Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 24 July 2026
Revised 28 Aug 2026
Accepted 15 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (20s): 191–215

KEYWORDS:
multi-crop disease detection, plant pathology, deep learning, Vision Transformer, YOLO, explainable AI, edge AI, PRISMA, domain generalizations

Efficient Multi-Crop Disease Detection Systems: A PRISMA-Guided Systematic Review of Deep Learning, Vision Transformers, Explainable AI, and Edge-Deployable Models

Mr. Siddharth S. Mandwade[1]*, Mr. Ashish S. Kale[2], Dr. Aruna Varanasi[3]

[1]*Department of Computer Science & Engineering, Vikrant University, Gwalior, Madhya Pradesh, India ORCID: 0009-0000-9807-5913 Siddharthraje1988@gmail.com
[2]Assistant Professor Department of AI&DS Engineering, SNJB's LSKBJ COE, Nashik, Maharashtra ORCID: 0009-0009-7475-4108 ashishsk.8872@gmail.com
[3]Department of Computer Science & Engineering, Sreenidhi Institute of Science & Technology, Hyderabad ORCID: 0000-0002-2943-8861 varanasi.aruna2002@gmail.com

ABSTRACT: Automated crop-disease diagnosis have moved beyond the single-crop convolutional classifier to multi-crop systems which include lesion localization, hierarchical classification, severity estimation, explainability and edge deployment. However, much of the existing literature is fragmented: many models have been shown to be close to saturated accuracy when trained on curated leaf images; however few models have demonstrated stable performance under field environments (i.e. background), various devices, unseen crops/diseases or varying levels of computational resources. The most prevalent families of deep learning models used were CNN's and Transfer-Learning based models although Hybrid CNN-Transformer architectures, YOLO based lesion localisation algorithms, Ensemble/XAI systems including knowledge distillation and quantization, and semantic zero-shot approaches all experienced significant growth. Benchmark accuracy rates greater than 95% were commonly reported by many studies; yet, studies conducted in field oriented settings often found disease classification, severity grading, and cross domain transfer to be significantly more difficult than controlled environment background recognition. A total of 22 studies validated their findings with either field or external validation; 17 studies provided evidence related to latency and/or model size; 14 studies examined severity and/or localized lesions at the level of pixels; and 13 studies included some form of explanation. Based upon the synthesis of results an effective multi-crop architecture would consist of a crop aware hierarchical pipeline: image quality gating; a lightweight leaf/lesion localization algorithm; a shared CNN-MobileViT backbone architecture; crop and disease specific heads; calibrated uncertainty estimates; severity estimation; visual explanations; and a quantized edge model connected to a cloud based drift detection and active learning loop. The review concludes that advances will likely depend less on achieving another small margin increase in benchmark accuracy and more on demonstrating cross region validation, open set recognition, calibration, energy consumption measurements and clinically analogous reporting of failures.

1. INTRODUCTION

A challenge that requires extremely high levels of visual intelligence is the identification of plant diseases. Multiple infections can

induce changes in color and texture that are similar to one another, the same illness can appear differently in different cultivars and stages of growth, and images shot in the field can reveal shadows, occlusion, soil, tools, hands, water droplets, insect damage, and patterns of nutrient stress. It is therefore necessary for a decent multi-crop system to do more than simply connect a picture of a leaf that is centered to one of multiple labels. The first thing that it needs to do is determine whether or not the image can be used for diagnosis, locate the relevant plant organ, identify the crop or remain stable if the crop name is unknown, differentiate between disease and abiotic stress, estimate the severity of the problem, measure the uncertainty, and provide a response that is quick enough for mobile or edge use.

A discernible shift in methodology may be seen in the published works from the years 2023 to 2026. A significant amount of research continues to support the idea that traditional transfer learning is effective [3, 4, 9, 16]. On the other hand, more recent research has included the addition of field datasets [2], explicit controlled-to-field analysis [1], segmentation [6], severity quantification [7, 25, 28], real-time detection [8], lightweight meta-ensembles [11], quantization and distillation [15], semantic zero-shot learning [18], and explainability [10, 12, 24, 29, 30, 35, and 40]. This modification is significant since the question that is now being asked in the real world is not whether or not a CNN can differentiate between different disease classifications in a curated library. The question that is more significant is whether or not a system can work with different crops, places, acquisition devices, seasons, and disease signs that have never been seen before while still operating with the hardware that is used on the farm.

Despite the fact that there have been recent studies of deep-learning algorithms for diagnosing plant diseases [52–55], a focused combination is still required in order to develop an efficient multi-crop system. Many of the reviews that have already been produced do not differentiate between controlled and field evidence, nor do they discuss accuracy without giving latency, model size, calibration, severity, and open-set behavior the same amount of weight. The purpose of this evaluation is to fill that gap by utilizing PRISMA to examine fifty journal publications that were published by four major publisher organizations. It does this by arranging the data in a taxonomy that is system-oriented and by developing a reference architecture that blends the quality of recognition with the readiness for deployment. Therefore, the review begins with a system-level purpose for doing the review, then proceeds on to the formulation of clear research questions, the selection of studies to examine, the mapping of the evidence, the critical combination of the evidence, and lastly, the formulation of an architecture and research agenda based on the evidence. In this manner, each and every design concept may be connected to the material that was examined.

2. REVIEW OBJECTIVES AND RESEARCH QUESTIONS

This review was conducted with the intention of locating and contrasting novel models that contribute to the development of a system that is effective in identifying illnesses in a variety of crops. To give a valid, reliable, comprehensible, and computationally viable diagnosis across several crops or to add a transferable module to such a system, such as localization, severity estimate, domain alignment, or model compression, was the basic definition of efficiency. Efficiency was also defined as the ability to deliver a diagnosis that was correct, reliable, clear, and practicable.

Question	Operational focus
RQ1	Which model families and task formulations dominate recent crop-disease diagnosis?
RQ2	How often do studies evaluate multi-crop, field, cross-domain, severity, explainability, and deployment requirements?
RQ3	What performance patterns and methodological weaknesses emerge across publishers and years?
RQ4	Which architecture is best supported for an efficient multi-crop disease detection system?
RQ5	What research agenda is required to move from curated benchmarks to trustworthy field deployment?

Table 1. Research questions guiding the systematic review.

According to Table 1, the purpose of the research is split down into five different operational questions. While RQ1 and RQ2 provide a description of the evidence base, RQ3 investigates problems in both performance and methodology, and RQ4–RQ5 use the review results to make decisions regarding the architecture of the system and the research priorities available. Instead of focusing solely on ranking accuracy, this methodology ensures that the synthesis continues to be accurate with regard to the deployment of several crops in the actual world. In addition, these questions determine the data that is collected during the

screening process; hence, the subsequent steps are closely connected to the comparisons and analyzes that are presented in the sections titled Results and Critical Synthesis.

3. METHODOLOGY

During the review, the procedures for reporting that were outlined in PRISMA 2020 [51] were followed. Instead of attempting to estimate the entire number of articles that have been published in all journals around the world, searches were conducted in the publisher platforms of IEEE Xplore, SpringerLink, ScienceDirect, and Wiley/Hindawi. This was done because the objective was to concentrate on publishers and technology. Hand checks were performed on each and every record that was included to ensure that it contained the correct year, journal status, title, publisher, and DOI (stable publisher identifier). Despite the fact that their descriptions could still be located, records that were tagged as "retracted" were not included in the compilations.

3.1 Information Sources and Search Strategy

Source	Core query logic	Filters
IEEE Xplore	("plant disease" OR "crop disease" OR "leaf disease") AND (deep learning OR CNN OR transformer OR YOLO OR ensemble)	2023–2026; journals; image-based detection/classification/segmentation
SpringerLink	Same disease terms combined with Vision Transformer, transfer learning, severity, lightweight, explainable AI	Journal articles; 2023–9 Aug 2026
ScienceDirect	Plant/crop/leaf disease AND classification/detection/segmentation AND deep learning/attention/edge	Research articles; Elsevier journals; 2023–9 Aug 2026
Wiley/Hindawi	Plant/crop leaf disease AND CNN/transfer learning/segmentation, with DOI prefix 10.1155	Research articles in Hindawi-origin journal titles; 2023–2026

Table 2. Publisher-specific search strategy.

As can be seen in Table 2, the conceptual search area was modified so that it corresponds with the language and indexing style of each publisher interface in the process. The fundamental concept of a plant or agricultural illness was preserved, and it was integrated with deep learning, detection, segmentation, transformers, ensembles, or efficiency. On the other hand, the corpus was restricted to only include peer-reviewed journal articles that were published during the review timeframe due to publisher-specific filtering. The establishment of this common search frame necessitated the establishment of a uniform eligibility rule. This was done to ensure that variations in the manner in which publishing interfaces were implemented did not result in various conclusions regarding what should be included in this process.

3.2 Eligibility Criteria

Decision	Criterion
Include	Peer-reviewed journal article; publication year 2023–2026; plant/crop disease detection, classification, localization, segmentation, or severity; technical model evaluation; IEEE, Springer Nature, Elsevier, or Hindawi/Wiley journal.
Exclude	Conference paper, book chapter, thesis, editorial, non-diagnostic agronomy study, pre-2023 article, retracted article, duplicate, or record lacking sufficient model/evaluation information.
Context only	Systematic reviews and PRISMA guidance were cited to frame the synthesis but were not counted among the 50 primary studies.

Table 3. Inclusion and exclusion criteria.

Evidence that was used for the primary synthesis is separated from material that is more general in Table 3, which can be found here. The distinction is significant due to the fact that background material derived from systematic reviews, opinions, conference papers, studies conducted prior to 2023, retracted articles, and papers that do not analyze models adequately can be of use to the discussion; but, they cannot directly contribute to the evidence map consisting of fifty independent studies. Following the establishment of the evidence limit, the review could go from the process of searching for and obtaining evidence

to the screening of evidence that could be repeated, the extraction of structured data, and the evaluation of evidence that was essential to the deployments.

3.3 Screening, Data Extraction, and Quality Appraisal

The search log contained 78 records that were known to be sources. A total of fifty records were included, five copies were removed, 73 titles and descriptions were examined, 62 records were put thru a comprehensive assessment by the publisher, and fifty records were included. 12 full-text/publisher records were discarded because they were deemed to be irrelevant or invalid. Four of the records were retracted, three of the recordings were from before the year 2023, three of the records did not have sufficient information regarding the model or evaluation, and two of the records were not journal articles. During the process of data extraction, the crop's scope, task, model, dataset configuration, key metric, external validation, severity/localization, explanation, and evidence of deployment were all included in process. There were six binary indications that were utilized in order to evaluate the quality of the product.

