognition commonly followed a modular pipeline: image acquisition, segmentation, feature extraction, and classification. Segmentation can isolate the leaf or symptomatic region using thresholding, clustering, colour spaces, contour analysis, or other image-processing techniques. Feature extraction can then summarize colour distributions, texture, shape, lesion geometry, edges, or local patterns. A classifier such as SVM, k-nearest neighbours, decision tree, random forest, or Naive Bayes operates on the resulting feature vector.

4.1 Colour Features

Colour features are attractive because many rice diseases change the visible colour of leaf tissue. RGB, HSV, Lab, and related colour spaces can be used to characterize hue, saturation, intensity, or channel statistics. However, colour is sensitive to illumination and camera processing. A model that relies strongly on absolute pixel values may therefore experience performance degradation when lighting conditions change.

4.2 Texture Features

Texture descriptors attempt to represent local spatial relationships. Local binary patterns, grey-level co-occurrence matrices, gradients, entropy measures, and related descriptors can capture lesion roughness, spot boundaries, or repetitive patterns. Texture features can be useful with relatively small datasets, but their representation is fixed by the feature design and may not adapt as flexibly as learned CNN representations.

4.3 Segmentation and Region-Based Features

Segmentation is particularly useful when the background is complex. K-means clustering, thresholding, and other approaches have been used to separate leaves or symptomatic areas from surrounding pixels. A two-stage system can then calculate disease-area ratios or other severity indicators. The limitation is that segmentation errors propagate to the classifier. Shadows, similar background colours, overlapping leaves, and irregular boundaries can make accurate segmentation difficult.

4.4 Classical Classifiers

SVMs are widely used in agricultural image classification because they can perform effectively in high-dimensional feature spaces and can be trained on relatively small datasets. Decision trees and random forests offer straightforward feature-based interpretation, while Naive Bayes can be computationally inexpensive when its feature assumptions are reasonably satisfied. K-nearest neighbours is simple but can become inefficient as the training set grows. These methods remain useful as baselines because they provide a reference against which learned representations can be evaluated.

4.5 Limitations of Handcrafted Pipelines

The main limitation of classical pipelines is the dependence on feature engineering. A feature that works well under one camera or lighting condition may be less stable under another. Separate segmentation and classification stages can also make the total system sensitive to upstream errors. Nevertheless, classical methods should not be dismissed. They are valuable for small datasets, embedded systems, explainable baselines, and controlled comparisons. A rigorous deep-learning study should ideally compare against at least one transparent non-deep baseline when the research question concerns the benefit of representation learning.

Evaluation should include more than a single aggregate accuracy when the class distribution is uneven. Per-class precision, recall, F1-score, confusion matrices, and where appropriate macro-averaged metrics can reveal whether a model performs consistently across disease categories. A high overall accuracy can coexist with weak recognition of a minority disease class. For an agricultural decision-support system, that class-specific behaviour may be more

informative than the aggregate percentage alone.

Data augmentation is frequently used to increase variation in the training set. Rotations, translations, flips, crops, and controlled changes in brightness can make a classifier less dependent on a single acquisition geometry. Augmentation must nevertheless be biologically plausible. A transformation that creates an unrealistic symptom orientation or alters the disease signal can change the learning problem rather than improve robustness. Augmentation must also be applied only to training data when the purpose is to preserve an untouched evaluation set.

ResNet is relevant because residual connections make it easier to optimize deeper networks. MobileNetV2 is relevant to deployment because depthwise separable operations and inverted residual structures reduce computational cost. EfficientNet provides another perspective by balancing depth, width, and input resolution through compound scaling. These architectures should not be treated as interchangeable labels; each makes different computational and representational trade-offs.

The choice of network depth and width should be considered together with dataset scale. Large models contain many parameters and can express complex functions, but they can also overfit small datasets. Transfer learning changes this trade-off by providing parameters learned from a much larger image corpus. Nevertheless, the source of pretrained weights, the input preprocessing associated with those weights, the layers that were frozen, and the number of fine-tuned epochs should be documented. These details affect the learned representation and can materially change the result.

CNN-based approaches replace much of the handcrafted feature-design stage with learned representations. During training, convolutional filters are optimized so that combinations of local patterns become useful for the final classification objective. Early features may respond to edges and colour transitions, while deeper features can represent more complex spatial arrangements. Pooling or strided operations reduce spatial resolution and can increase the effective receptive field, although excessive downsampling can also remove small lesions that are important for disease recognition.

V. CONVOLUTIONAL NEURAL NETWORKS AND TRANSFER LEARNING

CNNs learn spatially structured representations using convolutional filters, nonlinear activation functions, pooling or strided operations, and progressively higher-level feature maps. Early layers often respond to edges and simple textures, while deeper layers can capture combinations of shapes and disease-related patterns. This hierarchical representation reduces the need to manually specify every relevant visual feature.

5.1 Residual Networks

ResNet introduced residual connections that allow a network to learn a transformation relative to its input rather than requiring every layer to learn a complete new representation [1]. Residual connections made deeper networks easier to optimize and have influenced many later architectures. In rice disease recognition, residual-family networks are relevant because they can provide strong image representations while benefiting from pretrained weights.

