Deep Learning Approaches for Apple Disease Classification and Detection: A PRISMA 2020 Systematic Literature Review

Brahim Ouben Hssain[1], Khalil Ladrham[2], Noureddine El Barbri[3], Rachid El Ayachi[4]
Laboratory of Science and Technology for the Engineer (LaSTI), ENSA,
Sultan Moulay Slimane University, Khouribga, Morocco[1, 3]
Intelligent Processing and Security of Systems-Faculty of Sciences, Mohammed V University, Rabat, Morocco[2]
TIAD Laboratory-Department of Computer Science-Faculty of Sciences and Technology,
Sultan Moulay Slimane University, Morocco[4]

Abstract*—Apple diseases cause great losses in fruit yield and quality, while manual diagnosis is time-consuming, subjective, and difficult to scale across orchards. Deep learning has thus become a staple in image-based disease recognition, but evidence remains scattered across classification and object-detection paradigms, heterogeneous datasets, and inconsistent evaluation protocols. This systematic literature review integrates 30 peer-reviewed studies published between 2020 and 2026, selected from 180 database records using a PRISMA 2020-aligned protocol. The final corpus includes 22 image-classification studies and eight object-detection studies. One non-peer-reviewed preprint was kept as contextual evidence only and was excluded from all corpus counts and comparative analyses. Architectures, datasets, preprocessing and augmentation strategies, training configurations, evaluation metrics, evidence of generalization, and deployment characteristics are discussed separately for the two task paradigms. Reported classification accuracies vary from 91.0% to 99.99%, whereas detection studies report mAP@0.5 values from 82.1% to 99.99%; these ranges are descriptive and are not pooled estimates. These values are heavily dependent on dataset composition and evaluation design: controlled-background datasets often yield near-perfect scores, while field transfer can lead to a drop of almost 30 percentage points. CNNs still dominate the field, but hybrid transformer-based attention mechanisms, multi-scale feature fusion, class-imbalance-aware training, and lightweight YOLO variants are gaining traction. Long-standing limitations are the lack of reporting of hyperparameters, an over-reliance on accuracy, the lack of external validation, heterogeneity in annotation, the lack of uncertainty analysis, and limited reporting of latency, memory, and energy consumption. Accordingly, the research agenda emphasizes standardized real-orchard benchmarks, cross-dataset validation, calibrated and explainable predictions, reproducible experimental protocols, and deployment-aware model design.

Keywords*—Apple disease recognition; deep learning; convolutional neural networks; object detection; PRISMA 2020; systematic literature review; precision agriculture

I. Introduction

Apple (Malus domestica) is a high-value fruit crop, and its productivity is constrained by various fungal, bacterial, and viral diseases such as apple scab, powdery mildew, Alternaria leaf spot, Marssonina blotch, fire blight, and postharvest blue mold. Disease pressure reduces the marketable yield, accelerates defoliation, and may result in greater chemical use when diagnosis is late or inaccurate [1], [8]– [11]. Insect pests and mites such as codling moth, aphids, spider mites, and San Jose scale also affect orchard health and further complicate visual assessment and crop-management decisions [12]–[17]. Traditional diagnosis is based on field scouting and visual interpretation by farmers, agronomists, or plant pathologists. Expert inspection is still required but is labour-intensive and sensitive to symptom similarity, disease stage, illumination, occlusion, and observer experience. Image-based AI offers a non-destructive alternative that can learn discriminative patterns directly from leaf or fruit images. Therefore, convolutional neural networks (CNNs), vision transformers, hybrid CNN-transformer models, and object detectors have been used in precision-agriculture applications [6], [7].

The fast development of the field has created an evaluation problem. In classification studies, the objective is to assign one label to an image. Accuracy, precision, recall, and F1-score are the most common metrics reported in these studies. Detection studies also localize lesions, or diseased leaves, and are mostly evaluated in terms of mean average precision. Mixing these task types in one ranking is methodologically unsound because the outputs, annotation requirements, and error structures differ. In this review, recognition is used as an umbrella term for image classification and object detection rather than as a third task category. Dataset conditions may also be very different: sample counts are between a few hundred and over 200,000 images, and images can be taken in front of uniform laboratory backgrounds or in complex orchards with shadows, overlapping foliage, and different cameras [2], [18]–[29], [31]–[48]. Existing reviews give useful summaries on the development of apple disease recognition and bibliometrics [4], [5]. Nevertheless, the interaction between task paradigm, dataset realism, class balance, augmentation, training configuration, evaluation completeness, and deployment constraints has received little attention. A review is needed that is replicable for purposes of comparison.

The main contributions can be summarized as follows:

- A PRISMA 2020-aligned synthesis of 30 peer-reviewed studies published between 2020 and 2026 is presented,

with explicit separation of 22 classification and eight detection studies. Five research questions address architecture families, datasets and augmentation, performance evaluation, reproducibility, and deployment evidence.

- Dataset scale, training hyperparameters, acquisition settings, validation design, and reported metrics are compiled to reduce misleading cross-study interpretation.
- The gap between controlled benchmarks and real-orchard performance is examined together with class imbalance, annotation quality, explainability, and edge deployment.
- An evidence-based research roadmap is derived for standardized datasets, external validation, calibrated predictions, reproducible pipelines, and deployment-aware model design.

The rest of the study is organized as follows. Section II discusses apple disease imaging and its associated deep learning architecture families. The review protocol, research questions, search strategy, selection criteria, quality assessment, and data-extraction process are presented in Section III. Results by research question are presented in Section IV. Section V examines cross-study implications, unresolved gaps, and threats to validity. The review ends with Section VI.

II. Background and Related Reviews

A. Apple Disease Imaging and Diagnostic Complexity

Visible symptoms are often expressed as lesions, discolouration, powdery growth, necrotic areas, deformation, or premature defoliation. Apple scab is characterized by olive green or dark lesions on leaves and fruits. Powdery mildew causes pale or powder-like growth. Alternaria and Marssonina diseases cause spots whose appearance varies with severity. Fire blight can attack leaves, shoots, blossoms, and fruit [1], [8]-[11]. Similar patterns of color and texture across diseases create a fine-grained classification problem, while small overlapping lesions introduce an additional localization problem for object detectors. Representative disease manifestations considered in this review are illustrated in Fig. 1, showing the visual overlap in lesion color, texture, and morphology that complicates image-based diagnosis.

