INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 10 ISSUE: 11 | NOV - 2026 SJIF RATING: 8.659 ISSN: 2582-3930

A Comparative Review of Deep Learning-Based CNN Architectures for Crop Leaf Disease Detection: Datasets, Performance and Real-World Challenges

Pranav Patil[1], Sarthak Burande[1], Sakshi Madake[1]
Department of Information Technology, Nutan Maharashtra Institute of Engineering & Technology, Pune, Maharashtra, India
Project Guide: **Ms. Bhagyashree Phulmali**
Corresponding author: popthefact@gmail.com

Abstract— Crop diseases reduce agricultural productivity and can threaten food quality and farm income. Image-based diagnosis using convolutional neural networks (CNNs) has become an active research direction because CNNs learn visual features from labelled leaf images. This review examines CNN-based crop leaf disease classification with emphasis on dataset characteristics, architecture selection, transfer learning, preprocessing, evaluation design and deployment constraints. It discusses conventional CNNs, VGG, residual networks, MobileNet, EfficientNet and DenseNet, notes transformer-based and hybrid alternatives, and considers the different roles of controlled datasets such as PlantVillage and field-oriented collections such as PlantDoc. Published studies report high accuracy on curated benchmarks, but results obtained under different class definitions, image sources and split protocols are not directly comparable. Evidence that a model performs well on a controlled test set is not, by itself, evidence of reliable farm-level diagnosis. The review therefore analyses dataset shift, class imbalance, duplicate-image leakage, label quality, explainability, uncertainty estimation and mobile inference. It proposes a practical evidence hierarchy progressing from internal random splits to grouped evaluation, cross-dataset testing and prospective field assessment. The central conclusion is that model selection should balance predictive performance with generalisation, computational cost and the ability to recognise uncertain or unsupported inputs.

Index Terms— *Crop disease classification, convolutional neural networks, PlantVillage, PlantDoc, transfer learning, field generalisation*

I. INTRODUCTION

Agriculture is closely connected to food availability, rural livelihoods and economic development. Plant diseases can affect leaves, stems, fruits and roots, reducing both the quantity and quality of agricultural production. Leaf symptoms are often among the first visible indicators of disease, which makes leaf-image analysis a practical entry point for computer-assisted crop-health monitoring. In practice, disease recognition may be performed by farmers, extension workers or plant pathologists.

The reliability of visual assessment depends on the observer's experience, disease stage, crop variety and environmental conditions. Symptoms such as chlorosis, necrotic spots, mildew-like growth and leaf deformation can overlap across diseases and can also arise from pests, nutrient deficiencies, drought or chemical injury [24]. Earlier image-analysis systems often relied on manually designed colour, texture and shape descriptors followed by a conventional classifier. Deep learning changes this workflow by learning task-relevant representations directly from labelled examples.

Convolutional neural networks (CNNs) apply learned filters to local image regions and combine low-level patterns into increasingly abstract representations. This capability has made CNNs widely used for image-based plant disease classification. Mohanty, Hughes and Salathé demonstrated the feasibility of CNN-based plant disease recognition using 54,306 images from 14 crop species and 38 crop-disease classes [1], [7]. Ferentinos subsequently evaluated CNN models using 87,848 images across 58 plant-disease combinations [2].

These studies helped establish the potential of deep learning in this application area; broader surveys of deep learning in agriculture and plant-disease detection are given in [25], [26]. However, strong performance on a curated dataset does not necessarily translate into reliable performance on photographs collected in farms. Field images may contain soil, weeds, hands, overlapping leaves, shadows, motion blur, variable illumination and more than one visible symptom.

A model trained primarily on centred leaves against uniform backgrounds may exploit background or acquisition cues that correlate with labels rather than disease-specific visual evidence. This creates a gap between benchmark classification and practical decision support. This review focuses on CNN-based image classification of crop leaf health and disease. It compares architecture families and datasets, examines training and evaluation practices, and identifies factors that affect real-world reliability.

Its contribution is a critical synthesis connecting three decisions that are often discussed separately: the architecture selected, the dataset used to train and evaluate it, and the evidence required before deployment claims are justified. The review does not claim that one architecture is universally best, nor does it present itself as a completed systematic review. In practical terms, the review treats dataset realism and validation strength as conditions on architecture claims: a model is not considered preferable merely because it reports a higher score on a different benchmark. The comparison therefore distinguishes architectural capability, the difficulty and representativeness of evaluation data, and evidence of suitability for the intended operating environment.

II. REVIEW METHODOLOGY AND SCOPE

This article is a focused narrative review. It draws on foundational architecture papers, primary plant-disease classification studies, dataset publications and recent review articles, and it covers image-based classification of crop leaf health or disease using CNN-based methods, with a short discussion of transformer-based and hybrid alternatives. A source was included when it described a dataset, architecture, training method, evaluation procedure, deployment issue or research synthesis relevant to this task. Primary experimental studies are used for numerical performance claims,

architecture papers for network design, and dataset papers for image collection and annotation.

The selection is illustrative rather than exhaustive. The review is not a registered systematic review, and no database-specific search date, screening count or exclusion log is reported. Each representative study is read for its task, data source, model family, reported result, evaluation setting and any limitation relevant to generalisation. The synthesis is qualitative: reported scores are not pooled or recalculated and are not treated as a head-to-head ranking, because the number of classes, class balance, image source, split design, preprocessing, pretrained weights and test independence differ between studies. A future systematic update should record the search date, queries, databases, year range, screening decisions and reasons for exclusion. For the purpose of synthesis, the sources were examined against a common set of questions: (i) what crop, disease classes and image sources were represented; (ii) whether images were captured under controlled or field conditions; (iii) which architecture and transfer-learning strategy were used; (iv) how the data were partitioned and which metrics were reported; and (v) whether external generalisation or deployment constraints were assessed. This common reading framework supports a structured qualitative comparison while avoiding the false precision of pooling results from incompatible experiments. Because source selection was narrative and illustrative, the review should not be interpreted as an exhaustive census of every model or dataset.

