Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Risk-Calibrated Conformal Prediction and Cost-Sensitive Spray Decision Support for Cross-Crop Foliar Disease Diagnosis

Innocent Nyalala

innocent@iitmz.ac.in

Indian Institute of Technology Madras, Zanzibar

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11847600/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Risk-Calibrated Conformal Prediction and Cost-Sensitive Spray Decision Support for Cross-Crop Foliar Disease Diagnosis

Innocent Nyalala[1,2*]

[1*]Indian Institute of Technology Madras Zanzibar, Zanzibar, Tanzania.
[2]Wadhwani School of Data Science and Artificial Intelligence, Indian Institute of Technology Madras, Chennai, India.

Corresponding author(s). E-mail(s): innocent@iitmz.ac.in;

Abstract

Background: Deep crop disease classifiers reach high accuracy on curated laboratory benchmarks, yet automated spray decisions require calibrated confidence estimates and cost-aware triage under field imaging shift. Standard top-1 argmax predictions ignore asymmetric yield losses and expose underrepresented smallholder crops and unfamiliar plant species to silent miscalibration.

Results: Across eight vision models spanning 79,782 to 134,416,230 parameters evaluated on a deduplicated 35,420-image eight-crop testbed, six physical corruptions, 2,226 wild-field PlantDoc images, and 696 open-set unseen-crop images, laboratory accuracy failed to predict field safety. Although in-distribution expected calibration error remained low (0.58% to 1.29%), 32.19% to 75.61% of test errors carried confidence of at least 0.90, and uncalibrated field calibration error rose to 41.53%–70.68%. Under pooled conformal calibration, coverage dropped to 50.00%–77.78% on minority rice; crop-conditional calibration restored worst-crop coverage to 90.92%–93.15% at 0.952 to 0.966 classes per set. Zero-shot laboratory thresholds collapsed on PlantDoc (6.29% to 19.90% coverage), whereas calibrating on 200 field images restored 95% coverage with 8.97 classes for CropGuardNet-RawRGB against 30.10 to 34.65 classes for unconditioned networks. Standard out-of-distribution scores separated unseen field crops from laboratory test leaves (up to 0.9902 area under the receiver operating characteristic curve), but dropped to 0.4969–0.6301 when distinguishing

unseen field species from field in-domain leaves. Mapping 38 diseases onto four farm actions via Action-Consistent Conformal Triage reduced wild-field spray excess regret by 69.8% versus blanket fungicide and 66.9% versus naive argmax ($\boldsymbol{p = 3.89 \times 10^{-18}}$), outperforming matched confidence deferral (431.24 versus 508.22 regret per 1,000 leaves) with 1.97% unnecessary spray and 0.02% missed infections.

Conclusions: Risk-calibrated action unanimity and few-shot field temperature scaling turn overconfident plant disease models into dependable agricultural decision support, whereas naive point predictions and laboratory-only checks fail silently under field shift.

Keywords: conformal prediction, confidence calibration, plant disease diagnosis, fungicide decision support, out-of-distribution detection, domain shift, precision agriculture

1 Background

Crop pathogens destroy an estimated 20% to 40% of annual global agricultural production, threatening farm livelihoods across smallholder and commercial cropping systems [1, 2]. Climate variability and shifting pathogen ranges intensify outbreak pressure just as regulatory and resistance constraints tighten chemical pesticide use [3, 4]. Field meta-analyses show that targeted decision support systems (DSS) can halve fungicide applications relative to calendar-based blanket spraying without increasing disease risk [5]. Deep convolutional neural networks (CNNs) and hybrid attention models have therefore attracted wide attention as automated front-end diagnostic engines capable of identifying foliar pathology rapidly from proximal leaf photographs [6–9].

Despite high headline classification accuracies above 97% on curated laboratory benchmarks [6, 10, 11], three downstream barriers prevent benchmark accuracy from translating into safe on-farm spray recommendations [12–15]. First, modern deep neural networks exhibit severe probability miscalibration and produce high-confidence errors when imaging conditions depart from training distributions [16, 17]. Under field illumination shifts, sensor noise, background clutter, weakened foliar lesion signals, motion blur, or unseen crop species [18–20], an overconfident point prediction

(arg max) can trigger needless chemical spraying on healthy leaves or miss destructive infections entirely. Second, standard split conformal prediction guarantees only marginal coverage across pooled datasets [21–24], masking severe under-coverage on underrepresented smallholder crops and failing immediately when laboratory calibration thresholds face covariate shift in natural field scenes [25, 26]. Third, existing agricultural uncertainty studies stop at per-class set sizes without mapping pathogen uncertainty onto asymmetric economic consequences [27, 28]. In field agronomy, mistakenly withholding intervention on late blight or bacterial wilt incurs far greater yield loss than an unnecessary preventive spray, and applying a fungicide to a bacterial or viral infection incurs both chemical waste and unchecked crop loss.

Recent methodological advances address pieces of this reliability puzzle in isolation. On the calibration front, temperature scaling [16] rescales logits with a single scalar parameter, whereas Expected Calibration Error (ECE) [29], Adaptive Calibration Error (ACE) [30], and Brier scores [31] quantify confidence reliability across probability bins. For out-of-distribution (OOD) detection, maximum softmax probabilities (MSP) [32] and free-energy scoring [33] separate unfamiliar inputs without retraining, and selective classification trades automated coverage against residual error risk [34]. Within plant pathology, Farag et al. [27] recently examined inductive conformal prediction for crop production decisions, confirming its promise as a distribution-free wrapper. Even so, no unified study has audited probability calibration, minority-crop conformal equity, few-shot field recalibration budgets, open-set crop species rejection, and asymmetric spray-regret decision support jointly across lightweight edge models and heavyweight CNN baselines.

To close this benchmark-to-decision gap, this study evaluates eight trained vision architectures spanning 79,782 to 134,416,230 parameters across a leakage-free eight-crop laboratory benchmark, six physical imaging corruptions, wild-field PlantDoc

imagery [18], and seven unseen crop species, and introduces Action-Consistent Conformal Triage (ACCT) for risk-bounded spray advisory. The primary contributions are summarized as follows:

1. We conduct a systematic multi-architecture calibration and high-confidence error audit across eight crop disease classifiers (spanning lightweight CNN–Transformer hybrids, mobile depthwise models, and deep residual networks) under clean evaluation, six physical imaging corruptions, and wild-field domain shift.
2. We quantify in-distribution conformal efficiency across Least Ambiguous Set-Valued Classifiers (LAC), Adaptive Prediction Sets (APS), Regularized Adaptive Prediction Sets (RAPS), and Crop-Conditional Mondrian conformal calibration to resolve under-coverage in data-scarce minority crops such as rice.
3. We measure zero-shot conformal failure under wild-field PlantDoc shift alongside few-shot field recalibration curves ($n_{\text{cal}} = 50$–$1{,}200$) and CLAHE-preprocessed against raw RGB image pipelines under both illumination and noise perturbations.
4. We evaluate four training-free open-set OOD detectors (MSP, negative predictive entropy, energy score, and conformal p-value) for flagging seven unseen host crop species outside the eight-crop training taxonomy.
5. We formulate ACCT by projecting 38 disease classes onto four management actions (no spray, fungicide, bactericide, and vector control or roguing) and benchmarking automated versus human-deferred triage against blanket spraying, naive $\arg\max$, and matched-deferral confidence thresholding under asymmetric yield-loss regret.

2 Methods

2.1 Study design and leakage-free multi-crop benchmarks

The experimental pipeline evaluates fixed, independently trained disease classifiers across three diagnostic regimes: curated laboratory-to-smallholder benchmarks, physical camera corruptions, and unconstrained wild-field imagery (Fig. 1). All eight-crop in-distribution models were evaluated on a deduplicated 35,420-image benchmark spanning 38 mutually disjoint classes across eight host crops curated from four public sources: five crops (23 classes) from the PlantVillage repository [35, 36], Mango (8 classes) from MangoLeafBD [37], Corn/Maize (4 classes) [38], and Rice (3 classes) [39]. Before data splitting, grouping by specimen Universally Unique Identifier (UUID) removed 80,238 offline-augmented duplicate copies from the merged PlantVillage training and validation directories so that every physical leaf specimen resided strictly in a single partition. Stratified partitioning with fixed random seed 42 yielded 24,793 training images (70.0%), 5,313 calibration/validation images (15.0%, denoted ID-Val), and 5,314 held-out test images (15.0%, denoted ID-Test), including only 18 test samples in the minority Rice cohort.

