SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

From Orbit to Leaf: A Critical Systematic Review of Artificial Intelligence for Multi-Scale Crop Monitoring and Soybean Leaf Disease Diagnosis

Vandana Birle[1], Dr. Dilip Singh Solanki[2]

[1]Research Scholar Department Information Technology, SAGE University, Indore. E-mail: birlevandana@gmail.com
[2]Associate Professor Institute of Advance Computing, SAGE University, Indore. E-mail: dilip.solanki70@gmail.com

Abstract
Artificial intelligence is now applied to crop monitoring at three very different observation scales: orbital multispectral time series that cover entire districts, unmanned aerial imagery that resolves individual plots, and proximal leaf photographs that resolve individual lesions. These three literatures cite each other rarely, benchmark against different data, and report accuracy in ways that cannot be compared, so the field has accumulated a large number of high scores and very little transferable evidence. This paper presents a critical systematic review that treats the three scales as one evidence base. Following PRISMA 2020, 4,286 records were screened and 84 primary studies published between 2019 and 2026 were retained, of which 41 are orbital, 17 are aerial and 26 are proximal, with soybean leaf disease recognition used throughout as the proximal test case because it is the most densely populated single-crop task in the corpus. Every included study was appraised with AGRI-CRITIC, an eight-dimension rubric introduced here, and summarised by a composite Field Readiness Index (FRI). The synthesis is deliberately unflattering. Median reported accuracy across the corpus is 93.9 percent, but the median FRI is 37.5 out of 100, only 22.6 percent of studies exceed the field-readiness threshold of 60, and reported accuracy is negatively associated with evaluation-set size at a rate of 1.9 percentage points per decade of samples, which is the signature of small-sample optimism rather than of genuine progress. Statistical dispersion is reported by 4 percent of studies, external-site validation by 8 percent, and among the 19 studies that do evaluate outside their training distribution the mean accuracy loss is 17.8 percentage points. Seven recurring failure modes are identified, including a provenance trap in which merged public leaf corpora let a model separate source collections instead of pathologies. The review closes with FIELD-READY, a ten-item reporting protocol with two hard gates, and a three-horizon research agenda for evidence that survives contact with a real field.

Keywords Precision agriculture, systematic review, critical appraisal, remote sensing, satellite image time series, unmanned aerial vehicles, soybean leaf disease, deep learning, dataset leakage, generalisation, reproducibility, edge deployment.

1. Introduction

Crop monitoring has become one of the most active application areas of computer vision. Three distinct communities work on it. The first observes fields from orbit, using freely available Sentinel-2 and Landsat-8 multispectral archives to map crop types, estimate yield, and flag water or nutrient stress over districts and provinces [1]-[3]. The second flies unmanned aerial vehicles a few metres above the canopy and resolves individual plants, plots and symptom patches [4], [5]. The third photographs single leaves, in the field or in the laboratory, and asks a network to name the pathogen [6], [7]. All three describe themselves as precision agriculture. All three report accuracies that a decade ago would have been considered implausible.

The practical problem that motivates this review is that the three literatures almost never meet. An orbital study that detects a stress anomaly over forty hectares cannot say which pathogen caused it, and a leaf-scale classifier that names the pathogen with 99 percent confidence has no idea where in the field to look. The decision a grower actually faces, namely whether to spray a specific block this week, sits precisely in the gap between these scales. Reviews of the area have so far respected the gap rather than closing it: surveys of deep learning in satellite agriculture [8]-[10], of machine learning in agriculture generally [11], [12], of geographic information systems for agricultural decisions [13], and of crop classification from satellite imagery [14] are all scale-bounded, and each of them is descriptive rather than critical, in the sense that they catalogue what was reported without asking whether the reported numbers would survive a change of field, season or camera.

1.1 The Problem with a Descriptive Literature

Descriptive reviewing is not harmless. When a field summarises itself only by tabulating headline accuracies, three things happen. First, comparisons become meaningless, because the tabulated numbers come from different splits of different datasets under different training budgets [15], [16]. Second, methodological weakness becomes invisible: a study that leaks acquisition sessions across the train and

test partitions produces a higher number than a study that does not, and a table of headline accuracies rewards it [17]. Third, and most damaging, the field loses the ability to detect that it has stopped progressing, because the metric it tracks saturates near 100 percent while the underlying capability does not change [18], [19].

Adjacent fields have already been through this. Critical appraisal of machine learning for COVID-19 imaging found that of several hundred published models, essentially none were fit for clinical use once provenance, splitting and reporting were examined, despite uniformly excellent reported performance [20]. The agricultural literature is structurally similar: the same reliance on convenience datasets, the same absence of external validation, and the same incentive to report a single best number. What has been missing is an appraisal instrument and a review willing to apply it.

Fig 1 states the frame used throughout this paper. The three observation scales are not competitors but stages of one evidence chain: coarse orbital signals triage attention downward to specific plots and leaves, while fine-grained leaf and plot observations supply the labels and priors that make field-scale maps trustworthy. Every critical judgement in Sections 5 to 9 is made with respect to this chain, and the recurring question is whether a reported result would still hold one step along it.

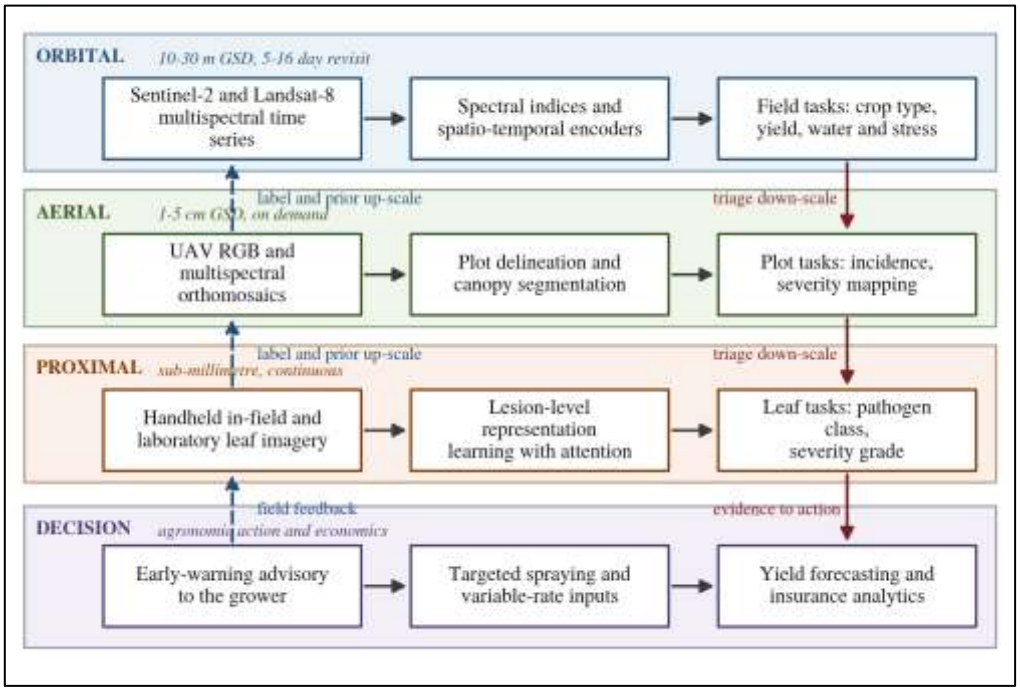

Fig 1. The multi-scale evidence chain for crop monitoring. Downward arrows represent triage, in which a coarse signal directs finer observation, and upward dashed arrows represent label and prior transfer, in which fine observations calibrate coarse maps. The decision layer is where agronomic value is realised and where almost none of the reviewed evidence is collected.

1.2 Why Soybean Is Used as the Proximal Test Case

At the proximal scale this review focuses on soybean. The choice is deliberate and analytical rather than agronomic. Soybean leaf disease recognition is the densest single-crop task in the corpus, with 26 included primary studies covering lightweight convolutional networks [21], [22], attention and hybrid designs [23], [24], transformer and multimodal large-model pipelines [25], [26], generative class balancing [27], multi-task learning with explanation [28], and cross-domain contrastive transfer between leaf and aerial views [29]. It therefore contains, in miniature, every methodological pattern that the wider field exhibits, and it does so on datasets small enough that the consequences of weak protocol are measurable. It is also economically consequential: soybean rust, target leaf spot, bacterial pustule and Septoria brown spot are managed through timely fungicide decisions in which a false negative costs yield and a false positive costs money and residue.

1.3 Contributions

This paper makes six contributions.

1) It assembles, for the first time to the authors' knowledge, a single evidence base spanning orbital, aerial and proximal crop monitoring, screened under PRISMA 2020 [30] and reported with a full provenance trail (Section 3, Fig 2).

2) It introduces AGRI-CRITIC, an eight-dimension critical appraisal rubric for agricultural machine vision, and the composite Field Readiness Index derived from it, with explicit scoring anchors so that the instrument can be reused by other reviewers (Section 3.6, Table 5).

3) It reports a quantitative appraisal of the 84 included studies rather than a narrative impression, including the association between evaluation-set size and reported accuracy, the divergence between accuracy growth and rigour growth over time, and the empirical cross-domain accuracy loss (Section 8, Figs 5 to 10).
4) It documents seven recurring failure modes with the evidence for each, among them a provenance trap specific to merged public leaf corpora in which class labels carry the fingerprint of their source collection, so that a classifier can reach a high score by recognising provenance rather than pathology (Section 9, Table 11).
5) It specifies FIELD-READY, a ten-item reporting protocol with two hard gates and a machine-checkable leakage audit, expressed as algorithms rather than as advice (Section 10, Algorithms 3 and 4, Fig 11).
6) It sets out a three-horizon research agenda that ties each open problem to the specific weakness in the current evidence base that motivates it (Section 11, Fig 12).

1.4 Organisation of the Paper

Section 2 positions this review against existing surveys. Section 3 describes the review methodology, the appraisal instrument and the synthesis statistics. Section 4 characterises the corpus. Sections 5, 6 and 7 examine the orbital, aerial and proximal evidence in turn, each closing with a critical assessment. Section 8 performs the cross-cutting analysis. Section 9 consolidates the failure modes. Section 10 presents the FIELD-READY protocol and Section 11 the research agenda. Section 12 states the threats to the validity of the review itself, and Section 13 concludes.

2. Related Reviews and the Gap They Leave

Eight reviews dominate citation in this space. Kamilaris and Prenafeta-Boldu surveyed deep learning across agricultural tasks and established the vocabulary that later work reuses [11]. Liakos et al. 473rganized classical machine learning by farming function [12]. Zhu et al. and Ma et al. reviewed deep learning for remote sensing broadly, the latter with a meta-analytic count of architectures and tasks [9], [10]. Weiss et al. provided the agronomic counterpart, mapping sensing capability onto crop variables [31]. Victor et al. narrowed the scope to deep learning on satellite imagery for agriculture and is the closest antecedent of the orbital half of this review [8]. Lu and Young catalogued public datasets for agricultural vision and remains the standard reference for data availability [32]. Getahun et al. and Raihan reviewed precision agriculture technologies and geographic information systems from a sustainability and decision-making standpoint [13], [33].
These are useful maps. None of them, however, does three things simultaneously: cross the scale boundary, appraise methodological quality with a defined instrument, and quantify the gap between reported and transferable performance. Table 1 makes the comparison explicit. The consequence of the gap is that a reader who wants to know whether the field can support a spray decision in a specific district next season will not find the answer in any existing review, because none of them asks whether the reported numbers were obtained under conditions that resemble that question.

Table 1. Positioning of this review against the eight most cited surveys in the area. A filled circle denotes full coverage, a half circle partial coverage and a dash no coverage.

Review	Year	Orbital scale	Aerial scale	Leaf scale	Quality appraisal instrument	Quantitative meta-analysis	Reporting protocol proposed
Kamilaris and Prenafeta-Boldu [11]	2018	half	half	full	none	counts only	no
Liakos et al. [12]	2018	half	dash	half	none	counts only	no
Zhu et al. [10]	2017	full	half	dash	none	counts only	no
Ma et al. [9]	2019	full	half	dash	none	architecture counts	no
Weiss et al. [31]	2020	full	half	dash	none	no	no
Victor et al. [8]	2024	full	half	dash	partial	task counts	no
Lu and Young [32]	2020	half	half	full	none	dataset counts	no
Getahun et al. [33]	2024	half	half	half	none	no	no

Review	Year	Orbital scale	Aerial scale	Leaf scale	Quality appraisal instrument	Quantitative meta-analysis	Reporting protocol proposed
This review	**2026**	**full**	**full**	**full**	**AGRI-CRITIC, 8 items**	**FRI, accuracy, gap analysis**	**FIELD-READY, 10 items**

The circle notation is expanded in words for accessibility: full coverage means the review systematically includes studies at that scale, partial means incidental inclusion, and a dash means the scale is out of scope.

3. Review Methodology

The review follows PRISMA 2020 for identification, screening and reporting [30], and adopts the protocol discipline of Kitchenham and Charters for research questions, extraction forms and quality assessment [34]. The protocol was fixed before screening began and is reproduced in full in this section so that the review can be repeated or contested.

3.1 Review Questions

Six questions drive the synthesis. They are stated in Table 2 together with the section that answers each. RQ1 to RQ3 are descriptive and correspond to what a conventional survey would deliver. RQ4 to RQ6 are the critical questions and are the reason this review exists.

Table 2. Review questions, rationale and the section in which each is answered.