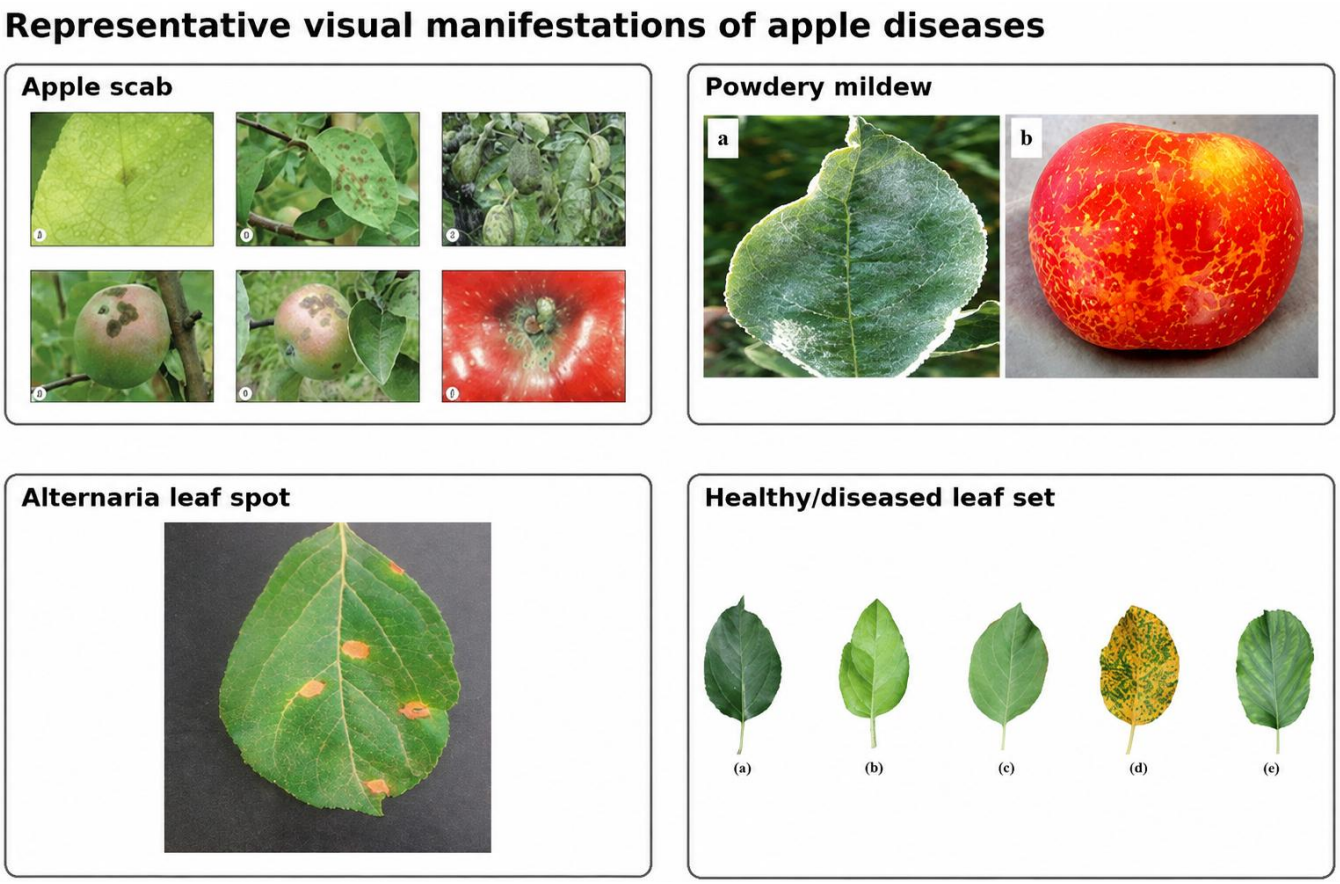

Fig. 1. Representative apple-disease symptoms illustrating scab, powdery mildew, Alternaria leaf spot, and healthy/diseased leaf conditions; disease descriptions are supported by [1], [8], and [10].

These visual ambiguities make dataset acquisition conditions crucial. Centered leaves and homogeneous backgrounds simplify segmentation in controlled datasets, but can cause models to learn background shortcuts. Natural-scene datasets have branches, fruits, soil, sky, mixed illumination, occlusion, and multiple symptoms, which makes the evaluation more difficult but operationally meaningful. Expert-annotated field imagery has been made significantly more accessible by

datasets such as Plant Pathology 2020 and 2021 [2], while self-collected datasets like MSALDD have enabled more research on real-time detection [48].

B. Deep-Learning Architecture Families

Convolution, non-linear activation, pooling, and classification layers enable a CNN to learn hierarchical image features. The early layers are sensitive to edges and local texture, and the deeper layers learn disease-specific shape and semantic patterns [6], [7]. Residual connections, dense feature reuse, depthwise separable convolution, squeeze-and-excitation attention, and feature pyramids are employed to enhance gradient flow, parameter efficiency, and multi-scale recognition. The hierarchical feature-extraction sequence underlying CNN-based image classification is illustrated in Fig. 2, from convolutional feature learning to the final class prediction.

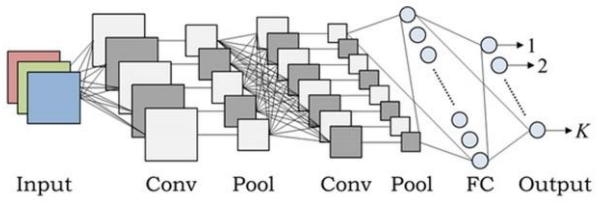

Fig. 2. Conceptual CNN processing pipeline for image classification; the architectural principles are discussed in [7].

The reviewed corpus adopted the classification architectures of ResNet, DenseNet, EfficientNet, Inception, RegNet, Xception, lightweight bilinear CNNs, vision transformers, and hybrid CNN-transformer architectures [18]–[29], [31]– [40]. Improved Faster R-CNN and several YOLO derivatives are detection architectures with feature pyramids, attention modules, lightweight convolution, normalized Wasserstein distance, or auxiliary detection heads [41]–[48]. Transformer-based models provide long-range context, whereas CNN branches retain local texture sensitivity. DBCoST and CTPlantNet are hybrid systems trying to combine both properties [23], [24].

C. Limitations of Earlier Reviews

Doutoum and Tugrul [4] have provided a systematic review, which proposed a comprehensive taxonomy of apple leaf disease classification and detection. Bonkra et al. [5] performed a bibliometric analysis to quantify publication and collaboration trends. These works show the maturity of the field but leave three issues open. First, the results of classification and detection are not always consistently separated. Second, the relationship between dataset realism, hyperparameter reporting, and reported performance is underexplored. Third, deployment evidence (model size, memory, inference latency, power consumption and field validation) is rarely synthesized as a unified dimension. The present review addresses these limitations using a multidimensional extraction framework.

III. Materials and Methods

A. Review Design and Research Questions

The systematic literature review was designed and reported in alignment with the PRISMA 2020 statement [3]. The protocol comprised six stages: identification, deduplication, title/abstract screening, full-text eligibility assessment, methodological-evidence appraisal, and structured qualitative synthesis. Research questions, eligibility criteria, appraisal domains, and data-extraction variables were defined before full-text extraction. No prospective registration or standalone protocol was available.

Study screening and data extraction were conducted by one lead reviewer and subsequently checked by a second author. Disagreements were resolved through discussion and consensus. This procedure is considered a potential source of selection bias and is addressed explicitly in the threats-to-validity analysis. The five research questions and analytical purposes are summarized in Table I.

TABLE I. Research Questions

ID	Research question	Analytical purpose
RQ1	Which deep-learning architectures and task paradigms are used for apple disease recognition?	Establish the taxonomy of classification, detection, hybrid, lightweight, and transformer-based models.
RQ2	Which datasets, acquisition conditions, preprocessing methods, and augmentation strategies are used?	Assess data realism, class balance, sample scale, and the role of preprocessing or synthetic data.
RQ3	How are models evaluated, and what performance patterns are reported?	Interpret accuracy, precision, recall, F1-score, and mAP within comparable experimental contexts.
RQ4	How completely are training settings and reproducibility information reported?	Examine optimizers, learning rates, batch sizes, epochs, validation protocols, and missing details.
RQ5	How close are current models to reliable field deployment, and which research gaps remain?	Evaluate robustness, external validation, interpretability, efficiency, latency, memory, energy, and future priorities.

B. Information Sources and Search Strategy

Scopus, Web of Science Core Collection, and IEEE Xplore were selected for their cross-disciplinary coverage of peer-reviewed research in artificial intelligence, computer vision, agriculture, and engineering. The final searches were conducted on 30 June 2026 and covered January 2020-June 2026. Only English-language peer-reviewed journal and conference articles were eligible as primary studies; the exact executed search expressions are reported in Table II. Specialized agricultural databases such as CAB Abstracts and AGRICOLA were not searched, and this restriction is addressed in Section V-C.

TABLE II. DATABASE SEARCH STRATEGY

Database	Field	Complete search expression	Limits
Scopus	TITLE-ABS-KEY	("apple" OR "Malus domestica") AND ("leaf disease" OR "fruit disease" OR "apple disease") AND ("deep learning" OR CNN OR transformer OR YOLO OR "Faster R-CNN") AND (classification OR detection OR recognition)	Jan. 2020-Jun. 2026; English; journal/conference
Web of Science	TS=	(("apple" OR "Malus domestica") AND ("leaf disease" OR "fruit disease" OR "apple disease") AND ("deep learning" OR CNN OR transformer OR YOLO OR "Faster R-CNN") AND (classification OR detection OR recognition))	Jan. 2020-Jun. 2026; English; article/proceedings
IEEE Xplore	All Metadata	(("apple disease" OR "apple leaf disease") AND ("deep learning" OR CNN OR "object detection" OR YOLO OR "Faster R-CNN"))	Jan. 2020-Jun. 2026; English; journals/conferences

Backward citation screening was performed using the reference lists of included studies and relevant reviews; no additional eligible peer-reviewed report was added. Full texts were retrieved from publisher repositories, institutional subscriptions, and open-access archives. Database exports, deduplication decisions, screening decisions, full-text decisions, and the extraction form constituted the audit record. The executed database strings relied mainly on generic apple-disease terms; disease-specific and segmentation terminology was not exhaustively represented and is therefore treated as a search-sensitivity limitation in Section V-C.

