

# Data Leakage in Agricultural Object Detection Studies: A Critical Analysis, Recipes, and Recommendations

Mohieddine Jelali

`mohieddine.jelali@th-koeln.de`

TH Köln - University of Applied Sciences

---

## Research Article

**Keywords:** Object Detection, Machine Learning Models, Data Augmentation, Data Leakage, Inflated Performance Metrics, Reproducibility, PlantDoc

**Posted Date:** October 15th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11368884/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Data Leakage in Agricultural Object Detection Studies: A Critical Analysis, Recipes, and Recommendations

Mohieddine Jelali<sup>1\*</sup>

<sup>1</sup>Cologne Laboratory of Artificial Intelligence and Smart Automation (CAISA), Institute of Product Development and Engineering Design (IPK), Technische Hochschule Köln—University of Applied Sciences, Betzdorfer Str. 2, Cologne, 50679, Germany.

Corresponding author(s). E-mail(s): [mohieddine.jelali@th-koeln.de](mailto:mohieddine.jelali@th-koeln.de);

## Abstract

Data leakage is a widespread yet frequently overlooked pitfall in agricultural object detection research. It silently inflates reported performance metrics and undermines the practical reliability of machine learning (ML) models deployed in real-world farming conditions. This paper examines the characteristics of agricultural data that make these studies vulnerable to leakage. These characteristics include temporal and spatial autocorrelation, batch effects from data collection, sequential video frames, limited annotated data, image-level near-duplication, and merging datasets without checking for duplication.

Drawing on best practices from ecology, geostatistics, and ML reproducibility research, we have developed a systematic, 12-criterion (C1–C12) recipe for identifying, assessing, and mitigating leakage risk. The recipe assigns each paper a score based on the applicable criteria (with a variable maximum) and maps the score to one of five bands: Strong, Minor Gaps, Material Gaps, Serious, or Critical. This process separates genuine paper-specific leakage signals from field-level reporting norms that no paper in the corpus can currently satisfy.

We applied the recipe to 33 peer-reviewed journal articles published between 2023 and 2026 that used the PlantDoc dataset for plant disease detection or classification. The distribution is centered on the middle band: 7 papers are rated "Strong," 7 are rated "Minor Gaps," 15 are rated "Material Gaps," 4 are rated "Serious," and none are rated "Critical."

The scope of augmentation is the dominant discriminator: papers that explicitly restrict augmentation to the training split average approximately 89% of their applicable maximum. Those that leave the scope ambiguous average

approximately 57%, and those that apply augmentation before splitting average approximately 31%. Reported performance is inversely correlated with documentation quality at the high end. The three highest PlantDoc mAP@0.5 values among papers with C8 ("Augmentation and resampling restricted to training")  $\leq 2$  in the corpus are 97.4%, 96.57%, and 84.5%. Only 5 of the 33 papers (15.2%) provide both public code and a public dataset link, and only two perform a within-dataset near-duplicate audit (with one performing a cross-dataset check). We conclude that the field is advancing faster in reported performance than in leakage safeguards. We provide practical recommendations for authors, reviewers, and dataset maintainers. Adopting a single convention — an explicit augmentation scope statement — would substantially raise the corpus mean and elevate the 4 papers rated "Serious" to the "Material Gaps" band.

**Keywords:** Object Detection, Machine Learning Models, Data Augmentation, Data Leakage, Inflated Performance Metrics, Reproducibility, PlantDoc

## 1 Introduction

Agricultural object detection has become one of the most active application areas of deep learning (DL). Models for plant disease detection, crop harvesting, and growth-stage monitoring are now routinely trained on public datasets such as PlantVillage and PlantDoc, and reported accuracies frequently exceed 90%. Yet a growing body of evidence across adjacent fields — computer security, medical imaging, autonomous driving, soil science — shows that such headline numbers are often inflated by a silent and systematic pitfall: data leakage [1–6]. This paper argues that the same pitfall is widespread in agricultural object detection, that it is concentrated in a small number of repeated practices, and that it can be diagnosed with a lightweight, reproducible protocol.

### 1.1 Data leakage: a field-level problem, not a paper-level accident

Data leakage is the introduction of information into the model training process that would not be available at deployment time. It produces a spurious relationship between predictors and target that is not present in the distribution underlying the model's intended use, and it consequently overestimates model performance [1]. In the standard inductive learning setting — the setting we consider here, in which the goal is generalization to unseen data instances — leakage is a validity threat, not merely a technical nuisance.

Apicella et al. [2] describe the conditions under which leakage spreads as a "push the button" phenomenon: as ML tooling becomes easier to use, practitioners without the methodological background to detect common pitfalls reproduce them at scale. The largest cross-domain survey to date, by Kapoor et al. [1], examined 17 fields and identified pitfalls affecting 294 papers. Agricultural object detection has not yet

been the subject of a comparable systematic audit, but the field exhibits precisely the structural characteristics that make leakage likely.

## 1.2 Why agricultural object detection is particularly exposed

Agricultural datasets differ from general-purpose benchmarks such as ImageNet or COCO in ways that matter for leakage. ImageNet and COCO images are independently sampled from a broad distribution. Agricultural images, by contrast, are typically collected from a limited number of fields, over a short time window, using a small number of cameras, and often include repeated observations of the same plants. This creates strong dependence structures — temporal, spatial, and hierarchical — that violate the independence assumptions underlying standard random train/test splitting.

Two pressures compound this vulnerability. First, agricultural annotation requires specialist expertise, so labeled data are scarce [7]. Second, the field has adopted augmentation as a standard remedy: it is used to expand effective sample size, to balance classes, and to emulate the variability of real deployment conditions [8–11]. When augmentation is applied before the train/test split — as is common in practice — augmented versions of the same image can appear in both training and test sets, producing leakage that is invisible in the reported metrics [12].

## 1.3 A reproducibility problem with field-level causes

The consequences of leakage are not confined to individual papers. Reported accuracies of 99% or higher on agricultural datasets should be treated with caution unless the evaluation protocol explicitly addresses the dependence structures described above [1, 13]. Models that perform well during development fail in deployment, particularly in agriculture, where a missed disease detection can lead to significant crop losses [14]. Downstream research builds on findings that cannot be reproduced, and practitioners lose trust when they adopt leaky models and experience failures [15].

Current reporting conventions do not require authors to state whether augmentation was applied before or after the split, whether the split preserves class distribution, whether the test set was checked for near-duplicates, or where normalization statistics were computed. Without these statements, reviewers cannot determine whether a reported result is a genuine estimate of generalization performance or an artifact of a particular preprocessing sequence.

## 1.4 Contribution: a diagnostic protocol and a first field audit

This paper makes three contributions:

1. We synthesize the mechanisms by which data leakage arises in agricultural object detection, drawing on the ecological and geostatistical literature that has long recognized the inadequacy of random cross-validation for spatially and temporally structured data [16–21], as well as recent ML-reproducibility studies that extend these concerns to non-spatial leakage pathways [22, 23]. We identify six recurring causes: temporal and spatial autocorrelation, batch effects from collection

conditions, sequential video frames, limited annotated data, image-level near-duplication across merged public datasets, augmentation applied before splitting, and dataset-level duplication without a deduplication audit.

2. We introduce a diagnostic recipe consisting of 12 criteria, each of which is scored on a categorical scale. Most criteria are scored on a scale of 0/0.5/1, while the augmentation-scope criterion, which carries the largest weight, is scored on a scale of 0/2/4. The criteria are mapped to five bands: Strong, Minor Gaps, Material Gaps, Serious, and Critical. The recipe is deliberately lightweight; it only requires information that a paper can reasonably be expected to report. It also separates paper-specific leakage signals from field-level reporting norms that no paper in the current literature can satisfy. Many PlantDoc-derived releases lack plant- or plot-level identifiers, making strict group-aware splitting difficult without additional annotation effort. Rather than penalizing papers for factors beyond their control, the recipe excludes N/A criteria from the denominator and awards points to papers that surpass the field baseline.
3. We apply the recipe to 33 peer-reviewed journal articles published between 2023 and 2026 that used the PlantDoc dataset for plant disease detection or classification. We selected PlantDoc because it is the most widely used, field-realistic benchmark. Its web-scraped origin and lack of plant-level metadata make it the ideal test bed for the identified leakage mechanisms. Additionally, its structural characteristics—small lesions, complex backgrounds, and class imbalance—make it a challenging test case for leakage detection. The audit produces a risk profile of the field, quantifies criterion-level gaps, and identifies four papers that can serve as templates for future documentation.

## 1.5 Roadmap

The scope of the analysis is deliberately limited to journal articles using PlantDoc, because short conference papers typically lack the methodological detail needed for leakage assessment. Section 2 reviews the mechanisms and causes of leakage in agricultural studies. Section 3 presents the recipe, the field-baseline adjustment, and the audit results. Section 4 distills the lessons and offers recommendations for authors, reviewers, and dataset maintainers. Section 5 concludes the work and gives an outlook on future work.

## 2 Background and related work

This section reviews the mechanisms by which data leakage arises in machine learning (ML) and DL studies, with particular attention to the structural characteristics of agricultural data that make such studies vulnerable. The section is organized into three parts: forms and sources of leakage, causes specific to agricultural object detection, and consequences for validity and reproducibility.

## 2.1 Forms and sources of data leakage

We adopt the tripartite taxonomy of Apicella et al. [2], which distinguishes three sources of leakage:

1. **Data-induced leakage** occurs when spurious information is embedded in a dataset during its construction. This arises when a dataset contains duplicates or near-duplicates of the same image that appear in both the training and test splits, whether the dataset is newly constructed or reused from a public repository. Data-induced leakage is particularly difficult to detect because it is a property of the data, not of the pipeline. It persists regardless of how carefully the split is performed.
2. **Preprocessing-related leakage** occurs when a preprocessing step is applied without respecting the separation between training and evaluation data. Operations such as imputation, oversampling, undersampling, normalization, and augmentation are frequent sources. When they are computed over the *full* dataset — rather than over the training split only — their parameters encode information from the test set. Augmentation applied before splitting is the most consequential instance in agricultural computer vision.
3. **Split-related leakage** occurs when information leaks between training and evaluation sets during the split itself. Two subtypes are relevant. *Similarity leakage* arises when the two sets share identical or near-identical instances; this is the failure mode we focus on in the present study. *Distribution leakage* arises when the split ignores heterogeneous distributions in the data — for example, when the split is random over a dataset that contains strong field-level or temporal structure.

All three sources are relevant to agricultural object detection. The empirical part of this study introduces criteria for all these leakage-source types.

## 2.2 Causes of data leakage in agricultural studies

Agricultural object detection exhibits a set of structural characteristics that make it particularly vulnerable to leakage. We identify six recurring causes.

### 2.2.1 Temporal and spatial autocorrelation

Agricultural data are inherently non-independent. The same plant, tree, or field is often imaged repeatedly over time — at different growth stages, under varying weather conditions, or across multiple days. When images of the same plant or field appear in both the training and test sets, the models learn spatial and temporal cues instead of general botanical features.

This phenomenon, known as spatial autocorrelation, systematically produces over-optimistic performance estimates. The ecological and geospatial literature has long recognized that standard random cross-validation is inappropriate for structured data of this kind [16, 17]. Roberts et al. [16] demonstrated that random cross-validation produces severely biased estimates when data exhibit temporal, spatial, hierarchical, or phylogenetic structure, and recommended blocked or leave-one-group-out strategies. Meyer and Pebesma [18] further showed that spatial and temporal prediction tasks require special cross-validation designs, and provided implementations through the

blockCV R package [19]. Ploton et al. [20] and Schratz et al. [21] documented the same effect in forest biomass mapping and spatial hyperparameter tuning, respectively. These findings apply directly to agricultural object detection, where the underlying data structure is comparable.

### 2.2.2 Batch effects from data collection

Researchers frequently collect data in a single field on a single day using one camera setup. Random splitting of such images causes leakage because images taken in the same row, under identical lighting conditions, or within a few meters of each other share highly correlated backgrounds, shadows, and soil conditions. Samples within a group are far more similar to each other than to samples in other groups, violating the independence assumption underlying random cross-validation [16, 21].