External wild-field domain shift and open-set species rejection were evaluated using the full PlantDoc repository [18], which consists of real-world field photographs collected under uncontrolled foliage backgrounds and natural illumination. Exactly 21 PlantDoc disease and healthy folders map biologically onto the 38-class taxonomy across six overlapping host species (Tomato, Grape, Apple, Potato, Bell Pepper, and Corn/Maize), comprising 2,033 canonical training images and 193 canonical test images (2,226 wild-field ID-overlapping images in total, denoted PlantDoc-ID; Additional file 1). Seven remaining PlantDoc crop categories representing species absent

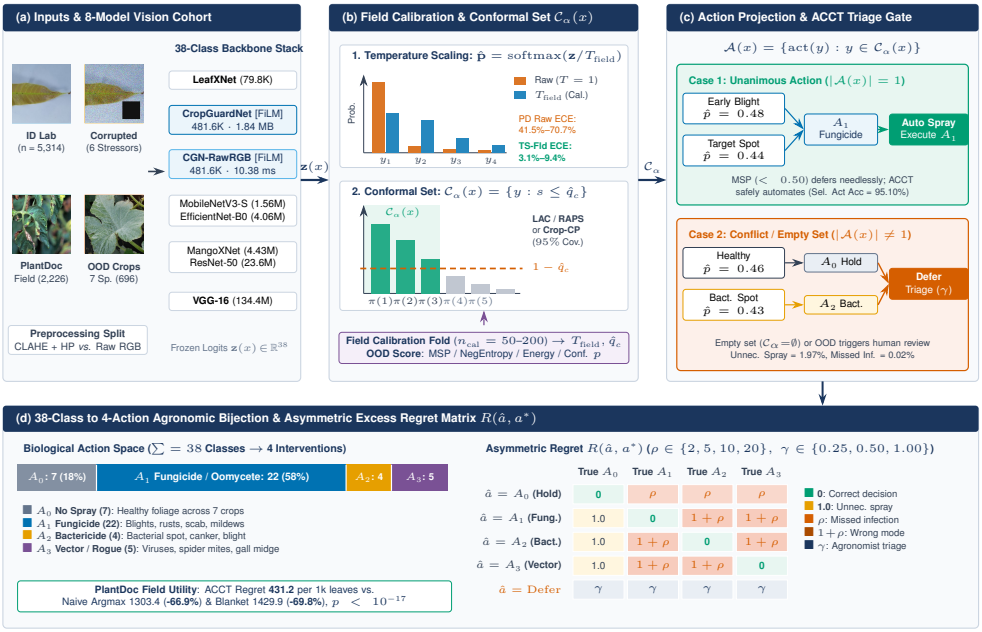

Fig. 1 Multi-panel Action-Consistent Conformal Triage workflow across diagnostic regimes and management actions. Panels (a)–(d) link leaf vision logits and few-shot field calibration to unanimous automation versus agronomist deferral and asymmetric excess regret.

from training (Blueberry, Cherry, Peach, Raspberry, Soyabean, Squash, and Strawberry, totaling 637 training plus 59 test = 696 images) form a strictly open-set evaluation cohort (PlantDoc-OOD).

2.2 Evaluated model cohort and preprocessing pipelines

To examine whether parameter efficiency or architectural family governs uncertainty quality, eight pre-trained 38-class architectures spanning three orders of magnitude in capacity were benchmarked without retraining (Table 1):

1. *CropGuardNet* (481,574 parameters; 1.84 MB FP32): a half-width MobileNetV3-Small ($\alpha = 0.50$) stem coupled with a three-branch dilated multi-scale pyramid, depthwise separable refinement, squeeze-and-excitation attention, a two-head Transformer encoder, dual-path pooling, and Feature-wise Linear Modulation (FiLM) crop conditioning [40, 41].

2. *CropGuardNet-RawRGB* (481,574 parameters; 1.84 MB FP32): an identical architecture trained directly on unpreprocessed 224 × 224 RGB images to isolate preprocessing effects on calibration and shift stability.
3. *LeafXNet* (79,782 parameters; 0.30 MB FP32): an ultra-lightweight four-stage CNN with a multi-head self-attention block [10].
4. *MobileNetV3-Small* (1,556,806 parameters; 5.94 MB FP32): standard mobile depthwise architecture [40].
5. *EfficientNet-B0* (4,056,226 parameters; 15.47 MB FP32): compound-scaled mobile inverted bottleneck CNN [42].
6. *MangoXNet* (4,426,374 parameters; 16.89 MB FP32): full-width MobileNetV3-Small hybrid CNN–Transformer [11].
7. *ResNet-50* (23,585,894 parameters; 89.97 MB FP32): deep residual convolutional backbone [43].
8. *VGG-16* (134,416,230 parameters; 512.76 MB FP32): classic deep sequential convolutional network [44].

Seven models operated on a deterministic three-stage preprocessing pipeline applied after resizing to 224×224 pixels: Contrast Limited Adaptive Histogram Equalization (CLAHE; clip limit 2.0, 8×8 tile grid on the CIELAB luminance channel L^*) [45], 3×3 Gaussian smoothing ($\sigma = 0.5$), and unsharp high-pass sharpening ($\alpha_{\text{hp}} = 0.3$ over a 5×5 Gaussian baseline), followed by ImageNet channel normalization. CropGuardNet-RawRGB bypassed CLAHE, smoothing, and high-pass filtering and received ImageNet-normalized raw RGB tensors directly. For both crop-conditioned networks on PlantDoc-OOD, logits were recorded under uniform random grower crop selection (seed 42) and across all eight crop slots ($696 \times 8 \times 38$) to expose worst-case overconfidence.

Table 1 Model cohort specifications and baseline accuracy across 38 foliar classes.

Model	Params	Size (MB)	Prep	ID Acc (%)	ID F1	PD Acc (%)	Corr-6 (%)
LeafXNet	79,782	0.30	CLAHE	97.25	0.9554	13.93	70.31
CropGuardNet	481,574	1.84	CLAHE	99.15	0.9873	45.87	**83.71**
CropGuardNet-RawRGB	481,574	1.84	Raw	**99.28**	**0.9883**	**48.61**	82.28
MobileNetV3-Small	1,556,806	5.94	CLAHE	99.23	0.9740	18.37	79.36
EfficientNet-B0	4,056,226	15.47	CLAHE	99.23	0.9835	21.74	82.08
MangoXNet	4,426,374	16.89	CLAHE	99.15	0.9746	18.19	78.59
ResNet-50	23,585,894	89.97	CLAHE	99.13	0.9751	22.28	79.24
VGG-16	134,416,230	512.76	CLAHE	97.97	0.9704	16.31	68.04

Note: Params denote trainable 32-bit weights; Prep indicates CLAHE plus Gaussian and high-pass filtering versus raw RGB; ID Acc and macro F1 cover ID-Test ($n = 5{,}314$); PD Acc covers PlantDoc-ID ($n = 2{,}226$); Corr-6 averages accuracy across six physical corruptions; bold values mark the best column result.

2.3 Physical imaging corruption protocol

To simulate realistic smartphone and sensor degradation without artificial post-preprocessing artifacts, six physical corruptions were injected onto clean 224×224 8-bit unsigned integer RGB images of ID-Test ($n = 5{,}314$) prior to CLAHE preprocessing and ImageNet normalization: (1) additive zero-mean Gaussian sensor noise ($\sigma = 0.12$ in $[0, 1]$ intensity space, clipped to $[0, 255]$, deterministic seed 42); (2) defocus blur via a 7×7 Gaussian kernel ($\sigma = 1.4$); (3) overexposure scaling pixel intensities by $\times 1.8$ (clipped to $[0, 255]$); (4) underexposure scaling pixel intensities by $\times 0.4$; (5) random rectangular foliage occlusion masking a 67×67-pixel (30% side length) black patch per image to approximate proximal foliage occlusion [46] (seed 43); and (6) lossy JPEG re-encoding at quality factor $Q = 15$.

2.4 Confidence calibration and high-confidence error metrics

Given pre-softmax logits $\mathbf{z}(x) \in \mathbb{R}^K$ ($K = 38$), post hoc temperature scaling [16] divides logits by a scalar $T > 0$ via $\hat{p}_y(x; T) = \exp(z_y(x)/T) / \sum_{j=1}^{K} \exp(z_j(x)/T)$, where T minimizes negative log-likelihood (NLL) on a calibration fold over $T \in [0.05, 20.0]$ using bounded Brent optimization. Two calibration regimes were evaluated:

laboratory temperature scaling (TS-Lab) fitted on ID-Val (n = 5,313), and few-shot field temperature scaling (TS-Field) fitted on field calibration partitions.

Reliability was quantified using five complementary metrics: 15-bin equal-width Expected Calibration Error (ECE) [29], 15-bin equal-mass Adaptive Calibration Error (ACE) [30], Maximum Calibration Error (MCE), NLL, and multi-class Brier score $\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(\hat{p}_k(x_i)-\mathbb{I}[y_i=k])^2$ [31]. Because uncalibrated agricultural systems fail most destructively when confident predictions are wrong, two error-conditional diagnostics were computed: mean confidence on errors ($\overline{\hat{p}_{\max} \mid \hat{y} \neq y}$) and high-confidence error rate ($\Pr(\hat{p}_{\max} \geq 0.90 \mid \hat{y} \neq y)$).

2.5 Split and crop-conditional conformal prediction

Let $\mathcal{D}_{\text{cal}} = \{(x_i, y_i, c_i)\}_{i=1}^{n}$ denote a labeled calibration split with host crop $c_i \in \{0, \ldots, 7\}$. Split conformal prediction constructs a prediction set $\mathcal{C}_\alpha(x) \subseteq \{1, \ldots, K\}$ satisfying $\Pr(Y \in \mathcal{C}_\alpha(X)) \geq 1-\alpha$ under exchangeability [21]. Three nonconformity score formulations were compared across $\alpha \in \{0.01, 0.05, 0.10\}$:

1. *Least Ambiguous Set-Valued Classifiers (LAC)* [22]: defines $s_{\text{LAC}}(x, y) = 1 - \hat{p}_y(x)$ and forms $\mathcal{C}_\alpha(x) = \{y : 1 - \hat{p}_y(x) \leq \hat{q}\}$, where $\hat{q}$ is the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile of calibration scores.
2. *Adaptive Prediction Sets (APS)* [23]: orders probabilities descending $\hat{p}_{\pi(1)} \geq \cdots \geq \hat{p}_{\pi(K)}$ and scores candidate y at rank $r(y)$ with uniform tie-breaking $u \sim \text{Uniform}(0, 1)$ via Eq. (1):

$$V_{\text{RAPS}}(x, y, u; \lambda, k_{\text{reg}}) = \sum_{j=1}^{r(y)-1} \hat{p}_{\pi(j)}(x) + u\,\hat{p}_y(x) + \lambda \max(r(y) - k_{\text{reg}},\, 0)\,, \qquad (1)$$

   where $\lambda = 0$ yields standard APS and $\lambda > 0$ yields Regularized Adaptive Prediction Sets (RAPS) [24] (evaluated at $k_{\text{reg}} = 2$, $\lambda = 0.05$, plus a sensitivity sweep over $\lambda \in \{0.01, 0.05, 0.10\}$ and $k_{\text{reg}} \in \{1, 2, 5\}$).