C. Eligibility Criteria

Eligibility criteria were applied at the title/abstract and full-text stages. A study was required to report an apple-specific deep-learning experiment and quantitative model evaluation. General plant-disease papers were included only when apple-specific results could be separated. Non-peer-reviewed sources were excluded from the primary synthesis. The D-KAP preprint [30] was retained only as contextual evidence and was excluded from corpus counts, performance ranges, figures, methodological appraisal, and comparative tables. The eligibility criteria are summarized in Table III.

TABLE III. INCLUSION AND EXCLUSION CRITERIA

Type	ID	Dimension	Criterion
Inclusion	IC1	Domain	Apple leaf, fruit, or apple-plant disease classification/detection.
Inclusion	IC2	Method	CNN, transformer, hybrid deep network, transfer learning, or deep object detector.
Inclusion	IC3	Evidence	Quantitative experimental evaluation using accuracy, precision, recall, F1-score, AUC, mAP, or deployment metrics.
Inclusion	IC4	Publication	Peer-reviewed journal or conference article in English, 2020–2026.
Inclusion	IC5	Accessibility	Full text available for methodological and numerical extraction.
Exclusion	EC1	Scope	General agriculture or plant disease research without separable apple experiments.
Exclusion	EC2	Method	Handcrafted image processing or shallow machine learning without a deep-learning component.
Exclusion	EC3	Evidence	No experimental validation or no quantitative results.
Exclusion	EC4	Publication	Editorials, tutorials, theses, non-English papers, and non-peer-reviewed items were excluded from the primary synthesis. The D-KAP preprint [30] was retained only as contextual evidence.
Exclusion	EC5	Redundancy	Duplicate database records or duplicate versions of the same study.

D. Quality Assessment and Data Extraction

The former binary quality score was replaced by a domain-based methodological-evidence appraisal because reporting completeness and risk of bias are not equivalent. Eight domains were considered: dataset provenance and label validity, evaluation independence and leakage control, class imbalance, generalization strength, metric adequacy and uncertainty, reproducibility, agronomic context, and transparency of limitations. Each domain was interpreted on a three-level basis (0 = insufficient/unclear, 1 = partial, 2 = strong evidence) when information was available. No aggregate study score or numerical weighting was used; the domains informed qualitative sensitivity interpretation. The appraisal domains are defined in Table IV, and the extraction variables are summarized in Table V.

TABLE IV. QUALITY ASSESSMENT CRITERIA

ID	Assessment domain	Purpose
Q1	Dataset provenance and diagnostic-label validity	Assess traceable data sources, labels, and expert/validated diagnosis.
Q2	Evaluation independence and leakage control	Assess train-test separation, source grouping, and augmentation-lineage leakage.
Q3	Class distribution and imbalance management	Assess class counts, balancing strategy, and minority-class evaluation.
Q4	Generalization and external validation	Distinguish holdout/CV from cross-dataset, orchard, or field validation.
Q5	Metric adequacy and statistical uncertainty	Assess task-appropriate, classwise, and uncertainty reporting when available.
Q6	Experimental reproducibility	Assess optimizer, learning rate, batch size, epochs, resolution, seeds/code, and validation details.
Q7	Biological and agronomic context	Assess cultivar, season, severity, acquisition conditions, expert diagnosis, and disease co-occurrence.
Q8	Transparency, limitations, and failure modes	Assess explicit threats, implementation constraints, and acknowledged failure modes.

TABLE V. DATA EXTRACTION FRAMEWORK

Attribute	Extracted information	Analytical use
Bibliographic profile	Authors, year, venue	Publication trend and traceability.
Task paradigm	Classification, detection, hybrid	Avoid invalid cross-task comparison.
Dataset	Source dataset size, experimental subset, augmented count, classes, acquisition environment, label provenance	Assess scale, biological independence, class balance, realism, and label validity.
Preprocessing	Resize, enhancement, segmentation, normalization	Identify data-pipeline variation.
Augmentation	Geometric, photometric, GAN, class-aware	Assess robustness and synthetic-data use.
Architecture	Backbone, attention, transformer, feature fusion	Build the model taxonomy.
Training settings	Optimizer, learning rate, batch size, epochs, split/CV design, seeds when reported	Assess reproducibility and evaluation independence.
Metrics	Accuracy, precision, recall, F1, AUC, mAP, test size, fold variability/CI when reported	Compare evidence within task while avoiding unsupported uncertainty reconstruction.
Deployment	Parameters, FLOPs, model size, FPS/latency, memory, power, target hardware, field evidence	Assign qualitative deployment-evidence level (D0-D3).
Limitations	Generalization, annotation, imbalance, interpretability	Derive research gaps.
Agronomic context	Organ, cultivar, season, severity/stage, location, device, annotation authority, disease co-occurrence	Assess biological validity and orchard transferability.

E. Study Selection and Synthesis

The database search yielded 180 records: 90 from Scopus, 50 from Web of Science, and 40 from IEEE Xplore. After 18 duplicates were removed, 162 unique records underwent title/abstract screening, and 88 were excluded. All 74 reports sought for full-text review were retrieved. Forty-four full-text reports were excluded because apple-specific results could not be separated (n = 12), no deep-learning architecture was evaluated (n = 9), quantitative results were absent (n = 11), the publication was not peer-reviewed (n = 8), or the paper was not in English (n = 4). The final corpus comprised 30 peer-reviewed studies: 22 classification and eight detection studies. These correspond to 16.7% of all identified records, 18.5% of unique records after deduplication, and 40.5% of full-text reports assessed. The screening pathway is shown in Fig. 3.

Full-text exclusion categories are summarized in Table VI. The item-level full-text exclusion list is not reproduced because it is no longer available in a form that can be independently verified. To avoid retrospective reconstruction and possible misclassification, only the contemporaneously retained aggregate exclusion counts and primary reason categories are reported. Reason-level counts for the 88 title/abstract exclusions were not retrospectively reconstructed when they had not been recorded as quantitative categories. The D-KAP preprint [30] remained contextual evidence only.

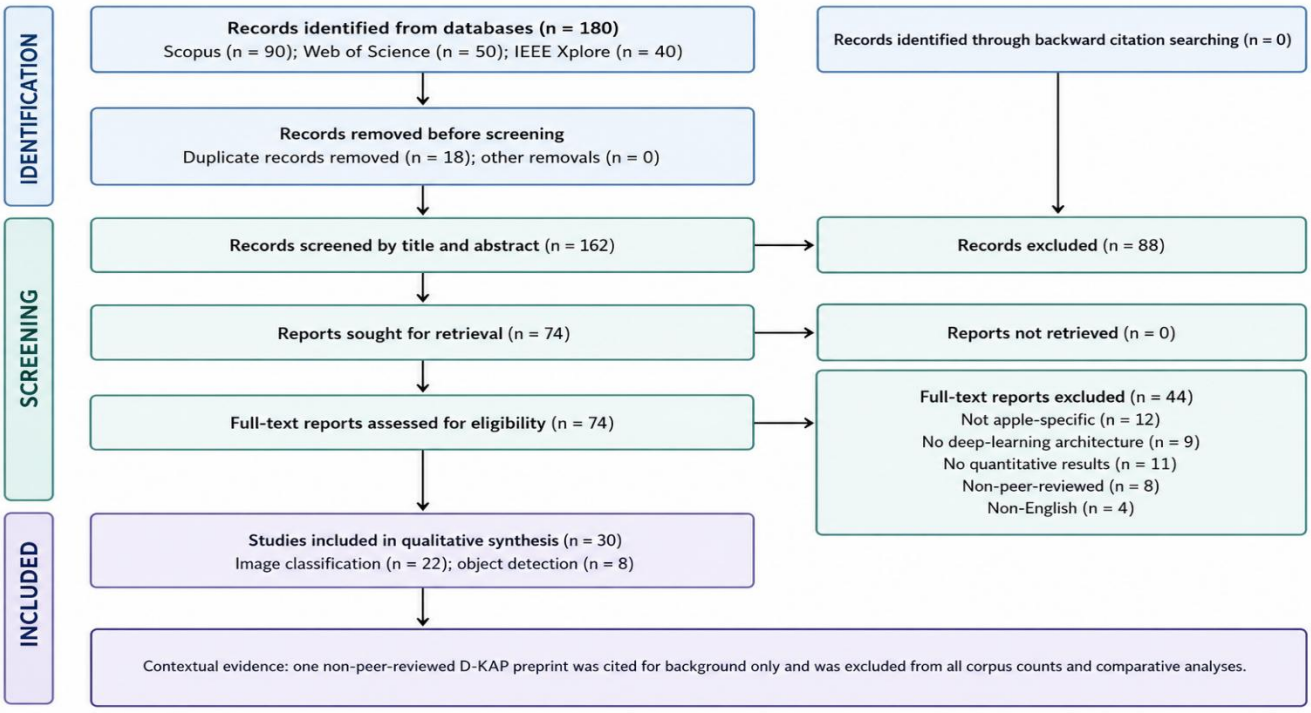

Fig. 3. PRISMA 2020 flow diagram for identification, screening, eligibility assessment, and inclusion.