Review questions

The review is organised around three questions: (1) what design and computational trade-offs distinguish the principal CNN architecture families used for crop-leaf disease classification; (2) how do dataset characteristics and evaluation protocols affect the interpretation of reported performance; and (3) what evidence is needed before a model can reasonably be considered suitable for field or mobile use? These questions define the comparison presented in the following sections and keep the discussion focused on the relationship between architecture, data and validation.

III. PUBLIC DATASETS FOR CROP LEAF DISEASE RESEARCH

Dataset design strongly influences what a model learns and how its performance should be interpreted. Plant disease datasets vary in crop coverage, disease categories, image quality, annotation type, acquisition environment and licensing. A dataset containing many images captured against uniform backgrounds may support high in-distribution accuracy while providing limited evidence of robustness in farms. Dataset size alone is therefore an incomplete measure of dataset quality [11]. PlantVillage is one of the most frequently used public resources.

The original study described 54,306 images representing 14 crop species and 38 crop-disease classes [1], [7]. Images are generally centred leaf photographs with relatively controlled backgrounds. This collection has enabled reproducible benchmarking and transfer-learning experiments. Its controlled acquisition conditions simplify classification and make it useful for initial model comparisons and ablation studies. However, results should be described as performance on PlantVillage rather than as general agricultural diagnostic accuracy.

Background, framing and repeated views of the same physical leaf can introduce shortcuts or leakage if the split is not carefully designed. PlantDoc was introduced to address the shortage of non-laboratory plant disease images. Singh et al. described a dataset containing approximately 2,598 images across 13 plant species, with up to 17 disease classes, collected from internet sources and annotated for visual plant disease research [3]. Compared with clean laboratory-style images, PlantDoc introduces clutter, occlusion, scale variation and illumination differences.

Such images may better resemble photographs encountered outside the laboratory, but they can also contain label ambiguity, inconsistent image quality and uneven class representation. Field-oriented does not automatically mean error-free or representative of every farm environment. Other resources include crop-specific collections, cassava leaf disease datasets, AI Challenger plant disease images and research groups' own field datasets. Crop-specific collections can provide detailed coverage for a target crop and may be more relevant to a local decision-support use case than a broad but controlled benchmark.

Their findings should nevertheless be generalised only to the crop, geography, disease classes and acquisition conditions represented. Custom datasets are harder to compare unless collection protocols, class definitions, annotation quality and split procedures are reported clearly. Dataset shift can occur at several levels. Covariate shift refers to changes in image appearance, such as lighting, camera, background or viewpoint. Label shift refers to changes in the prevalence of disease classes. Concept shift occurs when the visual relationship between an image and its label changes, for example because symptoms manifest differently across cultivars or growth stages. Cross-dataset testing can reveal some of these problems, but class taxonomies and label definitions must be aligned before metrics are calculated. A reliable dataset description should report image counts by class, crop species, source, collection setting, annotation procedure, licence and known limitations.

A. Dataset Curation, Label Quality and Annotation

Dataset construction is not simply an image-collection exercise. The label attached to each image defines the prediction task, and inconsistent labels place an upper limit on the reliability of any trained classifier. A dataset should therefore distinguish between a visually suspected disease, a diagnosis confirmed by an expert, and a label copied from an online caption or filename. These sources of evidence are not equivalent. When expert confirmation is unavailable, the annotation protocol should state that labels represent visible symptom categories rather than verified causal diagnoses. This distinction is especially important where nutrient deficiencies, pest damage and infectious diseases produce overlapping visual symptoms [24].

An annotation guide should define each class, list inclusion and exclusion criteria, and provide examples of ambiguous cases. Annotators should be able to select an uncertain or unusable label instead of being forced to assign every image to a disease class. For a subset of images, independent annotation by two or more reviewers can reveal disagreement. The study can report the number of disagreements and how they were resolved, for example through adjudication by a plant pathologist. If annotation confidence is recorded, it can later support sensitivity analyses that compare model performance on high-confidence and uncertain labels.

Image-level metadata are also important. Useful fields include crop species and cultivar where known, plant growth stage, disease stage, capture device, approximate acquisition setting, date or season, and whether the image was collected in a laboratory, greenhouse or open field. Metadata should be collected consistently and shared at a level that protects personal and farm information. Missing metadata should be explicitly marked rather than silently inferred. These fields make it possible to investigate whether errors cluster around particular devices, varieties or acquisition conditions.

The unit of annotation and the unit of data splitting should be considered together. A dataset may contain several images of one leaf, multiple leaves from one plant, or photographs from the same farm visit. If related images are distributed across training and test partitions, a model may benefit from repeated visual patterns without demonstrating generalisation to a new plant. A robust data record should therefore retain group identifiers, even if those identifiers are anonymised before release. Dataset documentation should also identify known duplicates, class imbalance, licensing restrictions and any transformations applied before publication.

TABLE I. SUMMARY OF PUBLIC AND REPRESENTATIVE CROP-LEAF DATASETS

Dataset	Scale and scope	Acquisition	Main caveat
PlantVillage [1], [7]	54,306 images; 14 crop species; 38 classes	Centred leaf photographs, controlled backgrounds	Background shortcuts; repeated views can leak across splits
PlantDoc [3]	Approx. 2,598 images; 13 species; up to 17 classes	Web-sourced images with clutter and varied illumination	Label ambiguity, uneven class representation, small size
Ferentinos [2]	87,848 images; 58 plant-disease combinations	Mixture of laboratory-style and field images	Results specific to this collection and protocol
Crop-specific collections	Varies by study	Often field-captured for a single crop	Conclusions hold only for the crop, geography and classes represented

Duplicate detection is important because multiple photographs of the same leaf or near-identical images can appear in different partitions. Where metadata permit, splitting should occur by leaf, plant, farm, acquisition session or geographic location rather than by individual image alone. Training, validation and test partitions should be independent at the grouping level relevant to the intended claim.