The problem is compounded when images from the same plant appear in multiple splits. A single tomato plant imaged from different angles on the same day yields images that are near-duplicates for model training. Splitting them across train and test sets produces leakage that no random-split protocol can detect.

### 2.2.3 Sequential video frames

Many studies employ video or continuous image bursts. Adjacent frames are nearly identical, differing only marginally in perspective or timing. Randomly splitting them — rather than splitting by time segment — guarantees leakage and artificially inflates performance metrics. This is a form of group leakage in which samples from the same temporal unit are distributed across training and test sets [17, 22].

Botache et al. [17] analyzed different splitting strategies for sequential data and found that random splitting systematically overestimates model performance compared to time-segment-based splitting. Figueiredo and Mendes [23] confirmed this effect specifically for object detection tasks using video-derived images. The implication for agricultural studies is direct: any study using video or burst imagery must split by temporal unit, not by frame.

### 2.2.4 Limited annotated data

Annotation in agriculture requires specialist expertise and is therefore expensive [7]. The scarcity of labeled data creates pressure to use every available sample, often without adequate separation between training and evaluation sets. Studies that repeatedly image the same plant specimen may inadvertently place near-identical images in both training and test sets, producing inflated accuracy estimates that do not reflect true generalization [24].

This cause interacts with the others. When data are scarce, the temptation to include every available image grows, and the same specimen is more likely to appear in multiple splits. The result is a pipeline that appears to be a legitimate held-out evaluation but is in fact leaky.

### 2.2.5 Augmentation applied before splitting

Data augmentation comprises a family of techniques that artificially expand training datasets by generating modified versions of existing samples through predefined transformations [8]. In computer vision, these include geometric manipulations (rotation, flipping, cropping), photometric adjustments (brightness, contrast, colour jittering), and more advanced techniques such as cutout [9], mixup, and randaugment [10].

Augmentation is applied in agricultural computer vision to enhance model generalization by exposing the learning algorithm to a broader spectrum of data variations, and to alleviate the scarcity of labeled data [11]. When it is applied before the train/test split — as is common in practice — the resulting expanded dataset contains near-duplicates of the same source image, and these near-duplicates can appear in both training and test sets. This is the single most consequential leakage mechanism in the PlantDoc literature, and it is the focus of Criterion C5 in our recipe.

We emphasize a distinction that is sometimes missed. Visual similarity between two images is not sufficient to constitute leakage. Leakage requires that the two images share a common source — the same leaf, the same plant, or the same capture session. Augmentation, by construction, guarantees such commonality: every augmented image is derived from a specific source image. This is why augmentation applied before splitting is a form of leakage by construction, not merely a risk.

### 2.2.6 Merging datasets without duplication checks

When multiple datasets are combined — for example, PlantVillage, PlantDoc, and locally collected field images — the risk of duplication or near-duplication increases substantially. PlantVillage itself contains images of the same leaf imaged under different conditions, and merging it with datasets that include the same or similar specimens creates opportunities for leakage. Without systematic checks for overlapping images or near-identical instances, the same visual content may appear in both training and evaluation sets.

This cause is particularly relevant to the corpus we analyze, in which a substantial number of papers merge PlantVillage with PlantDoc, EdenLibrary, or their own field collections. None of the papers we examined performs a cross-source duplication audit.

## 2.3 Consequences of data leakage

The consequences of data leakage are severe and multifaceted:

- **Inflated performance estimates:** Models appear to perform at levels that cannot be reproduced on independent test data. Reported accuracies of 99% or higher on agricultural datasets should be treated with caution unless the evaluation protocol explicitly addresses the dependence structures described above [1, 13].
- **Failed generalization:** Models that perform well during development fail in real-world deployment. This is particularly problematic in agriculture, where a model that fails to detect a disease in the field can lead to significant crop losses [14].

- **Wasted resources:** Research effort and computational resources are spent on models that cannot be deployed, and downstream studies build on findings that are not reproducible [1].
- **Erosion of trust:** Farmers and agricultural practitioners who adopt leaky models experience failures that undermine confidence in AI-based agricultural tools [15].

These consequences are amplified in agriculture because the domain combines high deployment stakes with low tolerance for failure. A leaky model in a benchmark study produces an unreliable number; a leaky model in a deployment context produces a missed detection.

## 3 Methods and results

This section presents the methodology used to assess leakage risk in the corpus, the recipe and its scoring rules, the field-baseline adjustment, and the results of the audit. The procedure is designed to be reproducible and to be applicable, with minor adaptation, to other applied ML domains.

### 3.1 Scope of the analysis

We selected 33 journal articles published between 2023 and 2026 that address plant disease detection or classification using the PlantDoc dataset, either partially or entirely. Short conference papers were excluded because they typically lack the methodological detail needed for leakage assessment.

PlantDoc is a suitable benchmark for our purpose. Unlike PlantVillage, which was captured under laboratory conditions, PlantDoc reflects the complexity of real-world agricultural environments with varying lighting, diverse backgrounds, and multiple leaves per image [25]. Figure 1 shows representative images from both datasets. The variability of PlantDoc makes it a demanding test case for leakage detection, and it has been widely reused across the corpus [26].

Singh et al. [24] released PlantDoc in 2020. It contains 2,598 images spanning 13 plant species and 27 to 30 classes, with approximately 8,851 bounding-box annotations. Most images depict plants in natural surroundings; a smaller subset was captured under laboratory-like conditions. Unlike PlantVillage, PlantDoc does not ship with a canonical train/validation/test split, nor with sample-level, plant-level, or field-level identifiers. This has two consequences for leakage assessment. First, authors must define their own split, which means the split protocol is paper-specific and must be documented by the authors. Second, criteria that depend on group or temporal identifiers (Criteria C3 and C5) cannot be satisfied in the strict sense on PlantDoc, because the identifiers required to construct groups or temporal orderings are not part of the dataset.

The PlantDoc dataset is publicly available on GitHub [27], Kaggle [28], and Hugging Face [29], with a standard train/test split of 2,342/236 images. Studies typically carve out a validation set (often 10%) from the training partition themselves [30]. Roboflow [31] hosts an improved and corrected version of PlantDoc. The relationship of the Roboflow version to the original split is nuanced: community members

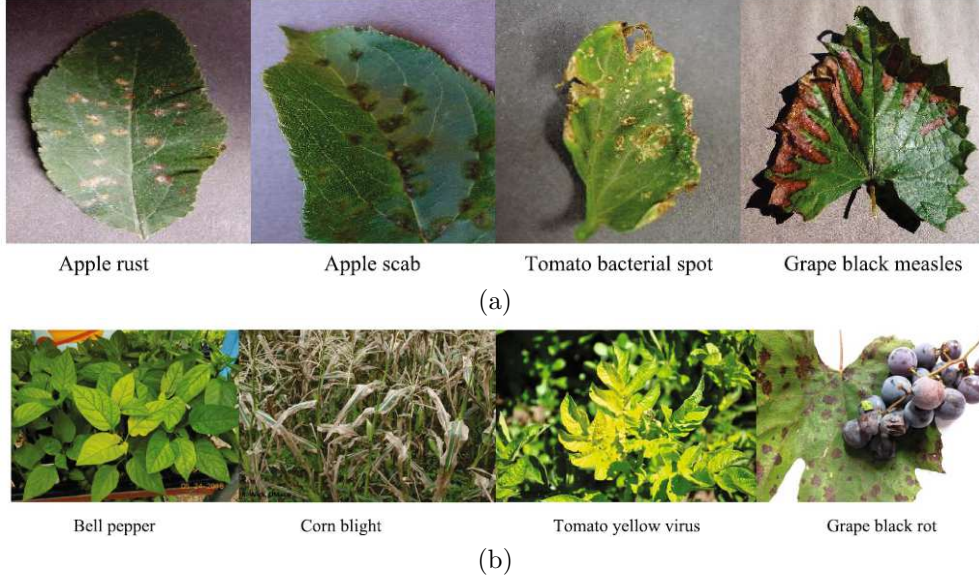


**Fig. 1:** Sample images of (a) PlantVillage and (b) PlantDoc datasets.

upload many PlantDoc variants, and each variant has its own split, configured during upload. For example, one community version uses a 90/8/2 split (5,778/513/128 images, train/valid/test) [32], but their data may include additional augmentations or different configurations that are not always described. We treat the original Singh et al. [24, 27] release as the reference dataset and note that many papers in the corpus use one of the Roboflow variants without specifying which one.

### 3.2 Criteria for data leakage-free studies

We specify 12 criteria, labeled C1 through C12. Table 1 lists the criteria and the diagnostic signals that identify compliance in a paper’s text. This is an extension of the 8-type leakage taxonomy by Kapoor and Narayanan [1], the most widely cited framework, which covers the most common data leakage pitfalls in ML-based science. Criterion C8 is deliberately weighted above the others because augmentation applied before the split is the most common and most damaging leakage mechanism (Section 2.2).

### 3.3 Scoring and quality labels

Each criterion is scored according to Table 2. For all criteria except C8, we propose a three-subscore assessment: 0 (not fulfilled), 0.5 (partially fulfilled), or 1 (fulfilled). Criterion C8 carries the highest weight (4 for explicit train-only, 2 for vague, 0 for absent or pre-split). The maximum total is 15 when all criteria are applicable; because some criteria are N/A for many papers (notably C5 for all 33), the applicable maximum varies between 11 and 13 in this corpus.

The total score maps to five quality labels: *Critical/fundamental failures*, *Serious integrity concerns*, *Material gaps*, *Minor gaps*, and *Strong leakage hygiene*. Table 3 provides the numeric mapping from score range to quality label, along with the qualitative definitions and interpretations that guide how to read the reported results.

**Table 1:** Criteria C1–C12 for the assignment of the quality label of data leakage.

| No. | Criterion                                      | Requirement for agricultural studies                                                                                                                                                                                                                                                                                                         | Status                                     |
|-----|------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------|
| C1  | Split design documented                        | Report how train, validation, and test sets were created across farms, plots, greenhouses, seasons, growth stages, or camera sessions. State whether the split is by image, by plant, by plot, by field, or by campaign, and whether it is temporal (e.g., season 1 $\rightarrow$ season 2) or random. Give per-class and per-split counts.  | Necessary                                  |
| C2  | Sample and group counts documented             | Report the number of images, unique plants, unique leaves, plots, and annotated lesions per species/disease/severity class, and per split. Include class imbalance metrics (long-tail ratio, effective coverage).                                                                                                                            | Necessary                                  |
| C3  | Appropriate grouping used                      | Keep all images of the same plant, same leaf, same plot, same field, or same acquisition session in the same split whenever independence at that level is required. Report any stratification by species, disease, severity, or geography separately.                                                                                        | Necessary when grouped data exist          |
| C4  | Duplicate and derivative overlap excluded      | Verify that web-scraped duplicates, re-photographed leaves, repeated frames from the same session, crops/patches, and augmented derivatives (flips, rotations, mosaics, GAN/diffusion outputs) do not cross the train/validation/test boundary. Document cross-dataset overlap (e.g., PlantVillage and PlantDoc, PlantDoc and Plants Final). | Necessary                                  |
| C5  | Temporal and prediction-time integrity ensured | Use only information available at the moment of in-field deployment (before or at image capture). Prevent future observations, harvest-time labels, weather post-event data, or retrospectively derived severity scores from entering features or training.                                                                                  | Necessary for temporal or prediction tasks |
| C6  | Target and post-outcome leakage excluded       | Exclude features that directly or indirectly encode the disease label, are derived from the annotation, or are generated after the disease event (e.g., lesion masks, severity grades, harvest outcomes, expert diagnoses used as inputs).                                                                                                   | Necessary                                  |

