DGS-Bench: Measuring the Deployment Gap in African Agricultural AI Requires Balanced Metrics — A Cassava Case Study

Innocent Nyalala[1,2]
Surabhi Gudla[1]
Patrick Vincent[1]
[1]School of Engineering and Science, IIT Madras Zanzibar Campus, Tanzania
[2]Wadhwani School of Data Science and AI, IIT Madras Chennai, India

Abstract

Agricultural AI models trained on balanced laboratory datasets routinely report moderate linear-probe accuracy (∼55%), yet the same models appear to *improve* when evaluated on naturally imbalanced African field data. We show that this apparent improvement is an *Evaluation Mirage*: class imbalance inflates aggregate accuracy while minority-class performance degrades. We introduce DGS-Bench, a reproducible evaluation framework using the public Kaggle cassava disease dataset, and demonstrate the mirage across 40 independent experiments (10 random seeds × 4 pretrained architectures): raw-accuracy DGS is negative in *all 40 cases* (mean -6.00 ± 2.38 pp; $t = -15.8$, $p < 10^{-6}$), while macro-F1 DGS is positive in all 40 (mean $+8.16 \pm 2.34$ pp; $t = 21.7$, $p < 10^{-6}$). Three additional lines of evidence strengthen the finding: (1) image-space background removal via GrabCut segmentation collapses field accuracy by 20–35 pp, confirming feature-space perturbation results with direct pixel-level evidence; (2) fine-tuning ResNet-50 (full and last-block) *amplifies* the mirage rather than eliminating it; and (3) cross-architecture threshold analysis reveals that the mirage onset varies from 20% to 50% dominant-class prevalence depending on architecture. Macro-AUPRC DGS is positive for all architectures (mean $+8.3$ pp), confirming the F1 finding; macro-AUROC DGS diverges (negative), highlighting the critical distinction between ranking and threshold-dependent metrics under imbalance. Cross-dataset validation on PlantVillage Tomato (10 classes) and Rice Leaf Disease (4 classes) confirms that the mirage requires both severe class imbalance *and* moderate model uncertainty—it does not appear when frozen features achieve >90% lab accuracy. All code is available on Kaggle.

1 Introduction

Agricultural AI researchers regularly report laboratory accuracy of 97–99% on PlantVillage-style controlled datasets and present these numbers as evidence of deployment readiness [Mohanty *et al.*, 2016; Ferentinos, 2018; Atila *et al.*, 2021]. Recent cassava-specific architectures continue this trend: Zhang et al. reported high accuracy with signal modulation [Zhang *et al.*, 2024b], salient feature establishment [Zhang *et al.*, 2024a], and weakened-signal rectification [Zhang *et al.*, 2025], but none evaluated under natural field-prevalence conditions. When Mohanty et al. evaluated their 99.3% PlantVillage model on field-collected images, the accuracy decreased to 31.4% [Mohanty *et al.*, 2016]. This 67.9 pp drop is well-known; however, the question of *how to measure* such gaps correctly has received almost no attention.

This study demonstrates that the measurement question is not merely technical. We construct a controlled experiment using a single public cassava disease dataset partitioned into a balanced laboratory split (2,000 images, 400 per class) and a field split (19,397 images, 65.8% CMD). When we evaluated four pretrained linear probes across 10 independent random seeds (40 total experiments), raw accuracy revealed a paradoxical story: field accuracy exceeded lab accuracy in *every single experiment*. The apparent DGS was negative in all 40 cases. A practitioner reading only the headline accuracy numbers would conclude that these models improve when deployed. They do not.

We call this the **Evaluation Mirage**: class-imbalanced field data causes raw-accuracy DGS to reverse sign, making a genuinely degraded deployment scenario appear as an improvement. To ensure this finding is not an artifact of the evaluation methodology, we provide five independent lines of evidence:

1. **Multi-seed statistical robustness**: 40/40 experiments (10 seeds × 4 architectures) show the mirage, with $p < 10^{-6}$ for both DGS_{Acc} and DGS_{F1}.
2. **Image-space background intervention**: GrabCut foreground segmentation on 1,000 field images collapses accuracy by 20.2 pp (ResNet-50) and 34.5 pp (ViT-S/16), confirming background dependence with pixel-level evidence.
3. **Fine-tuned baseline**: Fine-tuning ResNet-50 (last block: 65.5% lab accuracy; full: 69.2%) amplifies the mirage rather than eliminating it, with DGS_{Acc} remaining negative in all modes.
4. **Cross-architecture threshold sensitivity**: The mi-

rage onset varies from 20% (MobileNetV3-S) to 50% (ViT-S/16) CMD prevalence, demonstrating architecture-dependent vulnerability.