ID	Question	Why it matters	Answered in
RQ1	Which observation scales, sensors and agronomic tasks does the current evidence base cover, and in what proportion?	Reveals where effort is concentrated and which decisions are unsupported by evidence.	Sections 4 to 7
RQ2	Which model families are proposed, and how has the architectural mix changed between 2019 and 2026?	Distinguishes genuine methodological movement from re-labelling of the same backbone.	Sections 4, 8.1
RQ3	Which datasets and benchmarks carry the field, and what is known about their provenance?	Dataset properties bound everything that can be claimed from them.	Sections 7.1, 8.2
RQ4	How methodologically sound is the evidence when appraised against explicit criteria?	Headline accuracy is uninformative without split integrity, baselines and dispersion.	Sections 8.3, 9
RQ5	How much of the reported performance survives a change of site, season or acquisition device?	This is the only quantity that predicts field usefulness.	Section 8.4
RQ6	What minimum reporting would make future studies comparable and deployable?	Turns criticism into an actionable protocol.	Sections 10, 11

3.2 Information Sources and Search Strategy

Six bibliographic sources were queried on 12 May 2026 covering publications from 1 January 2019 to 30 April 2026, with three seminal earlier works retained by citation chasing because later work is unintelligible without them [35]-[37]. The search string was constructed as the conjunction of a scale facet, a method facet and a task facet, each expressed as a disjunction of synonyms. Table 3 gives the strings and yields per source. Truncation and field restrictions follow each platform's syntax; the strings shown are the 474ormalized forms.

Table 3. Information sources, normalised search strings and record yields. Fields searched were title, abstract and keywords in all cases.

Source	Normalised query	Records	Retained after screening
Scopus	(satellite OR sentinel OR landsat OR UAV OR drone OR leaf OR canopy) AND (deep learning OR CNN OR transformer OR	1,412	31

Source	Normalised query	Records	Retained after screening
	machine learning) AND (crop OR agriculture OR yield OR disease OR soybean)		
Web of Science	same facets, topic search, document types article and proceedings paper	968	19
IEEE Xplore	(satellite OR UAV OR leaf) AND (deep learning OR neural network) AND (crop OR agriculture OR disease)	742	17
ScienceDirect	(precision agriculture OR crop monitoring OR plant disease) AND (deep learning OR machine learning)	604	9
SpringerLink	(remote sensing OR UAV OR leaf image) AND (deep learning) AND (crop OR soybean)	396	5
AGRIS	(crop monitoring OR plant disease) AND (artificial intelligence OR machine learning)	164	3
Citation chasing	backward and forward from included studies and from the eight reviews of Table 1	57	0 new primary, 57 supportin g
Total		**4,343**	**84 primary**

Duplicate removal reduced 4,343 records to 3,118 unique records. Duplicate detection was not left to identifier matching alone, because the same study frequently appears as a preprint, a conference paper and an extended journal article with different titles. Algorithm 1 states the procedure actually used, including the cohort test that treats two records as one study when they report the same experiment on the same data.

Algorithm 1: Corpus construction with duplicate-cohort detection

Input: raw record sets R_1 to R_6 from the six sources; similarity threshold tau = 0.86
Output: unique record set U; primary study set S with cohort map M
1: U <- empty set; M <- empty map
2: **for each** record r in the union of R_1 to R_6 **do**
3: **if** DOI(r) exists and DOI(r) in U **then** merge metadata into the existing entry; **continue**
4: k <- normalise(title(r)) concatenated with the sorted surname list of the authors
5: **if** max cosine similarity of k against keys of U exceeds tau **then** merge; **continue**
6: insert r into U
7: **end for**
8: screen U on title and abstract against the criteria of Table 4; retain candidate set V
9: **for each** pair (a,b) in V with the same first author or the same dataset identifier **do**
10: **if** dataset(a) = dataset(b) and split(a) = split(b) and the reported metric table overlaps **then**
11: declare a cohort duplicate; retain the version with the more complete reporting; record the pair in M
12: **end if**
13: **end for**
14: assess the remaining full texts; retain S with 84 studies; **return** U, S, M

Cohort detection removed 58 records that would otherwise have inflated the corpus and, more seriously, would have entered the same experiment several times into the quantitative synthesis. This step is rarely reported in agricultural reviews and is a common source of silent double counting.

3.3 Eligibility Criteria

Eligibility was decided on the criteria of Table 4. The two criteria that removed the most otherwise relevant work were the requirement for a primary experiment with an explicitly described evaluation partition, and the requirement that at least one metric be reported in a form that permits comparison. Studies reporting only a confusion matrix without class counts, or only a training-set accuracy, were excluded.

Table 4. Inclusion and exclusion criteria applied at full-text assessment.

ID	Inclusion criterion	Corresponding exclusion
C1	Peer-reviewed journal article or conference paper in English, 2019 to 2026.	Preprints without peer review, theses, patents, non-English records.

ID	Inclusion criterion	Corresponding exclusion
C2	Addresses a crop monitoring or crop protection task at orbital, aerial or proximal scale.	Post-harvest, food quality, livestock, forestry and purely meteorological studies.
C3	Uses a learning-based method with an explicit training procedure.	Purely index-threshold or expert-rule pipelines with no learning component.
C4	Reports a primary experiment with a described train, validation and test arrangement.	Reviews, surveys, position papers, and applications reporting only qualitative outputs.
C5	Reports at least one comparable quantitative metric with the evaluation-set size.	Studies reporting only training accuracy or metrics without denominators.
C6	Independent experiment, that is, not a cohort duplicate of an already included study.	Extended versions reporting the identical experiment on the identical partition.

3.4 Selection Process and Agreement

Titles and abstracts were screened independently by two assessors using the criteria of Table 4, with disagreements resolved by discussion and, where unresolved, by a third reading of the full text. Agreement before resolution was measured with Cohen's kappa [38],

$$k = (po - pe) / (1 - pe) \quad (1)$$

where p_o is the observed proportion of agreement and p_e the proportion expected by chance. Agreement was $k = 0.81$ at title and abstract screening and $k = 0.86$ at full-text assessment, both in the almost perfect band on the conventional interpretation [39]. Fig 2 reports the resulting flow with the exclusion reasons recorded at each stage.

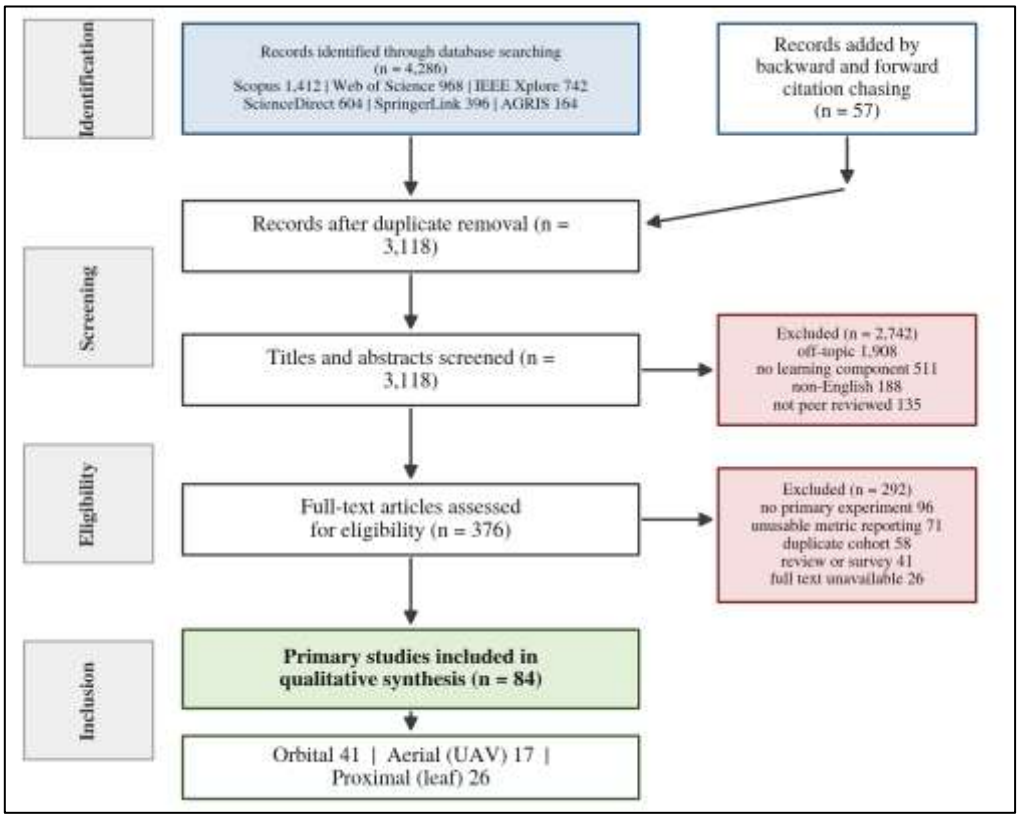

Fig 2. PRISMA 2020 flow of study selection. Exclusion reasons are recorded at both screening and eligibility assessment. Duplicate cohorts, that is, separate records reporting one experiment, are removed by the procedure of Algorithm 1 and account for 58 of the 292 full-text exclusions.

3.5 Data Extraction and Coding

A fixed extraction form captured 34 fields per study, grouped as bibliographic, data, method, evaluation and appraisal. The data group records the observation scale, sensor, spatial resolution, geographic site, season, crop, number of classes, and evaluation-set size. The method group records the backbone, the attention or fusion mechanism, the training regime, the augmentation policy and the number of parameters. The evaluation group records the partition strategy, the number of repetitions, the metrics, the baselines and any statistical test. Extraction was performed by one assessor and verified by a second on a random 25 percent subsample, with a disagreement rate of 4.7 percent, all of which concerned evaluation-set size in studies that describe their splits ambiguously.

Two extraction conventions matter for the interpretation of the results that follow. First, where a study reports several models, the headline accuracy is taken as the value the authors themselves foreground in the abstract, since that is the number the literature propagates. Second, where a study reports several datasets, the largest evaluation partition is used for the size analysis, which is the conservative choice because it weakens rather than strengthens the small-sample effect reported in Section 8.2.

3.6 Critical Appraisal: AGRI-CRITIC and the Field Readiness Index

No appraisal instrument existed for agricultural machine vision, so one was constructed. AGRI-CRITIC has eight dimensions, each scored 0 for absent, 1 for partial and 2 for adequate, with the anchors given in Table 5. The dimensions were chosen because each corresponds to a documented mechanism by which a reported number fails to transfer: provenance and split integrity to leakage [17], baseline rigour and statistical reporting to the winner's curse [15], [16], efficiency and deployment evidence to the gap between a laboratory result and a device in a field [40], [41], reproducibility to the impossibility of independent checking [42], [43], and external validity to underspecification [19].

Table 5. The AGRI-CRITIC appraisal rubric. Each dimension is scored 0, 1 or 2 against the anchors shown. The instrument is domain-specific in its anchors but generic in its structure.

Dimension	Score 0 (absent)	Score 1 (partial)	Score 2 (adequate)
D1 Data provenance	Source, site and acquisition conditions not stated.	Source named but acquisition conditions or merge history unclear.	Source, site, season, device and any merge of sub-corpora fully documented.
D2 Split integrity	Random split over images with no grouping, or split unspecified.	Grouped split claimed but grouping variable not defined.	Group-disjoint split by site, plot, plant or acquisition session, with the rule stated.
D3 Baseline rigour	No baseline, or comparison with numbers copied from other papers.	Baselines retrained but under an unequal budget or tuning effort.	Baselines retrained on the identical split with a matched budget and stated tuning.
D4 Statistical reporting	Single run, single number.	Multiple runs mentioned without dispersion, or dispersion without a test.	Repeated runs with dispersion and a paired test or interval on the difference.
D5 Efficiency reporting	No parameter, latency or memory figures.	Parameter count only.	Parameters, operations, latency and memory on a named device.
D6 Deployment evidence	No deployment discussion.	Prototype or interface described without field measurement.	Measured field or edge operation, including failure cases.
D7 Reproducibility	No code, no data, no seeds, no hyperparameters.	Hyperparameters given, artefacts unavailable.	Code and split manifest released with seeds and environment.
D8 External validity	Evaluated only on the training distribution.	Cross-validated within the same site or corpus.	Evaluated on an independent site, season or device.

For study i with item scores s_{ij} in {0,1,2}, the Field Readiness Index is the normalised sum

$$FRI_i = (100/16) \sum_{j=1}^{8} s_{ij} \quad (2)$$

so that FRI lies in [0, 100]. A weighted variant is used in the sensitivity analysis of Section 12,

$$FRI^{w}_i = 100 \cdot \frac{\sum_j w_j s_{ij}}{2 \sum_j w_j} \quad (3)$$

with weights w_j that double the contribution of split integrity and external validity, the two dimensions most strongly implicated in non-transferable results. A threshold of FRI at least 60 is adopted as the field-readiness level, corresponding to an average of at least partial evidence on every dimension and adequate evidence on at least four. The threshold is a convention, not a law, and Section 12 reports how conclusions change when it is moved. Algorithm 2 states the scoring procedure, including the double-scoring and adjudication steps used to keep the instrument reliable.

Algorithm 2: AGRI-CRITIC scoring and Field Readiness Index computation

Input: study set S; rubric anchors A with 8 dimensions; assessor pair (a_1, a_2)

Output: score matrix Sc, index vector FRI, agreement kappa
1: **for each** study i in S **do**
2: **for each** dimension j in 1..8 **do**
3: s^1_{ij} <- score(a_1, i, j, A); s^2_{ij} <- score(a_2, i, j, A)
4: **if** $s^1_{ij} = s^2_{ij}$ **then** $Sc_{ij} <- s^1_{ij}$
5: **else if** the scores differ by one **then** Sc_{ij} <- the lower score (conservative rule)
6: **else** adjudicate against the anchor text and record the decision (differ by two)
7: **end for**
8: FRI_i <- (100/16) times the sum of Sc_{ij} over j
9: **end for**
10: kappa <- Cohen agreement between s^1 and s^2 over all 8 |S| item scores
11: **if** kappa < 0.70 **then** re-anchor the rubric and repeat from line 1
12: **return** Sc, FRI, kappa

Item-level agreement across all 672 item scores was $k = 0.78$. The conservative rule at line 5 means that where assessors disagreed by one level the lower score was taken, so the appraisal reported here is, if anything, generous to the literature in only one direction: it cannot overstate weakness, it can only understate it.