**Table 1** (continued)

| No. | Criterion                                                 | Requirement for agricultural studies                                                                                                                                                                                                                                                                                                                | Status                                     |
|-----|-----------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------|
| C7  | Data-dependent preprocessing fitted on training data only | Fit normalization statistics, color balancing, segmentation thresholds, disease-region detectors, and any data-dependent preprocessing on the training split only. Confirm that validation and test images are not used to estimate these parameters.                                                                                               | Necessary                                  |
| C8  | Augmentation and resampling restricted to training        | Split first, then augment. Keep all derivatives of an original image (mirror, crop, rotation, mosaic, brightness jitter, GAN/diffusion output, copy-paste lesions, SMOTE-style resampling) in the same split. This includes any pre-augmented Roboflow/Kaggle release used as-is — the augmentation must have been applied after the split.         | Conditional                                |
| C9  | Synthetic-data generation restricted to training data     | Train, tune, and condition GANs, diffusion models, or other synthetic generators only on training images and training labels. Confirm that test-set observations or test-set labels are not used to condition, fine-tune, or filter synthetic samples.                                                                                              | Conditional                                |
| C10 | Model selection independent of the test set               | Do not use the held-out field/season/plot data for feature selection, hyperparameter search, architecture search, checkpoint selection, threshold tuning, calibration, seed selection, or experiment selection. Report the selection criterion and the split used to compute it.                                                                    | Necessary                                  |
| C11 | Cross-validation pipeline is leakage-safe                 | Perform all learned preprocessing, feature selection, augmentation, synthetic generation, and tuning inside each training fold. When the same CV process supports both tuning and final reporting, use nested CV. Use group-aware folds when the data are grouped by plant, plot, field, or session.                                                | Conditional                                |
| C12 | External-data provenance and overlap addressed            | Document external datasets, pretrained backbones (e.g., ImageNet, COCO, PlantVillage-derived weights), retrieval corpora, and public plant-disease repositories. Assess whether any of these overlap with the evaluation data (image overlap, class overlap, species overlap) and report the result. Provide a data/code release whenever possible. | Necessary when external resources are used |

### 3.4 Results of the audit

Tables 4 and 5 present the scores and quality labels for all 33 papers [33–65]. Figure 2 summarises the distribution. Of the 33 papers, 4 are rated Serious (score  $\leq 39\%$ ), 15 Material gaps (40–64%), 7 Minor gaps (65–84%), and 7 Strong (85–100%).

**Table 2:** Scoring rules for the leakage criteria C1–C12.

| No. | Criterion                                                 | Scoring                                                                                                                                                                                                                                                                 |
|-----|-----------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| C1  | Split design documented                                   | <b>1:</b> Proportions, split unit, temporal order (where relevant), and pre-defined vs. random stated explicitly.<br><b>0.5:</b> Some elements reported; missing proportions, split unit, or temporal ordering.<br><b>0:</b> No description of how splits were created. |
| C2  | Sample and group counts documented                        | <b>1:</b> Observation counts and unique-entity counts per split reported, including post-exclusion changes.<br><b>0.5:</b> Totals given but per-split or per-entity counts missing.<br><b>0:</b> No counts reported.                                                    |
| C3  | Appropriate grouping used                                 | <b>1:</b> Group-aware splitting applied and documented; stratification reported separately.<br><b>0.5:</b> Grouping acknowledged but implementation unclear.<br><b>0:</b> Grouped observations split across train/test.                                                 |
| C4  | Duplicate and derivative overlap excluded                 | <b>1:</b> Deduplication performed and described; derivatives (crops, frames, augmentations) kept within splits.<br><b>0.5:</b> Deduplication mentioned but scope or method unclear.<br><b>0:</b> No deduplication reported.                                             |
| C5  | Temporal and prediction-time integrity ensured            | <b>1:</b> Only prediction-time-available information used; temporal ordering enforced.<br><b>0.5:</b> Temporal split used but feature availability at prediction time unclear.<br><b>0:</b> Future or retrospectively derived information used.                         |
| C6  | Target and post-outcome leakage excluded                  | <b>1:</b> Features audited; target-derived and post-outcome features excluded and documented.<br><b>0.5:</b> Some feature auditing; exclusion rationale incomplete.<br><b>0:</b> No audit of target/post-outcome features.                                              |
| C7  | Data-dependent preprocessing fitted on training data only | <b>1:</b> All learned preprocessing (scaling, imputation, encoding, selection, reduction, batch correction) fit on training only.<br><b>0.5:</b> Some steps correctly scoped; others unclear or global.<br><b>0:</b> Preprocessing fit on full dataset.                 |

**Shape of the distribution.**

- Roughly unimodal, centered on the Material-gap band: 45.5% of papers.
- Right-skewed: a small cluster at the top (7 papers, 21.2%) and a long tail toward lower scores.
- No papers in the Critical band: the field’s reporting floor scores around 9.1%.
- Bimodal signal at the top: the Strong band contains two subgroups:
  - Very strong ( $\geq 95\%$ ) studies: Işık and Keskin [58], Chen et al. [55], and Zhao et al. [63]
  - Strong but isolated (85–92%) studies: Hamad et al. [42], Liu et al. [45], and [53]
- Serious cluster is small but consistent: 4 papers ([33], [40], [43], and [60]) share confirmed pre-split augmentation or unassessed cross-dataset overlap.

**Table 2** (continued)

| No.          | Criterion                                             | Scoring                                                                                                                                                                                                                                                                                |
|--------------|-------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| C8*          | Augmentation and resampling restricted to training    | <b>4:</b> Data split before augmentation/resampling; all derivatives confined to their split.<br><b>2:</b> Split-before-augmentation stated but derivatives or ordering partially unclear.<br><b>0:</b> Augmentation/resampling applied before splitting, or derivatives cross splits. |
| C9           | Synthetic-data generation restricted to training data | <b>1:</b> Generator trained, tuned, and conditioned without validation/test observations or labels.<br><b>0.5:</b> Generator scoped to training but conditioning or tuning unclear.<br><b>0:</b> Generator exposed to validation/test data or labels.                                  |
| C10          | Model selection independent of the test set           | <b>1:</b> Test set untouched for feature selection, tuning, architecture, checkpoints, thresholds, calibration, seeds, or experiment selection.<br><b>0.5:</b> Some selection steps isolated; others ambiguous.<br><b>0:</b> Test set used for any selection decision.                 |
| C11          | Cross-validation (CV) pipeline is leakage-safe        | <b>1:</b> All learned steps inside each training fold; nested CV for tuning + final estimation.<br><b>0.5:</b> CV used but some steps outside folds, or nesting absent when needed.<br><b>0:</b> Preprocessing/selection/tuning performed outside folds.                               |
| C12          | External-data provenance and overlap addressed        | <b>1:</b> External datasets, pretrained models, and retrieval corpora documented; overlap with evaluation data assessed.<br><b>0.5:</b> Resources listed but overlap not assessed.<br><b>0:</b> External resources undocumented.                                                       |
| <b>Total</b> | <b>C1–C12</b>                                         | <b>15</b> (max. when all criteria are applicable.)                                                                                                                                                                                                                                     |

\* In agricultural studies, where small datasets make augmentation the primary countermeasure, C8 is the highest-impact criterion, and the one most often failed. The 4× weight reflects this severity; the half-credit anchor (2) captures the common case in which training-scoped augmentation is claimed but split-before-augmentation is not demonstrated.

The 33-paper audit reveals a Material-gap-dominated distribution: a small, well-behaved top tier; a large middle tier, where most leakage safeguards are claimed but not demonstrated; and a small, Serious-gap tail, which is dominated by pre-split augmentation and unassessed cross-dataset overlap. While the field as a whole is not failing catastrophically, two-thirds of the papers have at least one unverified or confirmed leakage pathway.

#### Criterion-level statistics.

- C1 (Split design documented): 19 papers (57.6%) score 1.0; 14 papers (42.4%) score 0.5.
- C2 (Sample and group counts documented): 29 papers (87.9%) score 1.0; 4 papers (12.1%) score 0.5.
- C3 (Appropriate grouping used): C3 is applicable only to 4 papers (12.1%); all four score 0.5.

**Table 3:** Mapping from total score to quality label, with qualitative definitions.

| Quality label                        | Score range | Definition                                                                                                                                   | Interpretation                                                                                                |
|--------------------------------------|-------------|----------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|
| Critical/<br>fundamental<br>failures | 0–19%       | Fundamental methodological breakdowns render the reported results uninterpretable or invalid. Evaluation cannot be trusted for field claims. | The paper is not a valid scientific contribution in its current form; rejection or retraction is recommended. |
| Serious<br>integrity<br>concerns     | 20–39%      | Serious methodological concerns. Reported performance likely over-optimistic; major overlap or overlap-adjacent risks.                       | Results are unreliable; the paper requires major revision with a proper held-out evaluation.                  |
| Material<br>gaps                     | 40–64%      | Some gaps in safeguard documentation. Results are plausible but not fully verified. Results may not transfer to real fields.                 | Results should be interpreted with caution; independent replication is recommended.                           |
| Minor<br>gaps                        | 65–84%      | Strong safeguards; results are likely reliable. Minor omissions do not affect validity.                                                      | Results can be trusted as genuine estimates of generalization performance.                                    |
| Strong<br>leakage<br>hygiene         | 85–100%     | All or nearly all safeguards are documented and consistent. Ready for field deployment; leakage risks identified and controlled.             | Results are reliable; the paper can serve as a documentation template for the field.                          |

- C4 (Duplicate and derivative overlap excluded): Only 2 of 33 papers earn full credit; 19 papers (57.6%) score 0.0; 12 papers (36.4%) score 0.0.
- C5 (Temporal and prediction-time integrity ensured): C5 is N/A for all 33 papers because each paper solves a static-image classification, detection, or segmentation task, where the input image is the observation and the label describes the same image. There is no prediction-time gap, no future target, and no retrospective feature pathway.
- C6 (Target and post-outcome leakage excluded): All 33 papers score 1.0.
- C7 (Data-dependent preprocessing fitted on training only): 5 papers (15.2%) score 1.0; 25 papers (75.8%) score 0.5; 3 papers (9.1%) score 0.
- C8 (Augmentation and resampling restricted to training): 9 papers (27.3%) score 4.0; 20 papers (60.6%) score 2.0; 4 papers (12.1%) score 0.
- C9 (Synthetic-data generation restricted to training): C9 is applicable only to one paper, but this scores 0.
- C10 (Model selection independent of test set): 10 papers (30.3%) score 1.0; 23 papers (69.7%) score 0.5; no paper scores 0.
- C11 (Cross-validation pipeline leakage-safe): C11 is applicable only to 3 papers; two of them score 1.0, the other 0.5.
- C12 (External-data provenance and overlap addressed): 5 papers (15.2%) score 1.0; 21 papers (63.6%) score 0.5; 7 papers (21.2%) score 0.