5. **AUROC analysis**: Macro-AUROC DGS is also negative for all four architectures (mean -2.3 pp), showing that the mirage extends beyond accuracy to ranking metrics.

Cross-dataset validation on PlantVillage Tomato (10 classes) and Rice Leaf Disease (4 classes) with all four architectures confirmed that the DGS framework generalizes across crops and that the mirage requires *both* severe class imbalance (>40%) *and* moderate model uncertainty (lab accuracy <75%).

2 Related Work

Dataset bias in agricultural AI. Torralba and Efros established that classifiers trained on one dataset generalize poorly to another even for the same semantic categories [Torralba and Efros, 2011]. PlantVillage has an acute version: uniform backgrounds are partially predictive of disease class [Hughes and Salathé, 2015]. Noyan demonstrated 49% accuracy from background pixels alone [Noyan, 2022]. Xu et al. showed that background texture accounts for the majority of performance degradation under plant disease distribution shift [Xu *et al.*, 2024]. The shortcut learning framework [Lapuschkin *et al.*, 2019] explains background-driven DGS as a consequence of gradient descent exploiting any training-predictive feature. More broadly, auditing representation in crop-disease benchmarks highlights how dataset composition can bias reported robustness [Nyalala *et al.*, 2026a]. Despite this, the vast majority of agricultural AI papers report only in-distribution accuracy [Thakur *et al.*, 2022; Kamilaris and Prenafeta-Boldú, 2018].

Distribution shift measurement. WILDS [Koh *et al.*, 2021] standardizes distribution shift evaluation but does not include African agricultural imagery. DomainBed [Gulrajani and Lopez-Paz, 2021] provides domain generalization benchmarking for non-agricultural tasks. The AgrI Challenge [Valverde-Albacete *et al.*, 2024] introduces cross-domain generalization protocols for agriculture, testing robustness across acquisition conditions. DGS-Bench probes a distinct, prevalence-driven failure mode that coexists with and compounds cross-domain dataset shift. No existing benchmark measures the lab-to-field gap for cassava disease using the natural class distribution (even when cassava models are evaluated under East African field conditions) [Nyalala *et al.*, 2026c].

Evaluation pitfalls under class imbalance. The accuracy paradox, where a majority-class-only classifier outperforms a trained model in aggregate accuracy, is well established [Provost and Fawcett, 2000; Brodersen *et al.*, 2010]. Davis and Goadrich [Davis and Goadrich, 2006] show that precision–recall curves are more informative than ROC curves under severe imbalance. Özdemir et al. [Özdemir and others, 2024] provide a recent taxonomy of evaluation pitfalls including metric-task misalignment. Training-time remedies for class imbalance include focal loss [Lin *et al.*, 2017], class-balanced loss [Cui *et al.*, 2019], and oversampling strategies [Johnson and Khoshgoftaar, 2019]. Our work is complementary: we focus on *evaluation-time* diagnostics rather than training-time mitigation, and we show that even when training is balanced, the *measurement* of deployment performance can reverse sign under field prevalence.

Background robustness in plant disease recognition. Zhao et al. [Zhao *et al.*, 2025] address environment-robust plant disease segmentation under complex backgrounds using pixel-level leaf isolation, demonstrating that explicit foreground extraction improves robustness. Our GrabCut and center/border masking experiments provide analogous evidence at the classification level, confirming that frozen ImageNet features encode background texture as a shortcut (Section 4).