IV. CNN ARCHITECTURES USED IN LEAF DISEASE CLASSIFICATION

CNN architectures differ in depth, feature reuse, parameter count, computational cost and suitability for mobile inference. The best choice depends on the available data, hardware, response-time requirements and consequences of an incorrect prediction. Accuracy values from separate studies should not be ranked directly unless the datasets, class definitions, split protocols and evaluation procedures are sufficiently comparable. A conventional CNN typically consists of convolutional layers, nonlinear activation functions, pooling operations and one or more classification layers. Convolutional filters learn local patterns, while deeper layers combine them into higher-level representations. A task-specific network offers control over model size and design. A small custom model may underfit complex visual variation, whereas a model with excessive capacity may overfit when the training dataset is limited or insufficiently diverse. Regularisation, augmentation and validation design are therefore as important as layer count. VGG-style networks use repeated small convolution filters and a relatively simple sequential structure [14]. Their regular design makes them useful reference architectures and educational baselines. However, some VGG variants contain large fully connected layers and can be computationally and memory intensive.

Residual Networks (ResNet) introduced identity shortcut connections that allow a block to learn a residual mapping in addition to its input [4]. These connections support optimisation in deeper networks by providing paths through which information and gradients can propagate. ResNet-18 is a comparatively compact member of the family and is often used with ImageNet-pretrained weights for transfer learning. ResNet-50 is deeper and has greater representational capacity, but generally requires more computation. A deeper model is not automatically better for every crop dataset: performance depends on sample diversity, label quality, training strategy and test conditions.

MobileNet architectures were designed for efficient inference on mobile and embedded devices. A key design principle is depthwise separable convolution, which separates spatial filtering from channel-wise combination and reduces computation compared with standard convolution [5]. MobileNet variants are attractive for smartphone-based screening, offline use and low-power edge devices. Their practical value depends on whether acceptable class-wise performance is maintained after compression, quantisation or deployment on the target hardware. MobileNetV3 further explores architecture search and efficiency-oriented design [12].

EfficientNet uses compound scaling to balance network depth, width and input resolution [6]. Rather than increasing only one dimension, the approach scales these dimensions in a coordinated manner. EfficientNet variants can provide useful accuracy–efficiency trade-offs, but larger variants may be unsuitable for low-resource devices. DenseNet connects each layer to later layers within a dense block, encouraging feature reuse and providing short paths for information and gradients [13]. DenseNet has been used as a transfer-learning backbone in plant-disease classification. Too et al. reported 99.75% test accuracy for DenseNet in their PlantVillage comparison [10]. This figure should be interpreted within that study's dataset and split rather than as evidence that DenseNet will always outperform ResNet or MobileNet.

Transfer learning reuses features learned from a large source dataset, commonly ImageNet, and adapts the network to a plant-disease task. A typical workflow replaces the final classification layer, initially freezes some or all of the feature extractor, and then fine-tunes selected layers using labelled plant images. This can reduce training time and data requirements compared with training a deep network from random initialisation.

Beyond CNNs. Vision Transformers (ViT) divide an image into patches and model relationships among them with self-attention [21]; they generally benefit from large-scale pretraining. Hierarchical variants such as the Swin

Transformer compute attention within shifted local windows to reduce cost [22], while ConvNeXt modernises a standard ResNet using design choices inspired by transformers and remains a purely convolutional model [23]. These families are relevant to plant-disease work as alternatives or successors to the CNNs reviewed here, but they should be compared under the same grouped-split, cross-dataset and device-level protocols described in this paper before any claim of superiority is made.

Architecture selection is a multi-objective decision. A workstation experiment may favour a larger backbone when benchmark performance is the primary objective. A smartphone application may favour a compact network with low memory use and acceptable latency. A system intended to support a rare but damaging disease may prioritise minority-class recall and calibrated uncertainty over overall accuracy. Comparative studies should define the target use case and report predictive quality together with computational cost wherever possible.

A. Transfer Learning, Pretraining and Hybrid Design Choices

Transfer learning is often a practical starting point because agricultural datasets are usually smaller and more specialised than the large image collections used to pretrain general-purpose vision models. Early convolutional layers can provide reusable edge, colour and texture representations, while later layers are adapted to crop-specific symptoms. However, the usefulness of transferred features depends on the similarity between the source and target image distributions. A model pretrained on everyday objects may not represent subtle disease patterns equally well across all crops, so pretraining should be treated as an advantage to test rather than a guarantee of improved performance.

A controlled fine-tuning comparison can separate the contribution of the pretrained representation from the contribution of task-specific learning. One baseline can freeze the feature extractor and train only the classification head; a second can unfreeze the final blocks; and a third can fine-tune the complete network with a smaller learning rate. The comparison should use the same data partitions and validation procedure. If full fine-tuning improves the internal test score but reduces external performance, the result may indicate adaptation to dataset-specific cues rather than more transferable disease features.

Hybrid approaches combine convolutional feature extraction with attention mechanisms, transformer blocks or other modules intended to model broader spatial relationships. Such combinations may help when symptoms are distributed across a large leaf region or when context among multiple visual regions is informative. They also add parameters, design choices and training requirements. A hybrid model should therefore be justified by a clearly stated limitation of the baseline and evaluated with an ablation that isolates the added component. Merely adding attention does not establish that the model is more explainable or more robust.