5.2 MobileNetV2

MobileNetV2 was designed for computationally constrained environments using inverted residual blocks and depthwise separable convolutions [2]. Its relatively compact structure makes it relevant to smartphone and edge deployment. In rice disease applications, a lightweight architecture can be valuable when the model must run without a powerful cloud server. However, compactness does not automatically guarantee robustness; a small model still depends on representative training data and leakage-resistant evaluation.

5.3 EfficientNet

EfficientNet introduced a compound scaling strategy that balances network depth, width, and input resolution [3]. Its relevance to agricultural recognition is the possibility of improving the accuracy-efficiency trade-off without simply making the network deeper. When an EfficientNet variant is used in a rice disease study, the exact variant, input resolution, pretrained weights, fine-tuning procedure, and classifier head should be documented.

5.4 Transfer Learning

Transfer learning adapts a network pretrained on a large image dataset to a new classification task. A common strategy is to retain early feature-extraction layers, replace the final classifier with a task-specific head, train the new head, and optionally fine-tune some or all of the pretrained layers. The strategy is attractive for rice disease recognition because labelled agricultural datasets are often much smaller than general image datasets [9,11,14].

Transfer learning introduces its own reporting requirements. A study should state the source of pretrained weights, whether the weights were frozen, which layers were unfrozen, learning rates for different parameter groups if applicable, input size, normalization, augmentation, and stopping criteria. Without these details, two studies described as using the same architecture may in fact be using materially different training procedures.

5.5 Data Augmentation

Augmentation can increase training variability through transformations such as rotation, horizontal or vertical flipping where biologically appropriate, cropping, scaling, brightness adjustment, and mild geometric transformations. Augmentation must be applied only within the training process and must not allow information from the test set to influence model selection. The transformation should also be biologically defensible. For example, an arbitrary colour transformation may change the very symptom cue that the classifier needs to recognize.

5.6 Overfitting and Regularization

Small agricultural datasets can encourage overfitting, particularly when a large pretrained model is fine-tuned too aggressively. Dropout, weight decay, early stopping, data augmentation, and careful learning-rate scheduling are common countermeasures. Cross-validation can also help estimate variability when the dataset is limited, although the final test set should remain untouched until the evaluation stage.

5.7 Classification Metrics

Accuracy is intuitive but can be misleading under class imbalance. Precision measures the proportion of predicted positives that are correct, while recall measures the proportion of actual positives that are detected. The F1-score combines precision and recall through their harmonic mean. Macro-averaged metrics give each class equal weight and are useful when disease classes have different frequencies. A confusion matrix provides a direct view of which disease classes are being confused.

Explainable AI can provide additional evidence about model behaviour. Methods such as Grad-CAM produce visual maps indicating image regions associated with a prediction. Such maps are useful for qualitative inspection, but they are not proof that the model has learned the correct biological mechanism. Explanations should therefore be interpreted alongside expert review, counterfactual tests, or quantitative localization measures when available.

Detection and segmentation extend the task beyond whole-image classification. A detector can identify diseased regions or individual leaves within a field scene, while segmentation can estimate the spatial extent of lesions. These tasks require different labels and metrics from classification. Consequently, a detection score or segmentation overlap measure should not be compared numerically with a classification accuracy as if they quantified the same property.

Hybrid models are attractive because they can combine learned representations with complementary information. A CNN feature vector may be concatenated with handcrafted colour or texture descriptors, or features from several backbones may be fused before classification. Such designs introduce additional hyperparameters and therefore require ablation studies. Without an ablation, it is difficult to determine whether the proposed fusion mechanism actually contributes to the final result.

Ensemble methods can improve robustness by combining models that make different errors. Voting and probability averaging are simple examples, while more elaborate ensembles can combine heterogeneous backbones or training configurations. The evaluation must account for the additional computational cost of running multiple models. If an ensemble is compared with a single network, both predictive performance and resource requirements should be reported so that the comparison remains meaningful.

Model complexity becomes a practical issue when disease recognition is intended for field devices. A model that performs well on a workstation may require excessive memory or inference time on a smartphone or embedded processor. Deployment-oriented reporting should therefore include model size, parameter count, approximate computational cost, input resolution, and measured inference time on the target hardware where possible. Accuracy and

efficiency are complementary properties rather than substitutes for one another.

VI. LIGHTWEIGHT, ENSEMBLE, HYBRID, DETECTION, AND EXPLAINABLE MODELS

6.1 Lightweight Networks and Edge Deployment

For field use, a model may need to operate on a mobile device or an embedded processor. Predictive accuracy should therefore be accompanied by model size, number of parameters, computational complexity, memory requirement, inference latency, and, where possible, energy consumption. A model that is slightly less accurate but substantially faster may be useful for real-time screening, whereas a large ensemble may be more appropriate when computational resources are available.