3. *Crop-Conditional Mondrian Conformal Prediction (Crop-CP)* [25]: partitions $\mathcal{D}_{\text{cal}}$ by host species c and computes species-specific quantiles $\hat{q}_c$ at finite-sample level $\lceil (n_c+1)(1-\alpha) \rceil / n_c$ to guarantee balanced coverage across rare crops such as Rice ($n_{\text{val}} = 18$). When $\lceil (n_c+1)(1-\alpha) \rceil > n_c$, both the exact conservative threshold ($\hat{q}_c = 1.0$) and empirical maximum fallback ($\max_{i:c_i=c} s_i$) are reported.

Under PlantDoc wild-field shift, both zero-shot transfer of laboratory quantiles (ID-Val $\rightarrow$ PlantDoc-ID) and field recalibration across canonical splits (2,033 $\rightarrow$ 193) and 100 stratified Monte Carlo field calibration folds ($n_{\text{cal}} \in \{50, 100, 200, 400, 800, 1{,}200\}$) were evaluated.

2.6 Open-set out-of-distribution detection

Four training-free uncertainty scores were evaluated for rejecting the 696 open-set leaves in PlantDoc-OOD against both ID-Test (n = 5,314) and PlantDoc-Test (n = 193): (1) Maximum Softmax Probability, $S_{\text{MSP}}(x) = \max_y \hat{p}_y(x;T)$ [32]; (2) negative Shannon entropy, $S_{\text{Ent}}(x) = \sum_{y=1}^{K} \hat{p}_y(x;T) \log \hat{p}_y(x;T)$; (3) free-energy score, $S_{\text{Energy}}(x) = \log \sum_{y=1}^{K} \exp(z_y(x))$ [33]; and (4) maximum conformal p-value, $S_{\text{ConfP}}(x) = \left(1 + |\{i \in \mathcal{D}_{\text{cal}} : s_{\text{LAC}}(x_i, y_i) \geq 1 - \max_y \hat{p}_y(x;T)\}|\right) / (n_{\text{cal}} + 1)$. OOD separation was measured using Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the Precision–Recall curve (AUPR, treating unseen species as positive), and OOD false-acceptance rate at 95% in-distribution recall (FPR@95%ID-TPR), with 1,000-resample bootstrap 95% confidence intervals (CIs).

2.7 Agronomic action mapping and asymmetric regret decision support

To evaluate downstream intervention consequences, all 38 foliar classes were mapped bijectively onto four mutually exclusive farm management actions $\mathcal{M} = \{0, 1, 2, 3\}$ (full class mapping in Additional file 1): Action 0 (No Treatment / Healthy, 7 healthy

classes), Action 1 (Fungicide / Oomycete Control, 22 fungal and oomycete diseases), Action 2 (Bactericide / Copper Sanitation, 4 bacterial diseases), and Action 3 (Vector Control / Insecticide / Roguing, 5 arthropod pest and viral classes).

Let $a^* = \text{act}(y) \in \mathcal{M}$ be the true required action and $\hat{a}$ the automated policy decision, with chemical intervention cost normalized to $C_{\text{spray}} = 1.0$. Excess regret $R(\hat{a}, a^*)$ relative to an omniscient oracle was defined as:

$$R(\hat{a}, a^*) = \begin{cases} 0, & \text{if } \hat{a} = a^*, \\ 1.0, & \text{if } \hat{a} \neq 0 \text{ and } a^* = 0 \text{ (unnecessary spray)}, \\ \rho, & \text{if } \hat{a} = 0 \text{ and } a^* \neq 0 \text{ (missed active infection)}, \\ 1.0 + \rho, & \text{if } \hat{a} \neq a^*,\ \hat{a} \neq 0,\ \text{and } a^* \neq 0 \text{ (wrong chemical mode)}, \\ \gamma, & \text{if } \hat{a} = \text{DEFER (agronomist verification)}, \end{cases} \tag{2}$$

evaluated across yield-loss severity ratios $\rho \in \{2, 5, 10, 20\}$ and triage overhead costs $\gamma \in \{0.25, 0.50, 1.00\}$ (primary reporting at $\rho = 5$, $\gamma = 0.50$).

Action-Consistent Conformal Triage (ACCT) constructs $\mathcal{C}_\alpha(x)$ ($\alpha \in \{0.05, 0.10, 0.15, 0.20\}$) and projects candidate labels onto unique management actions $\mathcal{A}(x) = \{\text{act}(y) : y \in \mathcal{C}_\alpha(x)\}$. When $|\mathcal{A}(x)| = 1$, ACCT executes that unanimous action automatically; when $|\mathcal{A}(x)| \neq 1$ (multi-action conflict or empty conformal set), ACCT defers the image to agronomist review. ACCT was benchmarked against blanket fungicide spraying, zero intervention, naive arg max AI, fixed MSP confidence thresholding ($\tau \in \{0.50, 0.70, 0.85, 0.90, 0.95\}$), action-sum probability thresholding, and an equal-workload MSP baseline calibrated to match ACCT's exact deferral rate on every split.

2.8 Statistical analysis

All metrics report macro-averaged class summaries (using present classes on the 21-class PlantDoc-ID subset) alongside Matthews correlation coefficient (MCC) where applicable [47]. Bootstrap 95% CIs used 1,000 stratified or uniform resamples (seed 42; 500 resamples for multi-regime ECE curves). Paired differences in excess regret across the 100 Monte Carlo field splits were tested using two-sided Wilcoxon signed-rank tests [48], and clean classification accuracy differences were compared via McNemar's test [49].

3 Results

Across the clean eight-crop ID-Test benchmark ($n = 5{,}314$), all eight evaluated architectures achieved 97.25% to 99.28% classification accuracy and 0.9554 to 0.9883 macro F1 score (Table 1). Single-thread CPU inference latency spanned nearly two orders of magnitude, from 10.14 ms for CropGuardNet, 10.38 ms for CropGuardNet-RawRGB, and 10.66 ms for MobileNetV3-Small to 128.11 ms for ResNet-50 and 893.46 ms for VGG-16. However, clean benchmark scores separated widely under domain shift: on wild-field PlantDoc-ID ($n = 2{,}226$), CropGuardNet-RawRGB achieved 48.61% accuracy and significantly exceeded CLAHE-preprocessed CropGuardNet at 45.87% (exact McNemar test: $n_{01} = 228$, $n_{10} = 167$, $p = 0.0025$), whereas all six unconditioned baselines dropped to 13.93% to 22.28%. Removing FiLM crop conditioning at inference collapsed PlantDoc-ID accuracy from 45.87% to 1.44% for CropGuardNet and from 48.61% to 8.18% for CropGuardNet-RawRGB. By contrast, on ID-Test, the accuracy difference between CropGuardNet-RawRGB (99.28%) and CropGuardNet (99.15%) was not statistically significant (McNemar $p = 0.1892$).

3.1 Confidence calibration and high-confidence error collapse

Raw probability outputs appeared well calibrated on ID-Test, yielding 0.58% to 1.29% ECE across all eight models, yet masking extreme overconfidence on misclassified leaves (Table 2 and Fig. 2). Among ID-Test errors, 32.19% (LeafXNet) to 75.61% (MobileNetV3-Small) carried $\hat{p}_{\max} \geq 0.90$, with mean error confidences between 77.84% and 92.07%. Applying TS-Lab ($T = 1.454$ to 2.334) reduced ID-Test ECE to 0.28% to 0.42% and lowered average corruption ECE from 12.72%–26.88% to 7.74%–22.01%, leaving ID-Test HCE at 15.07% to 57.78%.

Under wild-field PlantDoc shift, confidence calibration collapsed across every uncalibrated network, raising PlantDoc-All ECE to 41.53% (CropGuardNet-RawRGB, 95% CI [39.75%, 43.62%]), 45.90% (CropGuardNet, [43.97%, 47.96%]), and 62.48% to 70.68% across unconditioned baselines, with HCE reaching 49.60% to 67.55%. Laboratory temperature scaling failed to transfer to field imagery, leaving PlantDoc-All ECE at 32.93% to 63.47% and HCE at 14.58% to 40.00%. Fitting TS-Field on 200 stratified PlantDoc images reduced PlantDoc ECE to 3.13% to 9.37% and cut HCE to 0.00% to 3.34%. Temperature diagnostics distinguished genuine probability ordering from confidence flattening: LeafXNet and VGG-16 saturated at the upper optimizer bound $T = 20.000$ (low ECE of 3.13% and 4.23% at flat NLLs of 3.420 and 3.658), whereas CropGuardNet-RawRGB ($T = 3.620$, NLL = 1.670, ECE = $8.35\% \pm 1.70\%$) and CropGuardNet ($T = 4.089$, NLL = 1.791, ECE = $9.37\% \pm 2.24\%$) preserved discriminative class probabilities.