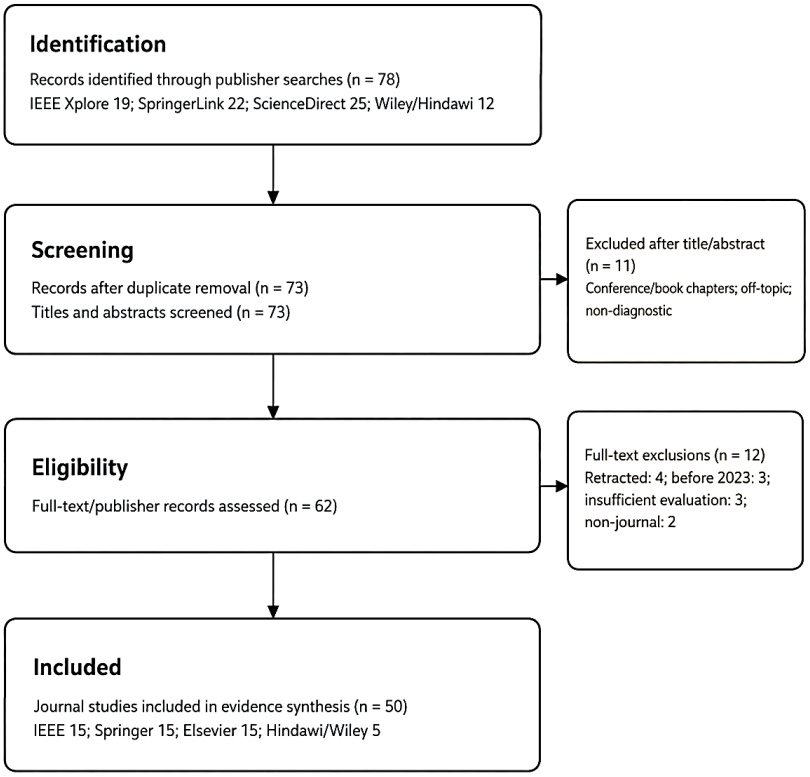


Figure 1. PRISMA-style flow of record identification, screening, eligibility assessment, and inclusion process

Figure 1 provides a clear illustration of the screening attrition that occurred between the two groups. The most significant decrease occurred prior to the incorporation of full-text. The majority of the final exclusions were made because of the status of the retraction, the release date, the absence of a technical evaluation, or the form of publishing because it was not a journal. With the new flow, the capability to monitor the search volume is maintained all the way thru to the conclusion of the evidence in the base. This screened collection of fifty studies serves as the foundation for all subsequent counts, percentages, tables, and visual comparisons throughout the rest of the research process. This enables the Results section to transition from the composition of the corpus to evidence at the model level without requiring any changes to the research base

4. RESULTS

4.1 Publication Profile

The final corpus consisted of 15 research published by the IEEE, 15 studies published by Springer Nature, 15 studies published by Elsevier, and 5 studies published by Hindawi/Wiley. In the year 2023, seventeen studies were published, five in the year

2024, fifteen in the year 2025, and three in the early part of the year 2026. Both Springer and Elsevier published a large number of papers in 2025 that discussed end-to-end systems, severity estimation, hybrid transformers, and real-world applications. These papers were published in a variety of journals. It is not appropriate to interpret the smaller subset for 2026 as an indication of a decrease in activity; and only records from that period of the year are included in the compilations.

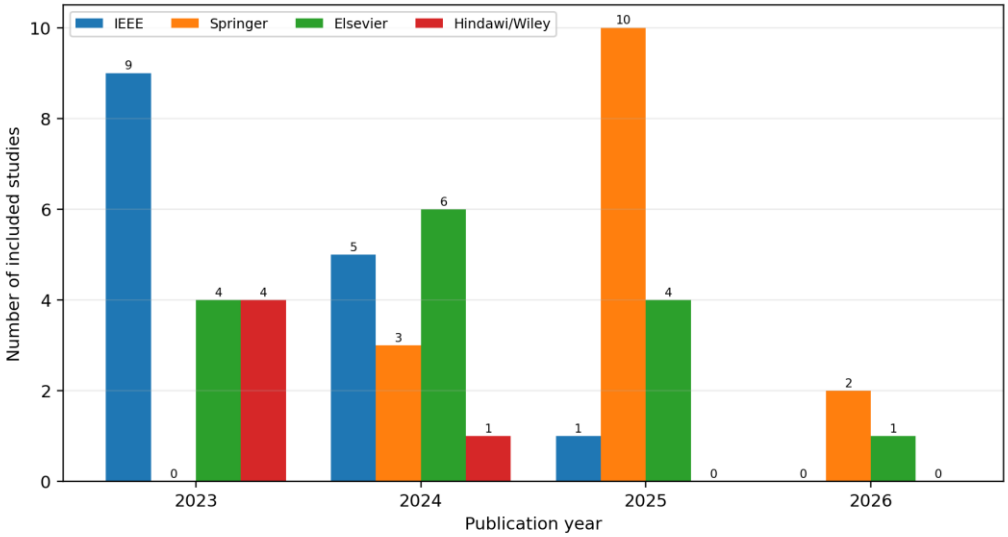


Figure 2. Included studies by publication year and publisher group.

Rather than a decrease in the amount of effort put into research, Figure 2 demonstrates that there has been a shift in the concentration of publishers. Between the years 2023 and 2026, Elsevier maintains a rather stable position, while the IEEE makes the most significant contribution in 2023. Springer takes a significant leap in 2025. Due to the fact that just three papers were recorded in 2026 due to a partial-year cutoff, this should not be interpreted as a sign that the number of papers that are being published is decreasing for this.

Table 4. Distribution of the 50 included journal studies by year and publisher group.

Year	IEEE	Springer Nature	Elsevier	Hindawi/Wiley	Total
2023	9	0	4	4	17
2024	5	3	6	1	15
2025	1	10	4	0	15
2026*	0	2	1	0	3
Total	15	15	15	5	50

A numerical representation of the time pattern depicted in Figure 2 is provided in Table 4. With regard to the corpus, there are fifteen studies that come from each of the three largest publisher groups. There are only five papers that Hindawi/Wiley contributes, and all of them were published between the years 2023 and 2024. The peak in the number of articles published by Springer in 2025 occurs at the same time that hybrid transformers, severity estimate, explainability, and end-to-end diagnostic systems are gaining popularity. The row that contains the year 2026 contains the records that can be seen in this process. However, temporal distribution is not sufficient to explain why the technical focus shifted in the first place. Because of this, the following section will restructure the same 50 research according to the primary methodological family that they employ in this process.

4.2 Primary Model Families

There was a major family assigned to each study, even when there was overlap between architectures. CNN and transfer-learning systems were the subject of fifteen different researches. After them came the detection and segmentation systems,

which accounted for ten, the CNN–Transformer hybrids, which accounted for eight, the ensemble/XAI systems, which accounted for eight, the lightweight/edge systems, which accounted for six, and the generalizations. The fact that CNNs appear to be the most popular should not be interpreted as a sign that scientific progress has achieved a plateau. Fuzzy optimization, knowledge distillation, attention, multiscale fusion, and hierarchical heads are some of the techniques that are utilized by many more recent CNN systems. The tendency that is more significant is functional convergence, which means that systems that function effectively combine an increasing number of families of classifiers rather than relying on just one in real time scenarios.

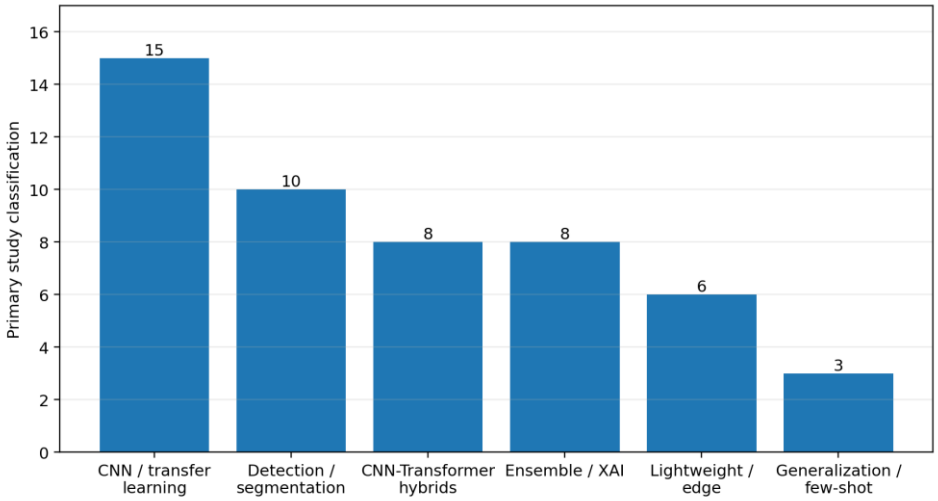


Figure 3. Primary methodological classification of the 50 included studies.

The data presented in Figure 3 demonstrates that there is not a single building family that can solve all of the issues. CNN and transfer-learning models account for thirty percent of the data, but seventy percent of the research already belong to one of the following categories: detection/segmentation, transformer-hybrid, ensemble/XAI, lightweight/edge, or generalization-oriented. This release makes it possible to construct a modular system in which local texture extraction, localization, global context, interpretability, compression, and domain robustness are all unified into a single option rather than being viewed as distinct choices that cannot be utilized together in process.

Table 5. Distribution, contribution, and principal limitation of the primary model families.

Primary family	Studies	Share	Principal contribution	Main system-level limitation
CNN / transfer learning	15	30%	Strong local texture learning; mature transfer learning	Dataset dependence and background shortcuts
Detection / segmentation	10	20%	Leaf/lesion localization; severity support	Higher annotation burden
CNN–Transformer hybrids	8	16%	Combines local lesion cues with global context	Attention cost can limit edge use
Ensemble / XAI	8	16%	Robust aggregation and interpretable decisions	Memory/latency can increase
Lightweight / edge	6	12%	Low-latency inference and reduced footprint	May trade representation depth for efficiency
Generalization / few-/zero-shot	3	6%	Addresses unseen domains/classes	Evidence base remains small

A design effect is generated based on the number of model families displayed in Table 5. For the majority of jobs, CNNs continue to be the most effective method for extracting features. On the other hand, detection and segmentation provide classification with more spatial context than simple classification used alone. By utilizing hybrid transformers, context modeling

can be improved, ensemble/XAI systems can strengthen interpretation and dependability, lightweight models can reduce the cost of deployment, and approaches that concentrate on generalization can bridge the gap between the laboratory and the field. The fact that their shortcomings complement one another is the reason why an integrated design is preferable to selecting only one "best" family. Following the completion of this comprehensive classification, the evaluation will proceed to the specific studies in order to verify the broad assertions made about families by comparing them to the crop scope, datasets, reported metrics, and deployment characteristics.