6.2 Ensemble Learning

Ensemble learning combines predictions from multiple models. The combination can occur through majority voting, probability averaging, weighted averaging, stacking, or other mechanisms. Ahad et al. evaluated multiple CNN configurations and reported strong performance for an ensemble framework [11]. Pai et al. used softmax probability averaging across selected CNNs and reported 96.81% accuracy for their ensemble [16]. These results illustrate the potential benefit of complementary models, but ensemble complexity must be justified through ablation studies and clear definitions of the combination rule.

6.3 Hybrid Feature Fusion

Hybrid architectures can combine features from different CNN backbones or combine deep representations with handcrafted features. Such a description is incomplete unless the feature flow is specified. A reproducible report should identify the exact source layers, tensor dimensions, concatenation or weighting operation, normalization or projection layers, classifier head, and training strategy. If a fusion model improves accuracy, an ablation study should show whether the improvement comes from the additional features, the larger parameter count, or a particular fusion operation.

6.4 Object Detection

Classification assumes that the image has already been associated with a target leaf or disease class. Field photographs may instead contain multiple leaves, lesions, soil, weeds, stems, and background objects. Object detection addresses this problem by predicting both the location and category of an object or lesion. YOLO-family detectors and other detection architectures have been investigated for rice disease recognition in complex environments [15,20]. Detection performance should be reported using metrics such as precision, recall, mean average precision, and inference time, rather than classification accuracy alone.

6.5 Segmentation and Disease Severity

Segmentation can identify the pixel-level extent of diseased tissue. This is useful when the objective includes disease severity estimation rather than only disease identity. However, segmentation requires more detailed annotation and may be more expensive to develop. A practical pipeline may combine detection or segmentation with classification, but each component should be evaluated separately so that the source of errors is visible.

6.6 Explainable Artificial Intelligence

XAI methods can help determine whether a model is focusing on disease-relevant regions. Grad-CAM can highlight image regions associated with a CNN prediction, while LIME and SHAP provide alternative approaches for local or feature-level explanation. Attention maps can also indicate where the network places emphasis. Such tools do not prove biological correctness, but they can reveal suspicious behaviour, such as strong attention to background regions or image borders. Explainability is therefore a useful diagnostic complement to quantitative metrics [17].

The literature also demonstrates that architecture names alone are insufficient to reproduce a result. Two studies can both report the use of a CNN while differing in input resolution, pretrained weights, augmentation, optimizer, learning rate, batch size, number of epochs, freezing strategy, and class weighting. These implementation choices can alter the outcome substantially. A reproducible comparison therefore requires protocol-level information, not merely model names.

When reading a published result, the first question should be what population the test set represents. If the test images come from the same source as the training images, the result primarily measures performance under that source distribution. If the test set comes from a different collection, location, device, season, or acquisition session, the result provides stronger evidence about generalization. The distinction should be reported explicitly rather than inferred from the numerical score.

Reported accuracy is particularly sensitive to dataset construction. A model evaluated on images that closely resemble its training distribution may obtain a high score without demonstrating strong environmental robustness. Conversely, a model tested on more varied field images may report a lower score while providing more demanding evidence. For this reason, the table in this article preserves study-specific values but does not combine them into a single ranking or pooled estimate.

The representative studies summarized in this review show a progression from handcrafted feature pipelines to transfer-learning and deeper CNN architectures. This progression reflects broader developments in computer vision, but the studies do not form a controlled sequence of experiments. They use different image collections, disease definitions, preprocessing pipelines, class balances, training procedures, and evaluation splits. The appropriate interpretation is therefore methodological synthesis rather than a league table of models.

VII. COMPARATIVE SYNTHESIS OF REPRESENTATIVE PUBLISHED STUDIES

Table 1 summarizes representative studies from the literature. The reported accuracy values are retained as published study-specific findings. They are not treated as a common benchmark because the studies use different disease classes, image sources, dataset sizes, and evaluation protocols.

Study	Method	Disease scope	Reported metric/result	Methodological note
Jiang et al. (2020) [7]	CNN feature extraction + SVM	Four rice diseases	96.8% recognition rate	CNN features combined with SVM; 10-fold cross-validation for SVM parameter selection
Krishnamoorthy et al. (2021) [9]	InceptionResNetV2 + transfer learning	Rice leaf diseases	95.67% accuracy	Transfer-learning classification study
Lu et al. (2022) [10]	WOACW-optimized SimpleNet	Five diseases	99.35% accuracy; 99.36% F1	Hyperparameter optimization with WOACW
Ahad et al. (2023) [11]	CNN transfer learning + ensemble	Nine rice diseases	98% accuracy	Comparison of six CNN architectures and an ensemble
Dutta et al. (2024) [14]	CNN / CAAR-U-Net / MobileNetV2	Four rice diseases	98% CNN accuracy	Field-image study; k-fold cross-validation reported
Gao et al. (2024) [15]	YOLO V5-EFFICIENT	Rice leaf disease detection	9.90% mAP improvement vs. YOLO V5	Object-detection study with attention, improved anchors, and ablation
Pai et al. (2025) [16]	CNN ensemble	Six rice diseases	96.81% accuracy; F1 = 0.967	Four-model softmax-probability averaging ensemble
Ismail et al. (2026) [18]	DenseNet121 + transfer learning	Seven rice diseases	97.9% accuracy	Independent test set; class-wise metrics reported

Upadhyay et al. (2026) [19]	Adaptive feature fusion	Rice leaf diseases	98.5% reported accuracy	Hybrid feature-fusion architecture
-----------------------------	-------------------------	--------------------	-------------------------	------------------------------------

Table 1. Representative published studies in rice leaf disease image recognition. The reported values are study-specific and are not directly comparable because the studies use different datasets, disease definitions, tasks, and evaluation protocols.