3.2 In-distribution conformal efficiency and minority-crop equity

Under TS-Lab calibration at $\alpha = 0.05$, every model tracked the distribution-free 95% coverage target closely (93.87% to 94.81% marginal LAC coverage; Table 3 and Fig. 3), with mean set sizes near 0.946 to 0.949 classes (5.14% to 5.42% empty

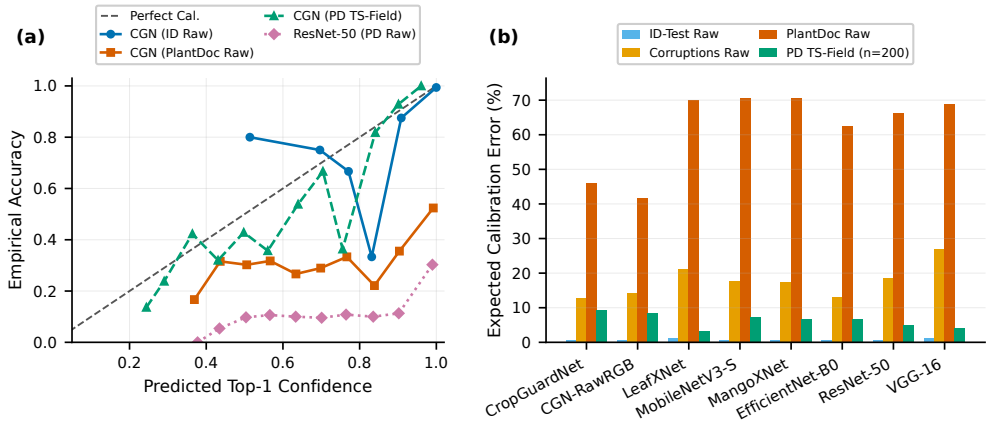

Fig. 2 Calibration reliability and expected calibration error across domain shifts. Panel (a) shows reliability curves before and after field temperature scaling; panel (b) compares ECE across eight models.

Table 2 Expected calibration error and high-confidence error rates across test regimes.

Model	ID ECE (%)	Corr ECE (%)	PD Raw ECE (%)	PD TS-Fld ECE (%)	ID HCE (%)	PD HCE (%)
LeafXNet	1.15	21.11	70.13	**3.13**	**32.19**	51.46
CropGuardNet	0.70	**12.72**	45.90	9.37	73.33	67.55
CropGuardNet-RawRGB	0.60	14.34	**41.53**	8.35	73.68	59.18
MobileNetV3-Small	0.62	17.76	70.68	7.39	75.61	62.41
EfficientNet-B0	**0.58**	13.18	62.48	6.59	56.10	**49.60**
MangoXNet	0.66	17.42	70.63	6.83	64.44	61.94
ResNet-50	0.62	18.56	66.19	4.86	47.83	60.17
VGG-16	1.29	26.88	68.74	4.23	50.93	54.75

Note: ECE denotes 15-bin Expected Calibration Error; HCE denotes fraction of misclassified samples with $\hat{p}_{\max} \geq 0.90$; ID ECE, Corr ECE, PD Raw ECE, ID HCE, and PD HCE report uncalibrated outputs; PD TS-Fld reflects 100-split field temperature scaling ($n_{\text{cal}} = 200$); bold values mark the lowest error per column.

sets). Because split conformal prediction guarantees aggregate coverage independently of backbone capacity, clean-domain marginal coverage does not rank model quality; instead, Table 3 demonstrates two essential reliability distinctions. First, pooled marginal LAC masked severe species-level inequity across every architecture: Rice coverage ($n = 18$) collapsed to 50.00% on MobileNetV3-Small, 55.56% on ResNet-50, 61.11% on MangoXNet and VGG-16, 72.22% on CropGuardNet, and 77.78%

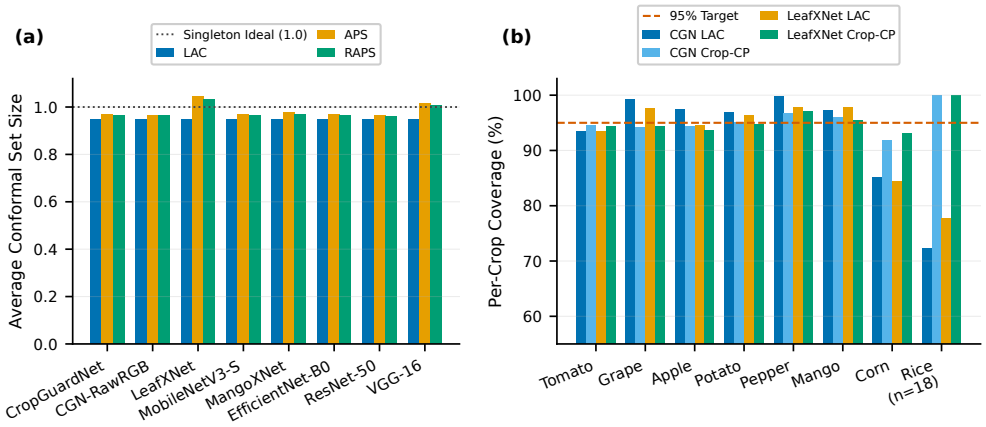

Fig. 3 In-distribution conformal set sizes and per-crop coverage equity ($\alpha = 0.05$). Panel (a) compares mean set size across scoring functions; panel (b) contrasts LAC against Crop-CP per species.

on CropGuardNet-RawRGB, LeafXNet, and EfficientNet-B0, whereas Corn coverage dropped to 79.30%–85.19%. Switching to RAPS ($k_{\text{reg}} = 2$, $\lambda = 0.05$) lifted Rice coverage to 88.89%–94.44% and worst-crop coverage to 88.89%–92.45% across $\lambda \in \{0.01, 0.05, 0.10\}$ and $k_{\text{reg}} \in \{1, 2, 5\}$, and Crop-Conditional Mondrian calibration (Crop-CP-Max) restored Rice coverage to 94.44%–100.00% and worst-crop coverage to 90.92%–93.15% without inflating mean set size (0.952 to 0.966 classes; Crop-CP-Exact reached 100.00% Rice coverage on all models at 1.074–1.085 classes). Second, true classifier sharpness separated under stricter risk tolerance ($\alpha = 0.01$) and downstream four-action accuracy: CropGuardNet and CropGuardNet-RawRGB produced compact high-assurance Crop-CP-Max sets of 1.010 and 1.017 classes with 99.98% action accuracy (matching ResNet-50 at 1.006 and 100.00%), whereas lower-accuracy LeafXNet (97.25% ID accuracy) inflated sets to 1.052–1.195 classes at $\alpha = 0.01$ with lower action accuracy (99.74%–99.78%, trailing every stronger model) and VGG-16 inflated sets to 1.042–1.110 classes.

Table 3 In-distribution conformal equity, high-assurance sharpness, and action accuracy across eight architectures.

Model	Method	Cov (.05)	Size (.05)	Size (.01)	ActAcc (%)	Worst (.05)	Rice (.05)
LeafXNet	LAC	93.87	0.947	1.052	99.74	77.78	77.78
	RAPS	94.67	1.034	1.195	99.78	88.89	88.89
	Crop-CP-Max	94.49	0.966	1.084	99.76	**93.15**	**100.00**
CropGuardNet	LAC	94.47	0.947	0.991	99.96	72.22	72.22
	RAPS	94.35	0.964	1.047	99.94	91.40	94.44
	Crop-CP-Max	94.47	0.952	1.010	99.98	**91.72**	**100.00**
CGN-RawRGB	LAC	94.58	0.946	0.992	100.00	77.78	77.78
	RAPS	94.39	0.963	1.041	99.94	**92.45**	**94.44**
	Crop-CP-Max	94.77	0.952	1.017	99.98	91.56	**94.44**
MobileNetV3-S	LAC	94.51	0.946	0.994	100.00	50.00	50.00
	RAPS	94.32	0.966	1.041	99.98	88.89	88.89
	Crop-CP-Max	94.69	0.954	1.011	100.00	**92.57**	**100.00**
EfficientNet-B0	LAC	94.81	0.949	0.988	100.00	77.78	77.78
	RAPS	94.51	0.965	1.032	99.96	**92.18**	94.44
	Crop-CP-Max	95.03	0.954	1.006	100.00	90.92	**100.00**
MangoXNet	LAC	94.51	0.946	0.993	99.98	61.11	61.11
	RAPS	94.43	0.968	1.052	99.96	92.45	**94.44**
	Crop-CP-Max	94.64	0.955	1.019	99.98	**92.68**	**94.44**
ResNet-50	LAC	94.54	0.946	0.989	100.00	55.56	55.56
	RAPS	94.34	0.962	1.027	99.94	**92.18**	94.44
	Crop-CP-Max	94.75	0.952	1.006	100.00	91.40	**100.00**
VGG-16	LAC	94.20	0.946	1.014	99.94	61.11	61.11
	RAPS	94.34	1.005	1.110	99.84	**92.18**	94.44
	Crop-CP-Max	94.22	0.952	1.042	99.94	91.72	**100.00**

Note: Evaluated on ID-Test (n = 5,314) with TS-Lab probabilities; CGN-RawRGB denotes CropGuardNet-RawRGB; Cov (.05), Worst (.05), and Rice (.05) report percentages targeting 95.0% coverage rather than maximization; Size (.05) and Size (.01) report mean set sizes at $\alpha = 0.05$ and $\alpha = 0.01$; ActAcc reports selective four-action accuracy at $\alpha = 0.05$; bold values mark the highest Worst-Crop and Rice equity within each architecture.