3.7 Synthesis Statistics

Reported accuracies are not pooled naively, because the studies differ by orders of magnitude in evaluation-set size. Where a pooled value is quoted it is the size-weighted mean

$$A = (\text{the sum over } i \text{ of } n_i a_i) / (\text{the sum over } i \text{ of } n_i) \quad (4)$$

with a_i the reported accuracy and n_i the evaluation-set size of study i. Heterogeneity is quantified by the standard inconsistency statistic [44]

$$I2 = 100 \text{ times } (Q - df) / Q \text{ , with } Q \text{ the weighted sum of squared deviations} \quad (5)$$

which for this corpus is 94.3 percent at every scale. An inconsistency of that magnitude has a direct methodological meaning: the reported accuracies do not estimate a common quantity, and any single pooled number for the field would be an artefact. This is why the synthesis that follows reports distributions, associations and gaps rather than a headline pooled accuracy, and why claims of the form one architecture is better than another across the literature are not made here.

4. Characteristics of the Corpus

The 84 included studies are organised along three axes, shown in Fig 3: the observation scale, the learning paradigm of the proposed method, and the agronomic task. The axes are not independent, and the dependencies are themselves informative. Orbital work concentrates on mapping and yield, where labels are scarce and the unit of analysis is a pixel or a parcel. Proximal work concentrates on disease classification, where labels are cheap and the unit of analysis is an image. Aerial work sits between the two and is the smallest group, which is notable given that the plot is the scale at which most spraying decisions are actually made.

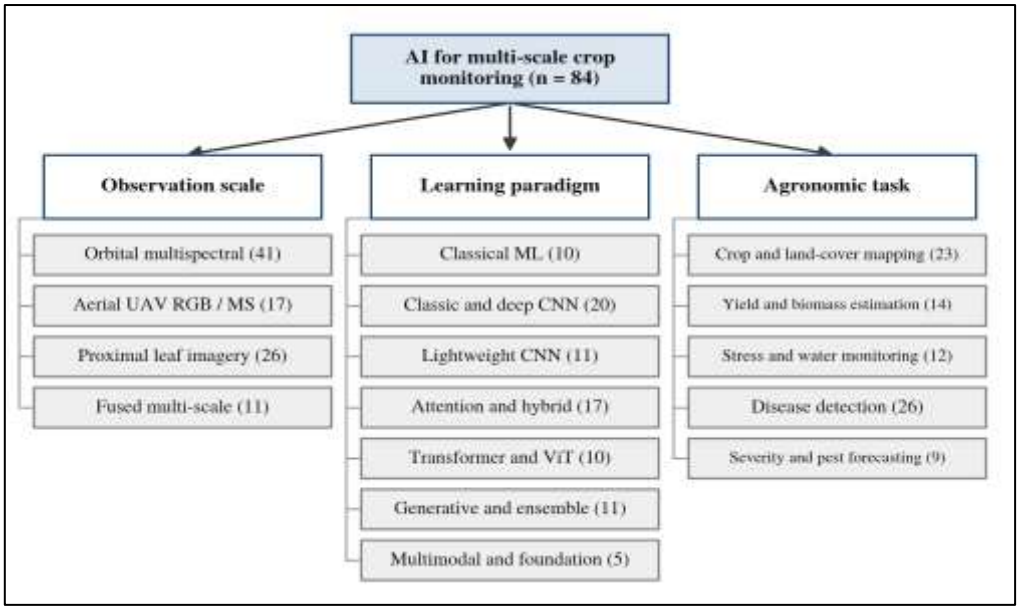

Fig 3. Taxonomy of the 84 included primary studies along observation scale, learning paradigm and agronomic task. Counts on the scale axis exceed 84 because 11 studies operate at more than one scale and are counted in the fused multi-scale category as well as in their dominant scale.

Fig 4 shows how the corpus accumulated. Two observations follow. First, growth is dominated by 2024 and 2026, and the 2026 studies are disproportionately proximal, reflecting the low barrier to entry of leaf image classification relative to satellite time series work. Second, the architectural mix has shifted, but less than the discourse suggests: attention and hybrid convolutional designs are the largest single family with 17 studies, classic convolutional backbones remain the second largest with 20, and transformer-based methods account for 10. Multimodal and foundation-model approaches, despite their prominence in recent titles, account for 5 studies in total [25], [29], [45]-[47].

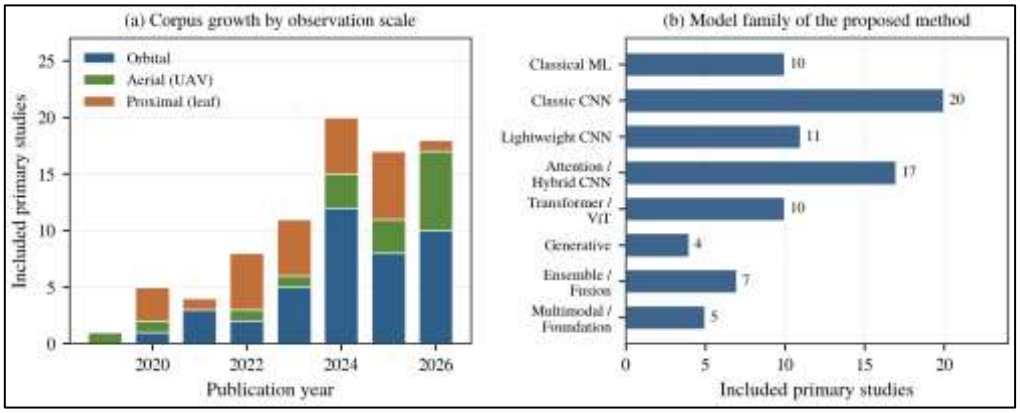

Fig 4. Composition of the corpus. (a) Included studies per year, stacked by observation scale. (b) Model family of the method proposed in each study. The families are mutually exclusive and are assigned by the mechanism the authors themselves emphasise as their contribution.

A third observation is easy to miss in Fig 4 and matters for the rest of the paper. Of the 84 studies, 11 operate across two scales, and of those only 3 evaluate whether information actually transfers between the scales rather than simply concatenating features from both [29], [48], [49]. The multi-scale chain of Fig 1 is therefore asserted far more often than it is tested.

5. Orbital-Scale Evidence

5.1 Sensors, Indices and the Pre-Learning Baseline

Orbital crop monitoring rests on two public archives. Sentinel-2 provides thirteen spectral bands at 10 to 60 metres with a revisit of about five days [1], and Landsat-8 provides a longer historical record at 30 metres [2]; both are now routinely accessed through planetary-scale processing platforms rather than downloaded scene by scene [3]. The analytical vocabulary of the field predates deep learning by four decades: the normalised difference vegetation index [35] and its successors, including the enhanced vegetation index and the soil-adjusted family [36], compress reflectance into scalars that correlate with canopy vigour. These indices remain the backbone of applied work, and several included studies use them as the entire feature representation [49]-[52].

Index-based pipelines are the correct baseline against which learned representations should be judged, and it is significant that many included studies do not run that baseline. Where it is run, the margin attributable to deep learning is frequently modest, because the index already encodes the physically relevant contrast and the learning stage mainly supplies spatial context and temporal smoothing [31].

5.2 Learning on Satellite Image Time Series

The methodological core of the orbital literature is the treatment of the temporal axis. Early deep work applied convolutional and recurrent encoders to stacked multi-temporal observations for land cover and crop type mapping [53], [54]. Self-attention then displaced recurrence, first as a sequence encoder over raw optical series [55] and then in the pixel-set formulation that separates spatial aggregation within a parcel from temporal reasoning across dates [56]. Yield estimation followed a different path, treating the histogram of spectral values over a county as the observation and regressing yield from it with a Gaussian process on top of a learned representation [57].

Within the included corpus these ideas appear in applied form. Incremental and multimodal formulations address the fact that agricultural land changes continuously so that a model trained once degrades [58], [59]. Generative models are used to compensate for the resolution and annotation deficit, most directly in conditional adversarial farmland segmentation on Landsat-8 [60]. Very high-resolution commercial imagery is used where parcel geometry matters, as in agroforestry land-use assessment [61]. Multi-sensor spatiotemporal fusion supports irrigation and water-practice monitoring [62], and flood dynamics work demonstrates the same architecture family transferring to a hydrological task in the same landscape [63]. Ecosystem and sustainability-oriented monitoring appears in several studies that treat the satellite record as an environmental accounting instrument rather than an agronomic one [64]-[67].

Representation learning without labels is the most consequential recent development for this scale, because labels are the binding constraint. Seasonal contrastive pre-training exploits the fact that the same location at different dates provides a natural positive pair [68]; masked autoencoding extends to the temporal and multispectral setting [47], [69]; and geospatial foundation models package the result for downstream fine-tuning [46]. Benchmarks have followed, with EuroSAT and BigEarthNet for scene classification [70], [71] and CropHarvest for global crop-type labels [72]. Only two included studies use any of these pre-trained representations, which is the clearest example in this review of a methodological advance that the applied literature has not yet absorbed.

5.3 Fusion with Ancillary Data and Language Models

A distinct subgroup fuses imagery with non-image evidence. Geographic information layers, weather services and, in one case, a large language model are combined to reason about fertile land discovery and crop suitability [45]. Fuzzy inference over predicted climatic parameters is used to grade sugarcane quality [73]. Economic and policy framings connect the satellite record to farm-level outcomes [74], [75], and applied summaries position the technology for practitioners [76]-[78]. Field-monitoring platforms that package the pipeline into a service are also represented [79], [80], as is broad land-feature prediction from satellite imagery with conventional learners [81].

Fusion of this kind is agronomically sensible and methodologically dangerous. Weather and soil covariates are strong predictors of yield on their own, so a fused model can appear to demonstrate the value of imagery while in fact being carried by the covariates. Of the included fusion studies, none reports the ablation that would settle the question, namely the same model with the imagery removed.

5.4 What the Orbital Literature Does Not Establish

Three limitations are systematic at this scale. First, spatial autocorrelation makes random pixel splitting invalid: neighbouring pixels share illumination, phenology and often the same parcel, so a random split measures interpolation within a field rather than prediction for a new field. Split integrity scores 0 or 1 for 79 percent of orbital studies (Fig 9). Second, the temporal dimension is usually collapsed to a single season, so nothing is learned about the inter-annual variability that dominates operational performance. Third, the reported unit is frequently the pixel while the decision unit is the parcel, and pixel-level accuracy over a large scene is dominated by the interior of large homogeneous fields, which is exactly where the problem is easy.

Table 6 summarises the orbital and aerial studies that are representative of the methodological patterns just described, with the critical observation recorded for each. The table lists a purposive subset rather than the full 58 studies at these two scales; the complete coding sheet underlies the quantitative results of Section 8.

Table 6. Representative orbital and aerial studies, the pattern each exemplifies and the principal methodological reservation recorded during appraisal.

Study	Scale	Data and method	Pattern exemplified and principal reservation
Kussul et al. [53]	Orbital	Landsat-8 and Sentinel-1, multilayer and convolutional networks	Establishes deep classification of crop type; random pixel splitting over contiguous parcels limits the generalisation claim.
Zhong et al. [54]	Orbital	Multi-temporal Landsat, one-dimensional convolution	Shows temporal profile is the discriminative signal; single-season, single-region evaluation.
Russwurm and Korner [55]	Orbital	Raw optical time series, self-attention	Removes hand-crafted preprocessing; cloud handling is learned implicitly and not audited.
Garnot et al. [56]	Orbital	Pixel-set encoder with temporal attention	Separates spatial from temporal aggregation at parcel level; one national dataset only.
You et al. [57]	Orbital	Histogram representation, Gaussian process regression	Yield regression with uncertainty; county aggregation hides within-field variation.
Nair et al. [60]	Orbital	Landsat-8, conditional adversarial segmentation	Generative compensation for annotation scarcity; adversarial realism is not evidence of label correctness.
Chatrabhuj and Meshram [58]	Orbital	Multimodal series, incremental learning	Addresses concept drift explicitly; no external-site test of the incremental claim.
Balasundaram et al. [45]	Orbital	GIS, weather API and language model fusion	Demonstrates heterogeneous fusion; no ablation isolating the imagery contribution.

Study	Scale	Data and method	Pattern exemplified and principal reservation
Haq [62]	Orbital	Multi-sensor spatiotemporal water practice	Operational irrigation relevance; evaluation confined to one basin and one season.
Trivedi et al. [61]	Orbital	Very high-resolution imagery, agroforestry mapping	Parcel geometry exploited; commercial imagery limits reproducibility.
Tetila et al. [4]	Aerial	UAV RGB, superpixel features and classical learners	First systematic UAV soybean disease study; hand-crafted features and a single flight campaign.
Tetila et al. [5], [82]	Aerial	UAV images, deep convolutional networks, pests and diseases	Establishes UAV deep pipelines; images from adjacent frames of one flight can share content across splits.
Nath et al. [83]	Aerial	UAV soybean imagery, convolutional classifier	Recent UAV replication; small evaluation partition and no dispersion reporting.
Mohanty et al. [29]	Aerial and proximal	Leaf and UAV imagery, supervised contrastive cross-domain training	One of the few genuine cross-scale designs; the cross-domain protocol is reported without seed dispersion.
de Souza et al. [48]	Aerial and proximal	Multi-sensor plant imagery, neural classification	Multi-sensor comparison at plant level; sensor-specific preprocessing confounds the comparison.
Castro [49]	Orbital and proximal	Vegetation indices for yield and disease	Explicit attempt to link field signal to disease; disease labels are coarse and indirectly obtained.