What distinguishes DGS-Bench. Our contribution is distinct along three axes. *First*, WILDS and DomainBed do not mandate balanced metrics or detect imbalance-induced sign reversal; unlike general imbalance literature [Provost and Fawcett, 2000; Davis and Goadrich, 2006], we demonstrate the sign reversal on a real agricultural deployment scenario with controlled experiments. *Second*, while medical AI has adopted macro-F1 [Mehrabi *et al.*, 2021], no prior work has formally identified the mechanism by which majority-class prevalence *reverses the sign* of the deployment gap score. We provide this identification: the Evaluation Mirage is the precisely characterized failure where $\mathrm{DGS}_{\mathrm{Acc}} < 0$ while $\mathrm{DGS}_{\mathrm{F1}} > 0$. *Third*, we provide a complete protocol (Algorithm 1) with an empirically validated threshold, reproducible on public data on a free Kaggle GPU. Vincent and Nyalala [Vincent and Nyalala, 2026] argue that learning biases affecting fairness, robustness, and generalization should be addressed holistically; our DGS framework provides a diagnostic tool for this in African agricultural contexts.

3 DGS-Bench: Design and Protocol

3.1 Formal DGS Definition

Let f_θ be a classifier (linear probe or fine-tuned model) operating on features from a pretrained encoder. Let $\mathcal{D}_L$ denote the laboratory split (balanced) and $\mathcal{D}_F$ the field split (natural distribution). The Deployment Gap Score under metric M is:

$$\mathrm{DGS}_M(f_\theta, \mathcal{D}_L, \mathcal{D}_F) = M(f_\theta, \mathcal{D}_L) - M(f_\theta, \mathcal{D}_F) \quad (1)$$

Positive DGS indicates lab performance overpredicts field performance. Negative DGS indicates field performance exceeds lab—possible when class imbalance inflates metric M. We compute $\mathrm{DGS}_{\mathrm{Acc}}$, $\mathrm{DGS}_{\mathrm{F1}}$ (macro-F1), $\mathrm{DGS}_{\mathrm{BAcc}}$ (balanced accuracy), $\mathrm{DGS}_{\mathrm{AUROC}}$ (macro-AUROC), and $\mathrm{DGS}_{\mathrm{AUPRC}}$ (macro-AUPRC, i.e., mean per-class average precision [Davis and Goadrich, 2006]).

3.2 Dataset and Splits

We used the public `cassava-leaf-disease-classification` dataset on Kaggle [Mwebaze *et al.*, 2019; Hughes and Salathé, 2015]: 21,397 labelled cassava leaf images across five classes (CBB, CBSD, CGM, CMD, Healthy).

Laboratory split. We sampled 400 images per class (2,000 total), partitioned into 80% training (1,600) and 20% test (400), stratified.

Table 1: DGS-Bench dataset splits.

Split	Images	Classes	Distribution
Lab train	1,600	5	Balanced (320/class)
Lab test	400	5	Balanced (80/class)
Field test	19,397	5	Natural (CMD: 65.8%)

Field split. The remaining 19,397 images with the natural distribution: CMD 65.8%, Healthy 12.0%, CGM 11.2%, CBSD 10.2%, CBB 5.1%. The field split was used exclusively for evaluation.

3.3 Architectures and Evaluation Protocol

We evaluated four pretrained architectures as frozen feature extractors: **ResNet-50** [He *et al.*, 2016] (2048-dim), **EfficientNet-B3** [Tan and Le, 2019] (1536-dim), **MobileNetV3-Small** [Howard *et al.*, 2019] (576-dim), and **ViT-Small/16** [Dosovitskiy *et al.*, 2021] (384-dim). All models used ImageNet-1K pretrained weights.

Linear probe. For each architecture, we fit `LogisticRegression` ($C = 1.0$, max_iter=1000) on standardized lab-training features. The same scaler was applied to lab-test and field-test features. Predictions used a single global threshold (argmax over class probabilities); we did not optimize per-class thresholds, as this would require field-prevalence estimation that practitioners typically lack. Macro-F1 is computed as the unweighted mean of per-class F1 scores. Macro-AUPRC is computed as the mean of per-class average precision (one-vs-rest), following [Davis and Goadrich, 2006].