3.3 Wild-field PlantDoc failure and few-shot recalibration

Transferring TS-Lab conformal quantiles directly onto PlantDoc-ID collapsed coverage across every model (Fig. 4). Zero-shot LAC coverage dropped to 19.90% for CropGuardNet (72.64% empty sets), 16.62% for CropGuardNet-RawRGB (77.31% empty sets), 8.81% for LeafXNet (51.66% empty sets), and 6.29% to 8.63% across the remaining baselines (73.50% to 92.50% empty sets). Zero-shot RAPS prevented complete set emptiness (0.63% to 1.75% empty sets) yet reached only 63.61% coverage

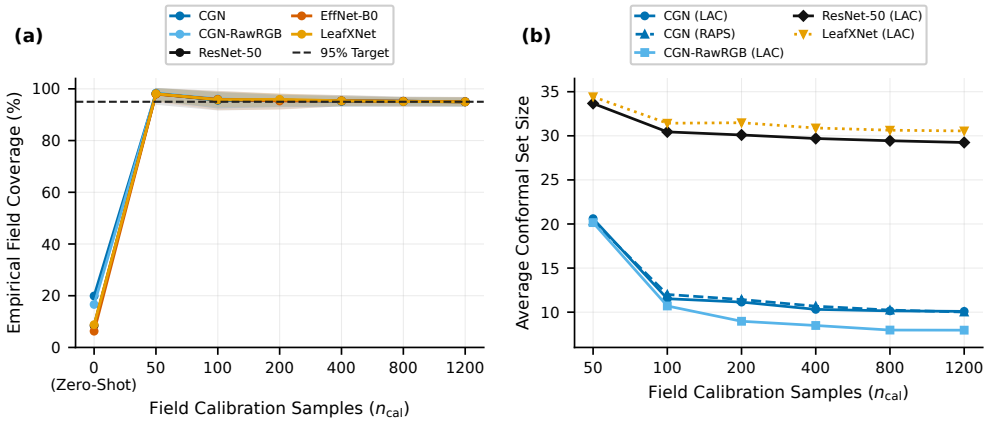

Fig. 4 Wild-field PlantDoc zero-shot coverage collapse and few-shot recalibration trajectories. Panel (a) shows mean coverage with 5th to 95th percentile intervals across 100 splits; panel (b) contrasts conformal set size.

for CropGuardNet-RawRGB, 60.56% for CropGuardNet, and 19.54% to 31.81% for unconditioned models.

Few-shot recalibration with field labels rapidly restored empirical coverage guarantees across 100 Monte Carlo splits (Fig. 4a). At $n_{\text{cal}} = 50$, TS-Field with LAC produced conservative coverage of 97.94% to 98.36%; by $n_{\text{cal}} = 100$, every architecture tracked the nominal 95% target closely (95.59% to 96.32%, SD 1.60% to 2.06%). However, conformal efficiency separated crop-conditioned models from standard architectures (Fig. 4b). At $n_{\text{cal}} = 200$, CropGuardNet-RawRGB converged to 8.97 ± 1.89 candidate classes (11.84% action-unanimous sets) and CropGuardNet to 11.14 ± 1.80 classes (11.07% action-unanimous), improving further at $n_{\text{cal}} = 1{,}200$ to 7.96 ± 0.59 and 10.09 ± 0.50 classes. Across all six unconditioned networks, restoring 95% coverage required 30.10 to 34.65 of 38 possible labels at $n_{\text{cal}} = 200$ (0.00% to 0.80% action-unanimous sets) and 29.24 to 34.43 labels at $n_{\text{cal}} = 1{,}200$. Canonical PlantDoc-Train ($n = 2{,}033$) to PlantDoc-Test ($n = 193$) evaluation replicated this split: CropGuardNet-RawRGB achieved 96.37% coverage with 7.89 classes and 13.99% action unanimity (96.30% selective action accuracy), whereas unconditioned networks yielded 29.21 to 34.71 classes with 0.00% to 1.55% action unanimity.

3.4 Physical corruption stress and unseen-crop OOD detection

Physical image corruptions exposed a distinct preprocessing trade-off between CLAHE filtering and raw RGB inputs (Fig. 5a). Under additive Gaussian sensor noise ($\sigma = 0.12$), CLAHE plus Gaussian smoothing in CropGuardNet preserved 44.84% accuracy (35.27% ECE, 63.04% RAPS coverage, 1.94× set inflation), outperforming CropGuardNet-RawRGB (28.43% accuracy, 52.74% ECE, 44.79% RAPS coverage) and preventing the collapse observed in ResNet-50 (6.40% accuracy, 91.65% ECE, 95.34% HCE, 7.51% RAPS coverage) and VGG-16 (7.68% accuracy, 8.73% RAPS coverage). Under underexposure (×0.4), CropGuardNet also outperformed CropGuardNet-RawRGB (98.08% accuracy and 0.29% ECE versus 96.63% and 1.03%). Conversely, under lossy JPEG compression ($Q = 15$), CropGuardNet-RawRGB outperformed CropGuardNet (89.97% accuracy, 2.32% ECE, and 93.83% RAPS coverage versus 82.35%, 6.85%, and 90.05%) and edged ahead under defocus blur (85.62% versus 83.78%). Random 30% foliage occlusion preserved 94.81% to 95.20% RAPS coverage and 98.89% to 99.17% accuracy across seven models (94.45% accuracy for LeafXNet). Across all six perturbations, RAPS prevented empty-set drops that caused fixed LAC coverage to plunge to 1.56%–30.28% under sensor noise (98.21% empty sets for EfficientNet-B0 LAC).

For open-set OOD detection separating the 696 PlantDoc-OOD images (seven unseen crops) from laboratory ID-Test ($n = 5{,}314$), all four post hoc scores achieved strong separation on standard backbones, led by EfficientNet-B0 with negative entropy at 0.9902 AUROC (95% CI $[0.9871, 0.9930]$, 3.59% FPR@95%ID), followed by MangoXNet (0.9761), CropGuardNet-RawRGB (0.9760, 12.50% FPR@95%ID), MobileNetV3-Small (0.9749), and CropGuardNet (0.9705), whereas LeafXNet trailed at 0.9069 with 48.13% FPR@95%ID (Table 4 and Fig. 5b). However, two sharp safety limits appeared. First, querying FiLM crop-conditioned models with adversarial worst-case crop slots lowered ID-vs-OOD AUROC from 0.9211 (Energy) and 0.9699 (MSP)

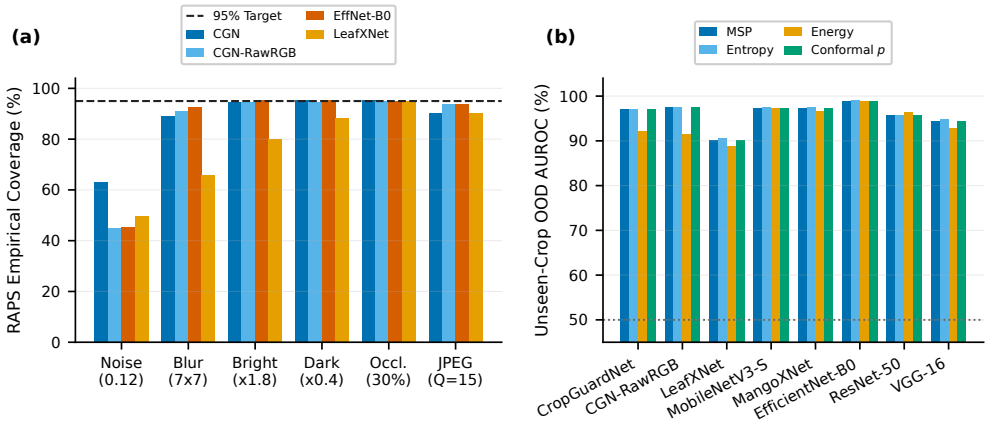

Fig. 5 Conformal coverage under physical imaging corruptions and unseen-crop OOD detection. Panel (a) compares RAPS coverage across six perturbations; panel (b) compares four OOD scores against ID-Test.

to 0.6970 and 0.8659 for CropGuardNet (96.84% Energy FPR@95%ID) and 0.6689 and 0.8822 for CropGuardNet-RawRGB (97.84% Energy FPR@95%ID). Second, when separating canonical wild-field PlantDoc-Test ($n = 193$) from wild-field PlantDoc-OOD ($n = 696$) under matched field imaging conditions, every detector across all eight architectures dropped near chance, spanning 0.4969 to 0.6301 AUROC with 88.36% to 96.12% FPR@95%ID.

3.5 Cost-sensitive spray decision support

Projecting 38 disease predictions onto four management actions transformed wild-field economic and chemical safety outcomes across 100 Monte Carlo splits ($n_{\text{cal}} = 200$, $\rho = 5$, $\gamma = 0.50$; Table 5 and Fig. 6). Calendar-style Blanket Fungicide achieved only 66.58% action accuracy, sprayed 100.00% of healthy leaves needlessly, selected wrong chemical modes on 21.92% of samples (missed bactericides and vector roguing), and incurred 1429.91 excess regret per 1,000 leaves; withholding all intervention incurred 4424.98 regret per 1,000 leaves. Unprotected Naive Argmax AI failed to rescue field utility, incurring 1272.26 ± 6.89 (CropGuardNet) to 1598.52 ± 9.90 (LeafXNet) regret

Table 4 Open-set out-of-distribution detection separating PlantDoc-OOD ($n = 696$) from in-distribution foliage.