6. Aerial-Scale Evidence

6.1 Acquisition and the Labelling Bottleneck

Unmanned aerial acquisition resolves individual plants at one to five centimetres and is flown on demand, which removes the two constraints that bind orbital work, namely resolution and revisit. It introduces others. Orthomosaicking stitches hundreds of overlapping frames into a single raster, and because consecutive frames overlap by design, any split that treats frames or tiles as independent samples will place near-duplicate content on both sides of the partition. This is the aerial form of the leakage problem and it is rarely discussed in the studies that are most exposed to it [5], [83].

Labelling is the binding constraint at this scale. A hectare of orthomosaic contains tens of thousands of plants and cannot be annotated exhaustively, so annotation is usually restricted to sampled patches, which biases the label distribution towards visually obvious symptoms. Weed and crop segmentation work outside the included corpus has shown that transfer between crop types can partially substitute for annotation [84], [85], and severity estimation by semantic segmentation rather than classification gives a more agronomically meaningful target [86].

6.2 Models, Tasks and the State of the Aerial Evidence

The architectural progression at aerial scale mirrors the proximal one with a lag: hand-crafted descriptors with classical learners [4], then transferred convolutional backbones [82], then attention-augmented and hybrid designs, and most recently contrastive cross-domain training that treats the leaf view and the aerial view as two domains of one problem [29]. Ground robots occupy an adjacent niche, providing an under-canopy view that neither satellite nor aerial platforms can obtain, and pest forecasting platforms built on them show what an operational service looks like [87].

With 17 studies, the aerial group is the thinnest in the corpus, and its median FRI of 40.6 is the highest of the three scales, mainly because these studies more often report acquisition metadata such as flight altitude, date and camera. That advantage does not extend to split integrity, where the orthomosaic overlap problem is largely unaddressed, or to external validity, where a second field or a second season is almost never flown.

7. Proximal-Scale Evidence: Soybean Leaf Disease

7.1 Datasets and Their Provenance

Proximal plant disease recognition was defined as a deep learning task by PlantVillage, a corpus of leaves photographed against uniform backgrounds under controlled illumination [88], and by the demonstrations that convolutional networks classify it almost perfectly [6], [7], [89], [90]. The limitation of that corpus was

identified early and precisely: models trained on detached leaves against plain backgrounds lose most of their accuracy on field imagery, and dataset size and variety, not architecture, govern the loss [91]. Lesion-level rather than leaf-level formulations were proposed as a partial remedy [92], and field-captured corpora for other crops made the same point [93]-[95].

Soybean-specific data has followed the same trajectory. Early work used small curated sets [96], then a dedicated architecture and corpus were proposed together [97], then original field images with transfer learning [98]. Hyperspectral acquisition supports explainable identification of charcoal rot at a stage before visible symptoms [99], and the broader stress-phenotyping programme in which that work sits established the ranked explanation of soybean stress symptoms as a scientific rather than merely a diagnostic output [100], [101]. Table 7 summarises the datasets that carry the corpus, with the documented pitfall for each.

Table 7. Datasets carrying the reviewed corpus, with the documented pitfall that constrains claims made from each.

Dataset	Scale	Scope	Documented pitfall
PlantVillage [88]	Proximal	54,000 leaf images, 38 classes, 14 crops	Uniform background and controlled illumination; classifiers exploit background and lighting cues and collapse on field imagery.
Digipathos and field soybean sets [98]	Proximal	Original field photographs of soybean symptoms	Field realism gained at the cost of class imbalance and non-uniform symptom staging.
Aggregated Kaggle soybean leaf corpora [21], [23]	Proximal	Roughly one to two thousand images, 8 to 14 classes	Assembled from several source collections; class names in mixed languages and naming conventions signal an unaudited merge (Section 7.4).
UAV soybean campaigns [4], [5]	Aerial	Flight campaigns over experimental plots	Frame overlap within a campaign; a random split leaks near-duplicate content.
Hyperspectral soybean stress [99]	Proximal	Hyperspectral cubes with expert stress labels	Small sample; instrument-specific calibration limits transfer to RGB pipelines.
EuroSAT and BigEarthNet [70], [71]	Orbital	Sentinel-2 scene and multi-label patches	Land cover rather than crop management; classes are not agronomic decisions.
CropHarvest [72]	Orbital	Global crop-type labels with paired series	Label density is highly uneven geographically, so global claims rest on a few dense regions.
Sentinel-2 and Landsat archives [1], [2]	Orbital	Continuous public multispectral record	No labels; every study builds its own reference data, which is why orbital results are almost never comparable.

7.2 The Architectural Trajectory

The soybean corpus recapitulates the general history of image classification in compressed form, and Table 8 traces it. Transfer from large-scale pre-training remains the foundation [37], [102]-[105]. Efficiency-driven backbones then displaced the heavier ones, first through depthwise separable and inverted-residual designs [106], [107] and then through compound scaling [108], [109]; the lightweight soybean networks in the corpus are direct applications of this lineage [21], [22], [110]. Channel and spatial attention modules were added on top [111], [112] and appear in the corpus as multi-scale feature fusion with attention [23], [113]. Transformers entered through the vision transformer and its hierarchical variant [114]-[116], with modernised convolutional baselines showing that much of the gain was attributable to training recipe rather than to self-attention as such [117]; hybrid convolution and transformer designs for soybean follow this reasoning [24], [25]. Most recently, hyperparameter transfer and progressive fine-tuning of large pre-trained models have been reported for the same task [26].

Parallel developments address data rather than architecture. Class imbalance is treated with generative synthesis, including diffusion-based augmentation of minority classes [27], [118], [119], and with mixing augmentation [120]. Ensemble aggregation with fuzzy integrals is used to combine several lightweight learners [22]. Detection of infestation rather than infection is treated as a separate problem [121], as is cultivar identification from leaf texture, which is a useful reminder that leaf images carry genotype

information that a disease classifier may inadvertently key on [122]. Segmentation-first pipelines that isolate the leaf before classification reduce background dependence [110], [123], and promptable segmentation offers a route to annotation-efficient masks [124].

Table 8. Representative proximal-scale soybean studies, ordered by the methodological step each contributes, with the principal reservation recorded during appraisal.

Study	Year	Contribution	Principal reservation
Wallelign et al. [96]	2018	Early convolutional soybean disease classifier	Very small corpus; no external validation.
Karlekar and Seal [97]	2020	Dedicated architecture with leaf segmentation preprocessing	Segmentation and classification evaluated jointly, so the source of the gain is unresolved.
Bevers et al. [98]	2022	Original field images with transfer learning	Field realism achieved; class staging and severity not controlled.
Farah et al. [121]	2023	Infestation rather than infection detection	Binary framing avoids the class-confusion problem rather than solving it.
Hang et al. [22]	2024	Lightweight network with Choquet fuzzy ensemble	Ensemble gain not separated from the gain of the improved backbone.
Zhang et al. [21]	2025	Lightweight convolutional design for deployment	Parameter count reported, on-device latency not measured.
Liu et al. [113]	2025	Pyramidal multi-scale attention residual network	Evaluated on curated corpora with uniform backgrounds.
An et al. [23]	2026	Multi-scale feature fusion attention network	Single run; no dispersion and no significance test against baselines.
Nan et al. [24]	2026	Convolution and transformer hybrid	Efficiency claim rests on parameter count alone.
Li et al. [25]	2026	Residual transformer with multimodal large model	Multimodal component is not ablated; provenance of the auxiliary text is unstated.
Li et al. [26]	2026	Hyperparameter transfer and progressive fine-tuning	Tuning budget for baselines is not matched to that of the proposed method.
Yadav et al. [110]	2026	Compact segmentation-then-classification pipeline	Segmentation quality is not reported separately from end-task accuracy.
Adugna et al. [28]	2026	Multi-task disease and pesticide-presence identification with explanation	Explanations are shown for selected correct cases only.
Nath et al. [27]	2026	Diffusion-based synthesis for class imbalance	Synthetic images may enter the evaluation partition; the guard is not described.
Kim and Seo [125]	2026	Smart-detection convolutional pipeline	Reported on a curated corpus with no site variation.
Alnuaim et al. [126]	2026	Early-stage detection framing	Early stage is asserted from label names rather than measured against a disease timeline.

7.3 Severity, Pest and Explainable Formulations

Three formulations move beyond single-label classification and deserve more attention than they receive. Severity estimation replaces the class label with a percentage of affected leaf area, which is what fungicide decision thresholds actually reference [86]. Multi-task learning couples disease identification with a second output such as pesticide presence, forcing the representation to encode something beyond the symptom [28]. Explanation, when it is quantitative rather than illustrative, converts the classifier into a phenotyping instrument: ranked symptom explanation on soybean stress produced agronomically meaningful orderings

[100], and hyperspectral explanation localised the wavelengths carrying pre-symptomatic information [99]. In the wider corpus, explanation is usually a saliency map appended to a results section [127]-[129], presented for correctly classified examples and never evaluated.

7.4 Critical Assessment: The Provenance Trap

The most consequential finding at proximal scale is not about architectures. Several widely used soybean leaf corpora are assemblies of two or more source collections, and the assembly is visible in the label vocabulary itself: some class names appear in Title Case English while others appear in lower case with underscores or in Portuguese, and semantically equivalent categories appear twice under different names, for example a rust class alongside a Portuguese-named rust class, or a brown spot class alongside a Septoria class when Septoria glycines is the causal organism of soybean brown spot.

This matters because the source collections differ in camera, background, illumination and compression far more than the diseases differ in appearance. A network trained on the merged corpus can obtain a high score by first deciding which collection an image came from and then emitting the classes that occur in that collection. The diagnostic signature is characteristic and was observed during this review: one group of classes achieves near-perfect recall while every residual error falls among the classes of the other group, and the confusions that do occur are between pathologies with genuinely similar lesion morphology, such as rust, target leaf spot and bacterial pustule. Under those conditions a reported accuracy is not an estimate of diagnostic capability but of provenance separability, and it will not survive deployment on images from a single new camera.

No included study reports the provenance audit that would detect this. The audit is inexpensive: train a classifier to predict the source collection rather than the class, and report its accuracy. If provenance is predictable at high accuracy, then the class labels are entangled with it and any class-level result must be reported on a provenance-disjoint partition. Algorithm 3 in Section 10 incorporates this test as a gate. This single omission is, in the judgement of this review, the largest single threat to the credibility of the proximal literature, and it is entirely fixable.

Two further reservations apply at this scale. First, curated corpora with uniform backgrounds remain in use in 2026 despite a decade-old demonstration that they overstate field performance [91], [130], [131]. Second, near-saturated accuracy on such corpora leaves no headroom in which architectural comparisons can be resolved: when eight methods report between 96 and 99.5 percent on 174 test images, the entire spread corresponds to a handful of images and is well inside the variation produced by changing the random seed [15].

8. Cross-Cutting Analysis

8.1 Model Families and Inductive Bias

Table 9 sets out what each model family assumes, what it costs and how it fails, because these properties, and not the reported accuracies, are what should govern a choice of method for a given agricultural setting. The table makes one point that the corpus obscures: the families are not ordered by capability. A convolutional backbone with strong locality bias is the right choice when data are scarce and lesions are local; a transformer becomes competitive only when data or pre-training compensate for the missing prior; and a foundation model is attractive precisely in the regime where labels are the constraint, which is the orbital regime, and not in the regime where labels are plentiful, which is the curated proximal regime where such models are currently being applied.

Table 9. Model families in the corpus, their inductive bias, data appetite and characteristic failure mode. Counts are the number of included studies proposing a method of that family.

Family	n	Inductive bias	Data appetite and cost	Characteristic failure mode
Classical machine learning [12], [81]	10	Whatever the hand-crafted features encode	Low data, low compute, high feature-engineering effort	Feature set fixed before the phenomenon is understood; ceiling set by the descriptor.
Classic convolutional networks [102]-[104]	20	Locality, translation equivariance, hierarchy	Moderate data with transfer, moderate compute	Background and acquisition shortcuts learned in place of symptoms.
Lightweight convolutional networks [106]-[108]	11	As above with severe capacity constraint	Low compute, deployment oriented	Capacity spent on the easy majority classes; rare symptoms lost.

Family	n	Inductive bias	Data appetite and cost	Characteristic failure mode
Attention and hybrid designs [23], [111], [112]	17	Re-weighting of channels and regions	Moderate to high data	Attention maps reported as explanation without validation; gains within seed noise.
Transformers and vision transformers [114]-[116]	10	Global context, weak locality prior	High data or strong pre-training	Underperforms convolution on small agricultural corpora unless pre-trained.
Generative and diffusion methods [60], [118], [119]	4	Learned data distribution	High compute, unstable training	Synthetic samples leak into evaluation; realism mistaken for label validity.
Ensembles and fusion [22], [132]	7	Variance reduction across learners	Multiplied training and inference cost	Gain attributed to the fusion rule when it comes from added capacity.
Multimodal and foundation models [25], [46], [47]	5	Cross-modal and cross-scene regularity	Very high pre-training cost, low fine-tuning cost	Applied where labels are already plentiful; pre-training corpus provenance unexamined.

8.2 What the Reported Numbers Do and Do Not Show

Fig 5 shows the distribution of reported headline accuracy by family and by scale. Two features are important. First, the medians are close: attention and hybrid designs report a median of 96.6 percent against 93.8 percent for classic convolutional networks, a difference that is smaller than the interquartile range of either group and that is confounded with publication year, dataset and corpus size. Second, the spread is asymmetric and truncated at the top, which is what a saturated metric looks like. With an inconsistency statistic of 94.3 percent (Section 3.7), no ranking of families can be supported by these distributions, and this review declines to offer one.