4.3 Evidence Map of Included Studies

The study-level extraction is divided down into four smaller evidence blocks based on the information presented in Tables 6–9. By doing so, it is possible to comprehend methodological tendencies without having to cram twenty-five studies into a single enormous table. Only when the numbers are actually present in the publication record or in the full-text metadata that has been confirmed are they displayed. In cases where they are not, the tables provide the methodological input that was presented without presenting a performance number that was made up. Due to the fact that datasets, crop counts, class counts, data splits, job descriptions, and field complexity are all quite diverse from one another, it is not possible to directly compare the accuracy values of different research. scenarios

Ref.	Year / publisher	Crop scope	Method	Dataset / evaluation	Main evidence	Readiness signal
[1]	2023 / IEEE	Multi-crop; controlled and field	Cross-domain CNN benchmarking	Controlled datasets plus field imagery	Large controlled-to-field degradation; domain robustness is the central finding	External/field validation
[2]	2023 / IEEE	Eight crop/disease groups; field	FieldPlant dataset and deep detectors	5,170 field images; 8,629 annotated leaves; 27 classes	Provides a field-oriented benchmark for detection and classification	Public field dataset
[3]	2023 / IEEE	Multi-crop	PlantDet multi-model ensemble	PlantVillage-style multi-class benchmark	Reported accuracy 99.81%	High benchmark accuracy; ensemble increases compute
[4]	2023 / IEEE	Multi-crop	TLMViT: transfer learning plus Vision Transformer	Multi-class leaf images	Transformer-assisted feature extraction improved classification over component baselines	Hybrid global/local features
[5]	2023 / IEEE	Maize	MaizeNet deep architecture	Maize leaf disease images	Average accuracy 97.89%; mAP 0.94	Crop-specific; strong detection accuracy
[6]	2023 / IEEE	Multi-class leaves	Cascading autoencoder + attention residual U-Net	Segmentation and classification datasets	Joint lesion segmentation and disease classification improved disease-region focus	Pixel-level localization
[7]	2023 / IEEE	3-D plant leaves	MULTINET: multi-agent DRL +	3-D leaf imagery	Combined identification and severity quantification	Severity-aware but computationally

			EfficientNet			complex
[8]	2023 / IEEE	Citrus; 6 classes	Light-SA YOLOv8	3,202 images	mAP@0.5 92.5%; 3.4 ms/image; 4.5 MB model	Strong real-time/edge profile
[9]	2023 / IEEE	Multi-crop	Deep CNN + local binary pattern feature fusion	Multi-class leaf disease benchmark	Fusion improved discriminability of texture-sensitive disease classes	Handcrafted + deep features
[10]	2024 / IEEE	13 plants; 38 classes	Deep prediction model + explainable AI	38-class benchmark	Accuracy 99.69%; precision 98.27%; recall 98.26%	XAI included; controlled-domain emphasis
[11]	2024 / IEEE	Multi-crop	Lightweight meta-ensemble	IoT-oriented evaluation	Maintained competitive accuracy with reduced deployment burden	Designed for IoT
[12]	2024 / IEEE	13 plant types; 38 diseases	VGG/ResNet/Inception voting ensemble + LIME	Multi-crop image benchmark	Ensemble improved reliability; LIME supplied local explanations	Explainable but multi-backbone cost
[13]	2024 / IEEE	Real-time multi-crop field data	Custom field dataset + deep detection	Real-time acquired images	Demonstrated the value of locally acquired data over curated-only training	Field data emphasis

Table 6. Included studies [1]–[13]: early multi-crop, field, localization, XAI, and real-time evidence.

The first section of the study corpus is presented in Table 6, and it consists primarily of work from the IEEE from 2023 to 2024. The FieldPlant dataset [2] and cross-domain benchmarking [1] are two examples of evidence that can be found. In addition, there are other instances such as high-accuracy ensembles [3,] transformer-assisted transfer learning [4,] lesion segmentation [6,] severity-aware learning [7], and the 4.5 MB Light-SA YOLOv8 detector [8]. From controlled classification to field collection, spatial localization, the ability to describe things, and low-latency inference, the most prominent trend is a transition away from controlled categorizations. This line of reasoning is continued in the subsequent piece of evidence, which includes the addition of UAV acquisition, compression, multitask transformers, zero-shot identification, and severity-aware modeling. The goals of efficiency and generalization are considered to be significant at this phase in the design process.

Ref.	**Year / publisher**	**Crop scope**	**Method**	**Dataset / evaluation**	**Main evidence**	**Readiness signal**
[14]	2024 / IEEE	UAV crop imagery	ShuffleNetV2 + fuzzy RBM + TCSSA optimization	Remote-sensing/UAV data	Optimization addressed noisy aerial features and class ambiguity	UAV and lightweight backbone
[15]	2025 / IEEE	Cross-domain multi-crop	Ensemble + knowledge distillation + quantization	Source and target domains	Deployment pipeline explicitly balances cross-domain accuracy and compressed inference	High deployment readiness
[16]	2024 / Springer	Multi-crop	Transfer-learning classification/detection	Public plant-disease repositories	Study reports improved disease recognition through transfer learning while noting overfitting risks	Sustainability-oriented
[17]	2024 / Springer	Multi-crop	Multitask Vision Transformer	Disease localization +	A shared transformer performs localization and	Multi-task design

				classification	class prediction jointly	
[18]	2024 / Springer	Unseen disease classes	Zero-shot learning with semantic attributes	Visual features + semantic descriptions	Demonstrates classification of unseen disease categories without class-specific training images	Open-set generalization
[19]	2025 / Springer	CRW crops	Hybrid deep learning models	Crop leaf images	Hybridization improves class separation across related crop diseases	Multi-crop comparison
[20]	2025 / Springer	Potato	Hybrid deep-learning ensemble	Potato disease datasets	Combines complementary feature learners for enhanced accuracy	Crop-specific
[21]	2025 / Springer	Tomato	E-TomatoDet global/local feature fusion	Tomato leaf images	Global context and local lesion features are fused efficiently	Efficiency-oriented feature fusion
[22]	2025 / Springer	Multi-crop leaves	EnConv enhanced CNN	Leaf disease datasets	Enhanced convolutional blocks improve fine-grained symptom representation	Compact CNN
[23]	2025 / Springer	Sugarcane	Transfer-learning deep neural network	Sugarcane leaf images	Transfer learning improves classification with limited crop-specific data	Crop-specific transfer
[24]	2025 / Springer	Plant disease and pests	Deep classification/detection + explainability	Public datasets	Links disease/pest classification with model explanations	XAI and broader stress scope
[25]	2025 / Springer	Large-scale multi-crop	Deep framework for detection and severity	Large multi-class image collection	Scales diagnosis from class label to disease severity	Severity-aware

Table 7. Included studies [14]–[25]: UAV, compressed cross-domain, transformer, zero-shot, and severity-oriented models.

The progression from recognition models that operate independently to designs that concentrate more on the structure as a whole is illustrated in Table 7. UAV imaging [14], knowledge distillation and quantization [15], multitask Vision Transformers [17], zero-shot semantic characteristics [18], compact feature fusion [21]–[22], disease/pest modeling that can be explained [24], and large-scale severity analysis [25] are some of the factors that contribute to an expansion of the operational scope. One of the most significant issues is that many models are only capable of testing one or two crops at a time. As a result, the stability of such crops across regions and devices is not completely demonstrated. In Table 8, which displays research that link recognition with suggestions, field-based validation, lightweight inference, and clear handling of tough visual situations, this improvement is demonstrated in a more clear and concise manner.

Ref.	Year / publisher	Crop scope	Method	Dataset / evaluation	Main evidence	Readiness signal
[26]	2025 / Springer	Rice	Twin CNN framework	Rice leaf disease images	Dual branches capture complementary disease features	Crop-specific dual-stream
[27]	2025 / Springer	Five datasets; multi-crop	PlantCareNet CNN + dual-mode recommendations	Five recognized datasets	Accuracy 82–97%; precision 97%; average inference 0.0021 s	Diagnosis plus preventive guidance

[28]	2025 / Springer	Rice; 10 diseases, 3 severity levels	Shared-backbone Vision Transformer	3,345 annotated images	Disease macro-F1 53.52%; severity macro-F1 77.79%; AUC 0.86	Realistic difficulty exposed
[29]	2026 / Springer	Four datasets; plant disease and pests	ViT and ResNet-9 hybrid + SHAP	15-class and three 3-class datasets	Accuracy 97.4%; precision 96.4%; recall 97.09%; F1 95.7%	Cross-dataset and explainable
[30]	2026 / Springer	Brinjal; 5 field classes	PD ViCo: ViT + patch dropout + CoCa	1,823 field-harvested images	Accuracy 99.12%; F1 97.76%; Grad-CAM/Grad-CAM++	Field dataset; lightweight/interpretable
[31]	2023 / Elsevier	Tomato, potato, pepper	DeepCrop CNN + web application	Multi-crop leaf images	Operationalizes image diagnosis in a browser-based application	Application deployment
[32]	2023 / Elsevier	Multi-crop	PPLC-Net + weather augmentation + multilevel attention	Complex-environment datasets	Weather-aware augmentation and attention improve robustness	Context-aware learning
[33]	2023 / Elsevier	Corn field imagery	Two-stage segmentation network	Real corn field images	Quantifies disease through staged leaf/lesion segmentation	Field quantification
[34]	2023 / Elsevier	Multi-crop	DLMC-Net lightweight CNN	Multi-class plant leaves	Deeper lightweight design targets accuracy with fewer resources	Edge-oriented
[35]	2024 / Elsevier	Three plant-disease datasets	DFN-PSAN multilevel fusion + attention	Three datasets	Average accuracy and F1 exceeded 95.27%; attention reduced parameters by 26%	Interpretable and parameter-efficient
[36]	2024 / Elsevier	Rice	DeepRice deep feature/model fusion	Rice disease subtype images	Compares and fuses deep representations for subtype discrimination	Fine-grained crop-specific
[37]	2024 / Elsevier	Multi-crop	Ensemble of deep architectures	Leaf disease benchmarks	Ensemble achieved stronger aggregate classification than individual networks	Accuracy–cost trade-off
[38]	2024 / Elsevier	Challenging environments	Attention-mechanism investigation	Variable-background plant images	Shows attention can improve localization under clutter, but benefits depend on attention type	Field robustness analysis

Table 8. Included studies [26]–[38]: dual-stream, recommendation, severity, explainable transformer, lightweight, and challenging-environment evidence.

As seen in Table 8, the practical usefulness is given a greater amount of weight. PlantCareNet [27] establishes a connection between disease recognition and preventive suggestions at the millisecond level. The rice Vision Transformer [28] illustrates how difficult it is to classify diseases and the severity of their symptoms. PD ViCo [30] combines images collected in the field with transformer regularization and visual explanations. Studies published by Elsevier [31–38] include online deployment,

weather-aware augmentation, field segmentation, lightweight CNNs, multilevel fusion, ensembles, and paying attention in challenging backgrounds. By concentrating on cross-domain alignment, mobile/web performance, energy awareness, and precise segmentation, the final piece of proof brings the methodological and chronological evolution to a successful conclusion for this process.