**Table 4:** Applicability and scoring of 12 criteria for assessing data leakage in the selected papers.

| Ref. | C1  | C2  | C3  | C4  | C5  | C6 | C7  | C8 | C9  | C10 | C11 | C12 | Raw sum | Appl. max |
|------|-----|-----|-----|-----|-----|----|-----|----|-----|-----|-----|-----|---------|-----------|
| [33] | 0.5 | 1   | N/A | 0   | N/A | 1  | 0   | 0  | N/A | 0.5 | N/A | 0.5 | 3.5     | 11        |
| [34] | 0.5 | 1   | N/A | 0   | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 0.5 | 6       | 11        |
| [35] | 1   | 1   | N/A | 0.5 | N/A | 1  | 0.5 | 2  | N/A | 1   | N/A | 0.5 | 7.5     | 11        |
| [36] | 0.5 | 1   | N/A | 0   | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 0   | 5.5     | 11        |
| [37] | 1   | 1   | N/A | 0.5 | N/A | 1  | 0.5 | 2  | N/A | 0.5 | 0.5 | 0.5 | 7.5     | 12        |
| [38] | 1   | 1   | 0.5 | 0.5 | N/A | 1  | 0.5 | 4  | N/A | 0.5 | N/A | 0.5 | 9.5     | 12        |
| [39] | 1   | 1   | N/A | 0.5 | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 0   | 6.5     | 11        |
| [40] | 0.5 | 1   | N/A | 0   | N/A | 1  | 0.5 | 0  | N/A | 0.5 | N/A | 0   | 3.5     | 11        |
| [41] | 0.5 | 0.5 | N/A | 0   | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 0   | 5       | 11        |
| [42] | 1   | 1   | N/A | 0.5 | N/A | 1  | 0.5 | 4  | N/A | 1   | N/A | 0.5 | 9.5     | 11        |
| [43] | 0.5 | 1   | N/A | 0   | N/A | 1  | 0   | 0  | N/A | 0   | N/A | 0   | 3       | 11        |
| [44] | 0.5 | 0.5 | N/A | 0   | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 0.5 | 5.5     | 11        |
| [45] | 1   | 1   | N/A | 0.5 | N/A | 1  | 1   | 4  | N/A | 0.5 | N/A | 0.5 | 9.5     | 11        |
| [46] | 1   | 0.5 | N/A | 0.5 | N/A | 1  | 0.5 | 4  | N/A | 1   | N/A | 0.5 | 9       | 11        |
| [47] | 0.5 | 0.5 | N/A | 0   | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 0.5 | 5.5     | 11        |
| [48] | 0.5 | 1   | N/A | 0.5 | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 0.5 | 6.5     | 11        |
| [49] | 1   | 1   | N/A | 0.5 | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 0.5 | 7       | 11        |
| [50] | 1   | 1   | N/A | 0.5 | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 0.5 | 7       | 11        |
| [51] | 0.5 | 1   | N/A | 0.5 | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 0.5 | 6.5     | 11        |
| [52] | 0.5 | 1   | N/A | 0   | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 0   | 5.5     | 11        |
| [53] | 1   | 1   | N/A | 0.5 | N/A | 1  | 0.5 | 4  | N/A | 1   | N/A | 0.5 | 9.5     | 11        |
| [54] | 1   | 1   | 0.5 | 1   | N/A | 1  | 1   | 4  | N/A | 0.5 | N/A | 1   | 11      | 12        |
| [55] | 1   | 1   | N/A | 0.5 | N/A | 1  | 1   | 4  | N/A | 1   | N/A | 1   | 10.5    | 11        |
| [56] | 0.5 | 1   | N/A | 0   | N/A | 1  | 0.5 | 2  | 0   | 0.5 | N/A | 0.5 | 6       | 12        |
| [57] | 1   | 1   | N/A | 0.5 | N/A | 1  | 0.5 | 2  | N/A | 1   | N/A | 0.5 | 7.5     | 11        |
| [58] | 1   | 1   | 0.5 | 1   | N/A | 1  | 1   | 4  | N/A | 1   | 1   | 1   | 12.5    | 13        |
| [59] | 1   | 1   | N/A | 0.5 | N/A | 1  | 0.5 | 2  | N/A | 1   | N/A | 0.5 | 7.5     | 11        |
| [60] | 0.5 | 1   | N/A | 0   | N/A | 1  | 0   | 0  | N/A | 0.5 | N/A | 0.5 | 3.5     | 11        |
| [61] | 0.5 | 1   | N/A | 0   | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 0   | 5.5     | 11        |
| [62] | 1   | 1   | 0.5 | 0.5 | N/A | 1  | 0.5 | 2  | N/A | 1   | 1   | 0.5 | 9       | 13        |
| [63] | 1   | 1   | N/A | 0.5 | N/A | 1  | 1   | 4  | N/A | 1   | N/A | 1   | 10.5    | 11        |
| [64] | 1   | 1   | N/A | 0.5 | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 1   | 7.5     | 11        |
| [65] | 1   | 1   | N/A | 0.5 | N/A | 1  | 0.5 | 2  | N/A | 0.5 | N/A | 0.5 | 7       | 11        |

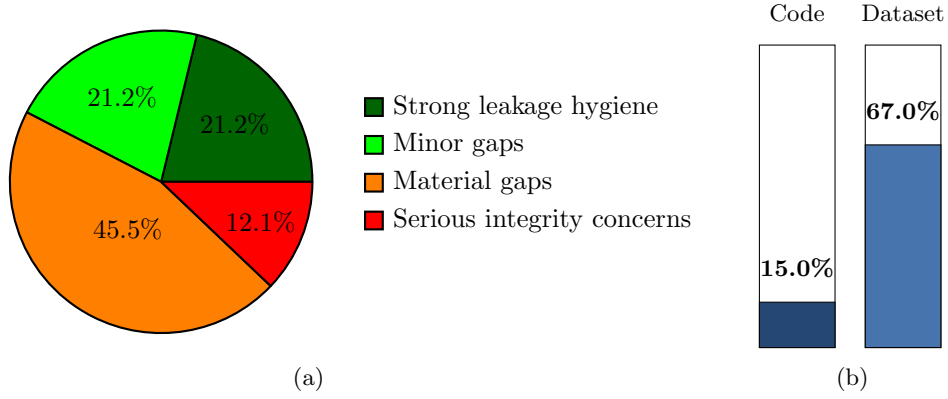


**Fig. 2:** Leakage risk assessment across the analyzed studies. (a) Quality label distribution. (b) Availability of code and datasets.

#### Key observations.

- C6 is the field’s universal strength: no study uses target-derived features.
- **C4 is the field’s systemic weakness:** 94% of papers fail or partially fail duplicate/derivative containment.
- **C8 is the field’s highest-impact criterion:** 60% the papers score 2/4, 12% score 0/4.
- C9 is a “trap” criterion: the one paper where it applies fails it.
- C3 and C11 are rarely tested but are handled correctly or near-correctly when they apply.
- C1, C2, and C6 are the field’s reporting strengths: documentation and count reporting are consistently strong.
- C7, C10, and C12 are the “almost there” criteria: a short explicit statement would upgrade most papers from 0.5 to 1.0.

### 3.5 Documentation versus performance

A striking pattern in the corpus is that reported performance and leakage documentation diverge at the top of the distribution. Some of the highest PlantDoc mAP@0.5 values in the corpus — 96.57% and 84.5% — all come from papers with Serious or Material gaps.

Conversely, several papers with lower reported metrics score (mAP@0.5 in the 60–85% range) in the Strong leakage hygiene band. These papers do not claim the highest metrics in the field, but their documentation discipline is far stronger. Their reported numbers are more credible. The one paper that combines a top-tier PlantDoc result (85.3% mAP@0.5) with top-tier documentation (score 95.5%, Strong leakage hygiene) demonstrates that strong results and strong documentation are compatible.

This divergence is consistent with the hypothesis that the scope of undocumented augmentation can inflate reported PlantDoc metrics. However, this is not proof. Some

**Table 5:** Data leakage risk assessment results of the selected papers on disease detection using ML/DL models.

| Ref. | Network architecture(s)                                              | Dataset(s) used                                         | Key issue(s)/finding(s)                                                                                                                       | Score / quality label                        |
|------|----------------------------------------------------------------------|---------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------|
| [33] | YOLOv7, YOLOv8 (default and modified with frozen backbone)           | PlantDoc, Plants Final                                  | Plants Final explicitly contains PlantDoc by construction, and overlap is unaddressed. Pre-augmented Roboflow version used. No deduplication. | 3.5/11 (31.8%)<br>Serious integrity concerns |
| [34] | Inception Convolutional Vision Transformer (ICVT)                    | PlantVillage, iBean, AI2018, PlantDoc                   | ImageNet pretraining and PlantDoc evaluation without overlap check, and PlantVillage PlantDoc overlap unaddressed.                            | 6.0/11 (54.5%)<br>Material gaps              |
| [35] | IBSA_Net (inverted bottleneck plus Shuffle Attention plus HardSwish) | PlantVillage plus PDDA plus PlantDoc++                  | PDDA is a derivative of PlantVillage, and PDDA PlantVillage overlap is unaddressed. Split-before-augmentation implied but not explicit.       | 7.5/11 (68.2%)<br>Minor gaps                 |
| [36] | YOLOv5, YOLOX, YOLOv7, YOLOv8, YOLO-NAS                              | Merged PlantVillage and PlantDoc (tomato subset)        | PlantVillage and PlantDoc merged without deduplication. No split proportions or per-split counts.                                             | 5.5/11 (50.0%)<br>Material gaps              |
| [37] | DM-YOLO (YOLOv8-based with MultiCat and ADC2f and C2fe)              | Merged PlantVillage and PlantDoc (tomato subset)        | Three-dataset merge with no cross-dataset dedup, and split-before-augmentation ordering unstated.                                             | 7.5/12 (62.5%)<br>Material gaps              |
| [38] | DAE-Mask (Mask R-CNN with DenseNet-CBAM-FPN and Edge Agreement Head) | MSWDD2022 (self-built), PlantDoc                        | Group structure (3 bases, 200+ varieties) but not enforced in split. Split-before-augmentation explicitly stated.                             | 9.5/12 (79.2%)<br>Minor gaps                 |
| [39] | ResNet50, ViT with PMF pipeline and Feature Attention module         | PlantVillage, PlantDoc repurposed for few-shot learning | Class-disjoint splits, but ImageNet pretraining and PlantDoc evaluation overlap unaddressed.                                                  | 6.5/11 (59.1%)<br>Material gaps              |
| [40] | FSM-YOLO (YOLOv8s with AFEM and SCAA and MKMC)                       | ALDD (custom), PlantDoc                                 | No test set (validation only), three-dataset merge without cross-dataset dedup, and ImageNet pretraining.                                     | 3.5/11 (31.8%)<br>Serious integrity concerns |
| [41] | Swin Transformer Small (ST-S) with four hybrid TL strategies         | PlantVillage, PlantDoc                                  | ImageNet-1K pretraining and PlantVillage/PlantDoc evaluation with no overlap assessment. PlantDoc augmentation scope ambiguous.               | 5.0/11 (45.5%)<br>Material gaps              |

**Table 5** (continued)

| Ref. | Network architecture(s)                                                                                        | Dataset(s) used                                                 | Key issue(s)/finding(s)                                                                                                                               | Score / quality label                        |
|------|----------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------|
| [42] | Pipeline: YOLO11 (leaf detection), SAM (segmentation), ResNet-50 (classification), and LIME (interpretability) | PlantVillage, PlantDoc (tomato subset), Roboflow leaf detection | PlantDoc PlantVillage overlap unaddressed, and ImageNet pretraining used. Training-only augmentation explicit.                                        | 9.5/11 (86.4%)<br>Strong leakage hygiene     |
| [43] | YOLOv8n                                                                                                        | PlantDoc (tomato subset)                                        | 9x dataset-level augmentation (737 to 6696) with no derivative containment, arithmetic error, and COCO pretraining plus PlantDoc overlap unaddressed. | 3.0/11 (27.3%)<br>Serious integrity concerns |
| [44] | YOLOv8n                                                                                                        | PlantDoc                                                        | Augmentation in preprocessing before split statement, no dedup, and PlantDoc overlap unaddressed.                                                     | 5.5/11 (50.0%)<br>Material gaps              |
| [45] | YOLO-BSMamba (YOLOv8s with HCMamba, SimAM, and BiFPN)                                                          | PlantVillage, PlantDoc, Taiwan tomato leaf                      | Augmentation employed exclusively on the training set, 80:10:10 split, and ablation thorough. Weakness is missing PlantDoc internal dedup.            | 9.5/11 (86.4%)<br>Strong leakage hygiene     |
| [46] | SerpensGate-YOLOv8 (with DySnakeConv, SPPELAN, and Super Token Attention)                                      | PlantDoc                                                        | No pretraining weights, mosaic disabled in last 10 epochs, and ablation without pretraining. Weakness is incomplete per-class counts (top-10 only).   | 9.0/11 (81.8%)<br>Minor gaps                 |
| [47] | YOLO-AgriNet (YOLOv8 with CBAM, ASPP, and additional Stage Layer 5)                                            | PlantDoc plus custom Cameroon images                            | PlantDoc web-scraped, pre-augmented Roboflow version used, no dedup, and split-before-augmentation unstated.                                          | 5.5/11 (50.0%)<br>Material gaps              |
| [48] | SoyaTrans (CNN plus Swin Transformer with random shifting)                                                     | Soybean (custom), PlantVillage, AI2018, PlantDoc, Embrapa       | No validation set, benchmark splits undocumented, and VGG16 ImageNet pretraining overlap unassessed.                                                  | 6.5/11 (59.1%)<br>Material gaps              |
| [49] | MobileNetV3Small, EfficientNetB0, DenseNet121                                                                  | PlantVillage, PlantDoc                                          | PlantVillage PlantDoc overlap unaddressed, augmentation ordering ambiguous, and web-sourced images added without deduplication.                       | 7.0/11 (63.6%)<br>Material gaps              |