Statistical testing should be matched to the comparison being made. If two models are evaluated on the same test images, paired comparisons may be more informative than treating their accuracies as independent samples. Confidence intervals or bootstrap estimates can also describe uncertainty. The essential requirement is that the statistical procedure follows the structure of the data rather than being added only to create a formal-looking result.

External validation is especially valuable for agricultural image recognition. A model trained on one dataset can be evaluated on images collected under a different set of conditions without retraining. This tests whether the learned representation transfers beyond the original source. External validation does not need to be perfect or universally representative to be useful; its value comes from making the distribution shift explicit and measurable.

Repeated runs and uncertainty estimates can strengthen conclusions when datasets are small. A single random seed can produce a result that is sensitive to initialization or partitioning. Reporting mean and standard deviation across repeated training runs, or using a clearly defined cross-validation procedure where appropriate, provides information about variability. The method should state what is repeated and what remains fixed so that the reported uncertainty is interpretable.

Class imbalance requires explicit reporting because it changes the interpretation of accuracy. If one disease class dominates the dataset, a classifier can achieve a high aggregate score while performing poorly on minority classes. Class weights, resampling, focal losses, or balanced sampling can address some forms of imbalance, but these interventions should be described because they change the effective training distribution.

A leakage-resistant split should be created before model development decisions are finalized. If repeated images of the same leaf appear in training and testing subsets, a network can learn nearly identical visual patterns. If preprocessing statistics are estimated from the complete dataset before splitting, information from the evaluation set can also influence the training process. The safest practice is to isolate the test set first and perform model selection using only the training and validation portions.

Dataset construction is the foundation of experimental validity. A study should define what constitutes one sample, how images were collected, how labels were assigned, and how related images were handled. For rice leaves, the unit of independence may be the image, leaf, plant, plot, field, or acquisition session. Randomly splitting images is not automatically valid when several images originate from the same biological source.

VIII. DATASET DESIGN AND EXPERIMENTAL VALIDITY

8.1 Train, Validation, and Test Partitions

The train/validation/test split is one of the most consequential decisions in an image-classification experiment. A conventional random image-level split may be acceptable when every image is independent. It becomes problematic when several images come from the same plant, plot, field, camera session, or source sequence. Related images can be visually similar enough that the model effectively sees part of the test distribution during training. Plant-level or source-level grouping is therefore preferable when such dependencies exist.

8.2 Data Leakage

Leakage occurs when information that should be unavailable at evaluation time influences model training or model selection. It can occur through duplicated images, near-duplicate images, preprocessing fitted on the full dataset, augmentation performed before splitting, or hyperparameter tuning repeatedly evaluated on the test set. Leakage is especially concerning in agricultural image datasets because images may be collected in batches and may share backgrounds, devices, or capture sessions.

8.3 Class Balance

If one class contains substantially more images than another, overall accuracy can hide poor performance on minority diseases. Researchers should report per-class counts and confusion matrices and should consider macro-averaged precision, recall, and F1-score. Class-weighted loss or balanced sampling can be used during training, but these choices should be documented.

8.4 Repeated Runs and Random Seeds

A single training run can be influenced by random initialization, data order, augmentation, and optimization dynamics. Repeated runs with documented random seeds can provide an estimate of variability. For small datasets, stratified cross-validation can provide additional information, although the final external or held-out test set should remain independent.

8.5 External Validation

External validation is one of the clearest ways to test generalization. An external set should differ meaningfully from the training distribution, ideally in acquisition environment, field location, device, cultivar, season, or other relevant factors. If external data are unavailable, authors should state this limitation directly rather than implying field readiness from a controlled test set.

8.6 Calibration and Uncertainty

A classifier may produce a probability-like score even when its predictions are poorly calibrated. Confidence calibration can therefore be valuable in agricultural decision-support systems. Methods such as temperature scaling, reliability diagrams, expected calibration error, or conformal prediction can provide additional information about uncertainty. A field tool should ideally be able to distinguish between confident predictions and cases that require human inspection.

8.7 Error Analysis

A strong paper should inspect incorrect predictions rather than reporting only aggregate accuracy. Error analysis can identify confusion between visually similar diseases, failures caused by shadows, background artefacts, small lesions, blurred images, or unusual symptom stages. Representative false-positive and false-negative examples can be more informative than an additional decimal place in accuracy.