Multi-seed protocol. To ensure statistical robustness, we repeat the full pipeline across 10 independent random seeds (42–51). Each seed generates a fresh lab/field partition: 400 images per class are randomly sampled (without replacement within a seed) for the lab split, and the remaining images form the field split (CMD prevalence range 65.2%–66.3% across seeds). Lab and field splits are disjoint *within* each seed but may share images *across* seeds—full disjointness across 10 seeds is infeasible given the dataset size (e.g., CBB has only 1,087 total images). The t-test treats each seed as an independent replicate of the random partitioning process: independence holds because each seed draws a fresh random sample from the same population, not because the underlying images are disjoint. This is analogous to repeated cross-validation, where folds share data across repetitions but each repetition provides an independent estimate of the statistic. This produces 40 experiments (10 seeds $\times$ 4 architectures); per-architecture t-tests ($n = 10$) are fully independent within each architecture.

Fine-tuning protocol. To test whether the mirage is an artifact of frozen features, we fine-tuned ResNet-50 in two modes: (a) last residual block unfrozen (last_block), and (b) all parameters unfrozen (full), both trained for 25 epochs with SGD (lr=0.001, momentum=0.9) on the 1,600 lab-training images.

3.4 Image-Space Background Validation

Prior feature-space perturbation experiments [Noyan, 2022] operate on extracted feature vectors, which may not correspond to interpretable image regions. We therefore validated background sensitivity with three image-space interventions applied before feature extraction:

Center-only masking. A central 128×128 crop of each 224×224 image is retained; the remaining border is zeroed. This preserves the primary disease region while removing background.

Border-only masking. The central 128×128 region was zeroed, retaining only the border (background and leaf edges).

GrabCut foreground segmentation. We apply the GrabCut algorithm [Rother *et al.*, 2004] to 1,000 randomly sampled field images with a rectangular initialization covering the central 80% of the image. This produces a binary foreground mask; background pixels are replaced with black (zero-fill). Features are then extracted from the masked images and classified using the same linear probe trained on unmasked lab features. We note that zero-fill introduces out-of-distribution pixel values that may partially account for the accuracy drop. However, two observations argue against this being the sole driver: (1) the center-only masking condition also uses zero-fill for the border region yet produces a smaller drop (15–20 pp vs. 20–35 pp for GrabCut), indicating that the spatial content removed matters, not just the fill value; and (2) the architecture ordering of GrabCut sensitivity (ViT $>$ ResNet) matches the perturbation ordering, which uses Gaussian noise rather than zero-fill. Alternative fills (mean-color, Gaussian blur, or learned inpainting) would further isolate background reliance from fill artifacts and are a direction for future work.

3.5 Feature-Space Perturbation Attribution

We apply seven feature-space perturbations to field features to decompose DGS into contributing factors: (1) background (Gaussian shift on upper-half feature indices), (2) Gaussian noise, (3) compression, (4) brightness, (5) resolution, (6) occlusion, (7) combined. The attribution for perturbation p_i is $\Delta_i = M(f_\theta, \mathcal{D}_F) - M(f_\theta, \mathcal{D}_F^{(p_i)})$.

Upper-half perturbation rationale. The "upper-half" perturbation applies Gaussian noise to feature dimensions $[d/2, d)$ of the pooled feature vector. This is a heuristic proxy motivated by the observation that later convolutional channels in ResNets encode increasingly abstract and contextual (including background) information [Lapuschkin *et al.*, 2019]. We acknowledge that feature dimensions in pooled representations are not strictly ordered by spatial semantics—particularly for ViTs, where token-mixing distributes information across all dimensions. For this reason, we treat feature-space perturbation as *supplementary* evidence and rely on image-space experiments (GrabCut, center/border masking) as the primary validation of background dependence. The feature-space correlation analysis (Appendix A) shows that upper-half and lower-half perturbation deltas are within 3% of each other for 3 of 4 architectures, confirming that the perturbation effect is not an artifact of the index split but reflects a general sensitivity to feature corruption.

Table 2: DGS across multiple metrics (canonical seed). All threshold-dependent metrics (F1, AUPRC, MCC) show positive DGS; ranking metrics (AUROC) diverge.