The comparison between CNNs and vision transformers is also affected by data scale and pretraining. Transformers can model long-range relationships through self-attention, but their performance is sensitive to architecture, image resolution and pretraining strategy [21], [22]. ConvNeXt demonstrates that modernised convolutional designs can remain competitive while retaining convolutional operations [23]. For crop-disease studies, these alternatives should be assessed under identical image partitions, label definitions and training budgets. Results from unrelated datasets should be described as complementary evidence, not a direct architecture ranking.

TABLE II. QUALITATIVE COMPARISON OF CNN ARCHITECTURE FAMILIES

Architecture family	Design characteristic and practical consideration
VGG	Sequential stacks of small convolutions. Simple baseline; some variants are memory-intensive.
ResNet	Identity shortcut connections. Supports optimisation of deeper networks and broad transfer-learning use.
MobileNet	Depthwise separable convolutions. Low-computation starting point for edge deployment; verify class-wise performance on target hardware.
EfficientNet	Coordinated depth, width and resolution scaling. Useful accuracy–efficiency trade-off; larger variants may exceed device budgets.
DenseNet	Dense feature connections and feature reuse. Encourages feature reuse; memory and inference cost depend on the chosen variant.
Custom CNN	Task-specific layer count and arrangement. Can be tailored to dataset size and hardware, but requires careful overfitting checks.

The table should be read as a model-selection aid rather than a league table. A defensible comparison trains candidate architectures with the same data partitions, augmentation policy, optimisation budget and stopping rule. Where pretrained weights differ in provenance or input normalisation, those differences should be documented because they can influence the outcome.

V. IMAGE PREPROCESSING AND TRAINING PRACTICES

CNNs generally require a fixed input tensor size. Images are resized or cropped to a resolution compatible with the chosen architecture, such as 224 × 224 pixels for many ImageNet-pretrained models. Resizing must preserve diagnostically useful details; excessive downsampling may remove small lesions. When pretrained weights are used, the corresponding channel normalisation should be applied consistently during training and inference. The preprocessing pipeline should be reported precisely enough to reproduce the experiment.

Data augmentation creates modified training examples through transformations such as horizontal or vertical flipping, rotation, random cropping, zooming, brightness adjustment and contrast adjustment. It can reduce overfitting and expose the model to some forms of image variation. Transformations must remain biologically plausible. Strong colour changes may alter symptoms that depend on colour, and aggressive cropping may remove the diseased region. Augmentation should be applied only to training data; validation and test images should receive deterministic preprocessing.

Common training choices include cross-entropy loss for multi-class classification and optimisers such as stochastic gradient descent or Adam/AdamW. Learning rate, batch size, weight decay, number of epochs and scheduler settings influence convergence. Early stopping and checkpoint selection should use a validation set, not the held-out test set. The test set should be evaluated only after model selection is complete. Random seeds, software versions and pretrained-weight sources should be recorded to improve reproducibility. Some disease classes may contain substantially fewer images than others.

A. Reproducible Training and Experimental Controls

A comparative experiment should begin by defining the prediction task before selecting an architecture. The authors should specify whether the model predicts one disease class per image, identifies several co-occurring conditions, distinguishes healthy from unhealthy leaves, or estimates disease severity. These tasks require different output layers, labels and evaluation measures. A single-label classifier is appropriate only when each image is expected to represent one mutually exclusive class. Where mixed symptoms are common, a multi-label formulation may be more faithful to the data, although it requires suitable annotations.

All candidate architectures should use the same partition of the dataset. If one model receives a more favourable split, higher input resolution or more extensive augmentation, the resulting comparison cannot isolate the effect of architecture. Training budgets should also be documented. These include the number of epochs, batch size, optimiser, learning-rate schedule, weight decay, early-stopping rule and whether the backbone is frozen or fine-tuned. When pretrained models are used, the source and version of the weights and the associated input-normalisation procedure should be stated.

Small agricultural datasets can produce unstable results because a few images may strongly affect class-wise metrics. Repeating training with several random seeds provides an indication of this variability. A paper should report the mean and standard deviation, or another clearly defined uncertainty summary, rather than selecting the single best run without explanation. Where the test set is small, confidence intervals for key metrics can help communicate sampling uncertainty. Such intervals do not correct dataset bias, but they make the limits of the estimate more visible.

Ablation experiments can help determine which parts of a pipeline contribute to performance. For example, a study may compare a pretrained backbone with and without fine-tuning, evaluate augmentation choices, or test the effect of class weighting. Each ablation should change one factor at a time where practical, while keeping the remaining settings fixed. The purpose is not to maximise the number of experiments but to support interpretable conclusions about why a system performs as reported.

Overall accuracy can conceal poor performance on minority classes. Possible responses include class-weighted loss, balanced sampling and carefully controlled augmentation. These techniques should be selected based on the training distribution and reported transparently. Methods such as Grad-CAM produce class-specific heat maps indicating image regions that influence a CNN prediction [15]. These visualisations can support error analysis and help identify suspicious reliance on backgrounds or irrelevant objects. A heat map does not prove that the model has learned a biologically valid disease feature.

VI. EVALUATION METRICS AND FAIR COMPARISON

Accuracy is the proportion of correctly classified examples, but it can be misleading when class frequencies are unequal. Precision measures the proportion of predicted positives that are correct, while recall measures the proportion of actual positives that are recovered. The F1-score is the harmonic mean of precision and recall. For multiclass tasks, macro-averaging gives each class equal weight, whereas weighted averaging weights classes according to their support. A confusion matrix reveals which disease categories are confused with one another.