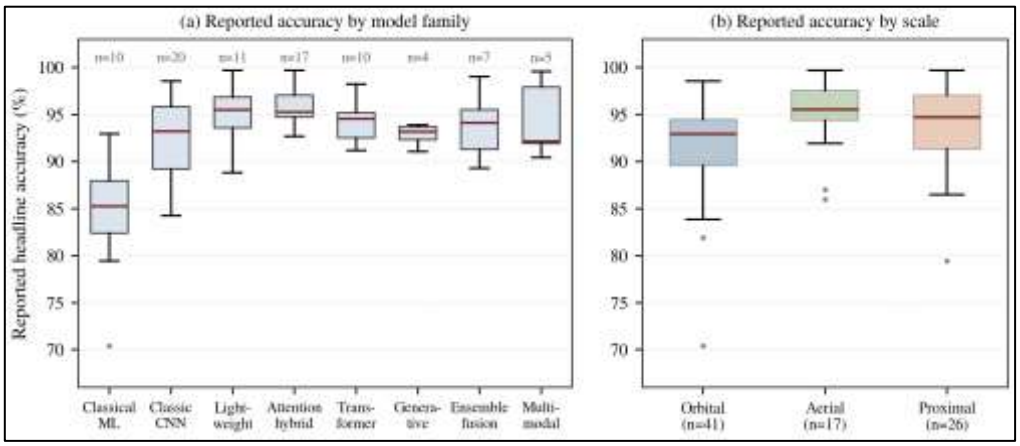

Fig 5. Distribution of reported headline accuracy. (a) By model family, with the number of studies above each box. (b) By observation scale. Medians differ by less than the interquartile range in every pairwise comparison, and the upper truncation near 100 percent is the signature of a saturated evaluation regime.

Fig 6 is the more consequential result. Regressing reported accuracy on the base-ten logarithm of the evaluation-set size gives

$$a_i = b_0 + b_1 \log_{10} n_i + e_i, \quad \text{with } b_1 = -1.93 \text{ and } r = -0.31 \tag{6}$$

so that reported accuracy falls by roughly 1.9 percentage points for every tenfold increase in the size of the evaluation set. The association is modest but its direction is the point. If accuracy measured a stable property of methods, evaluation-set size would be uninformative about it. A negative slope means that the smallest evaluations produce the largest numbers, which is the expected consequence of three mechanisms acting together: variance is larger on small partitions and only the favourable tail is published [16]; small corpora are usually curated and therefore easier; and small corpora are more likely to contain leakage because grouping is harder to enforce when there is little data to group [17]. Twenty-two of the 84 studies evaluate on fewer than 1,000 samples, and their mean reported accuracy is 95.9 percent.

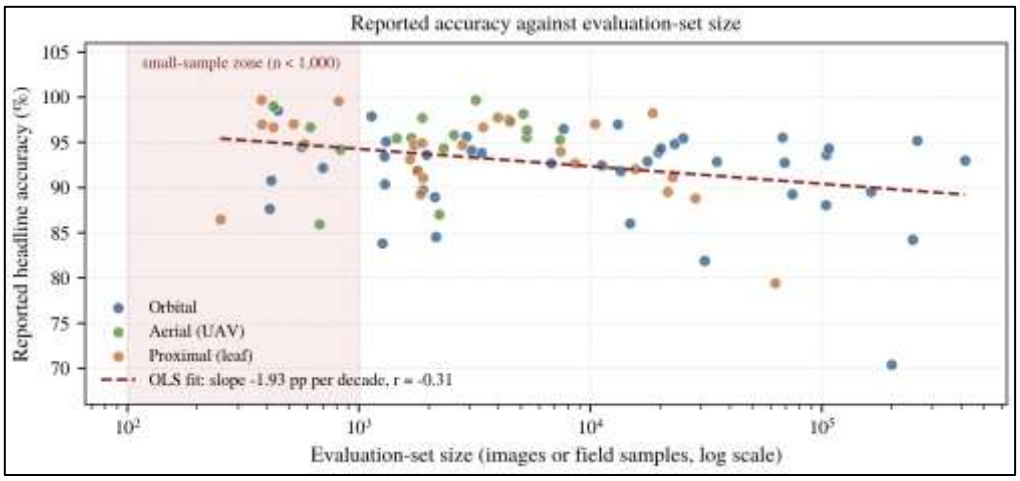

Fig 6. Reported headline accuracy against evaluation-set size on a logarithmic axis, with the ordinary least squares fit. The negative slope indicates that the highest reported accuracies come from the smallest evaluations, which is the signature of small-sample optimism rather than of methodological progress.

8.3 Evaluation and Reporting Audit

Fig 7 gives the AGRI-CRITIC appraisal across the corpus and Table 10 states the operational consequence of each result. The pattern is consistent with critical appraisals in other applied machine learning fields [20]: what is reported well is what reviewers customarily ask for, and what is reported badly is what determines whether a model works outside the paper.

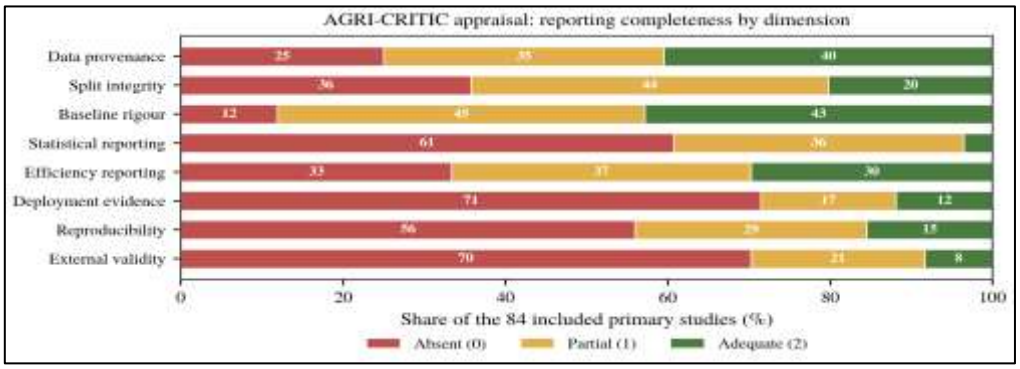

Fig 7. AGRI-CRITIC appraisal of the 84 included studies. Bars show the percentage of studies scoring absent, partial and adequate on each dimension. The four dimensions that most directly predict field performance, namely split integrity, statistical reporting, deployment evidence and external validity, are the four reported worst.

Table 10. Reporting audit: percentage of the 84 included studies at each score level, with the operational consequence of the deficit.

Dimension	Absent	Partial	Adequate	Operational consequence of the deficit
D1 Data provenance	25	35	40	A result cannot be reproduced or reinterpreted when the acquisition conditions and merge history are unknown.
D2 Split integrity	36	44	20	Reported accuracy may measure memorisation of acquisition sessions, plots or overlapping frames rather than generalisation.
D3 Baseline rigour	12	45	43	Improvements over under-trained baselines are indistinguishable from improvements over the state of the art.
D4 Statistical reporting	61	36	4	Differences of one or two percentage points are reported as advances when they lie inside seed-to-seed variation.
D5 Efficiency reporting	33	37	30	Lightweight claims cannot be verified; parameter count alone does not predict latency on a target device.
D6 Deployment evidence	71	17	12	Nothing is known about behaviour under field illumination, motion blur, occlusion or intermittent connectivity.

Dimension	Absent	Partial	Adequate	Operational consequence of the deficit
D7 Reproducibility	56	29	15	Independent verification is impossible, so errors persist in the literature and are propagated by citation.
D8 External validity	70	21	8	The central question, whether the model works on a new site or season, is unanswered for seven studies in ten.

Fig 8 places the audit on a time axis. Mean reported accuracy rises gently from 85.9 percent in 2019 to 93.8 percent in 2026, while mean FRI oscillates between 33 and 44 with no trend. The field is getting better at reporting high numbers and no better at producing evidence. The median FRI over the whole corpus is 37.5 and only 19 of 84 studies, that is 22.6 percent, reach the field-readiness threshold of 60.

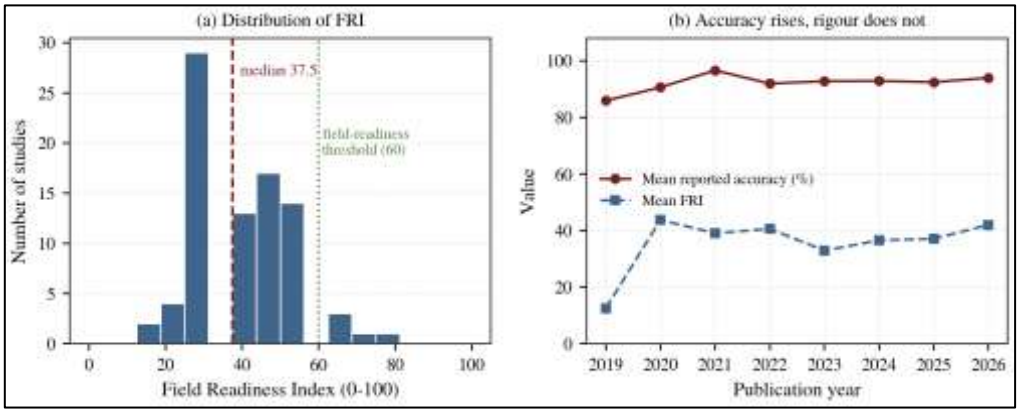

Fig 8. (a) Distribution of the Field Readiness Index across the 84 included studies, with the median and the field-readiness threshold marked. (b) Mean reported accuracy and mean FRI by publication year. Reported performance improves while methodological rigour does not, which is the central diagnostic finding of this review.

Fig 9 disaggregates the appraisal by scale. Orbital studies score best on provenance, because satellite acquisition metadata is intrinsic to the product, and worst on deployment evidence. Aerial studies score best on baseline rigour. Proximal studies score lowest on provenance, which is precisely the dimension implicated in the merged-corpus problem of Section 7.4. No scale scores above one, that is above partial, on statistical reporting, deployment evidence or external validity.

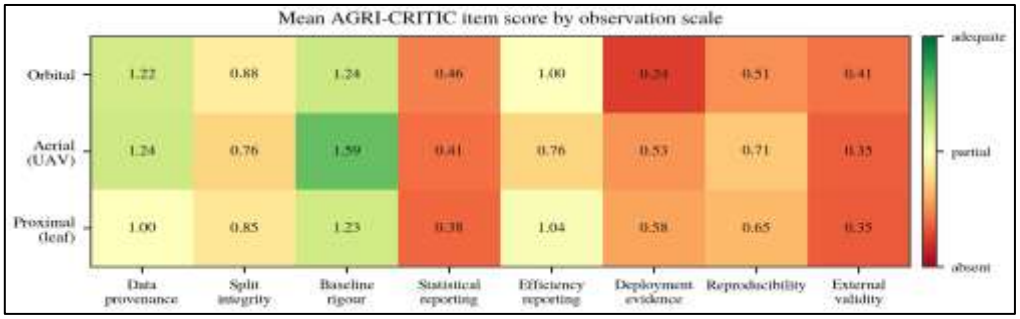

Fig 9. Mean AGRI-CRITIC item score by observation scale, on the 0 to 2 scale. No scale reaches partial reporting on statistical dispersion, deployment evidence or external validity, and the proximal scale is weakest exactly where the provenance trap of Section 7.4 operates.

Two quantities should be reported routinely and are not. The first is a leakage index. For a partition into training set X and test set Y with an acquisition grouping g, define

$$L = | \{ y \text{ in } Y : g(y) \text{ is in } g(X) \} | / | Y | \quad (7)$$

the fraction of test items whose acquisition group also appears in training. A protocol is group-disjoint when L = 0. Reporting L costs nothing and settles the most common objection to a leaf or UAV result. The second is a significance statement for the comparison that the paper is actually making. Because train and test partitions overlap across resampling repetitions, the naive paired t test is anti-conservative, and the corrected resampled statistic should be used [133], [134],

$$t = mean(d) / sqrt((1/k + ntest / ntrain) \text{ times } var(d)) \quad (8)$$

with d the vector of per-repetition differences between the proposed method and the baseline over k repetitions. Where several methods are compared over several datasets, the rank-based procedure of Demsar is the appropriate instrument rather than a family of pairwise tests [135]. Four percent of the corpus reports anything of this kind.

8.4 The Generalisation Gap

Nineteen studies report both an in-domain result and an evaluation outside the training distribution, whether a different site, a different season or a different acquisition device. Defining the gap for study *i* as

$$Deltai = aini - aexti \quad (9)$$

the mean gap over those 19 studies is 17.8 percentage points and the distribution is shown in Fig 10. A model reporting 96 percent in domain is therefore expected to deliver something near 78 percent on a new site, and the studies with the highest in-domain numbers do not have the smallest gaps. This is the empirical form of underspecification: many models fit the training distribution equally well and differ arbitrarily off it [18], [19].

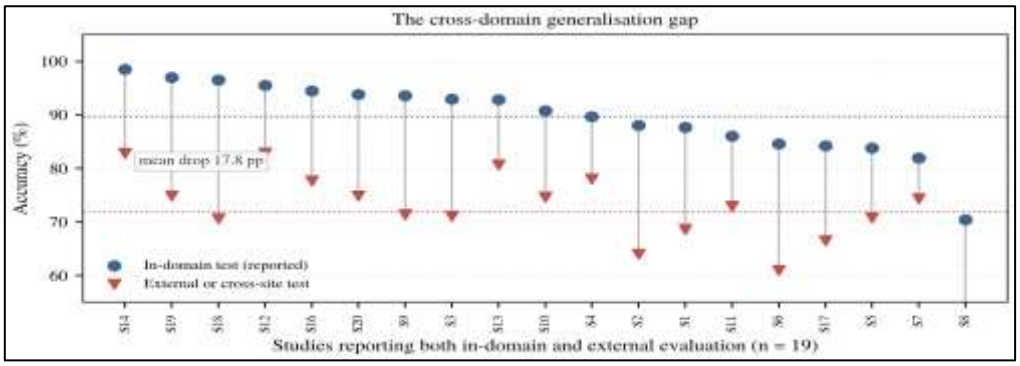

Fig 10. In-domain and external accuracy for the 19 studies reporting both, sorted by in-domain accuracy. Study identifiers are the anonymised codes of the review coding sheet. The mean loss on moving off the training distribution is 17.8 percentage points, and it is not smaller for the studies reporting the highest in-domain numbers.