Table 5 (continued)

| Ref. | Network architecture(s)                                                      | Dataset(s) used                                             | Key issue(s)/finding(s)                                                                                                                                                                                             | Score / quality label                     |
|------|------------------------------------------------------------------------------|-------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------|
| [50] | CAPNet (YOLOv8 with CGFM, PFEM, and ARFAhead)                                | TLDD (self-built), PlantDoc, DDCVP                          | No test set (validation only), ImageNet and PlantDoc overlap unaddressed, and split-before-augmentation unstated.                                                                                                   | 7.0/11 (63.6%)<br>Material gaps           |
| [51] | WD-YOLO (YOLOv10m with DSConv and WTConcat, and KD from YOLOv10l)            | PlantDoc                                                    | PlantDoc only, no deduplication, augmentation on 1,798 training images (scope partially explicit), and teacher model pretraining provenance unstated.                                                               | 6.5/11 (59.1%)<br>Material gaps           |
| [52] | ST-CFI (Swin Transformer with Convolutional Feature Interactions)            | PlantVillage, Plant Pathology 2021, PlantDoc, AI2018, iBean | ImageNet and COCO pretraining, PlantDoc PlantVillage overlap unaddressed, and class-disjoint split for AI2018 only.                                                                                                 | 5.5/11 (50.0%)<br>Material gaps           |
| [53] | Ensemble of InceptionResNetV2, MobileNetV2, and EfficientNetB3               | PlantVillage, PlantDoc, FieldPlant                          | One of the cleanest papers in the audit, the only one to test on three datasets, but PlantVillage–PlantDoc overlap unaddressed, normalization statistics scope unstated. Same universal issue as every other paper. | 9.5/11 (86.4%)<br>Strong leakage hygiene  |
| [54] | Ensemble of MobileNet-v3, ShuffleNet-v2, and SqueezeNet                      | PlantVillage, EdenLibrary, PlantDoc                         | Explicit split-before-augmentation, sample-wise dedup, group-aware CV by unique leaf ID, and crops excluded for unseen-crop test.                                                                                   | 11.0/12 (91.7%)<br>Strong leakage hygiene |
| [55] | YOLO11-ARL-PD (YOLOv11n with EMSCP, LADH, Wise-IoU, and LAFDP pruning)       | PlantDoc                                                    | Cross-dataset duplicate check performed, GitHub CALM-Aug release, and stratified 80:10:10 split. Only weakness is missing internal PlantDoc deduplication.                                                          | 10.5/11 (95.5%)<br>Strong leakage hygiene |
| [56] | MFVNet (YOLOv9-T with MEIE, Re-FCA, VBPN, and Stable Diffusion augmentation) | PlantDoc, Stable-PlantDoc, AI Challenger 2018               | Stable Diffusion synthetic generation without conditioning scope, split-before-generation unstated, and no code release.                                                                                            | 6.0/12 (50.0%)<br>Material gaps           |
| [57] | MobileNetV2, MobileNetV3s, ShuffleNetV2x0.5, YOLOv8n, YOLOv11n               | PlantDoc, Mango, Soybean Leaf Disease                       | Three-dataset evaluation, augmentation ordering ambiguous, pretraining provenance partial, and no code release.                                                                                                     | 7.5/11 (68.2%)<br>Minor gaps              |
| [58] | CA-SAUNet (SAUNet with Coordinate Attention)                                 | ISIC 2018, PlantDoc (pixel-annotated subset)                | Group-wise split by unique leaf ID, strict biological-sample separation, training-only normalization, 5-fold CV, and GitHub code release. Only weakness is ISIC random split.                                       | 12.5/13 (96.2%)<br>Strong leakage hygiene |

Table 5 (continued)

| Ref. | Network architecture(s)                                 | Dataset(s) used                                             | Key issue(s)/finding(s)                                                                                                                                   | Score / quality label                        |
|------|---------------------------------------------------------|-------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------|
| [59] | Optimized YOLOv11n                                      | PlantDoc (8-class subset)                                   | 80:10:10 stratified split and COCO pretraining disclosed. Weakness is augmentation scope not explicitly training-only.                                    | 7.5/11 (68.2%)<br>Minor gaps                 |
| [60] | YOLOv9                                                  | PlantDoc, Modified PlantDoc (with augmentation)             | Pre-augmented Roboflow Modified PlantDoc used as-is, mAP jumps 0.541 to 0.877 (+62%), and no split proportions.                                           | 3.5/11 (31.8%)<br>Serious integrity concerns |
| [61] | AD-DETR (RT-DETR with MSANet, SSAFF, and IP1oUv2)       | Crop Disease (self-built), PlantDoc                         | Self-constructed Crop Disease dataset un-released, PlantDoc 97.4% mAP (anomalous), and no split documentation.                                            | 5.5/11 (50.0%)<br>Material gaps              |
| [62] | VDD-Net (YOLOv10 with RFE, ACF, and GCA)                | PVDD (self-built), PlantDoc                                 | 5-fold stratified CV, mosaic disabled in final 10 epochs (training-only), and PVDD custom dataset. Weakness is no group-aware split.                      | 9.0/13 (69.2%)<br>Minor gaps                 |
| [63] | WGA-YOLO (YOLOv8n with WCR, PS-C2f, and DGAP)           | PlantDoc.boost (reconstructed), integrated external dataset | PlantDoc.boost publicly released, no pretrained weights, and explicit test-set-no-augmentation statement. Only weakness is PlantDoc PlantVillage overlap. | 10.5/11 (95.5%)<br>Strong leakage hygiene    |
| [64] | GBR-DETR (RT-DETR with MGAT-Net, BCPN, RFAM)            | M-TLD (self-built), PlantDoc                                | M-TLD dataset and code released, no pretraining, and 5-run seeds. Weakness is missing PlantDoc internal deduplication                                     | 7.5/11 (68.2%)<br>Minor gaps                 |
| [65] | YOLOv11n with SPPF-ASPP-asym and asymmetric convolution | PlantDoc                                                    | 3 seeds averaged and composite score transparent. Weakness is composite score used for both selection and reporting (circularity).                        | 7.0/11 (63.6%)<br>Material gaps              |

high-mAP papers may simply evaluate easier subsets or use non-standard protocols. However, the pattern is difficult to explain away entirely and reinforces the case for treating augmentation scope and including duplicate checks as primary reporting requirements.

### 3.6 Specific documentation–performance gap in agriculture research

Agricultural ML has three structural features that widen the documentation–performance gap:

- **Small, expensive datasets.** Field imagery is costly to collect and annotate. Augmentation is the primary countermeasure. When augmentation is applied before

splitting — as 4 of 33 papers do — the reported metric benefits from leaked near-duplicates. The gap between reported and true generalization can be 20–60 percentage points.

- **Public benchmark drift.** PlantVillage, PlantDoc, Plants Final, AI2018, and Embrapa overlap in undocumented ways. Papers that train and test across them without an overlap check routinely report cross-dataset metrics that partially reflect in-dataset memorization.
- **Pre-augmented third-party releases.** Roboflow and Kaggle provide ready-to-train versions of public plant disease datasets, often already augmented. Three papers in the audit use these directly [33, 43, 60]; a fourth paper [40] augments the original PlantDoc itself before splitting. The augmentation ordering is invisible to the authors, so the derivative-containment guarantee cannot be verified. The metrics are typically high; the C8 score is typically 0.

### 3.7 Selected assessments

To illustrate good examples, let’s go through the best assessments:

- **Işık and Keskin [58] — CA-SAUNet — Score 12.5/13 (96.2%).** This is the only paper in the audit with a full score for C7 (statement that it computes normalization statistics on the training set.) and C11 (Cross-validation pipeline leakage-safe). It is also one of only two papers with a full score C4, which includes a group-aware split by unique leaf ID and an explicit “separation of biological samples.” It trains on two structurally different domains (medical and agricultural) using the same disciplined pipeline. Moreover, it uses a paired t-test on the 5-fold CV to establish statistical significance; thus, it validates cross-domain robustness rather than merely claiming it. Its only weakness is the random split for the ISIC portion (C3 = 0.5), a minor blemish in an otherwise exemplary study.
- **Chen et al. [55] — YOLO11-ARL-PD/CALM-Aug — Score 10.5/11 (95.5%).** This is the only paper with zero flags, i.e., every safeguard it claims is demonstrated. It is the only paper to perform an explicit duplicate check between external test data and training data. It is also the only paper to release its augmentation code. The paper provides the most rigorous long-tail statistical analysis in the audit and cross-dataset evaluation on four external plant disease datasets. The only weakness is that the PlantDoc internal duplicate check was not performed (C4 = 0.5).
- **Zhao et al. [63] — WGA-YOLO — Score 10.5/11 (95.5%).** This is the first paper in the audit to earn a full C12 through dataset release rather than code release. It is the first paper to explicitly state that the test set contained no augmented samples and that there was no image overlap with the training or validation sets. It uses no pretrained weights, which eliminates the ImageNet/COCO overlap risk entirely. It includes the audit’s deepest data-cleaning contribution, correcting missing labels, resolving annotation inconsistencies, and removing unrealistic scenes and watermarks. It validates on three external datasets (corn, tomato, and apple) with 21,951 images and 32,436 bounding boxes. The only weakness is that the PlantDoc-PlantVillage overlap remained unaddressed (C4 = 0.5).

The three best papers share three properties:

1. Split precedes augmentation, which is stated explicitly rather than implied.
2. Evaluation data are structurally isolated from training data, either through group-aware splitting (Işık and Keskin [58]), explicit no-overlap guarantees (Zhao et al. [63]), or cross-dataset duplicate checks (Chen et al. [55]).
3. Reproducibility is genuine, and code, split files, or dataset releases are provided to allow for independent verification.

To illustrate bad examples, the following papers are considered:

- **Joshi et al. [43] — Hybrid Data Augmentation — Score 3.0/11 (27.3%).** The paper’s central contribution, hybrid data augmentation, is precisely where the leakage occurs. The augmentation pipeline is applied to the entire dataset before any splitting occurs, resulting in derivative images that can cross the training/testing boundary. The comparison table confuses epoch count (30 vs. 100). Since no split files or code were released, the reported 96.57% mAP cannot be verified. The +59% increase in mAP@50 is unbelievable. It fails both C7 and C8, the two most consequential safeguards.
- **Outaybi et al. [60] — YOLOv9 on Modified PlantDoc — Score 3.5/11 (31.8%).** The paper uses a pre-augmented Roboflow dataset, "Modified PlantDoc with Augmentations," whose augmentations were applied by a third party prior to any splitting. Therefore, the derivative containment guarantee is structurally absent. The reported mAP increases from 0.541 on the original PlantDoc to 0.877 on the Modified PlantDoc, which is a +62% relative improvement. This increase is consistent with derivative leakage rather than genuine augmentation value. There are no split proportions, counts, or validation set. The paper does not include a deduplication check. The central data augmentation contribution is precisely where the audit flags the leakage.
- **Singh et al. [33] — Real-Time Plant Disease Detection (YOLO comparison) — Score 3.5/11 (31.8%).** The paper confirms the overlap of the two datasets: it states that Plants Final combines PlantDoc with other sources. Both datasets are then used for evaluation without any overlap assessment. The pre-augmented Roboflow version ("640 and Augmentation") is used as is, with no split-before-augmentation ordering. Therefore, the comparison between the default and modified YOLO models is confounded by the potential overlap between the training and test datasets. The paper states that the training set was augmented from 2,354 to 7,062 images, but it does not state whether the split occurred before or after the augmentation. The authors did not release code or data to enable reproducing the results.