TABLE VI. FULL-TEXT EXCLUSION SUMMARY

Code	Primary reason for exclusion	Reports (n)	Interpretive note
FT1	Apple-specific experimental results could not be separated	12	Aggregate exclusion category retained from screening
FT2	No deep-learning architecture was evaluated	9	Aggregate exclusion category retained from screening
FT3	No quantitative experimental results were reported	11	Aggregate exclusion category retained from screening
FT4	Non-peer-reviewed publication	8	Includes D-KAP [30] as contextual evidence only
FT5	Non-English publication	4	Aggregate exclusion category retained from screening
Total	Full-text reports excluded	44	Aggregate count retained

A statistical meta-analysis was not performed due to large differences between included studies in dataset composition, class definitions, acquisition environments, training-validation-test splits, augmentation procedures, model implementations, and evaluation metrics. Classification accuracy and detection mAP were therefore not combined in a single effect-size model. A structured qualitative synthesis was conducted, and the numerical results were discussed in the framework of the original experimental settings.

IV. RESULTS

A. Descriptive Profile of the Included Corpus

Subsections IV-B to IV-F answer RQ1-RQ5 sequentially. The selected literature increased markedly after 2021: one included study was published in 2020, four in 2021, seven in 2022, four in 2023, six in 2024, six in 2025, and two in 2026. Classification accounts for 73.3% of the corpus (22/30), whereas detection accounts for 26.7% (8/30). The annual distribution is shown in Fig. 4.

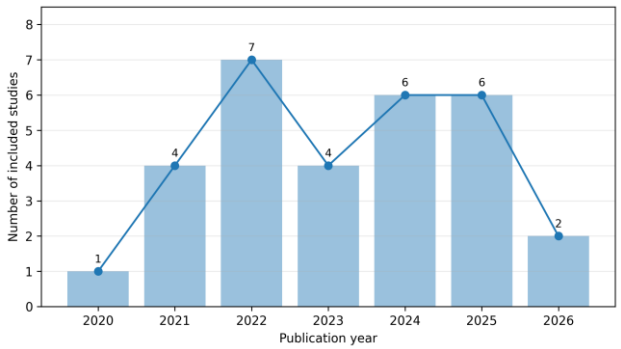

Fig. 4. Annual distribution of the 30 included peer-reviewed primary studies.

The corpus contains controlled PlantVillage images, datasets from the Plant Pathology competition, AppleLeaf9, special national or regional collections, self-constructed natural-scene datasets, and mixed datasets. The number of images varies over more than three orders of magnitude. Such heterogeneity rules out a simple model leaderboard and needs to be interpreted in the context of data conditions.

B. RQ1: Architectures and Task Paradigms

CNN-based classification remains dominant. EfficientNet, ResNet, DenseNet, Inception, RegNet, and Xception are commonly used because transfer learning can provide strong performance with limited training data [20]-[29], [31]-[40]. Dense and residual connections improve feature reuse and gradient propagation, whereas EfficientNet- and MobileNet-inspired designs reduce computational cost. RegNet-based studies also report strong performance across multiple disease classes. BLSENet [25] uses bilinear fusion and squeeze-and-excitation attention for fine-grained recognition.

A second major family comprises hybrid and transformer-based architectures. DBCoST utilizes local features from CNN and global context from Swin-Transformer [23]. CTPlantNet combines the outputs of SEResNeXt, EfficientNet-V2S, and Swin-Large [24]. To improve long-range representation, AppViT uses a lightweight transformer formulation [26]. Such models can improve the discrimination of diseases in complex backgrounds, but the computational burden is larger than that of compact CNNs.

Detection research is dominated by the YOLO family. YOLOX-ASSA-Nano replaces YOLOX-Nano with asymmetric ShuffleBlocks, attention, and separable convolution [48], Apple-Net improves YOLOv5 [47], and YOLOv5-Res adds multiscale residual blocks and lightweight feature fusion [45]. Recent detectors include auxiliary feature pyramids [41], RT-DETR components [42], attention modules [42], normalized Wasserstein distance [43], and specific small-object heads [43]. Improved Faster R-CNN is still relevant when two-stage localization is tolerable [44]. Fidan et al. [18] set up YOLOv11x for image-level classification, reporting classification metrics, but YOLO architectures are usually linked to detection; this is classified by experimental results, not architectural heritage. To summarize, the findings of RQ1 show that CNN classifiers are still the mainstream. However, hybrid transformer models and lightweight YOLO-based models are the two main emerging trends. Fig. 5 summarizes the task taxonomy.

C. RQ2: Datasets, Preprocessing, and Augmentation

Publicly available datasets dominate model development. PlantVillage provides a large, controlled benchmark, and Plant Pathology 2020/2021 provides expert-labeled field imagery with more variation in the illumination and background [2]. Specialized collections like AppleLeaf9 contain disease categories and imbalance patterns, while MSALDD was developed for lightweight detection in diverse natural scenes [48].

Image resizing and normalization are almost universal. Further pre-processing includes histogram equalization, contrast-limited adaptive histogram equalization, bilateral filtering, multi-scale retinex, edge enhancement, background suppression, and lesion segmentation [28], [29]. BAM-Net employs BF-MSRCR to highlight color and texture before classification [29]. Preprocessing can improve the contrast, but over-enhancement may mislead the natural distributions of color and reduce the transferability.

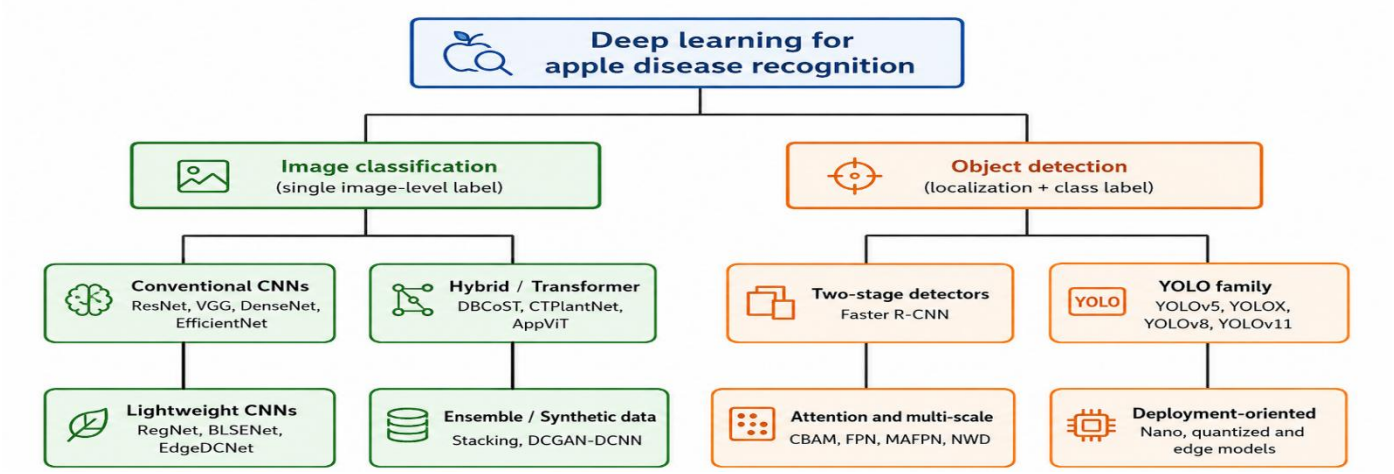

Fig. 5. Taxonomy of deep-learning approaches represented in the selected corpus.