A. Interpreting Metrics Under Class Imbalance and Unequal Error Costs

Metric selection should reflect the consequences of different errors. In a balanced benchmark, accuracy may provide a compact summary, but in an imbalanced dataset it can be dominated by common classes. Macro-F1 and macro-averaged recall give minority classes greater influence in the summary. Per-class precision and recall remain necessary because a single aggregate score can hide a disease category that is rarely detected. Reporting support counts beside each class metric helps readers distinguish a weak model from an estimate based on very few test examples.

False positives and false negatives may also have different practical consequences. A false positive can lead to unnecessary inspection or treatment, while a false negative may delay intervention. The relative cost depends on the crop, disease, local management practice and the action recommended by the system. A model intended for preliminary screening may reasonably favour sensitivity and refer uncertain cases for expert review. A system that recommends a costly intervention requires stronger evidence and careful control of false alarms. These operational choices should be stated rather than hidden inside a single threshold.

Confidence scores require particular care. A softmax output is a model score, not automatically a calibrated probability that the predicted disease is present. Calibration should be evaluated on data that are separate from model training, and any temperature scaling or threshold selection should be fitted without using the final test labels. Reliability diagrams and expected calibration error can reveal overconfidence, while selective-risk analysis can show how error changes when the system abstains on low-confidence cases [27], [28]. Calibration measured on one dataset may deteriorate under domain shift, so it should also be checked on external data where possible.

For a transparent report, the results section should include the confusion matrix, per-class support, macro and weighted metrics, and the chosen confidence or abstention policy. When computational suitability is claimed, latency should be measured on a named device with the input size, software runtime and batch size specified. Parameter count alone is insufficient because actual latency depends on hardware acceleration, memory transfer and implementation. A compact model that is theoretically efficient may still perform poorly on an older phone if the runtime does not support its operations.

For deployment-oriented studies, additional measures are valuable: parameter count, multiply-accumulate operations (MACs) or floating-point operations (FLOPs), model file size, peak memory and inference latency on specified hardware. Calibration measures indicate whether predicted confidence

corresponds to observed correctness. Where a model may abstain from uncertain predictions, coverage and selective risk should also be reported. In an agricultural setting, abstention may be preferable to an unsupported diagnosis. Calibration is particularly relevant when a farmer-facing interface displays confidence scores: a model can be accurate yet systematically overconfident. Expected calibration error, reliability diagrams and proper scoring rules can help assess whether confidence estimates correspond to observed correctness [27]. Selective classification provides a complementary option by allowing the system to abstain on uncertain inputs; its behaviour should be assessed through coverage and risk rather than accuracy alone [28]. Transferability also depends on the representation learned during pretraining, so the source of pretrained weights and fine-tuning protocol should be treated as experimental variables rather than implementation details [29].

A fair comparison should report the dataset release and version, class mapping, train/validation/test split, grouping strategy, image resolution, augmentation, pretrained weights, fine-tuning policy, hyperparameters, hardware, number of runs and uncertainty across runs. If two papers use different datasets or split protocols, their headline accuracy values should not be interpreted as a controlled head-to-head comparison. Runtime comparisons are meaningful only when hardware and software conditions are reported.

VII. COMPARATIVE EVIDENCE FROM REPRESENTATIVE STUDIES

Mohanty et al. reported 99.35% accuracy on their held-out PlantVillage test set, but the same model achieved 31.4% accuracy on 121 externally sourced images captured under different conditions [1]. This example demonstrates both the promise of CNN classification and the limitations of a controlled benchmark. It should not be interpreted as proof that every model will experience the same performance drop; rather, it illustrates why external evaluation is necessary before making field-readiness claims. Ferentinos reported 99.53% accuracy using 87,848 images across 58 plant-disease combinations [2]. Earlier work by Sladojević et al. demonstrated CNN-based recognition across multiple plant disease categories and reported class-level precision values between 91% and 98%, with an average of 96.3% on their collected image set [17]. Brahimi et al. focused on tomato leaf diseases and combined classification with symptom visualisation, reporting 99.18% accuracy on their dataset of 14,828 images [18]. These studies illustrate the progression from early task-specific CNN systems toward larger datasets and more interpretable symptom localisation, but their scores remain tied to their respective datasets and experimental protocols. Work on mobile capture devices further highlights that acquisition conditions are part of the classification problem, not merely a post-training deployment detail [19].

Too et al. compared fine-tuned VGG, Inception, ResNet and DenseNet architectures on PlantVillage and reported 99.75% test accuracy for DenseNet [10]. These high values show that CNNs can separate classes in curated datasets. They do not establish that the same models will retain performance under changes in crop variety, camera, location, illumination or symptom severity. The PlantDoc dataset provides a complementary evaluation setting because it includes web-sourced images with more varied backgrounds and visual conditions [3].

Recent review articles by Pacal et al. and Upadhyay et al. show that plant disease detection has already received substantial review coverage [8], [9]. Given this existing coverage, the present review takes a narrower comparative lens: how CNN architecture choices interact with dataset realism, evaluation design and deployment constraints. Richter and Kim provide a useful complementary benchmark, evaluating 23 model families across 18 open datasets with repeated transfer-learning experiments [16].

For agricultural decision support, evidence can be considered in increasing order of practical strength. First, an internal random split tests whether a model can learn patterns in a particular dataset. Second, a grouped split by leaf or plant reduces leakage from near-duplicate views. Third, a held-out dataset from a different collection session tests sensitivity to acquisition variation. Fourth, cross-dataset evaluation tests transfer to a different source. Fifth, prospective field evaluation with expert-confirmed labels tests the workflow under actual use. Many published studies report only the first or second level. A model should not be described as field-ready unless evidence at the later levels is available.