A reproducibility statement should distinguish between what is directly available and what is restricted. For example, a dataset may be proprietary while preprocessing code and model configuration remain shareable. Explicitly describing these boundaries is more informative than simply stating that data are 'available on request.'

Open materials are useful but should not be treated as a substitute for scientific description. A repository can disappear, access permissions can change, and dependencies can become obsolete. The paper itself should contain the essential information needed to understand the experiment. When code is shared, the authors should also record the software environment, package versions, and configuration files when feasible.

The evaluation pipeline should be documented separately from the training pipeline. The test set must remain untouched during model selection, and any threshold or hyperparameter chosen using validation data should not be tuned on the test set. Reporting the confusion matrix and class-wise metrics helps readers determine whether the aggregate score hides systematic errors.

Model configuration should be reported at a level that permits independent implementation. At minimum, this includes the architecture or backbone, input dimensions, pretrained-weight source when applicable, optimizer, learning rate or schedule, batch size, number of epochs, loss function, augmentation, regularization, early-stopping rule, and model-selection criterion. If a custom layer or fusion mechanism is introduced, its mathematical or architectural description should be sufficiently precise to remove ambiguity.

Reproducibility begins with a complete description of the data pipeline. A reader should be able to identify the dataset, understand the class labels, determine how images were divided, and reconstruct the major preprocessing operations. When a dataset cannot legally be redistributed, the paper can still provide identifying information, access conditions, version details, and a clear description of the collection process.

IX. REPRODUCIBILITY-ORIENTED REPORTING FRAMEWORK

The following framework converts the methodological lessons of the literature into a practical checklist for future experimental studies. The purpose is not to prescribe one architecture but to ensure that readers can understand what was done and, where possible, reproduce the result.

Table 2. Recommended minimum reporting information for reproducible rice leaf disease image studies.

Reporting item	Minimum information	Scientific purpose
Dataset provenance	Dataset name/version; source; collection method; access date; licence; class definitions	Enables source verification
Image acquisition	Camera/device; distance; lighting; field/lab setting; resolution	Defines image distribution
Annotation	Annotator expertise; label rules; disagreement handling	Clarifies label reliability
Partitioning	Exact counts; image/plant/source-level split; leakage controls	Protects test independence
Preprocessing	Resize; normalization; segmentation; augmentation	Defines model inputs
Architecture	Layer-level description; pretrained weights; classifier head; fusion	Enables replication
Training	Optimizer; learning rate; batch size; epochs; scheduler; stopping; seed	Defines optimization
Evaluation	Accuracy; precision; recall; F1; confusion matrix; uncertainty	Shows class-level behaviour
External validation	Independent field/device/source dataset	Tests generalization
Deployment	Parameters; FLOPs; memory; latency; hardware	Tests practical feasibility
Explainability	Grad-CAM/LIME/SHAP/attention; correct and incorrect cases	Checks visual evidence

9.1 Recommended Baseline Set

When proposing a new deep-learning architecture, a useful experimental baseline set should include at least one classical classifier, one conventional CNN, one pretrained transfer-learning model, and, where deployment is claimed, one lightweight model. The exact selection can vary with the research question. What matters is that the baseline models are trained and evaluated under the same data partition and preprocessing conditions.

9.2 Ablation Studies

If a model contains multiple components, an ablation study should remove or replace one component at a time. For example, a feature-fusion architecture can be compared with each backbone individually and with simple concatenation. An ensemble can be compared with its constituent models. An attention module can be evaluated against the same backbone without attention. Such experiments help establish whether the proposed component actually contributes to the reported improvement.

9.3 Open Materials

Where permitted by data licences and institutional rules, researchers should provide code, configuration files, preprocessing scripts, model definitions, random seeds, and trained weights. If raw images cannot be redistributed, the paper can still provide dataset identifiers, checksums or file lists where appropriate, and enough information for an independent researcher to access the same source.

Human-in-the-loop systems provide a realistic deployment model for agricultural diagnosis. Instead of forcing every image into a disease category, a system can report a confidence estimate, show relevant image regions, and route uncertain cases to a human expert. Such a design recognizes that image-based prediction is decision support and that biological diagnosis may require additional observations.

Domain adaptation and robustness testing are relevant because the image distribution can change between development and deployment. A model trained in one region may encounter different cultivars, backgrounds, weather conditions, or camera characteristics elsewhere. Future work can evaluate robustness through cross-domain testing, controlled perturbations, and calibration analysis rather than relying only on in-distribution accuracy.

Semi-supervised and self-supervised learning may help when expert labels are scarce. Large collections of unlabeled agricultural images can provide useful visual representations, after which a smaller labelled set can be used for disease-specific adaptation. Such approaches should still be evaluated with independent test data; reducing annotation cost does not eliminate the need for rigorous validation.

Multi-stage disease recognition is another important direction. Early symptoms can be visually subtle, while later symptoms can be more distinctive. A model trained only on mature lesions may therefore be poorly suited to early warning. Future studies can investigate stage-aware labels, ordinal severity categories, or multi-task learning in which disease identity and severity are predicted jointly.