Augmentation is widely used to address limited data and class imbalance through rotation, flipping, translation, scaling, cropping, photometric variation, blur, and noise. Tian et al. used CycleGAN to synthesize anthracnose and ring-rot symptoms [36], while Ayaz et al. combined DCGAN-generated samples with a deep CNN [35]. Fidan et al. showed that balancing effects are architecture-dependent: physically inspired augmentation benefited YOLOv11x, whereas EfficientNet-B1 performed best without the same intervention [18]. Augmented sample counts were not treated as equivalent to independent biological observations because lineage overlap can increase leakage risk.

The largest performance threat is the mismatch between development data and deployment conditions. Lin et al. reported a decline from 95.74% under controlled conditions to 66.18% in a real-field setting [32], a 29.56-point reduction that precludes treating benchmark accuracy as evidence of orchard readiness. Agronomic reporting is also uneven: optimized RegNet experiments include multiple diseases on the same leaf and complex acquisition conditions [27], whereas cultivar, season, disease severity, diagnostic authority, and co-occurrence are not consistently documented across the corpus. RQ2 therefore shows that dataset realism, biological annotation, class balance, and augmentation design are central determinants of transferability.

D. RQ3: Reported Performance and Evaluation Practice

Classification studies report accuracy values from 91.0% to 99.99%, whereas detection studies report mAP@0.5 values from 82.1% to 99.99%. These ranges are descriptive extrema from heterogeneous experimental settings and are not pooled estimates or confidence intervals. Differences in dataset realism, evaluation-set size, split design, and class composition preclude a universal architecture ranking. Validation design, test-set size, fold variability, and confidence intervals were considered when reported by the primary studies; unavailable uncertainty information was not reconstructed. Fig. 6 visualizes the reported distributions, while detailed classification and detection settings are summarized in Tables VII and VIII.

Accuracy is reported by 21 out of the 22 classification studies, but precision, recall, and F1-score are reported less consistently. Fig. 7 summarizes: All detection studies provide mAP@0.5. Only two studies provide an F1-score. This pattern is problematic for imbalanced data sets, because the overall accuracy can hide the minority-class failure. Besides the overall accuracy, the macro-averaged precision, recall, F1-score, per-class confusion matrices, balanced accuracy, AUC, and calibration measures should be reported. Detection studies should also report performance on small objects, classwise average precision, false-positive analysis, and mAP at multiple IoU thresholds. RQ3 therefore indicates that high point estimates are common, but metric completeness and evaluation context remain insufficient for direct cross-study ranking.

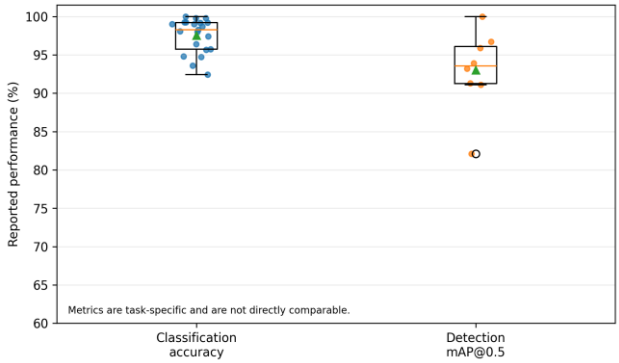

Fig. 6. Distribution of reported classification accuracy and detection mAP@0.5. Values are descriptive and not directly comparable across tasks or datasets.

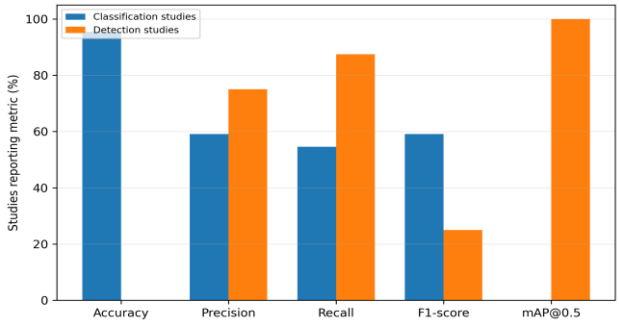

Fig. 7. Completeness of metric reporting across the 22 classification and eight detection studies.

TABLE VII. CLASSIFICATION STUDIES AND REPORTED EXPERIMENTAL SETTINGS

Ref.	Year	Model	Dataset/images	Training configuration	Main result (%)	Key observation
[18]	2026	YOLOv11x	AppleLeaf9; 14,582	SGD; LR 0.01; batch 64; 100 epochs; Momentum=0.9; WD=0.01	Acc 99.18; P 99.17;F1 99.17;R 99.18	Class-imbalance scenarios and physical augmentation.
[19]	2026	EdgeDCNet	PP2020/2021-derived set	Adam; 500 epochs; 10-fold CV	Acc 94.80	36.2 KB model; 162.2 KB RAM; 463.8 ms; 14 mW; edge quantization evaluated.
[20]	2025	Fine-tuned EfficientNet-B0	PlantVillage; 54,448	AdaMax; LR 0.001; batch 32; 20 epochs	Acc/F1/R 99.78; P 99.79	High accuracy with compact transfer learning.
[21]	2025	ResNet-9	TPD; 4,447	Adam; LR 0.011; batch 32; 100 epochs; WD=0.00018; Momentum = 0.4	Acc 97.41; P 96.40; F1 95.70; R 97.09	SHAP explanation and significance testing.
[22]	2025	Xception	Apple leaf subset; 12,000	Adam; LR 4×10^{-5}; batch 32	Acc/P/F1/R 99.0	Web application and visual explanations.
[23]	2024	DBCoST	PP2021; 18,632	AdamW; LR 10^{-4}; batch 128; 100 epochs ; WD=0.05	Acc 97.32; P 97.33; F1 97.36; R 97.40	Primary PP2021 result; additional AppleLeaf9/mixed-disease experiment reached Acc 98.06.
[24]	2024	CTPlantNet	PP2020/2021; 3,526/18,632	RectifiedAdam; LR 0.0007; batch 64	Acc 98.28 / 95.96 P 98.67/ 97.55	CNN-transformer ensemble; 5-fold validation.
[25]	2024	BLSENet	Apple leaf set; 14,582	Adam; LR 10^{-4}; batch 16; 200 epochs	Acc 93.58	Lightweight bilinear fusion and SE attention.
[26]	2024	AppViT	PP2021; 16,093	AdamW; LR 0.002→0.000125; batch 32; Epochs=65	Acc 96.40; P 96.70; F1 96.30; R 95.90	Lightweight vision transformer.
[27]	2024	Optimized RegNet	2,609 / 13,752	Ranger or Adam; batch 16; 50 epochs	Acc 99.23	Multiple diseases and complex backgrounds.
[28]	2023	InceptionV3 / VGG-16	PlantVillage; >18,000	Not completely reported	Acc 92.42 / 91.53	Comparison of pretrained CNNs.
[29]	2023	BAM-Net	Complex-background set; 3,500	AdamW; LR 10^{-3}; batch 64; 50 epochs	Acc 95.64; P 95.62 ; F1 95.25; R 95.89	BF-MSRCR, coordinate attention, multiscale refinement.
[31]	2023	Conv-3 DCNN	3,171	Adam; LR 10^{-4}; batch 50; 1,000 epochs	Acc 98.0; P 98.0 ;F1 97.0; R 97.0	Apple disease CNN with transfer comparison.
[32]	2022	BM + CMSFF + CA	PlantVillage and real-field AFD	SGD; LR 0.1→0.01; batch 128; Epochs=150	Acc 95.74 / 66.18	Strong controlled-to-field degradation.
[33]	2022	AFD-Net	PP2020	Adam; LR 5×10^{-6}→5×10^{-3}; 5-fold CV; Epochs=140	Acc 98.70; P 97.0; F1 89.0	Multiclass foliar disease classification.
[34]	2021	Stacking ensemble	10,000	Not fully reported	F1 98.05	Ensemble learning; incomplete training details.
[35]	2021	DCGAN-DCNN	319 expanded to 4,000	Adam-family; LR 0.001–0.002; batch 55; epochs 10 to 35	Acc 99.99	Synthetic-data benefit; redundancy risk.
[36]	2021	Multi-scale Dense Inception-ResNet-V2	12,714	LR 10^{-4}; batch 16; 4,000 epochs	Acc 94.74; P 94.71; F1 94.86; R 95.03	CycleGAN and multiscale dense features.
[37]	2022	Lightweight RegNet	8,510	Adam; LR 10^{-4}; batch 16; 50 epochs	Acc 99.23; P 99.24; R 99.16	Small and imbalanced dataset.
[38]	2022	CNN-C	200,000 augmented samples	16 epochs; dropout 0.25–0.5	Acc 99.20; R 99.70	Augmented sample count; not equivalent to 200,000 independent biological observations.
[39]	2021	EfficientNet-B7 / DenseNet	10,000 augmented samples	40 epochs	Acc 99.80 / 99.75	Edge/blur augmentation; augmented samples are not independent biological observations.
[40]	2020	ResNet-18/34/50	PlantVillage	Adam/SGD; LR=0.0001 to 0.01; Epochs=30 to 100; Momentum=0.9	Acc 99.0	Limited-data comparison across residual networks.