The three worst papers share three properties:

1. Pre-augmented or augmented-before-split data: the derivative containment guarantee is absent.
2. Dataset overlap is unaddressed, either confirmed (Singh’s [33] Plants Final with PlantDoc) or unassessed (Joshi et al. [43], Outaybi et al. [60]).
3. No verification artifacts are provided, such as split files, code, and dataset releases.

### 3.8 Code and data availability

Only 5 of 33 papers (15%) provide both public code and a public dataset link [42, 49, 52, 55, 58]. This is far below the norm in ML, where 30–50% of papers typically provide code. Dataset availability is more common: 22 of 33 papers (67%) provide a public dataset link, though most of these are reuse links for PlantDoc and PlantVillage rather than contributions of new datasets.

Two patterns are worth noting. First, none of the 4 Serious-rated papers provide code, and only one provides a public dataset link. Second, reproducibility practices and leakage documentation are correlated: papers that document their splits, augmentation, and stratification carefully also tend to share their code and data. This suggests that the same underlying discipline drives both practices, and that code and data availability can serve as a first-pass filter for reviewers.

## 4 Recommendations for clean modeling practice

### 4.1 Lessons from the audit

We summarise the main findings of the analysis:

1. **Augmentation scope is the field’s biggest blind spot.** Only 9 of the 33 papers (27.3%) explicitly state that augmentation is restricted to the training split. The remaining 25 papers either apply augmentation globally or leave its scope ambiguous: 20 papers (60.6%) use an ambiguous or implied training scope, and 4 papers (12.1%) apply augmentation before the split.
2. **Stratification is rare.** Only 4 of 33 papers (12.1%) document stratified splitting. On a class-imbalanced dataset such as PlantDoc, this matters for per-class reliability.
3. **Test independence is rarely operationalized.** Only 2 of 33 papers (6.1%) confirm it explicitly. A further 4 provide partial confirmation. 27 do not address it.
4. **Deduplication is absent across the board.** Only 2 of 33 papers (6.1%) perform a within-dataset near-duplicate audit; 1 of 33 also performs a cross-dataset duplicate check.
5. **Synthetic data are rare and undocumented.** Only one paper uses synthetic data. However, it does not earn full C9 credit because the paper does not document the generative model’s conditioning data, even though it states that the test set itself was not augmented.
6. **Performance and documentation are diverging.** Papers with strong documentation tend to report more conservative metrics; papers with weak documentation tend to report higher metrics. The trend is statistical, not deterministic.
7. **Co-availability of dataset and code is rare, and correlates with leakage hygiene.** Only 5 of 33 papers provide both public code and a public dataset link. None of the Serious-rated papers provides code.

## 4.2 Recommendations for authors

- **Document the split explicitly.** State the train/validation/test ratio and the per-split counts. If the split is stratified, say so and provide per-class counts across splits.
- **Keep the subsets independent.** Split first, then augment. Never apply augmentation or normalization to the whole dataset before splitting. The split-then-augment protocol is a precondition for any meaningful generalization claim.
- **State augmentation scope in one sentence.** A statement such as "Augmentation was applied only to the training set after splitting." upgrades a paper's C8 score from 0/4 to 4/4, which is the largest possible gain based on a single criterion. Transparent disclosure of the augmentation process would help ensure the methodology is clear and can spread through the community sustainably over the long term.
- **Perform a deduplication audit.** When two or more datasets are merged, or when the dataset is known to contain near-duplicates (as PlantDoc and PlantVillage do), audit for near-duplicates using perceptual hashing or embedding similarity, and report the number of duplicates found and removed.
- **Acknowledge limitations and avoid suspicious performance claims.** State which split drove checkpoint and threshold selection. Report the impact of pre-processing and augmentation separately. Provide ablation studies for the proposed algorithm. Be explicit about what the reported metrics do and do not support.
- **Release code and data.** Public code and data allow readers to verify the reported safeguards and reproduce the results. Only 5 of 33 papers in the current audit provide both. Their absence is a signal that methodological transparency is limited.

## 4.3 Recommendations for reviewers

- **Ask about augmentation scope first.** Before accepting any PlantDoc result, request an explicit statement of when augmentation was applied relative to the split. In the current audit, only 9 of 33 papers state this explicitly.
- **Ask for stratification documentation.** If the split is described only as a ratio, ask whether it was stratified by class. On a class-imbalanced benchmark such as PlantDoc, an unstratified split can distort per-class accuracy by tens of percentage points for minority classes.
- **Ask whether the test set was checked for near-duplicates.** Of the 33 audited papers, only two perform a within-dataset near-duplicate audit, and only one performs a cross-dataset check. This is not yet standard practice, but it should be.
- **Ask whether the evaluation dataset overlaps with the training dataset.** If PlantDoc, Plants Final, AI2018, or Embrapa are used as evaluation sets, request an image- or class-level overlap check against PlantVillage and the other public datasets. This is the single most common unaddressed pathway in the 33-paper audit.
- **Treat PlantDoc mAP@0.5 above 85% as requiring a deduplication audit.** In the current audit, every paper reporting above this level on PlantDoc has an unaddressed or unverified overlap risk. Without a deduplication audit, the number should not be taken at face value.

#### 4.4 Recommendations for dataset maintainers

- **Add leaf-level and plant-level identifiers to PlantDoc.** This is the structural change that would unlock Criterion C3 and make group-wise splitting possible.
- **Publish a canonical split.** A stratified train/validation/test split with a documented deduplication audit would give the field a common baseline.
- **Release a near-duplicate list.** Even a partial list would allow authors to audit their splits without needing to construct their own deduplication pipeline.

These are modest changes to the dataset infrastructure that would substantially raise the ceiling of what the field can verify.

### 5 Conclusions and outlook

This study examined data leakage in 33 peer-reviewed plant disease detection studies that use the PlantDoc dataset as a primary or secondary benchmark. We introduced a 12-criterion scoring recipe with a 0–15 total score mapped to five quality labels, and we applied it consistently to the corpus using a lenient-reading principle.

#### 5.1 Field-level conclusions

Our analysis yields four field-level conclusions:

1. **Augmentation scope is the field’s dominant leakage signal.** The single largest driver of score variation is C8 — whether augmentation is explicitly restricted to the training split. Papers with full C8 credit (4/4) average  $\approx 89\%$  of their applicable maximum; papers with partial credit (2/4) average  $\approx 57\%$ ; and papers with confirmed pre-split augmentation (0/4) average  $\approx 31\%$ . The pattern is not a statistical artifact of the scoring: pre-split augmentation creates near-duplicate images on both sides of the train/test boundary, which is the most common and most damaging leakage mechanism in the PlantDoc literature.
2. **Reported performance and leakage documentation diverge at the top of the distribution.** The three highest PlantDoc mAP@0.5 values in the corpus (97.4%, 96.57%, 84.5%) all come from papers with  $C8 \leq 2$ . Conversely, the papers that document augmentation scope most carefully ( $C8 = 4/4$ ) report PlantDoc mAP@0.5 in the 46–68% range. The pattern is consistent with the hypothesis that undocumented augmentation inflates reported metrics, although it does not prove it — some high-scoring papers may simply evaluate on easier subsets or use non-standard protocols. The papers that combine full C8 credit with strong PlantDoc results — for example, Hamad et al. (2025, 55.33% PlantDoc), Aarizou et al. (2026, 67.45% PlantDoc), and Zhao et al. (2026, 68.7% PlantDoc\_boost) — demonstrate that strong results and strong documentation are compatible.
3. **Several criteria are structurally unattainable on PlantDoc, and the recipe should treat them as baseline rather than penalty.** C3 (grouping by plant, plot, field, or session) applies only to papers with a genuine grouped structure — 4 of 33 in this audit — all four score 0.5. C9 (synthetic-data generation restricted to training) applies to only 1 paper, which does not earn full credit. C7

(normalization source) is fully satisfied by 5 of 33 papers. C2 (sample counts) and C6 (target leakage) are satisfied by the overwhelming majority of papers (29/33 and 33/33, respectively). The audit’s applicable-maximum rule already accounts for this: N/A criteria are excluded from both numerator and denominator. No additional adjustment is required.

4. **Reproducibility practices correlate with leakage hygiene.** Only 5 of 33 papers (15.2%) provide both a public code repository and a public dataset link. None of the Serious-rated papers provide code, and some of their dataset links point to secondary sources (Roboflow, PapersWithCode) rather than the primary release. Papers that document their safeguards carefully are also more likely to release their code and data. This correlation suggests that the same underlying discipline drives both practices, and that code and data availability can serve as a first-pass filter for reviewers: papers that release neither are more likely to have undocumented leakage safeguards.

## 5.2 Recommendations

We offer three concrete recommendations, ordered by expected impact:

- **For authors.** State augmentation scope explicitly. Using a split-then-augment protocol rather than a pre-split augmentation protocol upgrades a paper’s C8 score from 0/4 to 4/4, which is the largest possible gain based on a single criterion. Document the split ratio and per-split counts. State whether the split is stratified, and if possible provide a per-class count table. Perform a near-duplicate audit across the training and test sets and report the number of duplicates found and removed. Only 2 of 33 papers in the current audit do this, and only 1 performs a cross-dataset check. It is the largest remaining methodological gap in the field.
- **For reviewers.** Before accepting any (high) metric values, especially for real-field datasets such as PlantDoc, ask about the scope of augmentation. Ask for a stratification statement. Ask whether the test set was checked for near-duplicates against the training set. Treat PlantDoc mAP@0.5 values above 85% as requiring an explicit deduplication audit. In the current audit, every paper reporting above this level has an unaddressed or unverified overlap risk. Therefore, without a deduplication audit, these values should not be taken at face value. Also ask whether the evaluation dataset overlaps with the training dataset, since 25 of 33 papers do not address this.
- **For dataset consumers and benchmark curators.** Where feasible, add leaf-level and plant-level identifiers to PlantDoc-derived releases, or use group-aware splits that approximate them. Publish a canonical train/validation/test split with documented stratified sampling and a documented deduplication audit. Release a near-duplicate list for the community. These are modest changes to the dataset infrastructure that would substantially raise the ceiling of what the field can verify.

## 5.3 Closing remarks

The corpus reveals a field in which performance metrics are advancing faster than leakage safeguards. The best-documented papers in the corpus — Işık and Keskin [58] and Chen et al. [55] — show that strong PlantDoc results (85.3% mAP@0.5 in [55]) are

achievable with full documentation discipline. Zhao et al. [63] similarly reach 68.7% on PlantDoc.boost with C4/C8/C12 full credit. The Serious-rated papers show that high reported metrics can coexist with absent safeguards. The difference is not primarily the model or the dataset — it is the documentation. Papers should ideally report split dates or rules, sample counts, grouping variables, preprocessing order, tuning procedure, external data sources, and perform deduplication or a near-duplicate audit.

Adopting a single convention — an explicit augmentation-scope statement — would substantially raise the corpus mean and lift every Serious-rated paper to the Material-gaps band. The number of papers in the Serious band would drop from four to zero. If a second convention were adopted — a near-duplicate audit — the number of papers with full C4 credit would rise from two to several. Both changes would be inexpensive. Neither is currently standard practice. We hope this study will contribute to making them standard. Of course, the statement must be consistent with the methodology that has been implemented — that is, it must be true.

If the issues flagged in the papers stem from incomplete documentation of otherwise sound methodology, we hope that researchers will report their protocols more explicitly. If they stem from genuine methodological gaps, we hope that researchers will revise their pipelines accordingly. In either case, the goal is the same: reproducible, verifiable, and honestly-evaluated agricultural ML.