Ref.	Year / publisher	Crop scope	Method	Dataset / evaluation	Main evidence	Readiness signal
[39]	2024 / Elsevier	Multi-crop	Residual Vision Transformer + semantic segmentation	PlantVillage and lesion segmentation	Reported 97% accuracy with hierarchical classification and damage localization	Hierarchical diagnosis
[40]	2024 / Elsevier	Multi-crop	Ensemble transfer learning + explainable AI	Leaf disease datasets	Combines ensemble prediction with visual explanations	Interpretability included
[41]	2025 / Elsevier	Cross-crop healthy/unhealthy recognition	Generalized deep model	Multiple crop datasets	Explicitly studies cross-crop transfer rather than isolated crop classifiers	Generalized screening
[42]	2025 / Elsevier	Crop diseases across domains	Multi-granularity alignment	Cross-domain datasets	Aligns global and local disease representations to reduce domain shift	Domain adaptation
[43]	2025 / Elsevier	Multi-crop	RTR_Lite_MobileNetV2	Leaf disease datasets	Lightweight MobileNetV2 variant designed for efficient detection/classification	Mobile deployment
[44]	2025 / Elsevier	Plant leaves; web diagnosis	MobileViTv2, EfficientNet-B7 and hybrid model	Web Integrated dataset	MobileViTv2 reached 94% accuracy and was deployed in a web application	Practical application
[45]	2026 / Elsevier	Multi-crop	Energy-efficient deep learning framework	Agricultural consumer-electronics setting	Frames accuracy jointly with energy-efficient execution	Green/edge AI
[46]	2023 / Hindawi/Wiley	Tomato	Comparison of pretrained deep neural networks	Tomato leaf disease dataset	Identifies the most transferable pretrained architecture for tomato diagnosis	Transfer-learning benchmark
[47]	2023 / Hindawi/Wiley	Maize; multimodal	Deep transfer learning for multimodal images	Multiple image modalities	Multimodal transfer improves feature robustness over single View	Multimodal input

					classification	
[48]	2023 / Hindawi/Wiley	Tea grey blight	Improved deep CNN	Tea leaf images	Fourteen-layer-style deep feature learning and augmentation support grey-blight detection	Crop-specific; field relevance
[49]	2023 / Hindawi/Wiley	Paddy	Cross Validated ANN versus baseline classifiers	Paddy disease features/images	Cross Validation offers a more reliable estimate than a single train/test split	Validation emphasis
[50]	2024 / Hindawi/Wiley	Jasmine leaf spot	DCGAN augmentation + MobileNetV4-UNet	10,000 generated/curated segmentation images	mIoU 0.95; average pixel accuracy 0.91	Synthetic augmentation + precise segmentation

Table 9. Included studies [39]–[50]: hierarchical transformers, domain alignment, mobile/web deployment, energy-aware AI, and precise segmentation.

Table 9 presents a collection of characteristics that are essential for a multi-crop system that can be successfully implemented in any location. Directly addressing the issue of generality or operating expense can be accomplished thru a variety of approaches. The cross-crop screening [41], the multi-granularity domain alignment [42], the deployment of MobileNet/MobileViT [43–44], and the energy-aware inference [45] are some examples of these. Transfer-learning standards, multimodal inputs, cross-validation discipline, and MobileNetV4-UNet segmentation are all components that are included in the Hindawi and Wiley research [46–50]. These studies demonstrate that diverse types of evidence are useful for a variety of purposes rather than simply ensuring that the primary classification is accurate. When taken as a whole, Tables 6, 7, 8 & 9 make it abundantly evident that the objective of the research has evolved from the identification of disease classes to the compilation of the localization, robustness, interpretability, and efficiency functions that are required by a genuine system. This development is summarized in a single snapshot of the preparedness phase, which can be found in Figure 4 of this text.

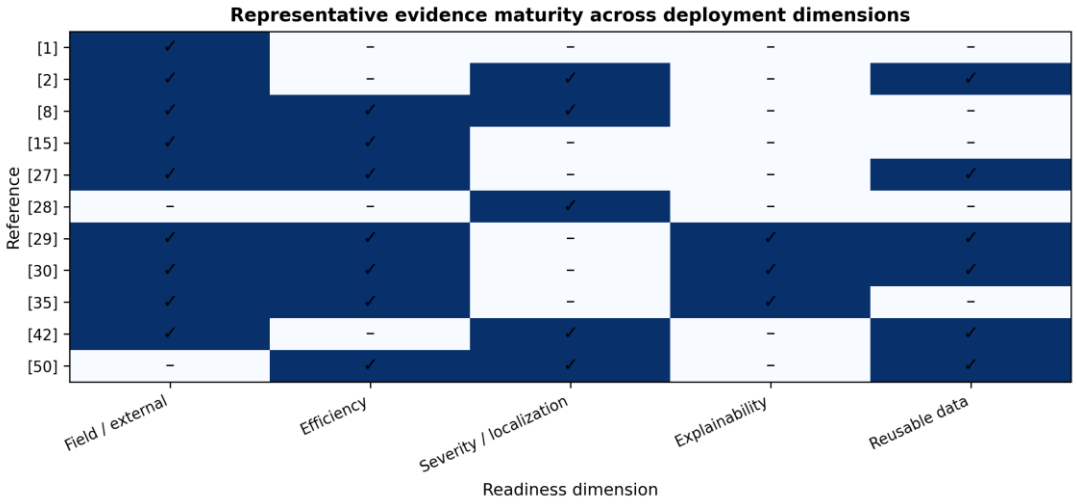


Figure 4. Representative evidence-maturity matrix across five deployment-readiness dimensions for this process

The illustration in Figure 4 illustrates one of the reasons why individual state-of-the-art findings do not necessarily result in a complete system. The combination of field/external validation and efficiency is only discovered in a limited number of sample studies simultaneously. Reusable data, explainability, and severity/localization are all found together even less frequently than they are found separately. A number of research, such as [30], investigate a more comprehensive range of readiness characteristics. The focus of other high-value articles, on the other hand, is on a single critical characteristic, such as domain

robustness [1, field benchmarking [2], real-time detection [8], or exact segmentation [50], which must be integrated with other modules that perform very similar functions. The next section offers a comparison of performance rather than explicitly rating the various accuracy, F1, mAP, latency, and segmentation scores. This is done in order to avoid any confusion that may arise. Due to the fact that readiness is dependent on a variety of skills that are reported using a variety of indicators, this is the case.

4.4 Performance Patterns and Their Interpretation

A number of controlled image classifiers, including PlantDet [3,] the explainable 38-class model [10], and recent hybrid/transformer systems [29, 30], have shown accuracy levels that are greater than 97%. The results of this study demonstrate that contemporary networks are capable of learning to differentiate between various diseases provided that the label quality and imaging settings are within acceptable parameters. As a standalone product, they do not prepare you for growing multiple crops. Accuracy is improved when backgrounds are consistent, classes are visually distinct, train and test distributions are closely matched, and each picture contains a symptomatic leaf that is centered. In light of this, the evaluation understands a high level of benchmark accuracy to be evidence of the capability to describe rather than evidence of the dependability of the field.

Studies that are conducted in the field are distinct from those that are conducted across domains. The controlled-to-field gap was demonstrated in a clear and convincing manner by Ahmad et al. [1], FieldPlant [2] offered a dataset that included natural scene complexity, the corn quantification model [33] made use of field imagery, the cross-domain compressed system [15] dealt with both transfer and deployment, and multi-granularity alignment [42] directly addressed domain mismatch. These studies are nevertheless more valuable for operational planning, despite the fact that they are not as accurate as the headlines suggest they appear to be. A bespoke dataset consisting of ten diseases yielded a disease macro-F1 of 53.52% and a severity macro-F1 of 77.79%, indicating that the rice severity research [28] is of great assistance. However, rather than being a negative finding, this reveals how difficult it actually is, which was concealed by benchmarks that were excessively stringent.

There are still inconsistencies in the evidence of efficiency. The fact that this real-time report from Light-SA YOLOv8 [8] had a mAP@0.5 of 92.5%, an inference time of 3.4 milliseconds, and a model size of 4.5 megabytes made it more transparent than others. PlantCareNet [27] reported that the average inference took 0.0021 seconds, but DFN-PSAN [35] reduced the number of factors by 26% while maintaining an average accuracy/F1 that was greater than 95.27%. With the help of lightweight meta-ensembles [11], systems based on MobileNet [43], and distillation/quantization [15], we are making progress in the right scopes. While, latency is typically measured on high-end or unidentified technology. While, energy, peak memory, thermal throttling, and the behavior of mobile devices when they are not connected are not typically recorded.

Ref.	Task	Reported performance	Deployment / evidence indicator	Interpretive value
[3]	Multi-crop classification	Accuracy 99.81%	Ensemble compute cost	Shows near-saturated controlled-benchmark performance
[8]	Citrus detection	mAP@0.5 92.5%	3.4 ms/image; 4.5 MB	Strongest explicit real-time edge profile
[10]	38-class explainable classification	Accuracy 99.69%; P 98.27%; R 98.26%	XAI included	High controlled-domain accuracy with interpretation
[27]	Multi-dataset diagnosis + recommendations	Accuracy 82–97%; precision 97%	0.0021 s average inference	Links recognition, efficiency, and advisory output
[28]	Rice disease + severity	Disease macro-F1 53.52%; severity macro-F1 77.79%; AUC 0.86	3,345 annotated images	Demonstrates realistic task difficulty
[30]	Field brinjal classification	Accuracy 99.12%; F1 97.76%	1,823 field images; Grad-CAM/++	Combines field data, performance, and XAI

[35]	Natural-field classification	Average accuracy/F1 >95.27%	26% parameter reduction; SHAP	Balances accuracy, efficiency, and interpretability
[39]	Hierarchical multi-crop classification	Accuracy 97.70%	13/38/51-class datasets	Supports hierarchical transformer design
[44]	Web Integrated diagnosis	Accuracy 94%; F1 0.94	Web deployment	Shows practical MobileViTv2 deployment
[50]	Jasmine lesion segmentation	mIoU 0.95; pixel accuracy 0.91	10,000 generated/curated images	Demonstrates precise segmentation with lightweight backbone

Table 10. Representative quantitative performance and deployment evidence from the reviewed corpus.

The original measurements that were specific to each job are preserved in Table 10, which does not require a cumulative ranking to be determined. In order to do a direct comparison between mAP [8, 28], macro-F1 [28], or mIoU [50], it is not possible to compare classification accuracy that is greater than 99% [3, 10, 30, 32]. [8] provides detailed information on model size and lag, [27] combines inference with suggestions, [30] includes field data and XAI, and [35] discusses both accuracy and parameter reduction. The comparison that is more informative has more than one dimension for comparative purposes. Within the context of evaluating anything, this pattern enables the utilization of capacity groups rather of a single scalar score. In light of this view, which is based on capabilities, a second question arises that goes beyond the quantity of performance: how frequently does the literature truly disclose the evidence that is required to support deployment claims? The answer to such question can be found in Section 4.5 of the corpus documentations.