TABLE VIII. DETECTION STUDIES AND REPORTED EXPERIMENTAL SETTINGS

Ref.	Year	Model	Dataset/images	Training configuration	Main result (%)	Key observation
[41]	2025	DMN-YOLO	6,301	AdamW; LR 0.1; batch 16; 300 epochs; Momentum=0.937; WD=0.0005	mAP50 93.2; P 88.8;F1 88.3; R 87.9	MAFPN, RT-DETR decoder, P2 head, NWD loss.
[42]	2025	ELM-YOLOv8n	13,218	SGD; LR 0.01; batch 16; 100 epochs	mAP50 96.7; P 94.5; F1 94.0; R 93.6	44.8% fewer parameters and 39.5% lower computation.
[43]	2025	E-YOLOv8	6,986	SGD; LR 0.01; batch 32; 300 epochs	mAP50 93.9; P 90.1; R 89.6	GhostConv, C3 fusion, CBAM, FPN; 1.8 M parameters.
[44]	2023	Improved Faster R-CNN	4,182	LR 0.0025; Soft-NMS	mAP50 91.3	Two-stage high-precision detection.
[45]	2024	YOLOv5-Res	4,863	Adam; LR 0.01; batch 2; 150 epochs	mAP50 82.1; P 81.4; R 76.9	Residual multiscale block; small batch may limit stability.
[46]	2022	DF-Tiny-YOLO	1,404	LR 0.001; batch 64; momentum 0.9; WD=0.0005	mAP50 99.99; R 100	Outlying result on a small, specific dataset.
[47]	2022	Apple-Net	12,500	Adam; LR 0.01; batch 4; 100 epochs	mAP50 95.9; P 93.1; R 94.4	Improved YOLOv5 with efficient feature fusion.
[48]	2022	YOLOX-ASSANano	MSALDD; 3,209	CIoU loss; incomplete hyperparameter report	mAP50 91.08; P 89.75 R 83.63	0.83 MB lightweight natural-scene detector.

E. RQ4: Hyperparameters, Validation, and Reproducibility

Training conditions vary substantially: learning rates range from approximately 10^-6 to 10^-1, batch sizes from 2 to 128, and training duration from 16 to 4,000 epochs. Adam, AdamW, AdaMax, Ranger, RectifiedAdam, and SGD are reported. Six studies (20.0%) omit at least one central setting such as optimizer, learning rate, batch size, or epoch count. Validation ranges from single holdout splits to five- or ten-fold cross-validation, and several studies do not clarify whether images from the same source tree or augmentation lineage were separated. Random image-level splits may therefore permit leakage when near-duplicate augmented samples occur in both training and testing. Orchard- or acquisition-session-level splitting, together with random seeds, software versions, pretrained weights, image resolution, early-stopping rules, augmentation probabilities, hardware, and code/data identifiers, would improve reproducibility. RQ4 therefore identifies incomplete configuration reporting and heterogeneous validation as persistent barriers to replication.

F. RQ5: Deployment Evidence and Operational Constraints

Deployment evidence was operationalized on four qualitative levels: D0, benchmark performance only; D1, efficiency-aware reporting such as parameters, FLOPs, model size, or theoretical speed; D2, prototype execution in a web, mobile, embedded, or hardware-relevant environment with at least one operational metric; and D3, target/resource-constrained hardware evaluation under natural or operational conditions with practical efficiency metrics and field validation. EdgeDCNet provides D3-level evidence through a 36.2 KB model, 162.2 KB runtime memory, 463.8 ms inference, and 14 mW power consumption [19]. YOLOX-ASSANano reports 122 FPS and a 0.83 MB model [48], while ELM-YOLOv8n and E-YOLOv8 provide strong D1-D2 efficiency evidence [42], [43].

Most classification studies do not evaluate field latency, battery consumption, network dependency, sensor variability, or failure modes. Web-based deployment [22] improves accessibility but may depend on connectivity and server resources. On-device inference can reduce latency and data transmission, yet quantization, pruning, and compression may disproportionately reduce minority-class sensitivity; classwise degradation should therefore accompany global metrics.

Similarly, evidence of explainability is limited. SHAP-based analysis is presented in [21] and visual explanations in a web system in [22] by Shafik et al. and Aksoy et al., respectively. “Saliency maps are not causally reliable per se. Scores should be given for localization, robustness and relevance to plant pathology. Calibrated confidence and uncertainty estimates should aid expert referral of ambiguous cases. RQ5 thus suggests that practical readiness is driven by hardware efficiency, field robustness, uncertainty handling, and clinically/agronomically meaningful explanations, not solely by predictive accuracy.

V. DISCUSSION

A. Cross-Study Interpretation

The results show that architectural innovation can help improve apple disease recognition, but the performance is determined by the complete experimental system, not by the network itself. Dataset curation, image acquisition, class balance, split design, augmentation, resolution, annotation, optimization, and metric selection. Thus, near-perfect performance on controlled datasets does not suffice to demonstrate field reliability. Classification as a benchmark task is mature, but limited in operation when multiple diseases or lesion locations need identification. Detection is more informative for targeted treatment because it localizes the symptoms, but it requires expensive bounding-box annotation and is still sensitive to small lesions and background clutter. Instead of treating them separately, a practical pipeline may combine fast detection, fine-grained classification, severity estimation, and temporal monitoring. The synthesis also indicates that, in general, larger datasets stabilize training, but size alone is not a decisive factor. A small data set can achieve a high score if the task is visually simple or the split is favorable, while a larger field data set can lead to lower but more credible performance. Sample count of the test set is less important than its diversity and independence. Similarly, synthetic

augmentation can also improve minority-class representation, but may inflate the results if the generated images are highly redundant or if the generator is trained on the same images that are subsequently used for evaluation.

B. Research Gaps and Future Directions

The principal evidence gaps and required validation steps are summarized in Table IX, and Fig. 8 links these gaps to a staged data-model-evaluation-deployment roadmap.

TABLE IX. RESEARCH GAPS AND RECOMMENDED DIRECTIONS

Gap	Current limitation	Recommended direction	Required validation
Real-world data	Controlled backgrounds and limited environmental diversity	Multi-orchard, multi-season, multi-device datasets with cultivar, severity, weather, and location metadata.	Cross-orchard testing and external validation.
Evaluation	Accuracy-dominated reporting and incomparable splits	Task-specific minimum metric sets, calibration, confidence intervals, classwise results, and robustness testing.	Standard benchmark protocol and public leaderboards.
Generalization	Large controlled-to-field performance decline	Domain adaptation, self-supervised pretraining, few-shot learning, and style-robust augmentation.	Unseen orchard, season, and device evaluation.
Reproducibility	Missing hyperparameters, seeds, code, and split details	Complete experimental checklists, public code, fixed dataset versions, and leakage-resistant splitting.	Independent reproduction studies.
Interpretability	Unvalidated saliency maps and overconfident predictions	Pathologist-informed explanations, uncertainty quantification, calibration, and reject options.	Human-in-the-loop diagnostic evaluation.
Deployment	Limited latency, memory, and energy reporting	Quantization-aware training, pruning, edge accelerators, offline inference, and energy benchmarking.	Accuracy-efficiency Pareto analysis on target hardware.
Disease scope	Single-label disease categories and weak severity modeling	Multi-label recognition, lesion segmentation, severity estimation, and progression tracking.	Decision support linked to treatment thresholds.
Data governance	Inconsistent annotations and unclear reuse conditions	Annotation guidelines, provenance, licensing, privacy, and FAIR dataset documentation.	Auditable data and model cards.