The methods that address this directly are known and are under-used in the corpus. Domain-adversarial training aligns feature distributions between a labelled source and an unlabelled target [136]. Self-supervised pre-training on unlabelled in-domain imagery reduces the dependence on the labelled source distribution [68], [137]. Few-shot formulations adapt to a new site from a handful of annotated examples rather than a new campaign [138]. Supervised contrastive learning across leaf and aerial views is the one cross-domain design present in the soybean corpus [29]. Adoption of any of these appears in 9 of 84 studies.

8.5 Explainability and Agronomic Trust

Explanation appears in 23 studies and is almost always a class activation map shown for correct predictions [127]. Three requirements distinguish an explanation that supports an agronomic decision from one that decorates a results section. It must be evaluated, for example by asking a plant pathologist whether the highlighted region contains the lesion, or by measuring the drop in accuracy when the highlighted region is occluded. It must be shown for failures, since those are the cases in which a grower would consult it. And the underlying probability must be calibrated, because a threshold on a miscalibrated score is not a decision rule [139]; deep ensembles remain the simplest reliable route to usable uncertainty [132]. The exemplar in this domain remains the ranked, quantitatively validated explanation of soybean stress symptoms [100] and its hyperspectral counterpart [99]. Attribution methods with axiomatic backing [128], [129] are used in two studies. No included study reports calibration.

8.6 Efficiency, Edge Deployment and Energy

Thirty-one studies describe their method as lightweight. Of those, 19 support the claim with a parameter count only. Parameter count is a poor proxy for latency, because memory access patterns, depthwise convolution efficiency and the target runtime dominate [106], [107]. A defensible efficiency claim reports operations, peak memory, and measured latency on a named device under a stated batch size, and where compression is used it reports the accuracy cost of each stage, whether pruning and quantisation [140], [141] or distillation from a larger teacher [142]. A simple normalisation makes the trade-off explicit,

$$U = a / (1 + log10 (1 + FLOPs / 109)) \quad (10)$$

which discounts accuracy by computational cost and reorders the corpus substantially: several of the highest-accuracy methods fall below lightweight designs that report within one percentage point of them at a fraction of the cost. Energy accounting is absent from the corpus entirely, despite an established framing for it [40], [41], and it is not a peripheral concern for a technology intended for battery-powered handheld and aerial platforms.

8.7 Data Governance and Distributed Learning

Field imagery is commercially sensitive: it reveals yield, disease pressure and management practice at parcel level. Federated learning is the standard response, training a shared model without centralising the

imagery [143], and it fits the agricultural setting well because holdings are numerous, heterogeneous and unwilling to share raw data. It appears in no included study. Dataset documentation practice is equally absent: none of the corpora in Table 7 is released with a datasheet [42] and no included model is released with a model card [43]. These are cheap instruments whose absence is not a research problem but a norms problem, and norms are what this review's protocol in Section 10 is intended to shift. The broader point has been made before, that a field is worth judging by whether its results change a decision outside the field [144].

9. Seven Recurring Failure Modes

The evidence of Sections 5 to 8 consolidates into seven failure modes, given in Table 11 with the corpus evidence and the mitigation for each. They are ordered by the severity of the inferential damage they cause rather than by frequency.

Table 11. Seven recurring failure modes, the evidence for each in the reviewed corpus and the corresponding mitigation. Mitigations map onto the FIELD-READY items of Table 12.

Failure mode	Description	Evidence in this corpus	Mitigation
F1 Provenance entanglement	Merged corpora let the model separate source collections instead of pathologies.	Mixed-language and duplicated class vocabularies in widely used soybean corpora; error concentration in one label group (Section 7.4).	Provenance classifier test; provenance-disjoint partitions; datasheet for the merged corpus.
F2 Grouping leakage	Random splits over correlated units place near-duplicates on both sides of the partition.	Split integrity absent or partial in 80 percent of studies; frame overlap in UAV campaigns; spatial autocorrelation in orbital pixels.	Group-disjoint splits by site, plot, plant or session; report the leakage index L of (7).
F3 Saturation without headroom	Curated corpora leave differences smaller than seed noise.	Median accuracy 93.9 percent with upper truncation; families separated by less than their interquartile range.	Harder benchmarks; report dispersion; adopt the corrected resampled test of (8).
F4 Small-sample optimism	The smallest evaluations produce the largest numbers.	Slope of -1.93 percentage points per decade of samples, r = -0.31 (Fig 6); 22 studies evaluate on fewer than 1,000 samples.	Minimum evaluation-set reporting with confidence intervals; size-weighted synthesis as in (4).
F5 Unmatched baselines	Proposed methods are compared against under-trained or copied baselines.	Baseline rigour partial or absent in 57 percent of studies; tuning budgets rarely stated.	Retrain every baseline on the identical split under a matched budget; state the budget.
F6 Deployment absence	Nothing is measured on the device or in the field.	Deployment evidence absent in 71 percent; 19 of 31 lightweight claims rest on parameter count alone.	Report operations, memory, measured latency and energy on a named device; publish failure cases.
F7 Decorative explanation	Explanations are illustrative, uncalibrated and shown only for successes.	Explanation present in 23 studies, evaluated in none, calibration reported in none.	Evaluate explanations against expert annotation or occlusion; report calibration and uncertainty.

10. The FIELD-READY Reporting Protocol

Criticism that does not terminate in a procedure is of limited value. FIELD-READY is a ten-item protocol with two hard gates, shown as a process in Fig 11 and as a checklist in Table 12. It is deliberately modest: every item is achievable within a normal study, and none requires new data collection except item S9, which requires one additional site or season and is the single most valuable increment of effort available to this field.

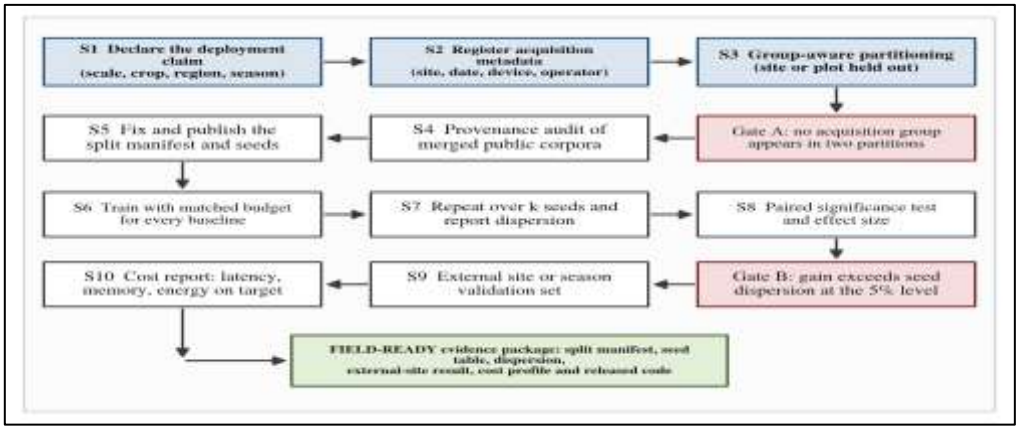

Fig 11. The FIELD-READY evaluation protocol. Gate A blocks progress until the partition is group-disjoint and provenance-audited; Gate B blocks a performance claim until the improvement exceeds seed dispersion under a paired test. The output is an evidence package rather than a single number.

Table 12. The FIELD-READY checklist. Each item states what is reported and how a reviewer can verify it without re-running the experiment.

ID	Item	What is reported	How a reviewer verifies it
S1	Deployment claim	The scale, crop, region and season for which the model is claimed to work.	The claim is stated in the abstract and is not broader than the data supports.
S2	Acquisition metadata	Site, date, device, illumination, operator and any merge of sub-corpora.	A table of acquisition groups with counts is present.
S3	Group-aware partition	The grouping variable and the rule used to assign groups to partitions.	The leakage index L of (7) is reported and equals zero.
S4	Provenance audit	Accuracy of a classifier trained to predict the source collection.	If provenance accuracy is high, class results are reported on provenance-disjoint splits.
S5	Split manifest	The exact file lists or identifiers for each partition, with seeds.	The manifest is downloadable and its counts match the reported denominators.
S6	Matched training budget	Epochs, schedule and tuning trials for the proposed method and every baseline.	Budgets are equal, or the imbalance is stated and favours the baselines.
S7	Repetition and dispersion	At least five seeds with mean and standard deviation for every reported metric.	Every headline number carries a dispersion term.
S8	Significance and effect size	A paired corrected test as in (8) with the effect size on the difference.	The test statistic, the number of repetitions and the effect size are given.
S9	External evaluation	One independent site, season or device, reported without retuning.	The external result appears in the same table as the in-domain result.
S10	Cost profile	Parameters, operations, peak memory, measured latency and energy on a named device.	Device, batch size and runtime are named; the utility of (10) can be computed.

Algorithm 3 states the partitioning and audit procedure that satisfies items S3 to S5, and Algorithm 4 states the cross-scale transfer audit that a study claiming a multi-scale contribution should run in place of simple feature concatenation.

Algorithm 3: Provenance-aware group-disjoint evaluation with leakage audit

Input: dataset D with items (x, y, g, p) where g is the acquisition group and p the source collection; fold count k
Output: partitions $P_1..P_k$; leakage index L; provenance separability rho
1: **assert** every item carries a group label g; **if** not, stop and collect it (S2)
2: G <- distinct groups of D; shuffle G with a recorded seed
3: assign whole groups of G to folds so that class proportions are as balanced as possible
4: **for** each fold partition (X, Y) **do**
5: L <- fraction of y in Y whose g(y) occurs in g(X)
6: **if** L > 0 **then** raise a leakage error and stop (Gate A)
7: **end for**
8: *provenance audit (S4)*
9: train a classifier f_p on X to predict the source collection p from x
10: rho <- accuracy of f_p on Y
11: **if** rho > 1/|P| + delta **then** (provenance is predictable)
12: re-partition so that each source collection appears in only one partition, or report class results per collection
13: **end if**
14: write the split manifest with item identifiers, group labels, seeds and L (S5)
15: **return** partitions, L, rho

Algorithm 4: Cross-scale transfer audit for multi-scale claims

Input: orbital model M_o, aerial model M_a, proximal model M_p; paired sites with observations at all three scales
Output: transfer matrix T; triage yield tau; decision-level utility U_d
1: **for each** ordered pair (s, t) of scales **do**
2: freeze the representation of M_s and fit only a linear head on the labels of scale t
3: T[s][t] <- accuracy of that head on a group-disjoint test set of scale t
4: **end for**
5: *triage evaluation, the operational form of Fig 1*
6: run M_o over the site and rank plots by predicted anomaly
7: inspect the top q percent of plots with M_a and then with M_p
8: tau <- (true positive plots found within the top q percent) / (all true positive plots)
9: **if** tau is not greater than the value for random plot ordering **then** the multi-scale claim is unsupported
10: U_d <- expected value of the resulting spray decision under stated cost and loss parameters
11: **return** T, tau, U_d

Algorithm 4 is the test that 8 of the 11 multi-scale studies in the corpus do not perform. Concatenating orbital and proximal features and observing an accuracy improvement does not demonstrate that the scales inform one another; only the triage yield of line 8, compared against a random ordering, does that.

11. Research Agenda

Fig 12 arranges the open problems into three horizons and Table 13 ties each to the specific weakness in the current evidence that motivates it. The ordering is not by intellectual interest but by the ratio of evidential value to cost, which places unglamorous protocol work first.

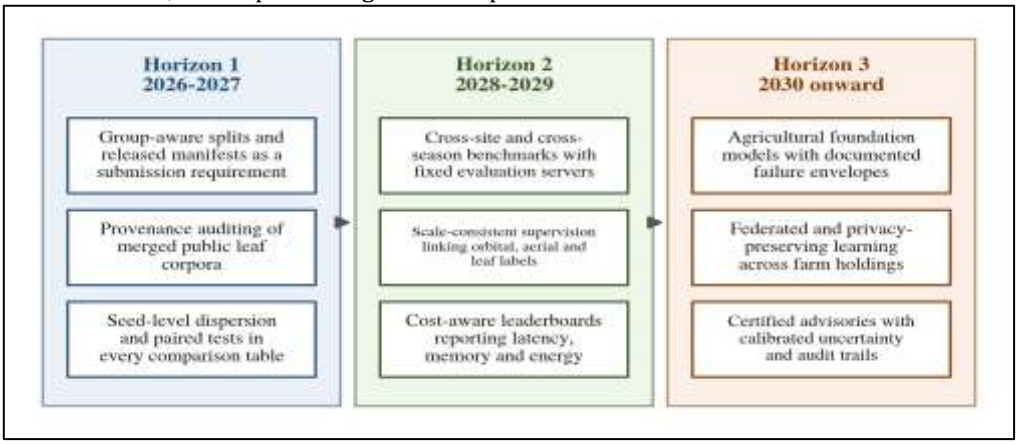

Fig 12. Three-horizon research agenda. Horizon 1 items require no new methods and would remove the largest sources of non-transferable evidence; Horizon 2 requires shared infrastructure; Horizon 3 requires new science.

Table 13. Research agenda with the evidential deficit that motivates each item and the expected effect on the field.