**Acknowledgements.** The author thanks Fabian Gerz for proofreading the paper and commenting on the discussion of the results.

**Author contributions.** The author is the sole contributor to this work.

**Funding.** No funding involved.

**Data availability.** The sources of the datasets mentioned in the paper are cited.

## Declarations

**Ethics approval and consent to participate.** Not applicable.

**Consent for publication.** Not applicable.

**Competing interests.** The author declares no conflict of interest.

## References

- [1] Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4(9), 100804 (2023) <https://doi.org/10.1016/j.patter.2023.100804>
- [2] Apicella, A., Isgrò, F., Prevete, R.: Don't push the button! exploring data leakage risks in machine learning and transfer learning. *Artificial Intelligence Review* 58(11) (2025) <https://doi.org/10.1007/s10462-025-11326-3>
- [3] Arp, D., Quiring, E., Pendlebury, F., Warnecke, A., Pierazzi, F., Wressnegger, C., Cavallaro, L., Rieck, K.: Dos and don'ts of machine learning in computer security.

**Table 6:** Overview of performance metrics documented in the selected papers and comparative studies on disease detection using ML/DL models.

| Ref. | mAP@0.5<br>(other metric)<br>on PlantDoc  | mAP@0.5<br>(other metric)<br>on other datasets                           | Comment                                                                                                                                                                                                            |
|------|-------------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [33] | 70.8%<br>(YOLOv8 XL)                      | 69.9%<br>(Plants Final)                                                  | –                                                                                                                                                                                                                  |
| [34] | N/A<br>(accuracy 77.54%)                  | N/A<br>(accuracy 99.94%<br>PlantVillage, 99.22%<br>iBean, 86.89% AI2018) | –                                                                                                                                                                                                                  |
| [35] | N/A<br>(accuracy 94.6% on<br>PlantDoc++)  | N/A<br>(accuracy 99.4%<br>PlantVillage, 98.3%<br>PDDA)                   | High classification accuracy for<br>PlantDoc++.                                                                                                                                                                    |
| [36] | N/A<br>(PlantDoc merged<br>into training) | 93.1%<br>(YOLOv5s on merged<br>dataset)                                  | High mAP credible because of<br>PlantVillage-dominated dataset, but does<br>not transfer to PlantDoc or to field<br>conditions.                                                                                    |
| [37] | 59.8%                                     | 87.9%<br>(self-built cucumber<br>dataset)                                | –                                                                                                                                                                                                                  |
| [38] | 57.68%                                    | 96.02%<br>(MSWDD2022)                                                    | –                                                                                                                                                                                                                  |
| [39] | N/A<br>(accuracy 90.12%<br>with ViT)      | N/A                                                                      | High classification accuracy for PlantDoc<br>few-shot.                                                                                                                                                             |
| [40] | 59.7%                                     | 88.6% (ALDD),<br>97.5% (ALDOD)                                           | –                                                                                                                                                                                                                  |
| [41] | N/A<br>(accuracy 98.97%)                  | N/A<br>(accuracy 97.30%<br>PlantVillage)                                 | –                                                                                                                                                                                                                  |
| [42] | N/A<br>(accuracy 55.33%)                  | N/A<br>(accuracy 96.76%<br>PlantVillage)                                 | Paper with high score, i.e. good<br>documentation of leakage safeguards, but<br>not representative because PlantDoc used<br>only for cross-validation testing.                                                     |
| [43] | 96.57%                                    | N/A                                                                      | Non-credible jump in mAP@0.5 from<br>34.2% to 96.6%.                                                                                                                                                               |
| [44] | 73.9%<br>(mAP@0.5:0.95 =<br>65.9%)        | N/A                                                                      | Critical rating driven by absence of<br>documented leakage-relevant safeguards,<br>not by mAP value — 73.9% may be<br>credible; It sits within the mid-to-high<br>range of reported PlantDoc detection<br>results. |
| [45] | N/A<br>(PlantDoc merged<br>into training) | 86.7% (self-built<br>merged dataset with 30<br>classes)                  | –                                                                                                                                                                                                                  |
| [46] | 64.9%                                     | N/A                                                                      | –                                                                                                                                                                                                                  |

Table 6 (continued)

| Ref. | mAP@0.5<br>(other metric)<br>on PlantDoc                  | mAP@0.5<br>(other metric)<br>on other datasets                                                              | Comment                                                                                                                                                                                                                      |
|------|-----------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [47] | 84.5%                                                     | N/A                                                                                                         | Suspicious jump of mAP@0.5 from 65.2% (YOLOv8) to 84.5% (YOLO-AgriNet).                                                                                                                                                      |
| [48] | 76.0%                                                     | 92.0% (Embrapa),<br>98.0% (PlantVillage),<br>97.0% (AI2018), 94.0%<br>(soybean)                             | Paper reports accuracy values on four public datasets plus a self-collected dataset, achieves strong PlantDoc accuracy (76%), but lacks significant documentation of leakage safeguards.                                     |
| [49] | N/A<br>(multiclass accuracy<br>60.7%)                     | N/A<br>(multiclass accuracy<br>99.7% PlantVillage)                                                          | –                                                                                                                                                                                                                            |
| [50] | 59.7%                                                     | 83.6% (TLDD),<br>89.0% (DDCVP)                                                                              | –                                                                                                                                                                                                                            |
| [51] | 65.4%                                                     | N/A                                                                                                         | –                                                                                                                                                                                                                            |
| [52] | N/A<br>(accuracy 80.02%)                                  | N/A<br>(accuracy 99.96%<br>PlantVillage, 92.35%<br>Plant Pathology 2021,<br>87.85% AI2018, 99.22%<br>iBean) | –                                                                                                                                                                                                                            |
| [53] | N/A<br>(accuracy 60%)                                     | N/A<br>(accuracy 99.69%<br>PlantVillage, 83%<br>FieldPlant)                                                 | Good-practice example for using PlantDoc as a clean external test set, constructing unseen-crop test sets, and documenting augmentation scope.                                                                               |
| [54] | N/A<br>(accuracy 67.45%)                                  | N/A<br>(accuracy 94.46%<br>PlantVillage, 91.89%<br>EdenLibrary)                                             | –                                                                                                                                                                                                                            |
| [55] | 85.3%<br>(mAP@0.5:0.95 =<br>71.2%)                        | N/A                                                                                                         | Strongest paper in the corpus from two independent dimensions: 1) high Score for leakage safeguard documentation: 10.5 / 11; 2) mAP@0.5 = 85.3% on PlantDoc (highest credible full-benchmark PlantDoc result in the corpus). |
| [56] | 81.6%<br>(Stable-PlantDoc)                                | 90.3%<br>(AI Challenger 2018)                                                                               | The only paper in the corpus that uses synthetic data augmentation, but does not state the source of the generative model’s conditioning data; high cross-domain mAP.                                                        |
| [57] | N/A<br>(classification<br>accuracy 89.86%<br>MobileNetV2) | N/A<br>(classification accuracy<br>99.59% Mango, 96.49%<br>Soybean)                                         | –                                                                                                                                                                                                                            |
| [58] | N/A<br>(mIoU 0.863 and<br>Dice 0.912)                     | N/A<br>(mIoU 0.847 and Dice<br>0.911 on ISIC 2018)                                                          | –                                                                                                                                                                                                                            |
| [59] | 72.4%                                                     | N/A                                                                                                         | –                                                                                                                                                                                                                            |

Table 6 (continued)

| Ref. | mAP@0.5<br>(other metric)<br>on PlantDoc | mAP@0.5<br>(other metric)<br>on other datasets | Comment                                                                                                                         |
|------|------------------------------------------|------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| [60] | 54.1%<br>(mAP@0.5:0.95 =<br>38.80%)      | N/A                                            | Serious-rating driven by absence of documented leakage-relevant safeguards, not by mAP value — 54.1% is credible/honest result. |
| [62] | 76.5%<br>(cross-domain)                  | 95.2% (PVDD)                                   | Strong and plausible cross-domain result at minor gaps in leakage safeguards.                                                   |
| [63] | 68.71%<br>(PlantDoc_boost)               | 93.98%<br>(integrated external dataset)        | —                                                                                                                               |
| [64] | 48.60%<br>(cross-domain)                 | 78.91% (M-TLD)                                 | —                                                                                                                               |
| [65] | 60.54%<br>(mAP@0.5:0.95 =<br>46.81%)     | N/A                                            | —                                                                                                                               |

In: 31st USENIX Security Symposium (USENIX Security 22), pp. 3971–3988. USENIX Association, Boston, MA (2022)

- [4] Tampu, I.E., Eklund, A., Haj-Hosseini, N.: Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Scientific data* **9**(1), 580 (2022) <https://doi.org/10.1038/s41597-022-01618-6>
- [5] Babu, M., Pandey, S., Durisic, D., Koppisetty, A., Staron, M.: D-LeDe: A data leakage detection method for automotive perception systems. In: *Proceedings of the 11th International Conference on Vehicle Technology and Intelligent Transport Systems*, vol. 1, pp. 210–221 (2025). <https://doi.org/10.5220/0013476700003941>
- [6] John, K., Saurette, D.D., Heung, B.: The problematic case of data leakage: A case for leave-profile-out cross-validation in 3-dimensional digital soil mapping. *Geoderma* **455**, 117223 (2025) <https://doi.org/10.1016/j.geoderma.2025.117223>
- [7] Hughes, D.P., Salathé, M.: An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060* (2015)
- [8] Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. *Journal of Big Data* **6**(1) (2019) <https://doi.org/10.1186/s40537-019-0197-0>
- [9] DeVries, T., Taylor, G.W.: Improved Regularization of Convolutional Neural Networks with Cutout. <https://arxiv.org/pdf/1708.04552>

- [10] Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3008–3017. IEEE, Piscataway, New Jersey (2020). <https://doi.org/10.1109/CVPRW50498.2020.00359>
- [11] Wang, J., Perez, L.: The Effectiveness of Data Augmentation in Image Classification using Deep Learning. <http://arxiv.org/pdf/1712.04621>
- [12] Varga, D.: Can Large Language Models Detect Methodological Flaws? Evidence from Gesture Recognition for UAV-Based Rescue Operation Based on Deep Learning (2026). <https://arxiv.org/abs/2604.14161>
- [13] Shafay, M., Hassan, T., Owais, M., Hussain, I., Khawaja, S.G., Seneviratne, L., Werghi, N.: Recent advances in plant disease detection: Challenges and opportunities. *Plant Methods* **21**(1), 140 (2025) <https://doi.org/10.1109/ASET66891.2025.11428061>
- [14] Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N., Nelson, A.: The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution* **3**, 430–439 (2019) <https://doi.org/10.1038/s41559-018-0793-y>
- [15] Ferentinos, K.P.: Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture* **145**, 311–318 (2018) <https://doi.org/10.1016/j.compag.2018.01.009>
- [16] Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., *et al.*: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**(8), 913–929 (2017) <https://doi.org/10.1111/ecog.02881>
- [17] Meyer, H., Pebesma, E.: Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution* **12**(9), 1620–1633 (2021) <https://doi.org/10.1111/2041-210X.13650>
- [18] Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T.: Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software* **101**, 1–9 (2018) <https://doi.org/10.1016/j.envsoft.2017.12.001>
- [19] Valavi, R., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G.: blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution* **10**(2), 225–232 (2019) <https://doi.org/10.1111/2041-210X.13107>
- [20] Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., *et al.*: Spatial validation reveals poor predictive performance of large-scale ecological