Future datasets should deliberately sample the variability that a deployed model will encounter. This includes different farms or growing environments, camera devices, illumination conditions, disease stages, cultivars, and symptom severities. Diversity should be documented rather than represented only by a larger image count. Ten thousand near-duplicate photographs do not provide the same evidence as a smaller but genuinely diverse collection.

X. RESEARCH GAPS AND FUTURE DIRECTIONS

10.1 Generalization Across Environments

The most important gap is the difference between curated datasets and field conditions. Future datasets should include multiple acquisition environments and should document the factors that vary between them. Multi-location and multi-season evaluation can reveal whether a model has learned disease characteristics or environment-specific shortcuts.

10.2 Multi-Cultivar and Multi-Stage Data

Disease appearance can vary by cultivar and disease stage. A robust dataset should therefore include meaningful biological variation rather than collecting a large number of visually similar images from one source. Where disease progression is relevant, severity or stage labels can support more detailed modelling.

10.3 Self-Supervised and Semi-Supervised Learning

Large collections of unlabeled agricultural images may be easier to obtain than expert-labelled disease datasets. Self-supervised representation learning and semi-supervised learning could therefore reduce annotation requirements. Such approaches should still be evaluated using independent labelled data and should not be assumed to improve generalization without evidence.

10.4 Domain Adaptation and Robustness

Domain adaptation can be used when the source and target image distributions differ. Potential directions include feature alignment, augmentation strategies designed for domain variation, style-transfer methods, and domain-generalization objectives. Robustness testing should deliberately vary lighting, background, resolution, compression, and camera characteristics.

10.5 Multimodal Decision Support

Images provide visible information but may not capture environmental or agronomic context. Future systems could combine image evidence with weather, soil, location, cultivar, growth stage, and farmer observations. Multimodal systems should nevertheless make the contribution of each information source measurable and should avoid presenting statistical associations as definitive biological diagnoses.

10.6 Edge and Mobile AI

Mobile deployment is a practical direction because disease recognition may be needed where internet connectivity is limited. Model compression, quantization, pruning, knowledge distillation, and hardware-aware architecture search can reduce computational requirements. Evaluation should report actual device latency rather than relying only on desktop GPU inference.

10.7 Explainability and Human-in-the-Loop Systems

Future tools should be designed as decision support rather than automatic replacement of agricultural expertise. A system can present a predicted class, confidence or uncertainty, a visual explanation, and a recommendation to seek expert assessment when the input is outside the model's reliable domain. Human-in-the-loop evaluation can test whether explanations improve appropriate use rather than merely making a system visually attractive.

Finally, reproducibility should be treated as a design principle rather than a final editing step. If data provenance, partitioning, seeds, preprocessing, and evaluation criteria are decided and documented during study design, the resulting paper is easier to inspect and reproduce. This also reduces the risk that an apparently strong result depends on undocumented choices made after the experiment was run.

The reviewed methods also suggest that a practical rice disease system should be evaluated at multiple levels. At the image level, classification metrics measure label prediction. At the object or lesion level, detection and segmentation metrics measure localization. At the deployment level, latency, memory, energy use, and user interaction become relevant. A system intended for farmers or field technicians therefore requires a broader evaluation framework than a laboratory classification benchmark.

This perspective is particularly relevant when published studies report very high accuracy. High accuracy can be entirely plausible for a well-controlled classification task, especially when classes are visually separable. The scientific question is not whether a high number is possible; it is what the number demonstrates. The answer depends on dataset independence, class balance, source diversity, and whether the test set represents the intended deployment environment. The central finding of the reviewed literature is that algorithmic progress and evidentiary progress are not identical. Newer architectures can improve representation learning and computational efficiency, but the credibility of a reported result still depends on the relationship between the dataset, training procedure, and evaluation design. A sophisticated network evaluated on a weak split does not provide stronger evidence merely because the architecture is newer.

XI. DISCUSSION

The literature demonstrates substantial progress in rice leaf disease recognition. Classical image processing established the importance of colour, texture, lesion geometry, and segmentation. CNNs subsequently reduced dependence on handcrafted descriptors, while transfer learning made sophisticated models more accessible for moderate-sized agricultural datasets. Lightweight architectures address deployment constraints, ensembles seek complementary errors, detection models address complex scenes, and XAI methods provide tools for inspecting model behaviour [7–20].

At the same time, the literature reveals that architecture alone does not determine the scientific strength of a study. Dataset provenance, partitioning, and evaluation environment can influence results as strongly as the choice between two closely related CNNs. A model can exploit background, image border, acquisition device, or other non-biological cues when these are correlated with the class label. High accuracy can therefore coexist with weak external generalization.

This observation changes how comparative results should be interpreted. Rather than asking which published architecture has the largest accuracy value, researchers should ask whether the compared studies solve the same task under comparable conditions. Disease count, class balance, image source, test independence, field complexity, and evaluation metrics must be considered together. The literature does not support a universal ordering of model families across all rice disease settings.