4.5 Readiness Gaps

On the evidence map, there is a consistent unevenness that stands out. There were 23 studies that clearly looked at more than one crop or cross-crop, but only 22 of them used field or external validation, 17 of them reported a latency/model-size indicator, 14 of them estimated severity or lesion localization, 13 of them included explainability, and 18 of them provided a public dataset, code artifact, or sufficiently reusable data source. These two groupings are comprised of one another. By using a much smaller intersection, it is possible to fulfilll all of the conditions simultaneously. The literature contains a great number of robust models, but there are not a great number of comprehensive and dependable systems.

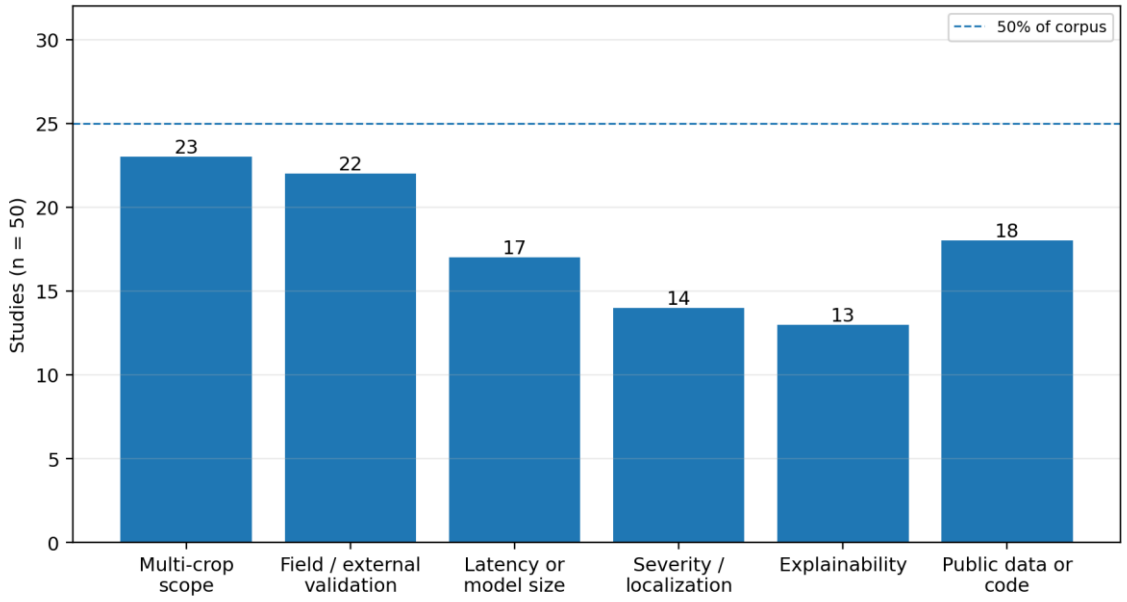


Figure 5. Frequency of deployment-relevant evidence in the included corpus.

The readiness imbalance is depicted in Figure 5, which is based on the corpus scale. Fourty-six percent of the fifty studies

examine the multi-crop scope, and forty-four percent examine the field or external evaluation. 34% of respondents report on latency and model size. In 28 percent of cases, there is a severity or localization, and in 26 percent of cases, there is explainability. Furthermore, it demonstrates that deployment evidence is still only utilized by a limited number of individuals, even in a new body of literature that contains a great deal of material that is technically advanced.

Readiness indicator	Studies	Coverage	Evidence deficit	System implication
Explicit multi-crop / cross-crop scope	23	46%	27 studies	Breadth across crops remains incomplete
Field or external validation	22	44%	28 studies	Laboratory-to-field generalization is under-tested
Latency or model-size evidence	17	34%	33 studies	Deployment cost is commonly under-reported
Severity or lesion localization	14	28%	36 studies	Many systems cannot quantify disease burden
Explainability	13	26%	37 studies	Decision transparency remains sparse
Public data/code or reusable source	18	36%	32 studies	Reproducibility and independent validation are limited

Table 11. Coverage and evidence deficit for deployment-relevant criteria across the 50-study corpus.

The readiness numbers are converted into percentages, and Table 11 displays the gaps that have been identified. Explainability and severity/localization are the two areas that present the greatest challenges. On the other hand, the problem with efficiency reporting, which accounts for 66% of the total, is equally as significant because real systems must function on hardware that is either mobile, embedded, or on the edge. In addition, the chart demonstrates that "multi-crop" and "deployment ready" are not synonymous terms. The crop width is only indicated in a little more than half of the papers that were examined. The connection between descriptive data and critical synthesis is made possible by these defects. It is not the question of which model family has the best headline score; rather, the question is which combination of families can compensate for the shortcomings that are not being tested out to their full potentials.

5. CRITICAL SYNTHESIS OF MODEL CLASSES

5.1 CNN and Transfer-Learning Models

The fact that local convolutions can mimic the shape of spots, streaks, necrotic edges, and variations in texture is one of the reasons why CNNs continue to be interesting. It is possible to achieve excellent results with transfer learning [46] when crop-specific samples are very small [4, 5, 16, 21, 23]. There is a reduction in the cost of implementation when using lightweight variants such as DLMC-Net [34], RTR_Lite_MobileNetV2 [43], and ShuffleNet-oriented UAV models [14]. Not the process of feature extraction itself is flawed; rather, the problem is in the degree to which it is dependent on the information. CNN journalists who have been educated on backgrounds that are identical to one another are able to learn color and frame shortcuts that are unique to a crop. As a result of this, the utilization of several crops requires a substantial background diversity, class-balanced sample, crop-aware brains, and evidence from credible outside sources. Moreover, this dependence on image framing explains why categorization alone is not sufficient in situations when there are no limitations of any kind. There is a clear connection between this and the subsequent two methods, which are segmentation and localization methods.

5.2 Object Detection and Segmentation

When there is more than one leaf on a frame, when there are minor lesions, when there is a complicated background, or when the severity needs to be determined, detection and segmentation are required. A few examples of techniques that are compatible with one another include the YOLO-based work [8], the attention residual U-Net [6], field corn segmentation [33], hierarchical residual ViT plus semantic segmentation [39], and the jasmine MobileNetV4-UNet [50]. Detector classification is typically superior to whole image classification when it comes to mobile field image samples. This is due to the fact that detectors

eliminate errors in disease categorization that are caused by errors in localization process. The area of the lesion can be determined using pixel-level segmentation, however it takes a significant amount of annotations. A suitable compromise would be to use masks or bounding boxes that are only lightly overseen. Once the areas that are diagnostically significant have been identified, the subsequent task is to demonstrate both the fine texture of the tumor as well as the larger picture of the accompanying symptoms. Transformer architectures and hybrid architectures are utilized for this set of reasons.

5.3 Vision Transformers and Hybrid Networks

Transformers are able to identify long-distance connections and differentiate between symptom patterns that are spread out, but pure ViTs typically require more data and processing resources than compact CNNs do. As a result of this, the majority of the recent research focuses on hybrid networks, including PD ViCo [30], MobileViTv2 [44], multitask ViT [17], residual ViT [39], and TLMViT [4]. One of its strengths is that they are able to blend the extraction of local textures with the context of the globe. There is a possibility that they will be overly complicated. In spite of the fact that a model may be referred to as "lightweight," it may nevertheless incur significant expenses when it is maintained at a high resolution. Therefore, accuracy should be linked with FLOPs, measured latency, memory, and battery energy. On its own, however, representational capacity is not sufficient to render a prediction dependable; model agreement, explanation quality, and calibrated uncertainty are the subsequent layers of evidence that are required to assist in the decision-making process.

5.4 Ensembles, Explainability, and Uncertainty

Ensembles make things more stable on account of the fact that they average out various inductive biases [3, 11, 12, 37, 40]. Notwithstanding this, naive ensembles result in increased memory and latency. Thru the process of distillation, a group of musicians can be reduced to a single student [15]. There has been an increase in the utilization of reasonableness tools such as LIME, SHAP, Grad-CAM, and Grad-CAM++ [10, 12, 24, 29, 30, 35, and 40]. In comparison to lesion comments or the opinion of an expert, visual explanations are only relevant in some situations. It is not sufficient evidence to have a vivid heatmap that depicts the leaf in a large field of view. Uncertainty is perhaps more fundamental. Multi-crop systems require probabilities that are calibrated, a choice between "abstain" and "unknown," and monitoring after deployment. This is because to the fact that a confident inaccuracy on a disease that has not yet been observed can result in the incorrect treatment being administered. In the event that the crop, region, device, or disease class undergoes a change, even a closed-set model that has been adequately explained and calibrated may fail. The reason for this is because eventually, robustness will need to be expanded to include domain shift and conditions that have never been encountered before.

5.5 Domain Generalization, Zero-Shot Learning, and Open-Set Diagnosis

In order to effectively manage the growth of multiple crops, it is not the most effective method to continually retrain a flat classifier. All of these things lead in the direction of something that can be employed on a broader scale: zero-shot semantic characteristics [18], cross-domain analysis [1], multi-granularity alignment [42], cross-crop screening [41], and field-focused datasets [2, 13, 30]. It is essential to have a thorough understanding of the nature of the ailment, the symptoms, the organ, and the type of crop that has to be treated. This type of system is able to communicate when a sample does not correspond to one of its training classes, identify potential diseases by making use of semantic descriptions, and request confirmation from an expert rather than imposing a closed-set name on the sample. Consequently, Sections 5.1–5.5 discuss skills that complement one another rather than competing with one another at the same timestamp sets. The comparison of various responsibilities within a single system design is presented in Table 12, and Figure 6 illustrates the relative strengths of each role in comparison to the others.

Model family	Primary strength	Compute cost	Edge suitability	Interpretability	Domain robustness	Best role in integrated system
CNN / transfer learning	Texture and lesion morphology	Low–moderate	High with compact backbones	Low unless XAI added	Weak–moderate	Shared local feature encoder
Detection / segmentation	Spatially grounded diagnosis	Moderate	Moderate–high	Moderate	Moderate	Leaf/lesion localization and severity

CNN–Transformer hybrid	Local + global context	Moderate–high	Moderate	Moderate	Moderate	Hierarchical diagnosis backbone
Ensemble / XAI	Prediction stability + explanation	High	Low–moderate	High	Moderate	Teacher model / explanation layer
Lightweight / edge	Fast compact inference	Low–moderate	Very high	Low–moderate	Moderate	On-device student model
Generalization / zero-shot	Unseen domain/class handling	Moderate	Moderate	Moderate	High	Unknown detection and cross-domain adaptation

Table 12. Critical comparison of model families for integrated multi-crop disease-detection system design.

The architectural choice matrix that is presented in Table 12 is a compilation of the components that came before it. Due to the fact that every family is in a unique position with regard to accuracy, localization, interpretability, efficiency, and strength, none of them are sufficient on their own. CNNs and hybrid transformers perform the best when used as shared encoders; detection and segmentation models are helpful when grounding in space; ensembles can be used as teachers or verification models; lightweight networks perform well when used for edge execution; and generalization methods are required to ensure safe behavior when domains shift or when new disease classes are observed. After that, Figure 6 illustrates the matrix in such a way that the trade-offs between recognition, localization, context modeling, interpretability, efficiency, and generalization may be viewed as a coordinated capability profile rather than as discrete model attributes.