- p mapping models.
- Nature Communications*
- 11**
- (1), 4540 (2020)
- <https://doi.org/10.1038/s41467-020-18321-y>
- [21] Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., Brenning, A.: Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling* **406**, 109–120 (2019) <https://doi.org/10.1016/j.ecolmodel.2019.06.002>
  - [22] Botache, D., Bethke, F., Harder, M., Jordan, L., Sick, B.: Unraveling the complexity of splitting sequential data for machine learning. *arXiv preprint arXiv:2306.08181* (2023)
  - [23] Figueiredo, R.B.D., Mendes, H.A.: Analyzing the effect of dataset splitting in object detection tasks with video-derived images. *IEEE Access* **12** (2024) <https://doi.org/10.1109/ACCESS.2024.3383047>
  - [24] Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., Batra, N.: PlantDoc: A dataset for visual plant disease detection. In: *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. CoDS COMAD 2020*, pp. 249–253. Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3371158.3371196>
  - [25] Khan, S.U., Alsuhaibani, A., Alabduljabbar, A., Almarshad, F., Altherwy, Y.N., Akram, T.: A review on automated plant disease detection: motivation, limitations, challenges, and recent advancements for future research. *Journal of King Saud University Computer and Information Sciences* **37**(3), 34 <https://doi.org/10.1007/s44443-025-00040-3>
  - [26] Jelali, M.: Deep learning networks-based tomato disease and pest detection: a first review of research studies using real field datasets. *Frontiers in Plant Science* **Volume 15** (2024) <https://doi.org/10.3389/fpls.2024.1493322>
  - [27] Kayal, P.: PlantDoc: A Dataset for Visual Plant Disease Detection. <https://github.com/pratikkayal/PlantDoc-Object-Detection-Dataset>
  - [28] Uddin, A.H.: Plant Doc Dataset. <https://www.kaggle.com/datasets/abdulhasibuddin/plant-doc-dataset>
  - [29] McCollam, G.: PlantDoc — full variant. <https://huggingface.co/datasets/geraldmc/plantdoc-full#2>
  - [30] Yao, J., Tran, S.N.: Multi-output deep-supervised classifier chains for plant pathology. In: *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8 (2023). <https://doi.org/10.1109/IJCNN54540.2023.10191302>. IEEE
  - [31] Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., Batra, N.: PlantDoc: A Dataset for Visual Plant Disease Detection. <https://github.com/pratikkayal/>

- [32] RF-DETR-plant-doc Computer Vision Model. <https://universe.roboflow.com/horta/rf-detr-plant-doc-mnwfn/dataset/1>
- [33] Singh, Y., Shukla, S., Trivedi, G.: Real-time plant disease detection: A comparative study. In: Agriculture-Centric Computation. Springer, Switzerland (2023). [https://doi.org/10.1007/978-3-031-43605-5\\_16](https://doi.org/10.1007/978-3-031-43605-5_16)
- [34] Yu, S., Xie, L., Huang, Q.: Inception convolutional vision transformers for plant disease identification. Internet of Things **21**, 100650 (2023) <https://doi.org/10.1016/j.iot.2022.100650>
- [35] Zhang, R., Wang, Y., Jiang, P., Peng, J., Chen, H.: IBSA\_Net: A network for tomato leaf disease identification based on transfer learning with small samples. Applied Sciences **13**(7), 4348 (2023) <https://doi.org/10.3390/app13074348>
- [36] Bellout, A., Zarboubi, M., Dliou, A., Latif, R., Saddik, A.: Advanced YOLO models for real-time detection of tomato leaf diseases. Mathematical Modeling and Computing **11**(4), 1198–1210 (2024) <https://doi.org/10.23939/mmc2024.04.1198>
- [37] Ding, J.-Y., Jeon, W.-S., Rhee, S.-Y., Zou, C.-M.: An improved YOLO detection approach for pinpointing cucumber diseases and pests. Computers, Materials & Continua **81**(3), 3989–4014 (2024) <https://doi.org/10.32604/cmc.2024.057473>
- [38] Mao, R., Zhang, Y., Wang, Z., Hao, X., Zhu, T., Gao, S., Hu, X.: DAE-Mask: a novel deep-learning-based automatic detection model for in-field wheat diseases. Precision Agriculture **25**(2), 785–810 (2024) <https://doi.org/10.1007/s11119-023-10093-x>
- [39] Rezaei, M., Diepeveen, D., Laga, H., Jones, M.G.K., Sohel, F.: Plant disease recognition in a low data scenario using few-shot learning. Computers and Electronics in Agriculture **219**, 108812 (2024) <https://doi.org/10.1016/j.compag.2024.108812>
- [40] Yan, C., Yang, K.: FSM-YOLO: Apple leaf disease detection network based on adaptive feature capture and spatial context awareness. Digital Signal Processing **155**, 104770 (2024) <https://doi.org/10.1016/j.dsp.2024.104770>
- [41] Chahbouni, A., El Manaa, K., El Alami, R.: Dynamic attention-based transfer learning strategies for efficient and scalable visual recognition in resource-constrained environments. Applied Soft Computing (2025) <https://doi.org/10.1016/j.asoc.2025.114014>
- [42] Hamad, L., Ananzeh, Q., Salameh, M., Salem, A., Alshboul, M., Al-Ayyoub, M.: Hierarchical multi-stage framework for robust and explainable tomato leaf disease

- p>identification.
- IEEE Access*
- 13**
- , 217913–217928 (2025)
- <https://doi.org/10.1109/ACCESS.2025.3648004>
- [43] Joshi, K., Hooda, S., Sharma, A., Sonah, H., Deshmukh, R., Tuteja, N., Gill, S.S., Gill, R.: Precision diagnosis of tomato diseases for sustainable agriculture through deep learning approach with hybrid data augmentation. *Current Plant Biology* **41**, 100437 (2025) <https://doi.org/10.1016/j.cpb.2025.100437>
  - [44] Kumari, L., Majumder, A.: Leaf disease detection using YOLOv8. In: *Recent Advances in Artificial Intelligence and Smart Applications. Lecture Notes in Networks and Systems*, pp. 71–83. Springer, Singapore (2025). [https://doi.org/10.1007/978-981-96-5822-0\\_7](https://doi.org/10.1007/978-981-96-5822-0_7)
  - [45] Liu, Z., Guo, X., Zhao, T., Liang, S.: YOLO-BSMamba: A YOLOv8s-based model for tomato leaf disease detection in complex backgrounds. *Agronomy* **15**(4), 870 (2025) <https://doi.org/10.3390/agronomy15040870>
  - [46] Miao, Y., Meng, W., Zhou, X.: SerpensGate-YOLOv8: an enhanced YOLOv8 model for accurate plant disease detection. *Frontiers in Plant Science* **15**, 1514832 (2025) <https://doi.org/10.3389/fpls.2024.1514832>
  - [47] Nkonjoh, A.N., Kaze, J.R.D., Nwokam, R.V., Njotsa, B.E., Kuate, A.F., Tchouta, A.S.M., Babagniack, S.B.B.: YOLO-AgriNet: A deep learning-based model for real-time plant disease detection in precision agriculture. *Journal of Computer and Communications* **13**(8), 181–204 (2025) <https://doi.org/10.4236/jcc.2025.138009>
  - [48] Sharma, V., Tripathi, A.K., Mittal, H., Nkenyereye, L.: SoyaTrans: A novel transformer model for fine-grained visual classification of soybean leaf disease diagnosis. *Expert Systems with Applications* **260**, 125385 (2025) <https://doi.org/10.1016/j.eswa.2024.125385>
  - [49] Shiyani, A.S., Kozlov, I.D., Baimuratov, I.R., Zhukova, N.A.: Recognizing plants and their diseases: Benchmarks for multiclass and multilabel classification. *Pattern Recognition and Image Analysis* **35**(2), 159–168 (2025) <https://doi.org/10.1134/S1054661825700087>
  - [50] Yan, C., Li, H.: CAPNet: tomato leaf disease detection network based on adaptive feature fusion and convolutional enhancement. *Multimedia Systems* **31**(2), 1–22 (2025) <https://doi.org/10.1007/s00530-025-01756-y>
  - [51] Yang, X., Wang, H., Zhou, Q., Lu, L., Zhang, L., Sun, C., Wu, G.: A lightweight and efficient plant disease detection method integrating knowledge distillation and dual-scale weighted convolutions. *Algorithms* **18**(7), 433 (2025) <https://doi.org/10.3390/a18070433>
  - [52] Yu, S., Xie, L., Dai, L.: ST-CFI: Swin transformer with convolutional feature

- p>interactions for identifying plant diseases.
- Scientific Reports*
- 15**
- (1), 25000 (2025)
- <https://doi.org/10.1038/s41598-025-08673-0>
- [53] Zubair, F., Saleh, M., Akbari, Y., Al Maadeed, S.: A robust ensemble model for plant disease detection using deep learning architectures. *AgriEngineering* **7**(5), 159 (2025) <https://doi.org/10.3390/agriengineering7050159>
  - [54] Aarizou, A., Merah, M., Mahiout, D.: Enhancing cross-crop plant disease detection on complex field images through dataset merging, transfer learning and ensemble averaging of convolutional neural networks. *Journal of Phytopathology* **174**(2) (2026) <https://doi.org/10.1111/jph.70264>
  - [55] Chen, L., Wu, Y., Meng, Y., Li, S., Tang, Q.: A lightweight plant disease detection model for long-tailed agricultural scenarios. *Plants* **15**(8), 1206 (2026) <https://doi.org/10.3390/plants15081206>
  - [56] Chen, Y., Gong, T., Wan, X., Li, Y.: MFVNet: a novel lightweight deep neural network for sustainable agriculture plant disease detection. *Cluster Computing* **29**(4), 1–18 (2026) <https://doi.org/10.1007/s10586-026-05968-3>
  - [57] Duhan, S., Gulia, P., Gill, N.S., Oliveira, T.A., Singh, P.K.: Evaluation of lightweight and efficient deep learning models for plant disease classification. *Discover Internet of Things* **6**(1), 1–21 (2026) <https://doi.org/10.1007/s43926-026-00310-0>
  - [58] Işık, G., Keskin, F.: CA-SAUNet: Coordinate-attention enhanced shape attentive UNet for medical and agricultural image segmentation. *Pattern Analysis and Applications* **29**(1), 6–1617 (2026) <https://doi.org/10.1007/s10044-025-01590-y>
  - [59] Nyawose, T., Maswanganyi, R.C., Khumalo, Z.P.: Multi-plant real-time disease detection and classification using YOLOv11. *IEEE Access* **14**, 95193–95208 (2026) <https://doi.org/10.1109/ACCESS.2026.3702738>
  - [60] Outaybi, S., Sabbar, H., Abbad, K.: Automated early leaf disease detection using YOLOv9. In: *Artificial Intelligence and Green Computing. Lecture Notes in Networks and Systems*. Springer, Switzerland (2026). [https://doi.org/10.1007/978-3-032-02312-4\\_9](https://doi.org/10.1007/978-3-032-02312-4_9)
  - [61] Wang, B., Zhou, H., Wang, Z., Chen, R.: AD-DETR: A real-time transformer with multi-scale alignment and spatial-spectral fusion for crop disease detection. *Sensors* **26**(10), 3206 (2026) <https://doi.org/10.3390/s26103206>
  - [62] Wang, X., Liu, J.: A feature-enhanced network for vegetable disease detection in complex environments. *Plants* **15**(8) (2026) <https://doi.org/10.3390/plants15081206>

- [63] Zhao, J., Liang, Y., Wei, G., Han, B., Li, J., Yu, C., Zeng, J., He, S.: A wavelet-based frequency-domain approach for accurate multi-crop disease detection. *Scientific Reports* **16**, 7099 (2026) <https://doi.org/10.1038/s41598-026-38476-w>
- [64] Zhuo, J., Dong, G., Huang, Q., Zhou, L., Zhao, F., Yuan, P., Yang, X.: Gbr-det: A real-time tomato leaf disease detection model for edge device deployment. *Sensors* **26**(10), 2950 (2026) <https://doi.org/10.3390/s26102950>
- [65] Yan, C.: Multi-granularity feature aggregation with asymmetric convolution for plant disease detection. In: *Communications in Computer and Information Science*, vol. 3039, pp. 284–295. Springer, Singapore (2026). [https://doi.org/10.1007/978-981-92-3554-4\\_24](https://doi.org/10.1007/978-981-92-3554-4_24)