Model	Detector	ID vs OOD AUROC	95% CI	FPR95 (%)	PD vs OOD AUROC
LeafXNet	NegEntropy	0.9069	[0.8957, 0.9179]	48.13	0.5279
	Energy	0.8890	[0.8758, 0.9024]	40.66	0.5137
CropGuardNet	NegEntropy	0.9705	[0.9638, 0.9768]	12.79	0.6192
	Energy	0.9211	[0.9093, 0.9317]	29.89	0.5700
CGN-RawRGB	NegEntropy	0.9760	[0.9714, 0.9804]	12.50	**0.6301**
	Energy	0.9145	[0.9027, 0.9257]	33.19	0.5822
MobileNetV3-S	NegEntropy	0.9749	[0.9686, 0.9803]	9.91	0.5368
	Energy	0.9728	[0.9660, 0.9790]	10.20	0.5228
EfficientNet-B0	NegEntropy	**0.9902**	[0.9871, 0.9930]	**3.59**	0.5558
	Energy	0.9899	[0.9859, 0.9932]	**3.59**	0.5278
MangoXNet	NegEntropy	0.9761	[0.9707, 0.9810]	10.49	0.5093
	Energy	0.9672	[0.9608, 0.9732]	13.22	0.4978
ResNet-50	NegEntropy	0.9586	[0.9511, 0.9653]	19.40	0.5939
	Energy	0.9639	[0.9566, 0.9701]	16.67	0.5155
VGG-16	NegEntropy	0.9479	[0.9394, 0.9558]	21.41	0.5008
	Energy	0.9276	[0.9182, 0.9369]	36.21	0.5299
CGN (Worst-Slot)	MSP	0.8659	[0.8524, 0.8790]	63.94	–
	Energy	0.6970	[0.6779, 0.7142]	96.84	–
CGN-Raw (Worst-Slot)	MSP	0.8822	[0.8713, 0.8931]	64.66	–
	Energy	0.6689	[0.6496, 0.6855]	97.84	–

Note: CGN denotes CropGuardNet; ID vs OOD compares ID-Test (n = 5,314) against PlantDoc-OOD (n = 696); PD vs OOD compares canonical PlantDoc-Test (n = 193) against PlantDoc-OOD; FPR95 denotes FPR@95%ID; Worst-Slot takes the maximum overconfidence logit across all eight FiLM crop slots per OOD image; bold values mark strongest column separation.

per 1,000 leaves, spraying 41.19% to 81.67% of healthy leaves needlessly and missing 1.62% to 7.07% of active infections.

Action-Consistent Conformal Triage lowered excess regret sharply and suppressed chemical waste and missed disease (Table 5). CropGuardNet-RawRGB with ACCT-LAC ($\alpha = 0.10$) achieved the lowest overall regret at 431.24 ± 5.80 per 1,000 leaves, representing a 69.8% reduction relative to Blanket Fungicide and a 66.9% reduction relative to Naive Argmax AI ($p = 3.89 \times 10^{-18}$). At identical agronomist deferral workload (71.50%), ACCT-LAC outperformed Matched MSP (508.22 ± 12.68, Wilcoxon $p = 3.89 \times 10^{-18}$), lifting selective action accuracy from 89.51% to 95.10%, cutting unnecessary spraying on healthy leaves from 3.79% to 1.97% (versus 41.19% for naive

arg max), and reducing missed active infections from 1.01% to 0.02%. Similarly, CropGuardNet ACCT-LAC ($\alpha = 0.10$) achieved 458.73 ± 11.59 regret versus 506.73 ± 25.71 for Matched MSP ($p = 3.89 \times 10^{-18}$), with 93.54% selective accuracy and 0.00% missed infections. Because all six unconditioned baselines produced diffuse conformal sets under PlantDoc shift, they deferred 98.51% to 99.99% of leaves under ACCT-LAC ($\alpha = 0.10$), yielding near-constant regret (492.56 to 504.18) anchored near $\gamma = 0.50$.

Sensitivity sweeps across $\rho \in \{2, 5, 10, 20\}$ and $\gamma \in \{0.25, 0.50, 1.00\}$ verified consistent ACCT dominance on wild-field PlantDoc (Fig. 6a–b). Under $\gamma = 0.50$, CropGuardNet-RawRGB ACCT-LAC ($\alpha = 0.10$) spanned 395.42 ($\rho = 2$) to 610.34 ($\rho = 20$) regret per 1,000 leaves, compared against 652.57 to 4557.48 for Naive Argmax and 772.46 to 4717.18 for Blanket Fungicide; varying triage cost γ from 0.25 to 1.00 at $\rho = 5$ yielded 252.48 to 788.76 regret. On canonical PlantDoc-Test ($n = 193$), CropGuardNet-RawRGB ACCT-LAC ($\alpha = 0.10$) similarly cut regret to 455.96 versus 1357.51 (Naive Argmax) and 1528.50 (Blanket Fungicide), with 93.65% selective accuracy and 0.00% unnecessary spray. Under physical corruptions (Corr-Mean), ACCT lowered regret to 210.51 (EfficientNet-B0 ACCT-RAPS) and 222.46 (CropGuardNet ACCT-LAC at $\alpha = 0.05$) compared against 331.45 and 383.67 for Naive Argmax. On clean ID-Test where baseline action accuracy already exceeded 99.85%, Naive Argmax achieved 3.58 to 6.21 regret per 1,000 leaves, whereas mandatory conformal deferral ($\alpha = 0.10$) incurred 25.69 to 57.11 triage overhead.

4 Discussion

High benchmark accuracy on curated leaf repositories (97.25% to 99.28%) neither guarantees calibrated probabilities nor protects farmers against costly spray errors under field shift. Only when crop-conditioned representations were paired with few-shot field calibration and ACCT did automated diagnostics become economically

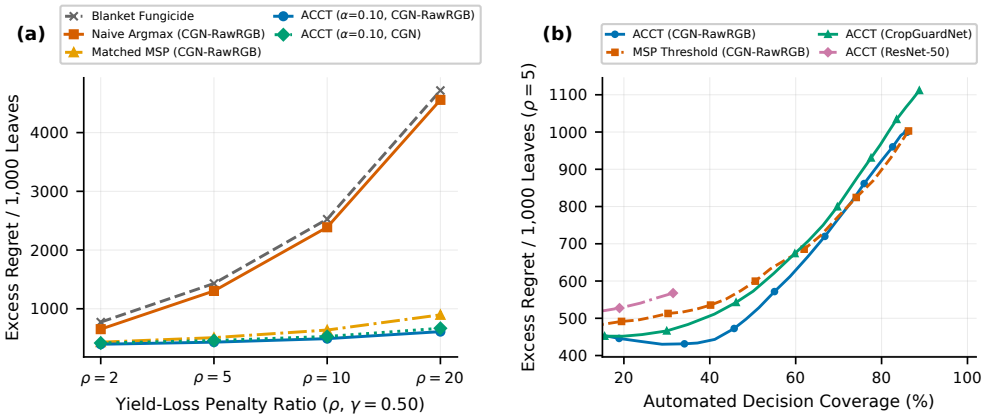

Fig. 6 Agronomic excess regret and risk-coverage trade-offs on wild-field PlantDoc imagery. Panel (a) sweeps yield-loss penalty ρ; panel (b) traces automated decision coverage against excess regret.

Table 5 Cost-sensitive spray decision outcomes on wild-field PlantDoc ($\rho = 5$, $\gamma = 0.50$).

Model	Policy	Defer (%)	Sel. Act Acc (%)	Unnec. Spray (%)	Missed Inf. (%)	Regret / 1k
System Baselines	Blanket Fungicide	0.00	66.58	100.00	**0.00**	1429.91
	No Intervention	0.00	11.50	**0.00**	100.00	4424.98
CGN-RawRGB	Naive Argmax AI	0.00	73.57	41.19	5.16	1303.39
	Matched MSP ($\alpha = 0.10$)	71.50	89.51	3.79	1.01	508.22
	ACCT-LAC ($\alpha = 0.10$)	71.50	95.10	1.97	0.02	**431.24**
	ACCT-LAC ($\alpha = 0.15$)	60.81	93.40	3.97	0.15	437.40
	ACCT-RAPS ($\alpha = 0.10$)	84.27	95.33	1.45	**0.00**	458.32
CropGuardNet	Naive Argmax AI	0.00	72.98	56.53	2.70	1272.26
	Matched MSP ($\alpha = 0.10$)	74.39	89.98	4.11	0.16	506.73
	ACCT-LAC ($\alpha = 0.05$)	88.93	98.48	0.28	**0.00**	455.90
	ACCT-LAC ($\alpha = 0.10$)	74.39	93.54	2.83	**0.00**	458.73
EfficientNet-B0	Naive Argmax AI	0.00	70.11	60.96	3.62	1410.84
	ACCT-LAC ($\alpha = 0.10$)	98.51	**100.00**	**0.00**	**0.00**	492.56
ResNet-50	Naive Argmax AI	0.00	70.84	70.12	1.62	1331.96
	ACCT-LAC ($\alpha = 0.10$)	99.80	99.92	**0.00**	**0.00**	499.01
LeafXNet	Naive Argmax AI	0.00	64.53	81.19	7.07	1598.52
	ACCT-LAC ($\alpha = 0.10$)	99.96	–	**0.00**	**0.00**	499.78

Note: CGN-RawRGB denotes CropGuardNet-RawRGB; mean outcomes across 100 stratified field calibration splits ($n_{\text{cal}} = 200$) on PlantDoc-ID ($n = 2{,}226$); Unnec. Spray reports unnecessary applications on healthy leaves; Missed Inf. reports missed treatments on diseased leaves; LeafXNet ACCT-LAC automated fewer than 0.05% of samples; bold values mark strongest column performance.

viable on PlantDoc, cutting excess spray regret by 69.8% relative to calendar fungicide applications and 66.9% relative to naive arg max outputs ($p = 3.89 \times 10^{-18}$)

and cutting unnecessary spraying on healthy foliage from 100.00% and 41.19% down to 1.97%. Without conformal deferral and action-space projection, deep neural models either saturated healthy leaves with unnecessary fungicide or missed destructive bacterial and viral outbreaks.