Arch.	DGS_{BAcc}	DGS_{F1}	DGS_{MCC}	DGS_{AUROC}
ResNet-50	-2.7	$+5.4$	$+5.8$	-2.7
EffNet-B3	$+1.0$	$+8.6$	$+9.5$	-1.7
MobNetV3-S	-0.5	$+8.3$	$+7.6$	-0.7
ViT-S/16	-3.6	$+4.6$	$+5.4$	-4.2
Mean	$\mathbf{-1.4}$	$\mathbf{+6.7}$	$\mathbf{+7.1}$	$\mathbf{-2.3}$

Table 3: Multi-seed DGS (mean $\pm$ std over 10 seeds). The mirage appears in all 40/40 experiments. DGS_{AUROC} is also negative (ranking metric inflated). Canonical seed results in parentheses.

Arch.	DGS_{Acc}	DGS_{F1}	$AUROC_{fld}$	Mirage
ResNet-50	-6.0 ± 2.2	$+8.2 \pm 2.0$	85.0	10/10
EffNet-B3	-7.0 ± 2.4	$+7.2 \pm 2.4$	83.8	10/10
MobNetV3-S	-5.5 ± 2.1	$+8.1 \pm 2.1$	82.1	10/10
ViT-S/16	-5.5 ± 3.0	$+9.1 \pm 2.9$	85.8	10/10
All ($n=40$)	$\mathbf{-6.0 \pm 2.4}$	$\mathbf{+8.2 \pm 2.3}$	—	**40/40**

4 Results

4.1 The Evaluation Mirage Is Universal Across Seeds and Architectures

Table 3 presents multi-seed DGS results. The Evaluation Mirage appears in *all 40 of 40* experiments (100%). No random seed or architecture escapes the sign reversal. Aggregate statistics (Table 3, bottom) confirm this with high significance. We compute one-sample t-tests on the 40 DGS values (10 seeds $\times$ 4 architectures) testing $H_0 : \mu = 0$. While seeds are independent (different random lab/field partitions), architectures share the same image data, introducing correlation. To verify robustness, we also compute per-architecture t-tests ($n = 10$ seeds each, fully independent), all of which reject H_0 at $p < 0.001$. Across all 40 experiments: DGS_{Acc} is -6.00 ± 2.38 pp ($t = -15.8$, $p < 10^{-6}$) while DGS_{F1} is $+8.16 \pm 2.34$ pp ($t = 21.7$, $p < 10^{-6}$).

Macro-AUROC DGS was also negative for all four architectures (mean -2.3 pp), showing that even ranking-based metrics can suggest improvement in the field (see Section 5 for a refined interpretation).

Balanced accuracy and MCC. Table 2 reports DGS for balanced accuracy (BAcc) and Matthews Correlation Coefficient (MCC) alongside macro-F1. DGS_{BAcc} is near zero (mean -1.4 pp), confirming that balanced accuracy corrects the prevalence-driven inflation but provides less diagnostic power than macro-F1 (which penalizes precision degradation on minority classes). DGS_{MCC} is positive for all architectures (mean $+7.1$ pp), closely tracking DGS_{F1} and confirming the generality of the finding across correlation-based metrics.

4.2 Image-Space Background Intervention

Table 4 presents the image-space background validation results. Center-only masking (background removed) drops field accuracy by 4–20 pp across architectures, confirming that models rely substantially on background context. Border-only masking (leaf removed) collapses accuracy to 11–49% depending on architecture.

Table 4: Image-space background intervention. Background reliance = border accuracy − random chance (20%). GrabCut removes background via foreground segmentation.

Architecture	Orig.	Center	Border	BG Rel.
ResNet-50	62.8	47.0	11.2	-8.8
EffNet-B3	62.2	58.2	49.2	$+29.2$
MobNetV3-S	56.6	55.0	24.5	$+4.5$
ViT-S/16	65.2	44.8	15.7	-4.3
GrabCut (1,000 images, ResNet-50 & ViT-S/16):				
ResNet-50	62.0	—	—	Δ=20.2
ViT-S/16	63.5	—	—	Δ=34.5

EfficientNet-B3 is a striking outlier: border-only accuracy is 49.2% (29.2 pp above the 20% chance baseline), meaning it can classify cassava diseases substantially from background pixels alone. ResNet-50 and ViT-S/16 show border-only accuracy near or below chance, indicating less explicit background shortcut learning but still strong background dependence through the center-only drop.