A. From Benchmark Classification to a Responsible Field Workflow

A practical crop-health assistant is more than a classifier. A complete workflow begins when a user captures an image and should include basic image-quality checks, supported-crop selection, model inference, uncertainty handling and a clear next step. A blurry image, a leaf that occupies only a small part of the frame, or a photograph taken in extreme shadow may not contain enough information for a reliable prediction. The interface should be able to request a clearer image or indicate that the input is outside the supported scope rather than always returning a disease name.

The output should distinguish a model prediction from a confirmed diagnosis. A useful explanation can state the predicted class, show the confidence or uncertainty in a carefully qualified way, identify whether the crop and disease are supported, and recommend consulting a local agricultural expert when symptoms are severe or ambiguous. Heat maps may help a reviewer inspect the regions associated with a prediction, but they should not be presented as proof of causation [15]. Farmer-facing language should avoid implying that a visual model can replace laboratory testing or professional plant pathology assessment.

Field evaluation should examine both technical and human outcomes. Technical measures include classification errors, abstention frequency, latency, failures under poor connectivity and performance across crops or locations. Human-centred measures include whether users can capture suitable images, understand the limitations, interpret the output and follow an appropriate referral pathway. A system may achieve a high image-level score yet fail in practice if users cannot frame leaves correctly or if the output is difficult to understand. Conversely, a cautious tool that abstains appropriately may be more useful than one that always produces a confident label.

Deployment also creates a maintenance requirement. New cultivars, seasonal conditions, camera changes and emerging diseases can alter the image distribution over time. A deployed system should therefore record version information, monitor error reports where privacy permits, and define when a model must be re-evaluated or updated. Updating a model with new data should not erase the ability to reproduce earlier results; dataset versions, model checkpoints and evaluation reports

should be retained. These operational controls connect the review's emphasis on dataset realism and validation with the ongoing reliability of a real decision-support service.

TABLE III. COMPARATIVE EVIDENCE FROM REPRESENTATIVE STUDIES

Study	Dataset / Scope	Reported result and interpretation
Mohanty et al. (2016) [1]	PlantVillage; external set of 121 images	99.35% on PlantVillage; 31.4% on external images. Shows risk of treating controlled accuracy as field performance.
Ferentinos (2018) [2]	87,848 images; 58 combinations	99.53% accuracy. High benchmark performance within the study protocol.
Too et al. (2019) [10]	PlantVillage; fine-tuned CNNs	99.75% for DenseNet. Strong result in that comparison, not a universal ranking.
Atila et al. (2021) [20]	PlantVillage; EfficientNet variants	EfficientNet among the strongest in that controlled-dataset comparison.
Richter & Kim (2025) [16]	23 model families across 18 datasets	Broad benchmark showing performance depends on dataset and protocol.

VIII. CHALLENGES AND RESEARCH GAPS

Domain shift occurs when the visual distribution of deployment images differs from the training data. Differences may arise from geography, crop cultivar, camera sensor, season, disease stage, lighting, soil background or management practices. A model trained on centred leaves may fail when a leaf is partly occluded or surrounded by other foliage. Cross-dataset testing, field-collected validation and geographically separated test sets are important steps toward measuring this problem.

Different diseases can produce similar spots, discolouration or leaf deformation. Conversely, one disease can appear differently depending on its stage and environmental conditions. Some visible symptoms may also be caused by nutrient deficiency, pests, drought or chemical injury. Image classifiers should therefore communicate uncertainty and avoid presenting a predicted class as a definitive laboratory diagnosis. Disease labels should be confirmed by trained plant pathologists or laboratory testing where possible.

Dataset bias can arise when certain classes are associated with particular backgrounds, image sources or acquisition devices. Class imbalance can cause minority diseases to be overlooked. Leakage can occur when duplicate images or photographs of the same leaf appear across different partitions. These issues can make evaluation results overly optimistic. Better dataset documentation, grouped splitting, duplicate detection and external validation are required. A model that performs well on a high-end GPU may not be practical on a low-cost smartphone or rural device. Mobile deployment requires attention to latency, memory, energy use, network connectivity and update mechanisms. A further concern is that image-level random splitting can overestimate performance when photographs from the same plant, farm visit or online source are distributed across training and test partitions. Near-duplicate detection and grouping by the unit of data collection should be considered before training. When the intended claim concerns transfer to a new farm, a farm-level or location-level holdout is more informative than a random image split. The grouping rule should be selected before examining test results and described clearly so that readers can judge whether the evaluation matches the intended deployment scenario. Another gap concerns the distinction between classification and diagnosis. A leaf classifier assigns one of the labels represented during training, but a field photograph may show multiple conditions, an unfamiliar disease, nutrient stress or a symptom that is not visible in the training taxonomy. Single-label softmax classification can force such an image into a known class even when none of the available labels is appropriate. Multi-label formulations, out-of-distribution checks and explicit abstention rules are therefore relevant when the application may encounter mixed or unsupported cases. Their effectiveness must be evaluated on deliberately collected examples of these cases rather than inferred from ordinary in-distribution test accuracy. Explainability methods also require careful interpretation. Saliency maps and Grad-CAM can help review whether predictions are associated with leaf regions, but visual plausibility alone does not establish causal reliance on disease symptoms. Explanations should be examined alongside controlled perturbation tests, background changes and error analysis. Similarly, segmentation may help isolate the leaf or lesion, yet it introduces an additional model or annotation requirement and can fail under occlusion or complex foliage. A deployment study should therefore report the complete pipeline, including any detection or segmentation stage, rather than presenting the classifier in isolation.