Probability miscalibration under distribution shift explained why raw model confidence misleads downstream agronomists. Even on ID-Test, where scalar ECE measured 0.58% to 1.29%, wrong predictions consistently carried excessive confidence (32.19% to 75.61% HCE), confirming earlier observations that deep networks emit peaked logits away from decision boundaries [16, 17]. Because PlantDoc introduces clustered field clutter and uncontrolled solar lighting, laboratory temperatures ($T = 1.454$ to 2.334) failed to restrain wild-field overconfidence, leaving ECE above 32.9% across every architecture. Fitting TS-Field on 200 field images revealed an important distinction between logit compression and ranking preservation: LeafXNet and VGG-16 drove T up to the 20.000 optimization ceiling, flattening probabilities to obtain 3.13% and 4.23% ECE at NLL values above 3.4, whereas FiLM-conditioned CropGuardNet and CropGuardNet-RawRGB settled near $T = 3.62$–4.09 with NLLs of 1.670–1.791. Consequently, scalar ECE alone cannot verify field readiness without NLL and ranking efficiency checks.

Standard marginal conformal calibration also masked sharp fairness failures against minoritized smallholder crops. Under pooled ID-Test quantiles, LAC achieved near-nominal 93.87%–94.81% aggregate coverage only by tolerating 50.00% to 77.78% coverage on Rice ($n = 18$) and 79.30%–85.19% on Corn, confirming theoretical warnings about conditional under-coverage under sample imbalance [24, 25]. Applying RAPS mitigated this gap substantially, and Crop-CP-Max restored Rice coverage to 94.44%–100.00% and worst-crop coverage to 90.92%–93.15% with virtually zero inflation in mean set size (0.952 to 0.966). Under zero-shot lab-to-field shift, LAC quantiles trained on ID-Test emitted empty prediction sets on 72.64% to 92.50% of

PlantDoc leaves, demonstrating how violations of exchangeability silently break unadjusted conformal guarantees [26]. Recalibrating on just 100 to 200 locally collected leaves immediately recovered nominal 95% field coverage, offering extension services a realistic regional calibration protocol [27]. Even after recalibration, however, unconditioned baselines inflated prediction sets to 30.10–34.65 of 38 classes because they wasted probability mass across taxonomically irrelevant host crops.

Physical stress testing and open-set evaluation exposed two practical failure mechanisms with direct engineering implications. First, preprocessing pipelines traded off resistance across distinct degradation modes: CIELAB CLAHE normalization paired with Gaussian smoothing protected CropGuardNet against additive Gaussian sensor noise (44.84% versus 28.43% RawRGB accuracy and 6.40% ResNet-50 accuracy) and underexposure (98.08% versus 96.63%), because histogram redistribution stabilizes dark regions and low-pass smoothing suppresses pixel-wise noise. Conversely, unsharp high-pass sharpening amplified block boundaries under $Q = 15$ JPEG compression (82.35% versus 89.97% RawRGB accuracy) and exaggerated natural soil and background textures in PlantDoc (45.87% versus 48.61%, McNemar $p = 0.0025$), confirming why separating weak foliar lesion cues from cluttered field foliage remains a primary challenge in agricultural computer vision [19, 20]. Second, negative OOD findings warn against evaluating open-set detectors across mismatched imaging environments. Although negative entropy and energy reached 0.9705 to 0.9902 AUROC separating PlantDoc-OOD from laboratory ID-Test [32, 33], separating PlantDoc-Test from PlantDoc-OOD reduced every detector across all eight architectures to 0.4969–0.6301 AUROC (88.36%–96.12% FPR@95%ID). Thus, standard post hoc scores separated studio versus natural background cues rather than unseen botanical species. Feeding mis-selected host crop indices to FiLM conditioning also degraded CropGuardNet Energy AUROC to 0.6689–0.6970, showing that crop-conditioned mobile tools require an upstream host verification step or multi-slot agreement check.

Projecting conformal sets onto management actions explains why ACCT consistently beat confidence thresholding at identical human review budgets (431.24 versus 508.22 regret per 1,000 leaves for CropGuardNet-RawRGB, $p = 3.89 \times 10^{-18}$). Scalar MSP thresholding rejects every image whose top-1 probability falls below τ, even when competing disease candidates (such as tomato early blight and target spot) both dictate fungicide application (Action 1) and can be safely automated together. Conversely, when moderate top-1 confidence spans healthy foliage (Action 0) and bacterial spot (Action 2), scalar thresholding may spray erroneously or miss bacterial control, whereas ACCT checks action-set cardinality $|\mathcal{A}(x)|$ directly and triggers agronomist triage. On clean ID-Test where baseline error stays below 0.15%, Naive Argmax incurs only 3.58–6.21 regret per 1,000 leaves because mandatory α-level conformal deferral adds avoidable review overhead (γ). In real farms subject to field shift and camera noise, however, CropGuardNet-RawRGB executes in 10.38 ms on a single CPU core within a 1.84 MB weight budget, enabling offline smartphone execution that reserves cloud or extension-officer review strictly for conflicted action sets.

Several methodological boundaries shape how these results should be translated into agricultural practice. PlantDoc lacks matched weather logs, growth-stage metadata, and longitudinal incidence series, restricting wild-field testing to six overlapping crop species rather than a controlled multi-season trial. Fixed economic regret ratios ($\rho \in \{2, 5, 10, 20\}$, $\gamma \in \{0.25, 0.50, 1.00\}$) parameterize general chemical-versus-yield trade-offs rather than crop-specific input prices and smallholder labor constraints, and the single-label 38-class taxonomy cannot represent leaves with simultaneous fungal and bacterial coinfections. In parallel, the Rice validation cohort ($n = 18$) bounds finite-sample Mondrian precision in underrepresented crops. Future work will extend ACCT to multi-label conformal risk control on coinfected leaves [28], integrate regional fungicide pricing and extension triage records across multi-season field deployments, and train lightweight host-species rejection heads before FiLM conditioning.

5 Conclusions

Evaluating eight foliar disease classifiers across clean, corrupted, wild-field, and open-set regimes demonstrates that headline benchmark accuracy hides severe probability overconfidence, minority-crop conformal inequity, and brittle cross-domain transfer. Resolving these vulnerabilities requires pairing domain-aware field calibration with crop-conditional conformal sets whose ambiguity collapses cleanly onto actionable spray categories before chemical decisions reach growers.

When disease uncertainty is translated directly into unanimous intervention sets through Action-Consistent Conformal Triage, lightweight crop-conditioned edge networks reduce excess agronomic regret, eliminate wasteful preventive spraying on healthy crops, contain missed active infections, and route genuinely ambiguous leaves to human agronomists. Grounding agricultural computer vision in action-level uncertainty bounds and transparent deferral criteria offers a practical pathway toward trustworthy, resource-conserving crop protection under real-world field variability.

List of abbreviations

ACE: Adaptive Calibration Error; ACCT: Action-Consistent Conformal Triage; APS: Adaptive Prediction Sets; AUPR: Area Under the Precision–Recall Curve; AUROC: Area Under the Receiver Operating Characteristic Curve; CI: Confidence Interval; CLAHE: Contrast Limited Adaptive Histogram Equalization; CNN: Convolutional Neural Network; Crop-CP: Crop-Conditional Mondrian Conformal Prediction; DSS: Decision Support System; ECE: Expected Calibration Error; FiLM: Feature-wise Linear Modulation; FPR: False Positive Rate; HCE: High-Confidence Error Rate; LAC: Least Ambiguous Set-Valued Classifiers; MCC: Matthews Correlation Coefficient; MCE: Maximum Calibration Error; MSP: Maximum Softmax Probability; NLL: Negative Log-Likelihood; OOD: Out-of-Distribution; RAPS: Regularized Adaptive Prediction Sets; TS: Temperature Scaling; UUID: Universally Unique Identifier.

Declarations

Ethics approval and consent to participate

Not applicable. This study uses exclusively public, anonymized plant leaf imagery and involves no human participants, clinical records, or animal subjects.

Consent for publication

Not applicable.

Availability of data and materials

All datasets supporting the conclusions of this article are publicly available in the PlantVillage and New Plant Diseases repositories [35, 36], MangoLeafBD [37], Rice Leaf Diseases [39], Corn or Maize Leaf Disease Dataset [38], and PlantDoc [18], and the complete 38-class to four-action biological taxonomy mapping is included within this article and its additional file [see Additional file 1]. Analysis scripts, precomputed logit extraction manifests, and conformal evaluation code are available from the corresponding author upon reasonable request.

Competing interests

The author declares no competing interests.

Funding

This research received no external grant funding from public, commercial, or not-for-profit agencies.

Authors' contributions

Innocent Nyalala conceptualized the uncertainty quantification and agronomic decision framework, wrote the logit-extraction, conformal calibration, and regret analysis

code, performed all statistical analyses, generated all figures, and drafted, audited, and approved the final manuscript.