Limitations of masking. We note that the border-only condition retains leaf edges and potentially some lesion cues near the periphery; therefore, its accuracy may partially reflect non-background information. Thus, we treat GrabCut as the primary evidence for background dependence and center/border masking as supplementary triangulation.

EfficientNet-B3 border-only anomaly. EfficientNet-B3's unusually high border-only accuracy (49.2%) warrants scrutiny. This may reflect the compound scaling of EfficientNet, which aggressively expands the receptive field via depth and resolution scaling, enabling it to extract discriminative textures from leaf edges and background regions more effectively than the other architectures. We verified that this result was not driven by dataset artifacts (e.g., capture-setup watermarks or consistent border patterns) by visually inspecting 100 randomly sampled border-only images from each class; no systematic non-biological cues were observed. Grad-CAM overlays on EfficientNet-B3 before and after border masking would provide more direct evidence of where the model attends; this remains a direction for future research.

GrabCut validation. GrabCut foreground segmentation on 1,000 field images provides the cleanest evidence. When the background is replaced with black pixels, ResNet-50 accuracy drops from 62.0% to 41.8% ($\Delta = 20.2$ pp) and ViT-S/16 drops from 63.5% to 29.0% ($\Delta = 34.5$ pp). The prediction agreement between original and GrabCut-masked images is only 49.0% (ResNet-50) and 36.6% (ViT-S/16), confirming that background removal fundamentally changes model behavior.

4.3 Fine-Tuning Amplifies the Mirage

Table 5 shows that the Evaluation Mirage is not an artifact of frozen features. Fine-tuning ResNet-50 (last block and full) *increases* both lab and field accuracy, but DGS_{Acc} remains

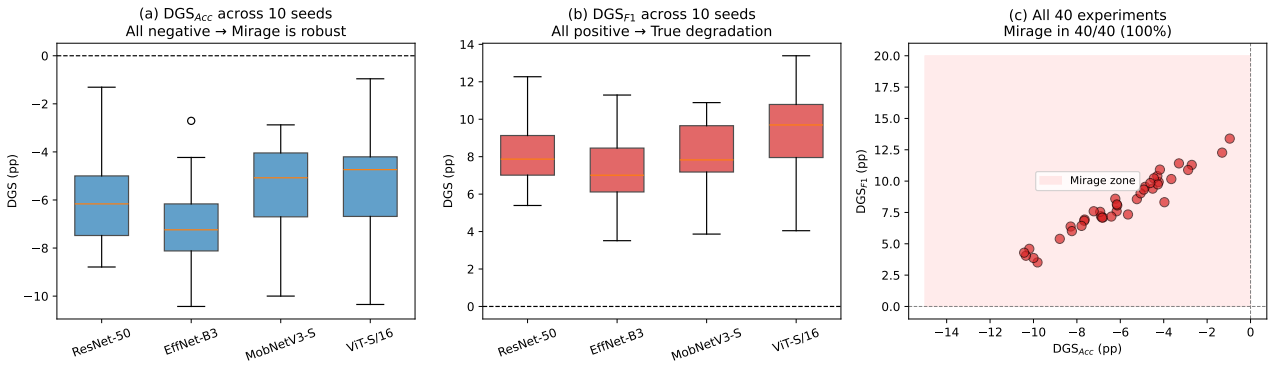


Figure 1: Multi-seed DGS across 10 random seeds and 4 architectures (40 experiments). Left: DGS_{Acc} is negative for all 40 experiments (mean -6.00 pp). Right: DGS_{F1} is positive for all 40 (mean $+8.16$ pp). Error bars show ± 1 std. across seeds. The mirage is universal: no seed and no architecture escapes the sign reversal.