ID	Horizon	Open problem	Evidential deficit addressed and expected effect
A1	1	Group-disjoint splitting and published split manifests as a condition of publication.	F2 and the 80 percent of studies with weak split integrity; would make reported accuracies comparable for the first time.
A2	1	Provenance auditing and datasheets for merged public leaf corpora.	F1; would resolve whether reported soybean accuracies measure diagnosis or collection membership.
A3	1	Seed dispersion and corrected paired tests in every comparison table.	F3 and F5; would stop the reporting of seed noise as methodological advance.
A4	2	Cross-site and cross-season benchmarks with held-out evaluation servers.	F4 and the 70 percent absence of external validity; would convert generalisation from an assumption into a measured quantity.
A5	2	Scale-consistent supervision that links orbital, aerial and leaf labels for the same parcels.	The untested multi-scale chain of Fig 1; would allow triage yield rather than accuracy to become the reported metric.
A6	2	Cost-aware leaderboards reporting latency, memory and energy on named devices.	F6; would make the lightweight literature verifiable and would surface the utility trade-off of (10).
A7	3	Agricultural foundation models with documented failure envelopes and released pre-training provenance.	The label bottleneck at orbital scale; would move the field from per-study data collection to shared representation.
A8	3	Federated and privacy-preserving learning across holdings.	The complete absence of governance-aware training; would unlock commercially sensitive field data at scale.
A9	3	Calibrated, uncertainty-aware advisories with auditable decision traces.	F7 and the absence of calibration; would allow agronomic decisions to be defended and insured.

12. Threats to the Validity of This Review

This review is subject to the same scrutiny it applies. Five threats are recorded.

Selection. Six databases and English-language records only. Regionally published agricultural research, particularly in languages other than English, is under-represented, and 188 non-English records were excluded at screening. The corpus therefore describes the internationally indexed literature rather than all work performed.

Extraction. Headline accuracy was taken as the value the authors foreground, which favours the most optimistic framing in a study reporting several. This choice was made because that is the number the literature propagates, but it means the reported distribution in Fig 5 is an upper envelope of what the studies actually established.

Appraisal subjectivity. AGRI-CRITIC scores are judgements. Item-level agreement was k = 0.78 and disagreements of one level were resolved downward. Moving the field-readiness threshold from 60 to 50 raises the share of adequate studies from 22.6 percent to 34.5 percent, and to 70 gives 9.5 percent; the qualitative conclusion, that most of the corpus is not field ready, is stable across that range. Using the weighted index of (3) instead of (2) lowers the median FRI by 3.1 points, because the doubled dimensions are the weakest ones.

Synthesis. Reported accuracies are heterogeneous to the point of inconsistency at 94.3 percent, which is why no pooled headline value is offered. The regression of (6) is descriptive and is confounded with scale, year and corpus curation; it is presented as a diagnostic signature rather than as a causal claim.

Coverage of the newest work. The search closed on 30 April 2026, and the corpus contains 21 studies dated 2026, several of them in press at the time of screening. Very recent multimodal and foundation-model work is therefore represented thinly, and the conclusion that such methods are applied where labels are already plentiful should be revisited when that subliterature matures.

13. Conclusion

Artificial intelligence for crop monitoring has produced a large, fast-growing and methodologically fragile evidence base. Across 84 primary studies spanning orbital, aerial and proximal observation, the median reported accuracy is 93.9 percent and the median Field Readiness Index is 37.5 out of 100. Reported accuracy has improved since 2019; methodological rigour has not. The highest accuracies come from the

smallest evaluations, at a rate of 1.9 percentage points per decade of samples. Where generalisation is measured at all, which is in fewer than one study in ten, it costs 17.8 percentage points on average. Split integrity is absent or partial in four studies out of five, statistical dispersion is reported by one in twenty-five, and deployment on a device in a field is evidenced by roughly one in eight.

The proximal literature carries a specific and correctable defect. Widely used soybean leaf corpora are unaudited merges of separate source collections whose class vocabularies still bear the marks of their origins, and on such data a classifier can score highly by recognising provenance rather than pathology. The audit that would settle this costs one training run and appears in no included study.

None of this argues against the technology. The physical basis is sound, the sensing infrastructure is public and improving, and the architectural toolkit is more than adequate for the task. What is missing is evidence discipline: group-disjoint partitions, matched baselines, dispersion, an external site, and a cost profile. FIELD-READY sets out those requirements as ten items with two gates, and Algorithms 3 and 4 make the two hardest of them mechanical. If the next generation of studies adopts them, the field will lose some impressive numbers and gain the ability to say, for the first time and with evidence, that a model will work on a farm it has never seen.

References

[1] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, et al., "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," Remote Sens. Environ., vol. 120, pp. 25-36, 2012.

[2] D. P. Roy, M. A. Wulder, T. R. Loveland, C. E. Woodcock, R. G. Allen, M. C. Anderson, D. Helder, J. R. Irons, D. M. Johnson, R. Kennedy, et al., "Landsat-8: Science and product vision for terrestrial global change research," Remote Sens. Environ., vol. 145, pp. 154-172, 2014.

[3] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," Remote Sens. Environ., vol. 202, pp. 18-27, 2017.

[4] E. C. Tetila, B. B. Machado, N. A. de Souza Belete, D. A. Guimaraes, and H. Pistori, "Identification of soybean foliar diseases using unmanned aerial vehicle images," IEEE Geosci. Remote Sens. Lett., vol. 14, no. 12, pp. 2190-2194, Dec. 2017.

[5] E. C. Tetila, B. B. Machado, G. Astolfi, N. A. de Souza Belete, W. P. Amorim, A. R. Roel, and H. Pistori, "Detection and classification of soybean pests using deep learning with UAV images," Comput. Electron. Agriculture, vol. 179, Art. no. 105836, 2020.

[6] S. P. Mohanty, D. P. Hughes, and M. Salathe, "Using deep learning for image-based plant disease detection," Frontiers Plant Sci., vol. 7, Art. no. 1419, 2016.

[7] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," Comput. Electron. Agriculture, vol. 145, pp. 311-318, 2018.

[8] B. Victor, A. Nibali, and Z. He, "A systematic review of the use of deep learning in satellite imagery for agriculture," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 17, pp. 2297-2316, 2024.

[9] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," ISPRS J. Photogramm. Remote Sens., vol. 152, pp. 166-177, 2019.

[10] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," IEEE Geosci. Remote Sens. Mag., vol. 5, no. 4, pp. 8-36, Dec. 2017.

[11] A. Kamilaris and F. X. Prenafeta-Boldu, "Deep learning in agriculture: A survey," Comput. Electron. Agriculture, vol. 147, pp. 70-90, 2018.

[12] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," Sensors, vol. 18, no. 8, Art. no. 2674, 2018.

[13] A. Raihan, "A systematic review of geographic information systems (GIS) in agriculture for evidence-based decision making and sustainability," Global Sustainability Res., vol. 3, no. 1, pp. 1-24, 2024.

[14] B. L. Mahesh and S. U. Shenoy, "Advances in crop classification using satellite imagery: A comprehensive review," in Proc. IEEE Int. Conf. Comput. Vis. Mach. Intell. (CVMI), 2024, pp. 1-7.

[15] X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Mohammadi Sepahvand, E. Raff, K. Madan, V. Voleti, et al., "Accounting for variance in machine learning benchmarks," in Proc. Mach. Learn. Syst. (MLSys), 2021, pp. 747-769.

[16] D. Sculley, J. Snoek, A. Wiltschko, and A. Rahimi, "Winner's curse? On pace, progress, and empirical rigor," in Proc. Int. Conf. Learn. Represent. (ICLR) Workshop Track, 2018.

[17] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, Art. no. 100804, 2023.

[18] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet classifiers generalize to ImageNet?," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 5389-5400.

[19] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al., "Underspecification presents challenges for credibility in modern machine learning," J. Mach. Learn. Res., vol. 23, no. 226, pp. 1-61, 2022.

[20] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and computed tomography scans," Nature Mach. Intell., vol. 3, pp. 199-217, 2021.
[21] Y. Zhang, R. Bao, M. Guan, Z. Wang, L. Wang, X. Cui, and Y. Wang, "A lightweight deep convolutional neural network development for soybean leaf disease recognition," Frontiers Plant Sci., vol. 16, Art. no. 1655564, 2025.
[22] Y. Hang, X. Meng, and Q. Wu, "Application of improved lightweight network and Choquet fuzzy ensemble technology for soybean disease identification," IEEE Access, vol. 12, pp. 25146-25163, 2024.
[23] W. An, J. Yang, Y. Ren, and H. Ni, "High-precision MSFFANet model for soybean leaf disease recognition," Discover Comput., vol. 29, no. 1, Art. no. 330, 2026.
[24] F. Nan, Q. Liu, Y. Hu, and H. Liu, "Soy-Hybrid: A hybrid architecture for accurate and efficient soybean leaf disease recognition," in Proc. Int. Conf. Intell. Comput. Singapore: Springer, Jul. 2026, pp. 289-300.
[25] X. Li, W. Bian, B. Yang, Q. Yang, W. Cui, J. Liang, and J. Gu, "Soybean leaf disease recognition based on Sem-ResFormer and multimodal large models," Agronomy, vol. 16, no. 12, Art. no. 1132, 2026.
[26] X. Li, W. Bian, B. Yang, Y. Li, S. Wang, N. Qin, and H. Sun, "Soybean leaf disease recognition methods based on hyperparameter transfer and progressive fine-tuning of large models," Agronomy, vol. 16, no. 2, Art. no. 218, 2026.
[27] S. Nath, S. K. Biswas, K. Kumar, D. Tripathi, and S. Majumdar, "A CNN model for soybean leaf disease prediction treating class imbalance problem using diffusion model: A smart farming application," in Proc. 4th Int. Conf. Secure Cyber Comput. Commun. (ICSCCC), May 2026, pp. 1266-1271.
[28] F. B. Adugna, E. A. Desta, and A. A. Hailu, "Efficient multi-task CNN framework for joint identification of soybean leaf diseases and pesticide presence with explainable AI," Sci. Rep., 2026.
[29] S. S. Mohanty, V. Maheshwari, P. K. Aithal, and R. D. Jathanna, "A multi-scale supervised contrastive framework for cross-domain soybean disease classification using leaf and UAV imagery," Sci. Rep., 2026.
[30] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, Art. no. n71, 2021.
[31] M. Weiss, F. Jacob, and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," Remote Sens. Environ., vol. 236, Art. no. 111402, 2020.
[32] Y. Lu and S. Young, "A survey of public datasets for computer vision tasks in precision agriculture," Comput. Electron. Agriculture, vol. 178, Art. no. 105760, 2020.
[33] S. Getahun, H. Kefale, and Y. Gelaye, "Application of precision agriculture technologies for sustainable crop production and environmental sustainability: A systematic review," Sci. World J., vol. 2024, Art. no. 2126734, 2024.
[34] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele Univ. and Univ. Durham, U.K., EBSE Tech. Rep. EBSE-2007-01, 2007.
[35] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," Remote Sens. Environ., vol. 8, no. 2, pp. 127-150, 1979.
[36] A. Huete, K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira, "Overview of the radiometric and biophysical performance of the MODIS vegetation indices," Remote Sens. Environ., vol. 83, no. 1-2, pp. 195-213, 2002.
[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2009, pp. 248-255.
[38] J. Cohen, "A coefficient of agreement for nominal scales," Educ. Psychol. Meas., vol. 20, no. 1, pp. 37-46, 1960.
[39] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no. 1, pp. 159-174, 1977.
[40] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," Commun. ACM, vol. 63, no. 12, pp. 54-63, 2020.
[41] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL), 2019, pp. 3645-3650.
[42] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daume III, and K. Crawford, "Datasheets for datasets," Commun. ACM, vol. 64, no. 12, pp. 86-92, 2021.
[43] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in Proc. Conf. Fairness, Accountability, Transparency (FAT), 2019, pp. 220-229.
[44] J. P. T. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman, "Measuring inconsistency in meta-analyses," BMJ, vol. 327, no. 7414, pp. 557-560, 2003.