Acknowledgements

The author acknowledges the creators and curators of the open-access PlantVillage, MangoLeafBD, Rice Leaf, Corn/Maize Leaf, and PlantDoc agricultural imaging datasets for making reproducible plant pathology benchmarking possible.

Authors' information

Innocent Nyalala holds academic appointments at the Indian Institute of Technology Madras Zanzibar campus (Zanzibar, Tanzania) and the Wadhwani School of Data Science and Artificial Intelligence, Indian Institute of Technology Madras (Chennai, India).

Additional files

- Additional file 1 (`Additional_file_1.csv`, comma-separated values): Complete biological taxonomy mapping assigning all 38 leaf disease and healthy classes across eight host crops to the four agronomic intervention actions and PlantDoc wild-field overlap indicators.

References

[1] Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A. The global burden of pathogens and pests on major food crops. Nat Ecol Evol. 2019;3(3):430–439. https://doi.org/10.1038/s41559-018-0793-y.

[2] Strange RN, Scott PR. Plant disease: a threat to global food security. Annu Rev Phytopathol. 2005;43:83–116. https://doi.org/10.1146/annurev.phyto.43.113004.133839.

[3] Corkley I, Fraaije B, Hawkins N. Fungicide resistance management: maximizing the effective life of plant protection products. Plant Pathol. 2022;71(1):150–169. https://doi.org/10.1111/ppa.13467.

[4] Lahlali R, Taoussi M, Laasli SE, Gachara G, Ezzouggari R, Belabess Z, et al. Effects of climate change on plant pathogens and host-pathogen interactions. Crop Environ. 2024;3(3):159–170. https://doi.org/10.1016/j.crope.2024.05.003.

[5] Lázaro E, Makowski D, Vicent A. Decision support systems halve fungicide use compared to calendar-based strategies without increasing disease risk. Commun Earth Environ. 2021;2:224. https://doi.org/10.1038/s43247-021-00291-8.

[6] Mohanty SP, Hughes DP, Salathé M. Using deep learning for image-based plant disease detection. Front Plant Sci. 2016;7:1419. https://doi.org/10.3389/fpls.2016.01419.

[7] Ferentinos KP. Deep learning models for plant disease detection and diagnosis. Comput Electron Agric. 2018;145:311–318. https://doi.org/10.1016/j.compag.2018.01.009.

[8] Liu J, Wang X. Plant diseases and pests detection based on deep learning: a review. Plant Methods. 2021;17:22. https://doi.org/10.1186/s13007-021-00722-9.

[9] Thakur PS, Khanna P, Sheorey T, Ojha A. Trends in vision-based machine learning techniques for plant disease identification: a systematic review. Expert Syst Appl. 2022;208:118117. https://doi.org/10.1016/j.eswa.2022.118117.

[10] Al Masud SMR, Eshita FS, Mahi ABS, Mowla TJ, Arnob NMK, Saha AK, et al. LeafXNet: a lightweight explainable model with integrated attention mechanism for fruit leaf disease detection. Plant Direct. 2026;10(10):e70201. https://doi.org/10.1002/pld3.70201.

[11] Subramanian RR, Ganesh M SR, Fauzi MA, Dewi DA. MangoXNet: lightweight hybrid deep learning with explainable AI for accurate mango leaf disease diagnosis in smart agriculture. Algorithms. 2026;19(11):947. https://doi.org/10.3390/a19110947.

[12] Barbedo JGA. Factors influencing the use of deep learning for plant disease recognition. Biosyst Eng. 2018;172:84–91. https://doi.org/10.1016/j.biosystemseng.2018.05.013.

[13] Chen J, Chen J, Zhang D, Sun Y, Nanehkaran YA. Using deep transfer learning for image-based plant disease identification. Comput Electron Agric. 2020;173:105393. https://doi.org/10.1016/j.compag.2020.105393.

[14] Haque MA, Marwaha S, Deb CK, Nigam S, Arora A, Hooda KS, et al. Deep learning-based approach for identification of diseases of maize crop. Sci Rep. 2022;12:6334. https://doi.org/10.1038/s41598-022-10140-z.

[15] Unal Y. Integrating CBAM and squeeze-and-excitation networks for accurate grapevine leaf disease diagnosis. Food Sci Nutr. 2025;13(6):e70377. https://doi.org/10.1002/fsn3.70377.

[16] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML); 2017. p. 1321–1330.

[17] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems. vol. 32; 2019. p. 13991–14002.

[18] Singh D, Jain N, Jain P, Kayal P, Kumawat S, Batra N. PlantDoc: a dataset for visual plant disease detection. In: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD; 2020. p. 249–253. Available from: https://www.kaggle.com/datasets/nirmalsankalana/plantdoc-dataset [cited 21 Nov 2026].

[19] Zhang J, Zhang B, Chen Z, Nyalala I, Chen K, Gao J. A salient feature establishment tactic for cassava disease recognition. Artif Intell Agric. 2024;14:115–132. https://doi.org/10.1016/j.aiia.2024.11.004.

[20] Zhang J, Zhang B, Nyalala I, Mecha P, Chen J, Chen K, et al. Rectifying the extremely weakened signals for cassava leaf disease detection. Comput Electron Agric. 2025;232:110107. https://doi.org/10.1016/j.compag.2025.110107.

[21] Shafer G, Vovk V. A tutorial on conformal prediction. J Mach Learn Res. 2008;9:371–421.

[22] Sadinle M, Lei J, Wasserman L. Least ambiguous set-valued classifiers with bounded error levels. J Am Stat Assoc. 2019;114(525):223–234. https://doi.org/10.1080/01621459.2017.1395341.

[23] Romano Y, Sesia M, Candès EJ. Classification with valid and adaptive coverage. In: Advances in Neural Information Processing Systems. vol. 33; 2020. p. 3581–3591.

[24] Angelopoulos AN, Bates S, Malik J, Jordan MI. Uncertainty sets for image classifiers using conformal prediction. ICLR. 2021;arXiv:2009.14193.

[25] Vovk V. Conditional validity of inductive conformal predictors. Mach Learn. 2013;92(3):349–376. https://doi.org/10.1007/s10994-013-5355-6.

[26] Tibshirani RJ, Foygel Barber R, Candès EJ, Ramdas A. Conformal prediction under covariate shift. In: Advances in Neural Information Processing Systems. vol. 32; 2019. p. 2530–2540.

[27] Farag M, Emam A, Leonhardt J, Roscher R. Enhancing decision support in crop production: analyzing conformal prediction for uncertainty quantification. Comput Electron Agric. 2025;237:110559. https://doi.org/10.1016/j.compag.2025.110559.

[28] Angelopoulos AN, Bates S, Fisch A, Lei L, Schuster T. Conformal risk control. ICLR. 2024;arXiv:2208.02814.

[29] Naeini MP, Cooper GF, Hauskrecht M. Obtaining well calibrated probabilities using Bayesian binning. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence; 2015. p. 2901–2907.

[30] Nixon J, Dusenberry MW, Zhang L, Jerfel G, Tran D. Measuring calibration in deep learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2019. p. 38–41.

[31] Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1–3. https://doi.org/10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

[32] Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. ICLR. 2017;arXiv:1610.02136.

[33] Liu W, Wang X, Owens JD, Li Y. Energy-based out-of-distribution detection. In: Advances in Neural Information Processing Systems. vol. 33; 2020. p. 21464–21475.

[34] Geifman Y, El-Yaniv R. Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems. vol. 30; 2017. p. 4878–4887.

[35] Hughes DP, Salathé M. An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing. arXiv preprint. 2015;arXiv:1511.08060.

[36] Bhattarai S. New plant diseases dataset [database on the Internet]. Kaggle. 2018 [cited 21 Nov 2026]. Available from: https://www.kaggle.com/datasets/vipoooool/new-plant-diseases-dataset.

[37] Ahmed SI, Ibrahim M, Nadim M, Rahman MM, Shejunti MM, Jabid T, et al. MangoLeafBD: a comprehensive image dataset to classify diseased and healthy mango leaves. Data Brief. 2023;47:108941. https://doi.org/10.1016/j.dib.2023.108941.

[38] Ghose S. Corn or maize leaf disease dataset [database on the Internet]. Kaggle. 2020 [cited 21 Nov 2026]. Available from: https://www.kaggle.com/datasets/smaranjitghose/corn-or-maize-leaf-disease-dataset.

[39] Prajapati HB, Shah JP, Dabhi VK. Detection and classification of rice plant diseases. Intell Decis Technol. 2017;11(3):357–373. https://doi.org/10.3233/IDT-170301.

[40] Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, et al. Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2019. p. 1314–1324.

[41] Perez E, Strub F, De Vries H, Dumoulin V, Courville A. FiLM: visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32; 2018. p. 3942–3951.

[42] Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML); 2019. p. 6105–6114.

[43] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770–778.

[44] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ICLR. 2015;arXiv:1409.1556.

[45] Zuiderveld K. Contrast limited adaptive histogram equalization. In: Heckbert PS, editor. Graphics Gems IV. San Diego: Academic Press; 1994. p. 474–485.

[46] Nyalala I, Okinda C, Chao Q, Mecha P, Korohou T, Yi Z, et al. Weight and volume estimation of single and occluded tomatoes using machine vision. Int J Food Prop. 2021;24(1):818–832. https://doi.org/10.1080/10942912.2021.1933024.

[47] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21:6. https://doi.org/10.1186/s12864-019-6413-7.

[48] Wilcoxon F. Individual comparisons by ranking methods. Biom Bull. 1945;1(6):80–83. https://doi.org/10.2307/3001968.

[49] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153–157. https://doi.org/10.1007/BF02295996.