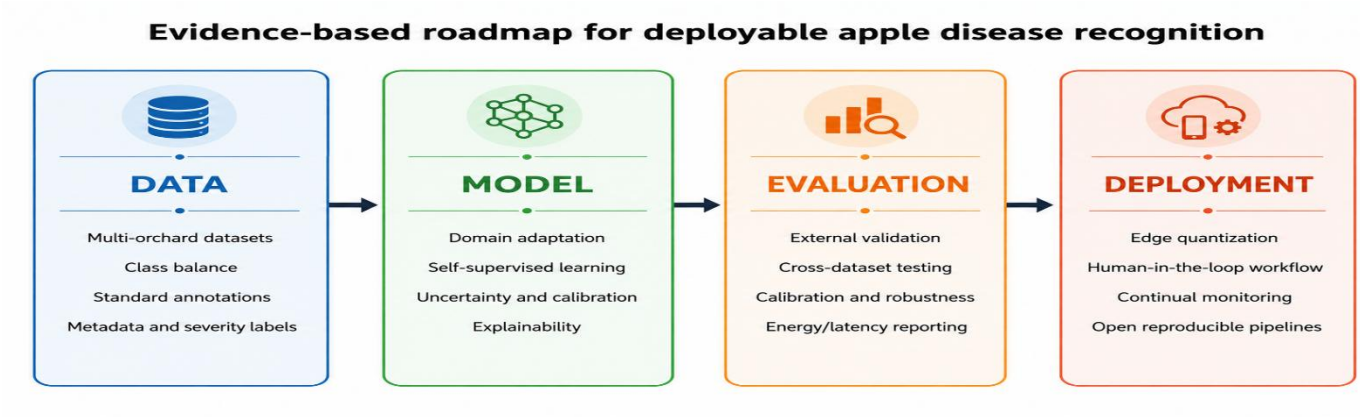

Fig. 8. Evidence-based roadmap for robust and deployable apple disease recognition.

C. Threats to Validity

There are still some threats to validity. The search was limited to English-language publications and generic apple-disease terminology and was conducted in Scopus, Web of Science Core Collection, and IEEE Xplore. Thus, CAB Abstracts, AGRICOLA, and exhaustive disease-specific or segmentation terms were not included, and relevant regional or otherwise indexed studies may have been missed. The review was not prospectively registered, which limits the ability to independently verify that there was no protocol drift during the development of the review. Screening and extraction were undertaken by a single lead reviewer and checked by a second author rather than by two independent teams. Results were extracted, not re-implemented, and dataset names, number of samples, definitions of splits, and reporting of uncertainty were inconsistent across studies. These risks were mitigated through the application of explicit eligibility criteria, separation of classification and detection, an audit record, domain-based methodological appraisal, and explicit documentation of missing information. The lack of a common benchmark, however, and the reporting of complete uncertainty preclude formal meta-analysis and causal claims of architectural superiority. In addition, the item-level list of full-text exclusions was no longer available in an independently verifiable form during revision; it was therefore not reconstructed retrospectively, and only the retained aggregate exclusion categories are reported.

VI. CONCLUSION

This systematic literature review synthesizes 30 peer-reviewed deep learning studies on apple disease recognition published between 2020 and 2026: 22 studies on image classification and eight studies on object detection. One preprint was included as contextual evidence only. The classification accuracy reported is in the range of 91.0% to 99.99%, and detection mAP@0.5 is in the range of 82.1% to 99.99%; these are descriptive and still depend on the realism of the dataset and the evaluation design. CNNs are still dominant, but the methodological landscape is enriched by transformer hybrids,

attention, multiscale fusion, class-imbalance-aware training, synthetic augmentation, and lightweight YOLO variants.

The primary takeaway is that good benchmark performance does not mean orchard reliability. Controlled background datasets can produce nearly ideal numbers, but there can be substantial degradation due to field transfer. Barriers to reproducibility and practical adoption remain in the form of partial hyperparameter reporting, heterogeneous validation, absence of uncertainty analysis, limited external testing, and scarce evidence of hardware-level deployment.

Future work should focus on standardized multi-orchard datasets, leakage-protected splits, complete experimental reporting, calibrated and interpretable outputs, cross-dataset evaluation, agronomic metadata, and efficiency metrics on target hardware. Robust, interpretable, and resource-aware decision support is more informative for real orchards than optimization of a single benchmark score.

REFERENCES

[1] A. Moinina, R. Lahlali, and M. Boulif, "Important pests, diseases and weather conditions affecting apple production: Current state and perspectives," Rev. Mar. Sci. Agron. Vét., vol. 7, no. 1, pp. 71–87, 2019.

[2] R. Thapa, K. Zhang, N. Snavely, S. Belongie, and A. Khan, "The Plant Pathology Challenge 2020 data set to classify foliar disease of apples," Appl. Plant Sci., vol. 8, no. 9, e11390, 2020, doi: 10.1002/aps3.11390.

[3] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, n71, 2021, doi: 10.1136/bmj.n71.

[4] A. S. Doutoum and B. Tugrul, "A systematic review of deep learning techniques for apple leaf diseases classification and detection," PeerJ Comput. Sci., vol. 11, e2655, 2025, doi: 10.7717/peerj-cs.2655.

[5] A. Bonkra, S. Pathak, A. Kaur, and M. A. Shah, "Exploring the trend of recognizing apple leaf disease detection through machine learning: A comprehensive analysis using bibliometric techniques," Artif. Intell. Rev., vol. 57, art. 21, 2024, doi: 10.1007/s10462-023-10628-8.

[6] L. Alzubaidi et al., "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," J. Big Data, vol. 8, art. 53, 2021, doi: 10.1186/s40537-021-00444-8.

[7] A. Hidaka and T. Kurita, "Consecutive dimensionality reduction by canonical correlation analysis for visualization of convolutional neural networks," Proc. 48th ISCIE International Symposium on Stochastic Systems Theory and Its Applications, pp. 160–167, 2017, doi: 10.5687/sss.2017.160.

[8] J. Cabrefiga, M. V. Salomon, and P. Vilardell, "Improvement of Alternaria leaf blotch and fruit spot of apple control through the management of primary inoculum," Microorganisms, vol. 11, art. 101, 2023, doi: 10.3390/microorganisms11010101.

[9] I. Vico, N. Duduk, M. Vasic, and M. Nikolic, "Identification of Penicillium expansum causing postharvest blue mold of apple fruit," Pestic. Fitomed., vol. 29, no. 4, pp. 257–266, 2014.

[10] M. Shuaibu, W. S. Lee, J. Schueller, P. Gader, Y. K. Hong, and S. Kim, "Hyperspectral band selection for apple Marssonina blotch detection," Comput. Electron. Agric., vol. 148, pp. 45–53, 2018.

[11] Y.-Q. Zhao et al., "Fire blight disease, a fast-approaching threat to apple and pear production in China," J. Integr. Agric., vol. 18, no. 4, pp. 815–820, 2019.

[12] E. Kamusiime, J. S. Nantongo, and C. Wacal, "Insect pests in apple (Malus domestica Borkh) gardens: Review," GSC Adv. Res. Rev., vol. 15, pp. 30–53, 2023.

[13] A. Khan, S. F. Honey, B. Babar, N. Jamil, and M. S. Mazhar, "Responses of Cydia pomonella (L.) reared on different artificial diets under laboratory condition," Sarhad J. Agric., vol. 35, 2019.

[14] D. Bosch, M. A. Rodríguez, and J. Avilla, "Monitoring resistance of Cydia pomonella Spanish field populations to new chemical insecticides and the mechanisms involved," Pest Manag. Sci., vol. 74, pp. 933–943, 2018, doi: 10.1002/ps.4791.

[15] A. Rousselin et al., "Harnessing the aphid life cycle to reduce insecticide reliance in apple and peach orchards: A review," Agron. Sustain. Dev., vol. 37, art. 38, 2017, doi: 10.1007/s13593-017-0444-8.

[16] P. A. Rode et al., "Phytophagous mites on apple trees: Population interactions and leaf injury," Res. Square, 2024.