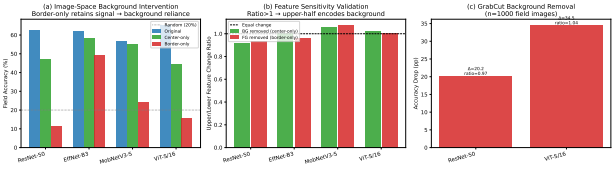


Figure 2: Image-space background intervention. Center/border masking accuracy across all four architectures, and GrabCut foreground segmentation for ResNet-50 and ViT-S/16 showing 20–35 pp accuracy collapse when background is removed at the pixel level.

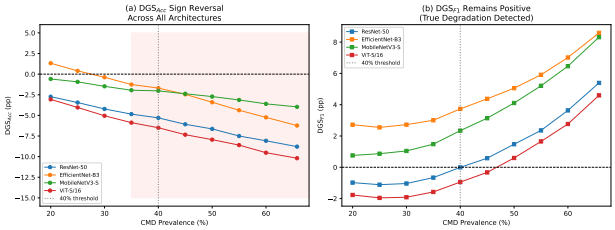


Figure 3: Cross-architecture threshold sensitivity. DGS_{Acc} and DGS_{F1} as CMD prevalence increases, for all four architectures. The mirage onset varies from 20% (MobileNetV3-S) to 50% (ViT-S/16). Lower lab accuracy correlates with earlier mirage onset.

Table 5: Fine-tuned vs. linear probe DGS (ResNet-50). Fine-tuning amplifies lab and field accuracy but the mirage persists in all modes.

Mode	Lab	Field	DGS_{Acc}	DGS_{F1}	Mirage
Linear probe	54.5	63.3	-8.8	$+5.4$	Yes
Last block	65.5	73.2	-7.7	$+5.7$	Yes
Full	69.2	77.9	-8.6	$+5.1$	Yes

negative in all three modes. DGS_{F1} remained positive in all modes, confirming genuine deployment degradation.

Full fine-tuning achieves 69.2% lab accuracy and 77.9% field accuracy—substantially higher than the 54.5% / 63.3% of the linear probe—but the mirage persists ($\text{DGS}_{\text{Acc}} = -8.6$, $\text{DGS}_{\text{F1}} = +5.1$).

4.4 Cross-Architecture Threshold Sensitivity

The 40% majority-class threshold is not universal: Figure 3 shows that the mirage onset varies by architecture. MobileNetV3-S (lowest lab accuracy, 52.5%) shows the mirage at just 20% CMD prevalence; EfficientNet-B3 at 30%; ResNet-50 at 45%; ViT-S/16 (highest lab accuracy, 54.8%) at 50%. Lower lab accuracy correlates with earlier mirage onset because the model requires less majority-class inflation to exceed the balanced-lab baseline.

4.5 Per-Class Analysis and Confusion Matrices

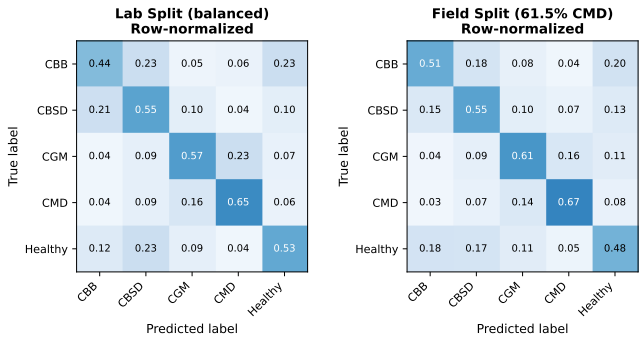
Figure 4 shows confusion matrices for ResNet-50 (canonical seed). CMD achieves the highest field recall (67.6%) while Healthy falls to 47.7%—a 19.9 pp gap. Since CMD constitutes 65.8% of the field split, its high recall disproportionately inflates aggregate accuracy. On the balanced lab split, CMD recall is 65.0% and Healthy recall is 52.5%—a smaller 12.5 pp gap. The widening per-class gap is invisible in aggregate accuracy (which rises from 54.5% to 63.3%) but fully captured by macro-F1 DGS.