The same principle applies to deployment claims. A model intended for a smartphone should be evaluated on the relevant hardware. A detector intended for field photographs should be tested on full scenes rather than only cropped leaves. A disease-severity system should report segmentation or regression metrics rather than only disease-class accuracy. A system intended for agronomic decision support should also characterize uncertainty and failure modes.

A second important issue is reproducibility. A paper can be conceptually strong but difficult to reproduce if the dataset

version, exact split, preprocessing, pretrained weights, hyperparameters, random seed, and model architecture are not reported. Reproducibility is particularly important in agricultural AI because datasets are often assembled from heterogeneous sources and may not be permanently available. Clear provenance and transparent experimental configuration therefore have lasting scientific value.

The field is also moving toward more realistic evaluation. Studies that include independent test data, field images, or multiple environmental conditions provide evidence that is complementary to controlled benchmarks. Such studies should not automatically be interpreted as universally superior either; rather, they answer a different and often more deployment-oriented question. The appropriate evaluation depends on the intended use case.

Taken together, the literature suggests that the next stage of rice disease recognition should emphasize evidence quality as much as architectural novelty. Better datasets, stronger experimental controls, independent validation, transparent reporting, and deployment-aware metrics can make incremental model improvements more meaningful and can help distinguish genuine generalization from dataset-specific optimization.

11.1 Practical Implications

- For researchers: document data and training procedures sufficiently for independent replication.
- For dataset developers: provide provenance, class definitions, collection conditions, and licence information.
- For model developers: evaluate accuracy together with class-wise metrics, uncertainty, computational cost, and robustness.
- For agricultural deployment: test on realistic field images and provide a mechanism for uncertain cases to be referred for human assessment.
- For reviewers and editors: interpret numerical results in the context of the dataset and test protocol rather than treating isolated accuracy values as interchangeable.

Visual examples are used only to provide contextual understanding of symptom appearance; quantitative conclusions are based on published study reports and their stated evaluation protocols.

The synthesis is also constrained by the reporting quality of the original studies. When a primary paper does not disclose dataset provenance, partitioning, or training details, the review cannot reconstruct those facts reliably. Such omissions are described as limitations of the evidence rather than silently supplemented with assumptions.

Because this is a narrative review, the literature selection does not constitute a formal systematic-search protocol or meta-analysis. The representative studies are used to explain methodological patterns rather than to estimate a pooled performance effect. Consequently, the review should not be interpreted as an exhaustive census of every rice disease image paper published in the field.

XII. LIMITATIONS OF THIS REVIEW

This article is a narrative review rather than a systematic meta-analysis. The selection of representative studies is therefore necessarily dependent on the defined search concepts and the availability of accessible scholarly records. Because rice disease studies use heterogeneous datasets and evaluation procedures, a pooled numerical estimate of model performance would not be scientifically appropriate for the present objective.

The review also depends on the methodological details reported by the original studies. When a paper does not provide exact partitioning, architecture, training configuration, or external validation information, the review cannot reconstruct those details reliably. Missing information is therefore treated as a reporting limitation rather than filled by assumption. Finally, the review focuses on image-based rice leaf disease recognition. It does not provide a complete treatment of molecular diagnosis, hyperspectral sensing, multispectral satellite imagery, pathogen genomics, or non-image agronomic diagnosis. These areas are relevant but outside the defined scope.

XIII. CONCLUSION

Image-based rice leaf disease recognition has developed from handcrafted feature and classical machine-learning pipelines toward CNNs, transfer learning, lightweight architectures, ensembles, hybrid feature-fusion systems, object detectors, segmentation methods, and explainable AI. The literature demonstrates that high predictive performance is achievable under a variety of study-specific conditions. However, reported accuracy values cannot be interpreted

independently of the dataset, disease classes, preprocessing, partitioning, and evaluation environment.

The most important methodological priorities for future studies are transparent dataset provenance, leakage-resistant partitioning, complete architecture and training descriptions, class-wise evaluation, independent testing, repeated experiments where appropriate, and deployment-aware measurements. External validation across fields, devices, cultivars, seasons, and disease stages can provide stronger evidence of generalization than a single controlled test set. XAI and uncertainty analysis can further help identify failure modes and support responsible use.

The review therefore supports a shift in emphasis from isolated accuracy values toward complete evidence chains: data provenance → experimental design → model architecture → training protocol → evaluation → error analysis → external validation → deployment. This approach can make rice disease recognition research more reproducible and can help translate computer-vision advances into practical agricultural decision-support systems.

Author Contribution and Data Availability Statements

Author contribution: The authors contributed to the interpretation and presentation of the reviewed literature and to the preparation of the narrative synthesis.

Data availability: No new experimental dataset is introduced in this review. The evidence discussed is derived from the published literature identified by the cited references.