IX. FUTURE RESEARCH DIRECTIONS

Future studies should prioritise diverse field datasets with expert-verified labels. Collection protocols should cover different locations, seasons, cultivars, devices and disease stages. Public release of images, metadata and annotation guidelines would improve reproducibility and enable independent evaluation. Evaluation should move beyond random image-level splits. Leaf-, plant-, farm- and location-level partitions can reduce leakage and provide stronger evidence of generalisation.

Lightweight architectures such as MobileNet and compact EfficientNet variants should be assessed using predictive and systems metrics together. Reporting latency and memory on actual target devices is necessary to support claims about mobile suitability. Explainability should be evaluated rather than merely illustrated. Uncertainty estimation and selective prediction can help a system abstain when an image is blurry, out of distribution or visually ambiguous. Such mechanisms may be more appropriate for decision support than forcing a confident label for every image.

Leaf photographs capture only part of the information relevant to plant health. Weather, soil moisture, crop variety, plant age, location and disease progression can affect the probability and appearance of infection. Future systems may combine images with meteorological data, field observations or time-series images. The practical success of a classifier depends on whether its output is understandable and useful to intended users. Farmer-facing evaluation should examine whether users can capture acceptable images, understand uncertainty messages and take appropriate next steps. Research on domain adaptation and domain generalisation should distinguish between adaptation using labelled target-farm images and

generalisation to a target environment for which no labels are available. These are different operational settings and should be reported separately. Where data governance permits, benchmark releases should include metadata describing crop variety, location at an appropriately non-identifying level, capture device, growth stage and annotation confidence. Such metadata can support subgroup evaluation without exposing sensitive farm details.

X. CROSS-STUDY SYNTHESIS AND MODEL-SELECTION FRAMEWORK

Evidence across the reviewed studies supports a cautious interpretation of headline accuracy. Taken together, these studies do not establish a universal winner. They show that reported performance is conditional on the dataset, class taxonomy, preprocessing, transfer learning strategy and evaluation split. A useful comparison therefore separates three questions. First, can the model discriminate among the classes represented in the training distribution? Second, does it retain performance when images come from a different plant, camera, location or acquisition session? Third, can it operate within the computational and human factors constraints of the intended use?

For a controlled benchmark with limited compute, a compact pretrained model such as ResNet-18 or MobileNet may be a reasonable baseline. Where accuracy is the main objective and hardware is available, a larger DenseNet or EfficientNet variant can be included. For a mobile screening tool, MobileNet-family models and compact EfficientNet variants deserve early evaluation because their design explicitly considers computation. These are starting hypotheses, not final recommendations: the selected model must be tested on the target crop, disease classes and device. A useful experimental design is to begin with a compact baseline, then add one or two higher-capacity alternatives under identical data partitions and training budgets. The comparison should report not only the best checkpoint but also variation across repeated runs, because small datasets can produce unstable estimates. If a candidate improves aggregate accuracy but reduces recall for a high-consequence disease class, that trade-off should be made explicit. Similarly, if a model performs well only after extensive fine-tuning on a narrow dataset, its apparent advantage may not transfer to a new crop, region or camera. These considerations make model selection a constrained decision rather than a search for the largest benchmark score.

TABLE IV. USE-CASE-DRIVEN MODEL SELECTION AND VALIDATION FRAMEWORK

Candidate model strategy	Evidence required before selection
Small custom CNN and pretrained ResNet-18	Grouped test split, macro-F1, per-class recall and repeated runs.
MobileNet variant or compact EfficientNet	On-device latency, memory use, energy/battery proxy and class-wise performance.
DenseNet/EfficientNet compared with ResNet	Identical data partitions, training budget, input resolution and external test set.
Best validated candidate with uncertainty/abstention	External field data, expert-confirmed labels, calibration and user assessment.

XI. LIMITATIONS OF THE REVIEW

B. Minimum Reporting Protocol for Future Comparative Studies

A useful benchmark should make it possible for another research group to reproduce the data split and understand the limits of the resulting claim. At minimum, a paper should identify the dataset version, download or access date, class mapping, image exclusions and any relabelling. It should describe whether multiple images can originate from the same leaf, plant, field or collection session, and state the group used to separate training, validation and test data. If a public dataset does not provide the metadata needed for grouped splitting, that limitation should be stated rather than assuming that an image-level random split is independent.

The experimental protocol should then document the input resolution, colour handling, normalisation, augmentation operations and the stage at which each operation is applied. For transfer learning, the backbone, pretrained-weight source, frozen layers and fine-tuning schedule should be specified. Hyperparameter selection should be based on the validation set, with the test set reserved for final evaluation. Reporting a compact configuration table can make these details easier to inspect without interrupting the main discussion.

Results should include overall accuracy alongside macro-F1, macro-recall, class-wise precision and recall, and a confusion matrix. The number of test examples in each class should be visible. Repeated runs should be summarised with a measure of variability, and external evaluation should be reported separately from internal validation. When a model abstains, the coverage and error among accepted predictions should be provided. When efficiency is claimed, the device, runtime, input size, batch size and measurement procedure should accompany latency and memory figures.

Finally, conclusions should be written at the same level as the evidence. A result on PlantVillage supports a claim about performance on the stated PlantVillage split; it does not, by itself, support a claim about performance on farms. A cross-dataset result supports a stronger transfer claim only when the class definitions and label quality are sufficiently aligned. Prospective field evaluation supports claims about practical use only when the image-capture workflow, expert reference labels and user-facing output are also assessed. This evidence-aligned reporting approach can reduce overstatement and make future studies more comparable.