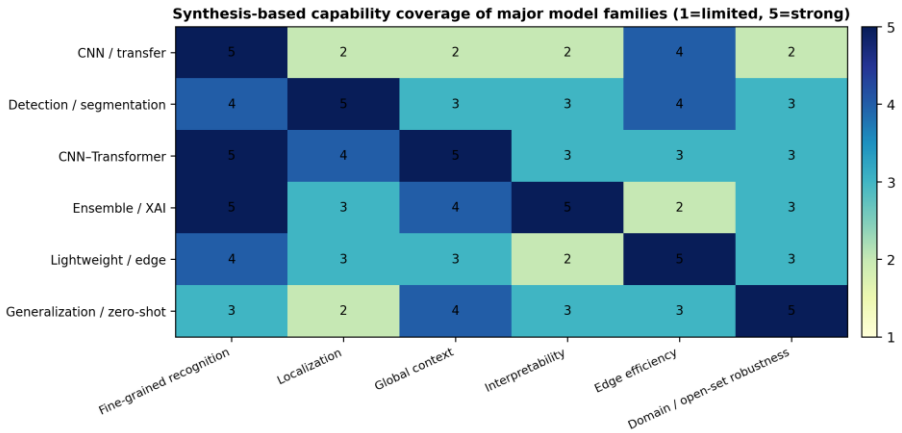


Figure 6. Synthesis-based capability matrix of the six dominant model families.

A graphic that corresponds to Table 12 is referred to as Figure 6 in this text. It's not just one architecture that has a lot of high scores; CNN/transfer models do the greatest job of fine-grained recognition, detection/segmentation models do the best job of localization, CNN–Transformer hybrids do the best job of dealing with global context, and ensemble/transfer models do the best job of dealing with global context."XAI models are the most effective when it comes to interpretability, lightweight models are the most effective when it comes to edge efficiency, and generalization and zero-shot approaches are the most effective when it comes to domain or open-set resilience. For this reason, the matrix is capable of supporting a modular, crop-aware architecture, in which each family is responsible for handling a distinct kind of failure. This complementarity provides us with the design rationale for Section 6: the reference architecture does not select one family as being superior to all others; rather, it places each feature on the stage where it can remedy a particular operational failure condition & process.

6. REFERENCE ARCHITECTURE FOR AN EFFICIENT MULTI-CROP DISEASE DETECTION SYSTEM

The evidence that supports a single end-to-end classifier is not as strong as the data that supports a five-stage crop-aware

designed process. The design depicted in Figure 7 is not a claim of performance data that were just taken; rather, it is a collection of those measurements. It eliminates the typical issues that impede down deployment while retaining the sections of the literature review that have the strongest supports.

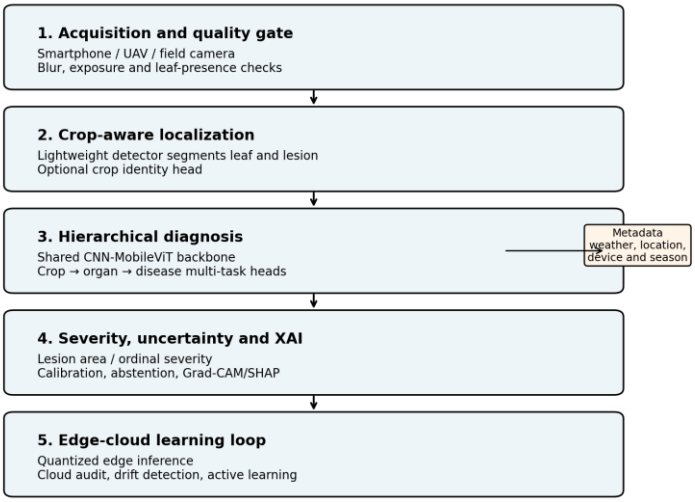


Figure 7. Evidence-derived reference architecture for efficient multi-crop disease detection.

By transforming the evidence synthesis into a sequence that may be utilized, Figure 7 is presented. Quality gating prevents unreliable inputs from entering the diagnosis process; crop-aware localization separates scene parsing from disease reasoning; the hierarchical backbone allows for the sharing of representations across crops; severity, calibration, abstention, and XAI transform prediction into a more secure decision; and the edge–cloud loop ensures that everything runs smoothly while allowing you to track drift and update the model in a controlled manner. Modularity is incorporated into the design of the architecture so that individual components can be modified independently as additional evidence from the field is gathered. There are unambiguous input-output agreements displayed in Table 13, which depicts the visual pipeline. Thru this process, the architectural flow is broken down into components that are capable of being trained, tested, altered, and audited independently for this.

Stage	Recommended component	Input	Output	Design rationale
1. Quality gate	MobileNet-class quality model; leaf-presence detector	Image, camera metadata	Accept/reject, quality score	Prevents confident diagnosis from unusable images
2. Crop-aware localization	Nano/small YOLO or RT-DETR; weakly supervised lesion mask	Accepted image	Leaf boxes, lesion regions, crop probability	Separates scene parsing from diagnosis
3. Hierarchical diagnosis	Shared EfficientNetV2/MobileViT backbone with crop→organ→disease heads	Crops/lesions + context	Disease probabilities and embedding	Reduces class confusion; shares features across crops
4. Severity and trust	Ordinal severity head; temperature scaling; conformal/entropy abstention; Grad-	Disease logits and lesion mask	Severity, confidence, unknown flag, explanation	Supports treatment decisions and safe deferral

	CAM/SHAP			
5. Edge–cloud loop	INT8 quantization; optional distillation; drift detector; active learning	Edge predictions and consented audit samples	Updated model and regional calibration	Maintains performance across seasons and devices

Table 13. Functional specification of the proposed reference architecture.

It is demonstrated in Table 13 that the input-output contract for each stage of the design process. This indicates that the concept can be put to the test. Everything from accepting or rejecting quality scores and locating lesions to determining illness probabilities, severity levels, unknown flags, explanations, and updates to the regional model can be monitored. This includes everything from accepting or rejecting quality ratings to located lesions. Instead of concealing all errors within a single end-to-end classifier, the staged architecture makes it simpler to determine what went wrong with the acquisition, localization, diagnosis, calibration, or deployment of the data samples. In order for these modular outputs to operate together while learning, the design needs to have a unified goal that strikes a compromise between crop recognition, illness discrimination, severity ordering, spatial anchoring, calibration, and compressions. A formal declaration of this link can be found in Section 6.1 of this text.

6.1 Hierarchical Objective

While the crop head restricts the diseases that can be discovered, the disease head performs fine-grained recognition, the ordinal head takes into consideration the order of severity, and the segmentation head provides a spatial basis for the choice. Calibration loss and distillation are two factors that are connected to the reliability and efficiency of deployment. A single accuracy value cannot indicate whether or not localization, calibration, open-set safety, or edge efficiency have been unsuccessful for this process.

6.2 Evaluation Protocol

Dimension	Required metrics	Purpose
Recognition	Macro-F1, balanced accuracy, per-crop/per-disease recall	Avoids dominance by common crops/classes
Localization	mAP@0.5:0.95, Dice, mIoU	Measures lesion/leaf localization separately
Severity	Quadratic weighted kappa, macro-F1, MAE	Respects ordinal severity and clinical-style agreement
Calibration	ECE, Brier score, risk–coverage curve	Evaluates whether confidence is trustworthy
Generalization	Leave-one-region/device/season-out testing; cross-dataset transfer	Measures real distribution shift
Efficiency	On-device latency P50/P95, RAM, model size, energy/image, throughput	Reports actual deployment cost
Safety	Unknown-disease AUROC, false-treatment rate, abstention utility	Tests open-set and decision risk

Table 14. Minimum evaluation protocol for a trustworthy multi-crop system.

Macro-F1 and balanced accuracy are used to protect against class imbalance; mAP, Dice, and mIoU are used to separate spatial errors; kappa and MAE are used to show the order of severity; calibration metrics are used to determine whether or not confidence can be trusted; leave-one-region/device/season-out experiments are used to determine how much the distribution actually changes; and P50/P95 latency, RAM, energy, and throughput are used to measure the cost of deployment. In the final step of the technique, open-set AUROC and abstention utility are utilized to determine the level of safety in situations where unknown pathogens are present. When this more stringent criteria for evaluation is compared to the data that was presented in the corpus that was evaluated, the methodological holes that have not yet been closed become apparent. The research strategy presented in Section 7 is derived from these gaps.

7. Research Gaps and Future Agenda

Gap	Research direction
Cross-region benchmarks	Curated datasets should be supplemented with prospectively collected images from multiple regions, devices, seasons, cultivars, and disease stages. A model should not be called generalized when train and test images share the same acquisition campaign.
Open-set and mixed-stress diagnosis	Future datasets must include unseen diseases, insect injury, nutrient deficiency, herbicide damage, and co Infections. Unknown detection and top-k differential diagnosis are safer than forced single-label output.
Severity as an ordinal and spatial task	Severity labels should be tied to lesion area or expert scales, with inter-rater agreement reported. Classification-only systems cannot directly support treatment intensity or yield-loss estimation.
Calibration and selective prediction	Confidence calibration, abstention, and risk–coverage analysis should become standard. The practical objective is not maximum accuracy at 100% coverage, but minimum harmful error at a useful coverage level.
Transparent efficiency	Studies should report device model, input resolution, batch size, quantization level, P50/P95 latency, peak RAM, model size, and energy per image. FLOPs alone are insufficient.
Causal/contextual integration	Weather, crop stage, location, and outbreak history can improve ranking but may also create spurious shortcuts. Context should be fused after visual evidence and evaluated through ablation and out-of-region tests.
Continual learning governance	Active learning and periodic updates require versioning, audit trails, consented data collection, rollback, and checks for regional or crop-specific performance drift.
Reproducibility and reporting	Code, class distributions, data splits, duplicate Image checks, augmentation policies, seeds, confidence intervals, and failure examples should be disclosed. Mean±SD across repeated runs is preferable to a single best run.

Table 15. Research agenda derived from the evidence synthesis.

The gaps that directly prevent reliable field use are given the most weight in Table 15, which can be seen here. Not only does the research agenda include the collection of new architectures, but it also includes the collection of datasets from different regions in the future, the handling of unknown classes, the linking of severity labels to spatial evidence, calibrated selective prediction, clear energy/latency reporting, controlled contextual fusion, continuous-learning governance, and the reporting of experiments that can be repeated. By achieving these objectives, the industry will transition from benchmark optimization to providing farmers with auditable decision support. In addition to this strategy for the future, Figure 8 provides an additional addition by returning to the observed corpus and demonstrating the degree to which the current reporting is not yet fully ready across the same deployment dimensions in this process.