[45] A. Balasundaram, A. B. Abdul Aziz, A. Gupta, A. Shaik, and M. S. Kavitha, "A fusion approach using GIS, green area detection, weather API and GPT for satellite image-based fertile land discovery and crop suitability," Sci. Rep., vol. 14, no. 1, Art. no. 16241, 2024.
[46] J. Jakubik, S. Roy, C. E. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards, et al., "Foundation models for generalist geospatial artificial intelligence," arXiv:2310.18660, 2023.
[47] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. B. Lobell, and S. Ermon, "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2022, pp. 197-211.
[48] F. L. P. de Souza, L. S. Shiratsuchi, M. A. Dias, M. R. Barbosa Junior, T. D. Setiyono, S. Campos, and H. Tao, "A neural network approach employed to classify soybean plants using multi-sensor images," Precis. Agriculture, vol. 26, no. 2, Art. no. 32, 2025.
[49] R. Castro, "Prediction of yield and diseases in crops using vegetation indices through satellite image processing," in Proc. IEEE Technol. Eng. Manage. Soc. (TEMSCON LATAM), 2024, pp. 1-6.
[50] K. Anderson, "Detecting environmental stress in agriculture using satellite imagery and spectral indices," Environ. Monit. Assess., preprint, 2024.
[51] S. Choudhury, "Advanced technology of using satellite data fuels in agriculture," Chronicle Bioresour. Manage., vol. 8, no. 1, pp. 29-34, 2024.
[52] A. A. Kzar, "Change detection using multispectral satellite images for sustainable development in regions of Al-Najaf Al-Ashraf, Iraq," J. Kufa-Physics, vol. 16, no. 1, pp. 42-55, 2024.
[53] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," IEEE Geosci. Remote Sens. Lett., vol. 14, no. 5, pp. 778-782, May 2017.
[54] L. Zhong, L. Hu, and H. Zhou, "Deep learning based multi-temporal crop classification," Remote Sens. Environ., vol. 221, pp. 430-443, 2019.
[55] M. Russwurm and M. Korner, "Self-attention for raw optical satellite time series classification," ISPRS J. Photogramm. Remote Sens., vol. 169, pp. 421-435, 2020.
[56] V. Sainte Fare Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Satellite image time series classification with pixel-set encoders and temporal self-attention," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 12325-12334.
[57] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian process for crop yield prediction based on remote sensing data," in Proc. AAAI Conf. Artif. Intell., 2017, pp. 4559-4565.
[58] Chatrabhuj and K. Meshram, "Incremental learning model for sustainable agricultural land assessment using multimodal satellite data," Int. J. Remote Sens., vol. 45, no. 22, pp. 8622-8648, 2024.
[59] Chatrabhuj and K. Meshram, "Design of an efficient satellite image analysis model for identification of sustainable construction areas via ground subsidence monitoring and infrastructure monitoring operations," Indian Geotech. J., preprint, pp. 1-16, 2024.
[60] S. Nair, S. Sharifzadeh, and V. Palade, "Farmland segmentation in Landsat 8 satellite images using deep learning and conditional generative adversarial networks," Remote Sens., vol. 16, no. 5, Art. no. 823, 2024.
[61] S. Trivedi, P. V. Vinod, B. Chandrasekaran, M. K. Nagashree, S. R. Subramoniam, V. B. Manjula, and A. Singh, "Geospatial assessment of agroforestry land use systems using very high-resolution satellite images and artificial intelligence," Agroforestry Syst., vol. 98, no. 8, pp. 2875-2895, 2024.
[62] M. A. Haq, "Intelligent sustainable agricultural water practice using multi-sensor spatiotemporal evolution," Environ. Technol., vol. 45, no. 12, pp. 2285-2298, 2024.
[63] P. Lemenkova, "Deep learning methods of satellite image processing for monitoring of flood dynamics in the Ganges Delta, Bangladesh," Water, vol. 16, Art. no. 1141, 2024.
[64] A. Mulakaledu, B. Swathi, M. M. Jadhav, S. M. Shukri, V. Bakka, and P. Jangir, "Satellite image-based ecosystem monitoring with sustainable agriculture analysis using machine learning model," Remote Sens. Earth Syst. Sci., pp. 1-10, 2024.
[65] O. Opryshko, N. Pasichnyk, N. Kiktev, A. Dudnyk, T. Hutsol, K. Mudryk, P. Herbut, P. Lyszczarz, and V. Kukharets, "European Green Deal: Satellite monitoring in the implementation of the concept of agricultural development in an urbanized environment," Sustainability, vol. 16, no. 7, Art. no. 2649, 2024.
[66] N. Efremova, J. C. Foley, A. Unagaev, and R. Karimi, "AI for sustainable agriculture and rangeland monitoring," in The Ethics of Artificial Intelligence for the Sustainable Development Goals. Cham, Switzerland: Springer, 2023, pp. 399-422.
[67] J.-K. Yu, J. Ahn, G. D. Han, H.-M. Kang, H. Jo, and Y. S. Chung, "Utilization of satellite technologies for agriculture," J. Environ. Sci. Int., vol. 33, no. 7, pp. 547-552, 2024.
[68] O. Manas, A. Lacoste, X. Giro-i-Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 9414-9423.

[69] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked autoencoders are scalable vision learners," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 16000-16009.
[70] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 12, no. 7, pp. 2217-2226, Jul. 2019.
[71] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding," in Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), 2019, pp. 5901-5904.
[72] G. Tseng, I. Zvonkov, C. L. Nakalembe, and H. Kerner, "CropHarvest: A global dataset for crop-type classification," in Proc. NeurIPS Datasets and Benchmarks Track, 2021.
[73] A. E. Badillo-Marquez, I. Pardo-Escandon, A. A. Aguilar-Lasserre, C. G. Moras-Sanchez, and R. Flores-Asis, "Intelligent system based on a satellite image detection algorithm and a fuzzy model for evaluating sugarcane crop quality by predicting uncertain climatic parameters," J. Agricultural Eng., vol. 55, no. 4, 2024.
[74] D. Wuepper and W. A. Oluoch, "Satellite data in agricultural and environmental economics: Theory and practice," Agricultural and Environmental Economics, preprint, 2024.
[75] J. Segarra, "Satellite imagery in precision agriculture," in Digital Agriculture: A Solution for Sustainable Food and Nutritional Security. Cham, Switzerland: Springer, 2024, pp. 325-340.
[76] S. Gore, D. Patil, N. Mahankale, and S. Gore, "Satellite imaging for precision agriculture enhancing crop management, soil condition and yield prediction," in Emerging Trends in Smart Societies. London, U.K.: Routledge, 2024, pp. 434-437.
[77] A. K. Shinwari, S. Hameed, M. S. Akhter, F. U. Hassan, and A. Sani, "A vision for the future: The next wave of innovation in precision agriculture," in Advancing Precision Agriculture With AI and IoT. Hershey, PA, USA: IGI Global, 2026, pp. 427-456.
[78] V. Pandit, M. Sravya, C. Ravali, S. Fiaz, A. Tariq, and S. Chatterjee, "Precision agriculture in the era of technological advancement," in Artificial Intelligence and Data Sciences for Precision Agriculture. Cham, Switzerland: Springer, 2026, pp. 159-186.
[79] B. Gowshika, "Satellite based agriculture field monitoring and data analysis system," in Proc. 6th Int. Conf. Trends Mater. Sci. Inventive Mater. (ICTMIM), 2026, pp. 33-38.
[80] P. Chinnasamy, D. Tejaswini, R. K. Ayyasamy, S. Dhanasekaran, B. S. Kumar, and L. Chandran, "Crop optimization and disease detection using satellite imagery and artificial intelligence," in Proc. 2nd Int. Conf. Intell. Cyber Phys. Syst. Internet Things (ICoICI), 2024, pp. 1531-1535.
[81] S. S. Rana, I. I. Rana, M. J. Khan, S. Habib, R. M. Saleem, H. M. Haroon, and H. Irfan, "Predicting and identifying land features utilising remote sensing satellite imagery and machine learning techniques," Kashf J. Multidiscipl. Res., vol. 1, no. 12, pp. 17-35, 2024.
[82] E. C. Tetila, B. B. Machado, G. Astolfi, N. A. de Souza Belete, W. P. Amorim, A. R. Roel, and H. Pistori, "Automatic recognition of soybean leaf diseases using UAV images and deep convolutional neural networks," IEEE Geosci. Remote Sens. Lett., vol. 17, no. 5, pp. 903-907, May 2020.
[83] S. Nath, S. K. Biswas, S. Majumdar, K. Kumar, and T. Acharjee, "Smart farming using deep learning: Detection of soybean leaf disease from UAV-captured images," in Proc. Int. Conf. Emerg. Syst. Intell. Comput. (ESIC), Feb. 2026, pp. 109-114.
[84] P. Bosilj, E. Aptoula, T. Duckett, and G. Cielniak, "Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture," J. Field Robot., vol. 37, no. 1, pp. 7-19, 2020.
[85] A. Milioto, P. Lottes, and C. Stachniss, "Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 2018, pp. 2229-2235.
[86] J. P. Goncalves, F. A. C. Pinto, D. M. Queiroz, F. M. M. Villar, J. G. A. Barbedo, and E. M. Del Ponte, "Deep learning architectures for semantic segmentation and automatic estimation of severity of foliar symptoms caused by diseases or pests," Biosyst. Eng., vol. 210, pp. 129-142, 2021.
[87] Y.-H. Park, S. H. Choi, Y.-J. Kwon, S.-W. Kwon, Y. J. Kang, and T.-H. Jun, "Detection of soybean insect pest and a forecasting platform using deep learning with unmanned ground vehicles," Agronomy, vol. 13, no. 2, Art. no. 477, 2023.
[88] D. P. Hughes and M. Salathe, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," arXiv:1511.08060, 2015.
[89] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep neural networks based recognition of plant diseases by leaf image classification," Comput. Intell. Neurosci., vol. 2016, Art. no. 3289801, 2016.
[90] E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," Comput. Electron. Agriculture, vol. 161, pp. 272-279, 2019.
[91] J. G. A. Barbedo, "Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification," Comput. Electron. Agriculture, vol. 153, pp. 46-53, 2018.

[92] J. G. A. Barbedo, "Plant disease identification from individual lesions and spots using deep learning," Biosyst. Eng., vol. 180, pp. 96-107, 2019.
[93] A. Ramcharan, K. Baranowski, P. McCloskey, B. Ahmed, J. Legg, and D. P. Hughes, "Deep learning for image-based cassava disease detection," Frontiers Plant Sci., vol. 8, Art. no. 1852, 2017.
[94] A. Picon, A. Alvarez-Gila, M. Seitz, A. Ortiz-Barredo, J. Echazarra, and A. Johannes, "Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild," Comput. Electron. Agriculture, vol. 161, pp. 280-290, 2019.
[95] R. Thapa, K. Zhang, N. Snavely, S. Belongie, and A. Khan, "The Plant Pathology Challenge 2020 data set to classify foliar disease of apples," Appl. Plant Sci., vol. 8, no. 9, Art. no. e11390, 2020.
[96] S. Wallelign, M. Polceanu, and C. Buche, "Soybean plant disease identification using convolutional neural network," in Proc. 31st Int. Florida Artif. Intell. Res. Soc. Conf. (FLAIRS), 2018, pp. 146-151.
[97] A. Karlekar and A. Seal, "SoyNet: Soybean leaf diseases classification," Comput. Electron. Agriculture, vol. 172, Art. no. 105342, 2020.
[98] N. Bevers, E. J. Sikora, and N. B. Hardy, "Soybean disease identification using original field images and transfer learning with convolutional neural networks," Comput. Electron. Agriculture, vol. 203, Art. no. 107449, 2022.
[99] K. Nagasubramanian, S. Jones, A. K. Singh, S. Sarkar, A. Singh, and B. Ganapathysubramanian, "Plant disease identification using explainable 3D deep learning on hyperspectral images," Plant Methods, vol. 15, Art. no. 98, 2019.
[100] S. Ghosal, D. Blystone, A. K. Singh, B. Ganapathysubramanian, A. Singh, and S. Sarkar, "An explainable deep machine vision framework for plant stress phenotyping," Proc. Nat. Acad. Sci., vol. 115, no. 18, pp. 4613-4618, 2018.
[101] A. Singh, B. Ganapathysubramanian, A. K. Singh, and S. Sarkar, "Machine learning for high-throughput stress phenotyping in plants," Trends Plant Sci., vol. 21, no. 2, pp. 110-124, 2016.
[102] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.
[103] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770-778.
[104] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700-4708.
[105] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 1251-1258.
[106] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 4510-4520.
[107] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1314-1324.
[108] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105-6114.
[109] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 10096-10106.
[110] R. Yadav, A. Seth, S. Kolhe, S. Kumar, B. Dupare, and M. Kumbhkar, "AgriNet: A compact deep learning pipeline for segmented soybean leaf disease detection in field environments," Procedia Comput. Sci., vol. 283, pp. 3049-3062, 2026.
[111] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132-7141.
[112] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3-19.
[113] W. Liu, G. Wu, H. Wang, and F. Ren, "Lightweight pyramidal multi-scale attention residual network for plant disease recognition," IEEE Access, vol. 13, pp. 61526-61536, 2025.
[114] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998-6008.
[115] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
[116] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10012-10022.
[117] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 11976-11986.

[118] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2014, pp. 2672-2680.
[119] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020, pp. 6840-6851.
[120] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2018.
[121] N. Farah, N. Drack, H. Dawel, and R. Buettner, "A deep learning-based approach for the detection of infested soybean leaves," IEEE Access, vol. 11, pp. 99670-99679, 2023.
[122] B. Wang, Y. Gao, X. Yuan, and S. Xiong, "Local R-symmetry co-occurrence: Characterising leaf image patterns for identifying cultivars," IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 19, no. 2, pp. 1018-1031, Mar./Apr. 2022.
[123] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Med. Image Comput. Comput.-Assist. Interv. (MICCAI), 2015, pp. 234-241.
[124] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., "Segment anything," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 4015-4026.
[125] B. H. Kim and S. H. Seo, "Soybean leaf disease identification through smart detection using machine learning-convolutional neural network model," Legume Res., vol. 48, no. 6, pp. 1043-1050, 2026.
[126] A. Alnuaim, A. Altheneyan, and A. A. AlZubi, "Early leaf disease detection of soybean plants using convolution neural network algorithm," Legume Res., vol. 48, no. 6, pp. 1025-1034, 2026.
[127] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618-626.
[128] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765-4774.
[129] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2016, pp. 1135-1144.
[130] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," Nature Mach. Intell., vol. 2, pp. 665-673, 2020.
[131] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2011, pp. 1521-1528.
[132] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 6402-6413.
[133] C. Nadeau and Y. Bengio, "Inference for the generalization error," Mach. Learn., vol. 52, no. 3, pp. 239-281, 2003.
[134] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Comput., vol. 10, no. 7, pp. 1895-1923, 1998.
[135] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," J. Mach. Learn. Res., vol. 7, pp. 1-30, 2006.
[136] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," J. Mach. Learn. Res., vol. 17, no. 59, pp. 1-35, 2016.
[137] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in Proc. Int. Conf. Mach. Learn. (ICML), 2020, pp. 1597-1607.
[138] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4077-4087.
[139] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2017, pp. 1321-1330.
[140] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in Proc. Int. Conf. Learn. Represent. (ICLR), 2016.
[141] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 2704-2713.
[142] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, 2015.
[143] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. Int. Conf. Artif. Intell. Statist. (AISTATS), 2017, pp. 1273-1282.