[17] K. Buzzetti, R. A. Chorbadjian, and R. Nauen, "Resistance management for San Jose scale (Hemiptera: Diaspididae)," J. Econ. Entomol., vol. 108, no. 6, pp. 2743–2752, 2015.

[18] E. Fidan, S. Aksoy, P. Demircioglu, and I. Bogrekci, "Class imbalance-aware deep learning approach for apple leaf disease recognition," AgriEngineering, vol. 8, art. 70, 2026, doi: 10.3390/agriengineering8020070.

[19] M. Grujev, D. Stojanovic, A. Milosavljevic, M. Ilic, and V. Prodanovic, "An efficient real-time apple disease classification approach using a novel lightweight neural network on an Arduino edge device," Internet Things, vol. 35, 101833, 2026, doi: 10.1016/j.iot.2025.101833.

[20] H. Ali, N. Shifa, R. Benlamri, A. A. Farooque, and R. Yaqub, "A fine-tuned EfficientNet-B0 convolutional neural network for accurate and efficient classification of apple leaf diseases," Sci. Rep., vol. 15, 25732, 2025, doi: 10.1038/s41598-025-04479-2.

[21] W. Shafik, A. Tufail, L. C. De Silva, R. A. A. H. M. Apong, and K.-H. Kim, "Deep learning technique for plant disease classification and pest detection and model explainability elevating agricultural sustainability," BMC Plant Biol., vol. 25, 1491, 2025, doi: 10.1186/s12870-025-07377-x.

[22] S. Aksoy, P. Demircioglu, and I. Bogrekci, "Web-based AI system for detecting apple leaf and fruit diseases," AgriEngineering, vol. 7, art. 51, 2025, doi: 10.3390/agriengineering7030051.

[23] H. Si et al., "A dual-branch model integrating CNN and Swin Transformer for efficient apple leaf disease classification," Agriculture, vol. 14, art. 142, 2024, doi: 10.3390/agriculture14010142.

[24] A. Ait Nasser and M. A. Akhloufi, "A hybrid deep learning architecture for apple foliar disease detection," Computers, vol. 13, art. 116, 2024, doi: 10.3390/computers13050116.

[25] T. Fang, J. Zhang, D. Qi, and M. Gao, "BLSENet: A novel lightweight bilinear convolutional neural network based on attention mechanism for apple leaf disease identification," J. Food Qual., vol. 2024, art. 5561625, 2024, doi: 10.1155/2024/5561625.

[26] W. Ullah et al., "Efficient identification and classification of apple leaf diseases using lightweight vision transformer (ViT)," Discov. Sustain., vol. 5, art. 116, 2024, doi: 10.1007/s43621-024-00307-1.

[27] B. Wang, H. Yang, S. Zhang, and L. Li, "Identification of multiple diseases in apple leaf based on optimized lightweight convolutional neural network," Plants, vol. 13, art. 1535, 2024, doi: 10.3390/plants13111535.

[28] B. Igried, S. AlZu'bi, D. Aqel, A. Mughaid, I. Ghaith, and L. Abualigah, "An intelligent and precise agriculture model in sustainable cities based on visualized symptoms," Agriculture, vol. 13, art. 889, 2023, doi: 10.3390/agriculture13040889.

[29] Y. Gao, Z. Cao, W. Cai, G. Gong, G. Zhou, and L. Li, "Apple leaf disease identification in complex background based on BAM-Net," Agronomy, vol. 13, art. 1240, 2023, doi: 10.3390/agronomy13051240.

[30] H. Sharma, D. Padha, and N. Bashir, "D-KAP: A deep learning-based Kashmiri apple plant disease prediction framework," TechRxiv, 2022, doi: 10.36227/techrxiv.21210320.v2.

[31] V. K. Vishnoi, K. Kumar, B. Kumar, S. Mohan, and A. A. Khan, "Detection of apple plant diseases using leaf images through convolutional neural network," IEEE Access, vol. 11, pp. 6594–6609, 2023, doi: 10.1109/ACCESS.2022.3232917.

[32] H. Lin, R. Tse, S.-K. Tang, Z. Qiang, and G. Pau, "Few-shot learning approach with multi-scale feature fusion and attention for plant disease recognition," Front. Plant Sci., vol. 13, art. 907916, 2022, doi: 10.3389/fpls.2022.907916.

[33] A. Yadav, U. Thakur, R. Saxena, V. Pal, V. Bhateja, and J. C.-W. Lin, "AFD-Net: Apple foliar disease multi-classification using deep learning on the Plant Pathology dataset," Plant Soil, vol. 477, pp. 595–611, 2022, doi: 10.1007/s11104-022-05407-3.

[34] H. Li, Y. Jin, J. Zhong, and R. Zhao, "A fruit tree disease diagnosis model based on stacking ensemble learning," Complexity, vol. 2021, art. 6868592, 2021, doi: 10.1155/2021/6868592.

[35] H. Ayaz, E. Rodríguez-Esparza, M. Ahmad, D. Oliva, M. Pérez-Cisneros, and R. Sarkar, "Classification of apple disease based on non-linear deep features," Appl. Sci., vol. 11, art. 6422, 2021, doi: 10.3390/app11146422.

[36] Y. Tian, E. Li, Z. Liang, M. Tan, and X. He, "Diagnosis of typical apple diseases: A deep learning method based on multi-scale dense classification network," Front. Plant Sci., vol. 12, 698474, 2021, doi: 10.3389/fpls.2021.698474.

[37] L. Li, S. Zhang, and B. Wang, "Apple leaf disease identification with a small and imbalanced dataset based on lightweight convolutional networks," Sensors, vol. 22, art. 173, 2022, doi: 10.3390/s22010173.

[38] S. Singh et al., "Deep learning based automated detection of diseases from apple leaf images," Comput. Mater. Contin., vol. 71, pp. 1849–1866, 2022, doi: 10.32604/cmc.2022.021875.

[39] V. V. Srinidhi, A. Sahay, and K. Deeba, "Plant pathology disease detection in apple leaves using deep convolutional neural networks," in Proc. 5th Int. Conf. Computing Methodologies and Communication (ICCMC), 2021, pp. 1119–1127, doi: 10.1109/ICCMC51019.2021.9418268.

[40] A. Afifi, A. Alhumam, and A. Abdelwahab, "Convolutional neural network for automatic identification of plant diseases with limited data," Plants, vol. 10, art. 28, 2021, doi: 10.3390/plants10010028.

[41] L. Gao, H. Cao, H. Zou, and H. Wu, "DMN-YOLO: A robust YOLOv11 model for detecting apple leaf diseases in complex field conditions," Agriculture, vol. 15, art. 1138, 2025, doi: 10.3390/agriculture15111138.

[42] G. Wang, W. Sang, F. Xu, Y. Gao, Y. Han, and Q. Liu, "An enhanced lightweight model for apple leaf disease detection in complex orchard environments," Front. Plant Sci., vol. 16, 1545875, 2025, doi: 10.3389/fpls.2025.1545875.

[43] C. Gupta et al., "An enhanced deep learning-based framework for diagnosing apple leaf diseases," Sci. Rep., vol. 15, 39699, 2025, doi: 10.1038/s41598-025-23272-9.

[44] X. Gong and S. Zhang, "A high-precision detection method of apple leaf diseases using improved Faster R-CNN," Agriculture, vol. 13, art. 240, 2023, doi: 10.3390/agriculture13020240.

[45] Z. Sun, Z. Feng, and Z. Chen, "Highly accurate and lightweight detection model of apple leaf diseases based on YOLO," Agronomy, vol. 14, art. 1331, 2024, doi: 10.3390/agronomy14061331.

[46] J. Di and Q. Li, "A method of detecting apple leaf diseases based on improved convolutional neural network," PLoS ONE, vol. 17, e0262629, 2022, doi: 10.1371/journal.pone.0262629.

[47] R. Zhu, H. Zou, Z. Li, and R. Ni, "Apple-Net: A model based on improved YOLOv5 to detect apple leaf diseases," Plants, vol. 12, art. 169, 2023, doi: 10.3390/plants12010169.

[48] S. Liu, Y. Qiao, J. Li, H. Zhang, M. Zhang, and M. Wang, "An improved lightweight network for real-time detection of apple leaf diseases in natural scenes," Agronomy, vol. 12, art. 2363, 2022, doi: 10.3390/agronomy12102363.