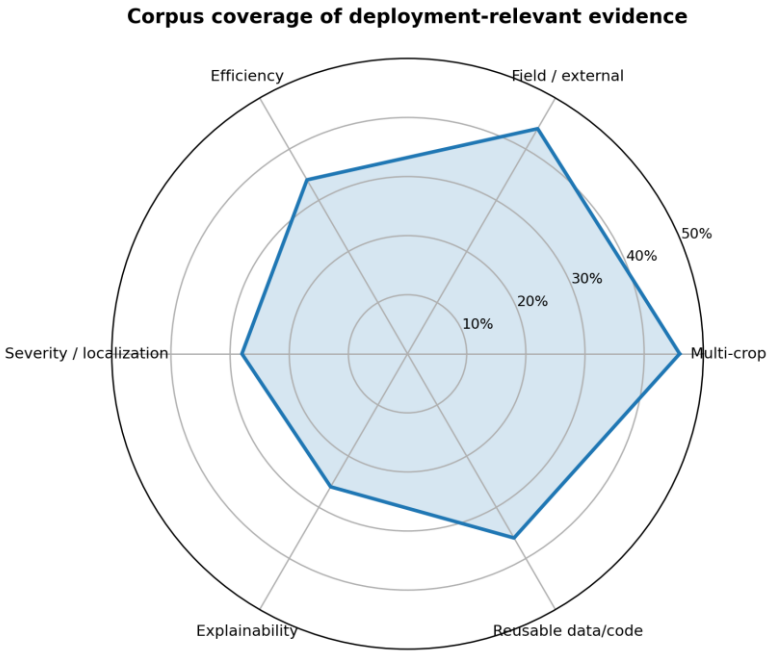


Figure 8. Radar profile of deployment-relevant evidence coverage across the reviewed corpus.

A different representation of the readiness counts is presented in Figure 8, which is a coverage map. The polygon remains contained within the 50% ring along all of the axes. In terms of explainability, it has the lowest coverage (26%), whereas in terms of severity and location, it has the highest coverage (28%). Only 34% of the points are awarded for demonstration of efficiency, while 36% are awarded for proof of data or code that may be used again. Even with field/external validation and explicit multi-crop scope, it is still not possible to cover more than half of the corpus. Due to this, the shortened radar version demonstrates the most important finding of the evaluation, which is that the capability of the model has increased more rapidly than the validation of the entire deployment. This review is helpful in pointing researchers in the proper path; yet, its strength must be recognized within the confines of the review itself for this process. In light of this, the following section will discuss the limitations of the review in terms of its scope and comparability before arriving at any conclusive findings.

8. LIMITATIONS OF THE REVIEW

The only publishing groups that are taken into consideration in this evaluation are just four. Neither Scopus, Web of Science, PubMed, arXiv, MDPI, nor Frontiers are included in its list of publications that it claims to have combed thru. As a result of the fact that publisher search tools have the ability to alter the ranking and indexing of articles, it is possible that some 2026 content will appear after the cutoff date of August 9, 2026. It was not possible to conduct a meta-analysis on the performance values because the datasets, splits, classes, tasks, and metrics were all distinct from one another. In the event that they were averaged, the pooled estimate that was produced would be inaccurate. One should not consider the quality map to be a proven risk-of-bias scale; rather, it is an evaluation of how well prepared the system is. In conclusion, there are some publisher pages that merely provide brief listings of writers. It is true that the synthesis is restricted in its extent and its capacity to be statistically compared; nonetheless, this does not alter the primary finding that it presents, which is that deployment-oriented evidence consistently lags behind advancements in recognition architecture across a confirmed corpus of recent studies & articles.

9. CONCLUSION

With PRISMA-based syntheses of relatively new peer-reviewed articles from 2023–2026, it becomes clear that multi-crop disease detection has moved beyond questions of if deep networks can identify pre-processed images of leaves. The real issue now is: Will a network continue to be useful when the crop being identified, the camera used to take the photo, the background in the image, where the plant is located geographically, and/or what type of computing resources are available all change at the same time? The research corpus establishes a way to measure this gap: Only 23 of 50 studies (46%), stated that they had either included multiple crops or multiple types of crops. Twenty-two (44%), utilized some form of field testing/external validation. Seventeen (34%), reported the latency or size of their models. Fourteen (28%), documented disease severity or location of

lesions. Thirteen (26%), integrated forms of explainability. Eighteen (36%) made their code, datasets, or enough information available so other researchers could reproduce their work. Therefore, there is a significant difference between achieving high levels of performance based on benchmarks and developing systems that have been tested and validated to function in real-world environments. While the literature has reached a level of maturity regarding feature extraction using deep networks, the literature remains uneven as to the amount of evidence needed to demonstrate trustworthy use.

This comparative review indicates that there is no single type of model that can resolve this challenge independently. CNNs, Transfer-Learning Models: These models are very good at extracting local features. Detection and Segmentation: These approaches allow researchers to quantify symptoms by providing the necessary spatial structure. CNN-Transformer Hybrid Models: These models bring symptom context into view globally. Ensemble/XAI Methods: These methods enhance both verification and interpretation. Lightweight Networks: These networks enable researchers to execute models at the edge. Generalization or Zero-Shot Methods: These models address novel domains/classes. Example results show why it would be better to consider these functionalities collectively instead of ranking them using a single metric: Light-SA YOLOv8 demonstrated 92.5% mean Average Precision @ 0.5 with an inference time of 3.4 ms and model size of 4.5 MB [8]. PlantCareNet showed an average inference time of 0.0021 seconds while integrating diagnosis with recommendations [27]. PD ViCo demonstrated 99.12% accuracy and 97.76% F1 score on field-harvested Brinjal images via Grad-CAM-based interpretation [30]. However, the study on Rice Severity reported disease macro-F1 = 53.52%, and severity macro-F1 = 77.79%, indicating how much harder diagnosing diseases is when you extend your diagnostic abilities beyond simply identifying curated classifications.

Together, the findings suggest a hierarchical, crop-specific, Edge-Cloud architecture in which quality of input images is evaluated before evaluating locations of leaves and lesions; after shared CNN-Mobile-ViT-type representations support crop/disease/severity heads, calibration, abstention, and visual explanations determine whether predictions should be followed-up or deferred. Optional Distillation and Quantization may enable the ability to move this capability onto constrained-resource devices. Cloud-Side Drift Detection and Active Learning may help maintain relevance across regions and seasons. Therefore, future improvements in multi-crop disease detection should be determined by Cross-Region and Cross-Device Generalization, Open-Set Recognition, Calibrated Risk/Coverage Behavior, Spatially Grounded Severity Reporting, Reproducible Evaluation Studies, and Measured On-Device Latency/Memory/Energy—Not Simply Marginally Better Performance on Familiar Splits. An actually effective multi-crop disease-detection system will be able to accurately detect diseases under favorable conditions; transparent about its uncertainty when those conditions are not; computationally feasible at the edge of farms; and auditable enough to inform agricultural decision-making responsibly for this process.

Declarations

Funding: No external funding was declared for preparation of this review.

Conflict of interest: The review reports no conflict of interest.

Data availability: The extracted evidence is contained in Tables 6–15 and the cited publisher records. No proprietary dataset was created.

Review date: Searches and verification were completed through 9 August 2026.

REFERENCES

[1] Ahmad, A., El Gamal, A., & Saraswat, D. (2023). Toward generalization of deep learning-based plant disease identification under controlled and field conditions. IEEE Access, 11, 9042–9057. https://doi.org/10.1109/ACCESS.2023.3240100

[2] Moupojou, E., Tagne, A., Retraint, F., Tadonkemwa, A., Wilfried, D., Tapamo, H., & Nkenlifack, M. (2023). FieldPlant: A dataset of field plant images for plant disease detection and classification with deep learning. IEEE Access, 11, 35398–35410. https://doi.org/10.1109/ACCESS.2023.3263042

[3] Shovon, M. S. H., Mozumder, S. J., Pal, O. K., Mridha, M. F., Asai, N., & Shin, J. (2023). PlantDet: A robust multi-model ensemble method based on deep learning for plant disease detection. IEEE Access, 11, 34846–34859. https://doi.org/10.1109/ACCESS.2023.3264835

[4] Tabbakh, A., & Barpanda, S. S. (2023). A deep features extraction model based on the transfer learning model and Vision Transformer 'TLMViT' for plant disease classification. IEEE Access, 11, 45377–45392. https://doi.org/10.1109/ACCESS.2023.3273317

[5] Masood, M., Nawaz, M., Nazir, T., Javed, A., Alkanhel, R., Elmannai, H., Dhahbi, S., & Bourouis, S. (2023). MaizeNet: A deep learning approach for effective recognition of maize plant leaf diseases. IEEE Access, 11, 52862–52876. https://doi.org/10.1109/ACCESS.2023.3280260

[6] Abinaya, S., Kumar, K. U., & Alphonse, A. S. (2023). Cascading autoencoder with attention residual U-Net for multi-class plant leaf disease segmentation and classification. IEEE Access, 11, 98153–98170. https://doi.org/10.1109/ACCESS.2023.3312718

[7] Chelloug, S. A., Alkanhel, R., Muthanna, M. S. A., Aziz, A., & Muthanna, A. (2023). MULTINET: A multi-agent DRL and EfficientNet assisted framework for 3D plant leaf disease identification and severity quantification. IEEE Access, 11, 86770–86789. https://doi.org/10.1109/ACCESS.2023.3303868

[8] Luo, D., Xue, Y., Deng, X., Yang, B., Chen, H., & Mo, Z. (2023). Citrus diseases and pests detection model based on self-attention YOLOv8. IEEE Access, 11, 139872–139881. https://doi.org/10.1109/ACCESS.2023.3340148

[9] Hosny, K. M., El-Hady, W. M., Samy, F. M., Vrochidou, E., & Papakostas, G. A. (2023). Multi-class classification of plant leaf diseases using feature fusion of deep convolutional neural network and local binary pattern. IEEE Access, 11, 62307–62317. https://doi.org/10.1109/ACCESS.2023.3286730

[10] Nigar, N., Faisal, H. M., Umer, M., Oki, O., & Lukose, J. M. (2024). Improving plant disease classification with deep-learning-based prediction model using explainable artificial intelligence. IEEE Access, 12, 100005–100014. https://doi.org/10.1109/ACCESS.2024.3428553

[11] Maurya, R., Mahapatra, S., & Rajput, L. (2024). A lightweight meta-ensemble approach for plant disease detection suitable for IoT-based environments. IEEE Access, 12, 28096–28108. https://doi.org/10.1109/ACCESS.2024.3367443

[12] Oad, A., Abbas, S. S., Zafar, A., Akram, B. A., et al. (2024). Plant leaf disease detection using ensemble learning and explainable AI. IEEE Access, 12, 156038–156049. https://doi.org/10.1109/ACCESS.2024.3484574

[13] Joseph, D. S., Pawar, P. M., & Chakradeo, K. (2024). Real-time plant disease dataset development and detection of plant disease using deep learning. IEEE Access, 12, 16310–16333. https://doi.org/10.1109/ACCESS.2024.3358333