This review has several limitations. First, it is a focused narrative review rather than a systematic review with a preregistered protocol, database-specific search strings, duplicate screening and a formal risk-of-bias assessment. The selection of representative studies is therefore illustrative and may omit relevant work. Second, the reviewed studies use different crops, disease taxonomies, image sources, class distributions, preprocessing pipelines and data-splitting procedures. Their reported accuracy values cannot be treated as directly comparable estimates of a common model effect. Third, this article does not independently retrain the architectures or recalculate published metrics; numerical results are reported as described by the cited studies. Finally, the review focuses primarily on image-based leaf classification. It does not provide a complete assessment of plant pathology diagnosis, multi-modal sensing, disease severity estimation or agronomic outcomes of interventions. These limitations define the scope of the conclusions and

indicate where a future systematic review or controlled benchmark would add stronger evidence.

XII. CONCLUSION

CNN-based methods have become an important approach to crop leaf disease classification. Conventional CNNs, VGG, ResNet, MobileNet, EfficientNet and DenseNet provide different trade-offs in feature learning, computational cost and deployment suitability. Transfer learning can make these architectures practical when labelled agricultural data are limited, while preprocessing and augmentation can improve robustness when used carefully. Public datasets such as PlantVillage have enabled large-scale benchmarking, whereas field-oriented resources such as PlantDoc expose additional visual complexity.

The central conclusion of this review is that high accuracy on controlled images is not sufficient evidence of reliable agricultural deployment. Dataset realism, leakage prevention, class-wise performance, external validation, explainability, calibration and mobile efficiency should be considered together. Model selection should be guided by the intended use case and the quality of evidence available, not by a single headline accuracy value. Future work should prioritise field data, transparent evaluation and human-centred decision support.

CNN classifiers can assist crop-health monitoring, but their predictions should be interpreted within the limits of the training data and confirmed by appropriate agricultural expertise when necessary.

ACKNOWLEDGEMENTS

The authors acknowledge the guidance of Ms. Bhagyashree Phulmali, Department of Information Technology, Nutan Maharashtra Institute of Engineering & Technology, Pune.

Funding: The authors received no external funding for this research.

Conflict of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions: All authors contributed to the conception and scope of the review, literature collection and analysis, and preparation and critical revision of the manuscript. All authors reviewed and approved the final manuscript.

REFERENCES

[1] S. P. Mohanty, D. P. Hughes, and M. Salathé, “Using deep learning for image-based plant disease detection,” Frontiers in Plant Science, vol. 7, Art. no. 1419, 2016.

[2] K. P. Ferentinos, “Deep learning models for plant disease detection and diagnosis,” Computers and Electronics in Agriculture, vol. 145, pp. 311–318, 2018.

[3] D. Singh et al., “PlantDoc: A dataset for visual plant disease detection,” in Proc. 7th ACM IKDD CoDS and 25th COMAD, 2020, pp. 249–253.

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE CVPR, 2016, pp. 770–778.

[5] A. G. Howard et al., “MobileNets: Efficient convolutional neural networks for mobile vision applications,” arXiv:1704.04861, 2017.

[6] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. ICML, 2019, pp. 6105–6114.

[7] D. P. Hughes and M. Salathé, “An open access repository of images on plant health to enable the development of mobile disease diagnostics,” arXiv:1511.08060, 2015.

[8] I. Pacal et al., “A systematic review of deep learning techniques for plant diseases,” Artificial Intelligence Review, vol. 57, Art. no. 304, 2024.

[9] A. Upadhyay et al., “Deep learning and computer vision in plant disease detection: A comprehensive review,” Artificial Intelligence Review, vol. 58, Art. no. 92, 2025.

[10] E. C. 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.

[11] J. G. A. 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.

[12] A. Howard et al., “Searching for MobileNetV3,” in Proc. IEEE/CVF ICCV, 2019, pp. 1314–1324.

[13] G. Huang et al., “Densely connected convolutional networks,” in Proc. IEEE CVPR, 2017, pp. 4700–4708.

[14] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv:1409.1556, 2014.

[15] R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE ICCV, 2017, pp. 618–626.

[16] D. J. Richter and K. Kim, “Assessing the performance of domain-specific models for plant leaf disease classification,” Scientific Reports, vol. 15, Art. no. 18973, 2025. doi: 10.1038/s41598-025-03235-w.

[17] S. Sladojevic et al., “Deep neural networks based recognition of plant diseases by leaf image classification,” Computational Intelligence and Neuroscience, vol. 2016, Art. no. 3289801, 2016. doi: 10.1155/2016/3289801.

[18] M. Brahimi, K. Boukhalfa, and A. Moussaoui, “Deep learning for tomato diseases: Classification and symptoms visualization,” Applied Artificial Intelligence, vol. 31, no. 4, pp. 299–315, 2017. doi: 10.1080/08839514.2017.1315516.

[19] A. 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.

[20] M. Atila et al., “Plant leaf disease classification using EfficientNet deep learning model,” Ecological Informatics, vol. 61, Art. no. 101182, 2021.

[21] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. ICLR, 2021.

[22] Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF ICCV, 2021, pp. 10012–10022.

[23] Z. Liu et al., “A ConvNet for the 2020s,” in Proc. IEEE/CVF CVPR, 2022, pp. 11976–11986.

[24] J. G. A. Barbedo, “A review on the main challenges in automatic plant disease identification based on visible range

images," Biosystems Engineering, vol. 144, pp. 52–60, 2016.

[25] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," Computers and Electronics in Agriculture, vol. 147, pp. 70–90, 2018.

[26] J. Liu and X. Wang, "Plant diseases and pests detection based on deep learning: A review," Plant Methods, vol. 17, Art. no. 22, 2021.

[27] C. Guo et al., "On calibration of modern neural networks," in Proc. ICML, 2017, pp. 1321–1330.

[28] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Advances in Neural Information Processing Systems 30, 2017.

[29] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?," in Proc. IEEE/CVF CVPR, 2019, pp. 2661–2671.