REFERENCES

- [1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [2] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520. <https://doi.org/10.1109/CVPR.2018.00474>
- [3] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, 6105–6114.
- [4] Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145, 311–318. <https://doi.org/10.1016/j.compag.2018.01.009>
- [5] Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture*, 161, 272–279. <https://doi.org/10.1016/j.compag.2018.03.032>
- [6] Barbedo, J. G. A. (2018). Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Computers and Electronics in Agriculture*, 153, 46–53. <https://doi.org/10.1016/j.compag.2018.08.013>
- [7] Jiang, F., Lu, Y., Chen, Y., Cai, D., & Li, G. (2020). Image recognition of four rice leaf diseases based on deep learning and support vector machine. *Computers and Electronics in Agriculture*, 179, 105824. <https://doi.org/10.1016/j.compag.2020.105824>
- [8] Rahman, C. R., Arko, P. S., Ali, M. E., et al. (2020). Identification and recognition of rice diseases and pests using convolutional neural networks. *Biosystems Engineering*, 194, 112–120. <https://doi.org/10.1016/j.biosystemseng.2020.03.020>
- [9] Krishnamoorthy, N., Narasimha Prasad, L. V., Pavan Kumar, C. S., Subedi, B., Abraha, H. B., & Sathishkumar, V. E. (2021). Rice leaf diseases prediction using deep neural networks with transfer learning. *Environmental Research*, 198, 111275. <https://doi.org/10.1016/j.envres.2021.111275>
- [10] Lu, Y., Zhang, X., Zeng, N., Liu, W., & Shang, R. (2022). Image classification and identification for rice leaf diseases based on improved WOACW_SimpleNet. *Frontiers in Plant Science*, 13, 1008819. <https://doi.org/10.3389/fpls.2022.1008819>
- [11] Ahad, M. T., Li, Y., Song, B., & Bhuiyan, T. (2023). Comparison of CNN-based deep learning architectures for rice diseases classification. *Artificial Intelligence in Agriculture*, 9, 22–35. <https://doi.org/10.1016/j.aiia.2023.07.001>

- [12] Mukherjee, R., Ghosh, A., Chakraborty, C., De, J. N., & Mishra, D. P. (2025). Rice leaf disease identification and classification using machine learning techniques: A comprehensive review. *Engineering Applications of Artificial Intelligence*, 139, 109639. <https://doi.org/10.1016/j.engappai.2024.109639>
- [13] Simhadri, C. G., Kondaveeti, H. K., Vatsavayi, V. K., Mitra, A., & Ananthachari, P. (2025). Deep learning for rice leaf disease detection: A systematic literature review on emerging trends, methodologies and techniques. *Information Processing in Agriculture*, 12(2), 151–168. <https://doi.org/10.1016/j.inpa.2024.04.006>
- [14] Dutta, M., Islam Sujan, M. R., Mojumdar, M. U., Chakraborty, N. R., Al Marouf, A., Rokne, J. G., & Alhaji, R. (2024). Rice leaf disease classification—A comparative approach using CNN, CAAR-U-Net, and MobileNet-V2 architectures. *Technologies*, 12(11), 214. <https://doi.org/10.3390/technologies12110214>
- [15] Gao, W., Zong, C., Wang, M., Zhang, H., & Fang, Y. (2024). Intelligent identification of rice leaf disease based on YOLO V5-EFFICIENT. *Crop Protection*, 183, 106758. <https://doi.org/10.1016/j.cropro.2024.106758>
- [16] Pai, P., Amutha, S., Patil, S., et al. (2025). Deep learning-based automatic diagnosis of rice leaf diseases using ensemble CNN models. *Scientific Reports*, 15, 27690. <https://doi.org/10.1038/s41598-025-13079-z>
- [17] Kondaveeti, H. K., & Simhadri, C. G. (2025). Evaluation of deep learning models using explainable AI with qualitative and quantitative analysis for rice leaf disease detection. *Scientific Reports*, 15, 31850. <https://doi.org/10.1038/s41598-025-14306-3>
- [18] Ismail, A., Hamdy, W., Ibrahim, A. H., & Awad, W. A. (2026). Classification of rice plant diseases using efficient DenseNet121. *Scientific Reports*, 16, 7482. <https://doi.org/10.1038/s41598-026-38078-6>
- [19] Upadhyay, S. K., Prasad, R., Vikas, & Vats, P. (2026). A hybrid deep learning model with adaptive feature fusion for automated rice leaf disease detection and classification. *Scientific Reports*, 16, 20345. <https://doi.org/10.1038/s41598-026-52534-3>
- [20] Hossen, M. K., Das, P. K., & Roy, R. (2025). Detection and classification of rice leaf diseases using OpenCV and deep learning. *Current Applied Science and Technology*, 25(5), e0260191. <https://doi.org/10.55003/cast.2025.260191>

Acknowledgment

The authors acknowledge the published literature that forms the basis of this narrative synthesis.

Declaration of Competing Interests

The authors declare no competing financial or commercial interests related to this review.

Ethical Statement

No human or animal subjects, participant-level data, field intervention, or laboratory experiment was conducted as part of this literature review.

Data Availability

No new experimental dataset is introduced in this review. The evidence discussed is derived from the cited published literature.