[14] Pajany, M., Venkatraman, S., Sakthi, U., Sujatha, M., & Ishak, M. K. (2024). Optimal fuzzy deep neural networks-based plant disease detection and classification on UAV-based remote sensed data. IEEE Access, 12, 162131–162144. https://doi.org/10.1109/ACCESS.2024.3488751

[15] Hasan, M. J., Mazumdar, S., & Momen, S. (2025). Deployable deep learning for cross-domain plant leaf disease detection via ensemble learning, knowledge distillation, and quantization. IEEE Access, 13, 140313–140336. https://doi.org/10.1109/ACCESS.2025.3595390

[16] Shafik, W., Tufail, A., De Silva Liyanage, C., & Apong, R. A. A. H. M. (2024). Using transfer learning-based plant disease classification and detection for sustainable agriculture. BMC Plant Biology, 24, 136. https://doi.org/10.1186/s12870-024-04825-y

[17] Hemalatha, S., & Jayachandran, M. (2024). A multitask learning-based Vision Transformer for plant disease localization and classification. International Journal of Computational Intelligence Systems. https://doi.org/10.1007/s44196-024-00597-3

[18] Kumar, P., Mathew, J., Sanodiya, R. K., Setty, T., & Bhaskarla, B. P. (2024). Zero shot plant disease classification with semantic attributes. Artificial Intelligence Review, 57, 305. https://doi.org/10.1007/s10462-024-10950-9

[19] Baiju, B. V., Kirupanithi, N., Srinivasan, S., Kapoor, A., Mathivanan, S. K., et al. (2025). Robust CRW crops leaf disease detection and classification in agriculture using hybrid deep learning models. Plant Methods, 21, 18. https://doi.org/10.1186/s13007-025-01332-5

[20] Sinamenye, J. H., et al. (2025). Potato plant disease detection: Leveraging hybrid deep learning models for enhanced accuracy. BMC Plant Biology. https://doi.org/10.1186/s12870-025-06679-4

[21] Sun, H., Fu, R., Wang, X., Wu, Y., Al-Absi, M. A., Cheng, Z., Chen, Q., et al. (2025). Efficient deep learning-based tomato leaf disease detection through global and local feature fusion. BMC Plant Biology, 25, 311. https://doi.org/10.1186/s12870-025-06247-w

[22] Thanjaivadivel, M., et al. (2025). EnConv: Enhanced CNN for leaf disease classification. Journal of Plant Diseases and Protection, 132, 32. https://doi.org/10.1007/s41348-024-01033-6

[23] Srinivasan, S., et al. (2025). Sugarcane leaf disease classification using deep neural network with transfer learning. BMC Plant Biology. https://doi.org/10.1186/s12870-025-06289-0

[24] Shafik, W., et al. (2025). Deep learning technique for plant disease classification and pest detection and model explainability elevating agricultural sustainability. BMC Plant Biology, 25, 1491. https://doi.org/10.1186/s12870-025-07377-x

[25] Elfouly, M. K., et al. (2025). A deep learning-based framework for large-scale plant disease detection and severity classification. Journal of Big Data, 12. https://doi.org/10.1186/s40537-025-01265-9

[26] Pai, P., et al. (2025). A twin CNN-based framework for optimized rice leaf disease classification. Journal of Big Data, 12, 89. https://doi.org/10.1186/s40537-025-01148-z

[27] Islam, M., Azad, A. K. M., Arman, S. E., Alyami, S. A., et al. (2025). PlantCareNet: An advanced system to recognize plant diseases with dual-mode recommendations for prevention. Plant Methods. https://doi.org/10.1186/s13007-025-01366-9

[28] Roy, P. S., & Kukreja, V. (2025). Vision transformers for rice leaf disease detection and severity estimation: A precision agriculture approach. Journal of the Saudi Society of Agricultural Sciences. https://doi.org/10.1007/s44447-025-00007-w

[29] Shafik, W., et al. (2026). Lightweight Vision Transformer and ResNet-9 models for real-time plant disease detection and pest classification with SHAP explainability. BMC Plant Biology. https://doi.org/10.1186/s12870-026-08667-8

[30] Bitto, A. K., et al. (2026). PD ViCo: An explainable AI-based contrastive captioner Vision Transformer with patch dropout for multi-class brinjal disease classification. BMC Plant Biology. https://doi.org/10.1186/s12870-026-09330-y

[31] Islam, M. M., Adil, M. A. A., Talukder, M. A., Ahamed, M. K. U., Uddin, M. A., Hasan, M. K., Sharmin, S., Rahman, M. M., & Debnath, S. K. (2023). DeepCrop: Deep learning-based crop disease prediction with web application. Journal of Agriculture and Food Research, 14, 100764. https://doi.org/10.1016/j.jafr.2023.100764

[32] Dai, G., Fan, J., Tian, Z., & Wang, C. (2023). PPLC-Net: Neural network-based plant disease identification model supported by weather data augmentation and multi-level attention mechanism. Journal of King Saud University - Computer and Information Sciences, 35(5), 101555. https://doi.org/10.1016/j.jksuci.2023.101555

[33] Divyanth, L. G., Ahmad, A., & Saraswat, D. (2023). A two-stage deep-learning based segmentation model for crop disease quantification based on corn field imagery. Smart Agricultural Technology, 3, 100108. https://doi.org/10.1016/j.atech.2022.100108

[34] Sharma, V., Tripathi, A. K., & Mittal, H. (2023). DLMC-Net: Deeper lightweight multi-class classification model for plant leaf disease detection. Ecological Informatics, 75, 102025. https://doi.org/10.1016/j.ecoinf.2023.102025

[35] Dai, G., Tian, Z., Fan, J., Sunil, C. K., & Dewi, C. (2024). DFN-PSAN: Multi-level deep information feature fusion extraction network for interpretable plant disease classification. Computers and Electronics in Agriculture, 216, 108481. https://doi.org/10.1016/j.compag.2023.108481

[36] Ritharson, P. I., Raimond, K., Mary, X. A., Robert, J. E., & Andrew, J. (2024). DeepRice: A deep learning and deep feature based classification of rice leaf disease subtypes. Artificial Intelligence in Agriculture, 11, 34–49. https://doi.org/10.1016/j.aiia.2023.11.001

[37] Ali, A. H., Youssef, A., Abdelal, M., & Raja, M. A. (2024). An ensemble of deep learning architectures for accurate plant disease classification. Ecological Informatics, 81, 102618. https://doi.org/10.1016/j.ecoinf.2024.102618

[38] Duhan, S., Gulia, P., Gill, N. S., & Narwal, E. (2024). Investigating attention mechanisms for plant disease identification in challenging environments. Heliyon, 10(9), e29802. https://doi.org/10.1016/j.heliyon.2024.e29802

[39] Vallabhajosyula, S., Sistla, V., & Kolli, V. K. K. (2024). A novel hierarchical framework for plant leaf disease detection using residual Vision Transformer. Heliyon, 10(9), e29912. https://doi.org/10.1016/j.heliyon.2024.e29912

[40] Raval, H., & Chaki, J. (2024). Ensemble transfer learning meets explainable AI: A deep learning approach for leaf disease detection. Ecological Informatics, 84, 102925. https://doi.org/10.1016/j.ecoinf.2024.102925

[41] Bouacida, I., Farou, B., Djakhdjakha, L., Seridi, H., & Kurulay, M. (2025). Innovative deep learning approach for cross-crop plant disease detection: A generalized method for identifying unhealthy leaves. Information Processing in Agriculture, 12(1), 54–67. https://doi.org/10.1016/j.inpa.2024.03.002

[42] Guo, G., et al. (2025). Multi-granularity alignment for crop diseases detection. Plant Phenomics, 7(1), 100001. https://doi.org/10.1016/j.plaphe.2025.100001

[43] Duhan, S., Gulia, P., Gill, N. S., & Narwal, E. (2025). RTR_Lite_MobileNetV2: A lightweight and efficient model for plant disease detection and classification. Current Plant Biology, 42, 100459. https://doi.org/10.1016/j.cpb.2025.100459

[44] Kheir, A. M. S., Koubaa, A., Kolluru, V., Mungara, S., & Feike, T. (2025). Smart plant disease diagnosis using multiple deep learning and web application integration. Journal of Agriculture and Food Research, 21, 101948. https://doi.org/10.1016/j.jafr.2025.101948

[45] Akram, T., et al. (2026). A sustainable approach to plant disease detection using a deep learning framework for energy-efficient agricultural consumer electronics. Information Processing in Agriculture. https://doi.org/10.1016/j.inpa.2026.05.003

[46] Bouni, M., Hssina, B., Douzi, K., & Douzi, S. (2023). Impact of pretrained deep neural networks for tomato leaf disease prediction. Journal of Electrical and Computer Engineering, 2023, 5051005. https://doi.org/10.1155/2023/5051005

[47] Balaji, V., Anushkannan, N. K., et al. (2023). Deep transfer learning technique for multimodal disease classification in plant images. Contrast Media & Molecular Imaging, 2023, 5644727. https://doi.org/10.1155/2023/5644727

[48] Pandian, J. A., Nisha, S. N., Kanchanadevi, K., Pandey, A. K., & Rima, S. K. (2023). Grey blight disease detection on tea leaves using improved deep convolutional neural network. Computational Intelligence and Neuroscience, 2023, 7876302. https://doi.org/10.1155/2023/7876302

[49] Malathi, V., Gopinath, M. P., Kumar, M., Bhushan, S., Jayaprakash, S., & Jan, S. U. (2023). Enhancing the paddy disease classification by using cross Validation strategy for artificial neural network over baseline classifiers. Journal of Sensors, 2023, 1576960. https://doi.org/10.1155/2023/1576960

[50] Shwetha, V., Bhagwat, A., Laxmi, V., & Shrivastava, S. (2024). A custom backbone UNet framework with DCGAN augmentation for efficient segmentation of leaf spot diseases in jasmine plant. Journal of Computer Networks and Communications, 2024, 5057538. https://doi.org/10.1155/2024/5057538

[51] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

[52] Pacal, I., Kunduracioglu, I., Alma, M. H., Deveci, M., Kadry, S., Nedoma, J., Slany, V., & Martinek, R. (2024). A systematic review of deep learning techniques for plant diseases. Artificial Intelligence Review, 57, 304. https://doi.org/10.1007/s10462-024-10944-7

[53] Demilie, W. B. (2024). Plant disease detection and classification techniques: A comparative study of the performances. Journal of Big Data, 11. https://doi.org/10.1186/s40537-023-00863-9

[54] George, R., Thuseethan, S., Ragel, R. G., Mahendrakumaran, K., Nimishan, S., Wimalasooriya, C., & Alazab, M. (2025). Past, present and future of deep plant leaf disease recognition: A survey. Computers and Electronics in Agriculture, 234, 110128. https://doi.org/10.1016/j.compag.2025.110128

[55] Ahmad, A., Saraswat, D., & El Gamal, A. (2023). A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools. Smart Agricultural Technology, 3, 100083. https://doi.org/10.1016/j.atech.2022.100083