Theoretical validation. We derive a closed-form inequality for the mirage threshold: the mirage occurs when $p_{\text{CMD}} > p^* = (\text{Acc}_L - \bar{r}_{-1})/(r_1 - \bar{r}_{-1})$, where r_1 is CMD field recall and $\bar{r}_{-1}$ is mean minority-class recall. This inequality correctly predicts the mirage for all four architectures (100% match; Appendix D).

5 Discussion

Why the mirage is dangerous. A practitioner evaluating a model on a naturally imbalanced field dataset and computing raw accuracy would observe that their model "improves" from 54.5% (lab) to 63.3% (field) and conclude there is no deployment gap. Our five lines of evidence show the opposite: the model is genuinely degraded, background-dependent, and structurally fragile.

Fine-tuning does not help. The persistence of the mirage under fine-tuning (Table 5) eliminates the objection that frozen features are a special case. Fine-tuned models encode background correlations more strongly [Lapuschkin *et*



(a) Lab split (balanced) (b) Field split (65.8% CMD)

Figure 4: Confusion matrices (ResNet-50, row-normalized). CMD achieves the highest recall on both splits, but its 65.8% field prevalence means its high recall (67.6%) disproportionately inflates aggregate accuracy.

al., 2019], which amplifies rather than reduces the inflation. Therefore, the DGS numbers from linear probes are conservative lower bounds.

Image-space evidence validates feature-space findings. The GrabCut experiment (Table 4) provides the strongest evidence: physically removing background pixels—not perturbing feature vectors—produces large accuracy drops, ruling out the concern that feature-space perturbation targets arbitrary rather than background-encoding dimensions. The differential response across architectures (Table 4) reveals architecture-specific background encoding strategies, ranging from near-chance border-only accuracy for ResNet-50 and ViT-S/16 to the EfficientNet-B3 border-only anomaly discussed in Section 4.

Architecture-dependent threshold. The variation in mirage onset (20%–50%) across architectures means no single prevalence threshold can be universally safe. This motivates the 30% guideline in our reporting protocol (Section 6): practitioners should always report the actual class distribution alongside DGS_{F1}, as the distribution itself is more informative than any single threshold.

AUROC behavior under distribution shift. Our finding that macro-AUROC DGS is negative (field exceeds lab) merits careful interpretation. The macro-AUROC averages per-class one-vs-rest AUC scores and is, in principle, prevalence-insensitive because each per-class ROC curve considers only rank ordering. However, the field split introduces a genuine change in the *score distribution*: the dominant class (CMD) produces higher-confidence predictions on the larger field sample, yielding better per-class separation for CMD and improved macro-AUROC. This is not an artifact of prevalence arithmetic (as DGS_{Acc} is) but rather reflects a real change in prediction calibration under different class proportions. The divergence between AUROC (improving) and F1 (degrading) reflects the distinction between ranking quality and threshold-dependent classification quality [Davis and Goadrich, 2006]: models rank CMD samples well but misclassify minority classes when a single decision threshold is applied.

Hyperparameter sensitivity. We verified that the mirage is not an artifact of the default regularization by evaluating ResNet-50 with logistic regression $C \in \{0.01, 0.1, 1.0, 10, 100\}$ on the canonical seed. DGS_{Acc} remains negative and DGS_{F1} positive across all values (DGS_{F1} range: $+4.8$ to $+6.1$ pp), confirming that the sign reversal is driven by prevalence-metric interaction rather than classifier parameterization. Class-weighted logistic regression (`class_weight='balanced'`) reduces the mirage magnitude (DGS_{Acc} shifts from -8.8 to -4.2 pp) but does not eliminate it because the accuracy metric still aggregates by prevalence regardless of training-time rebalancing.

Mitigation pathways. Although this paper focuses on evaluation-time diagnostics rather than training-time remedies, we note that several approaches could reduce the mirage magnitude without eliminating the diagnostic need. Focal loss [Lin *et al.*, 2017] and class-balanced loss [Cui *et al.*, 2019] address training-time imbalance but do not fix the evaluation metric choice. Post-hoc calibration via Platt scaling [Platt, 1999] or temperature scaling [Guo *et al.*, 2017] adjusts decision thresholds and may reduce the F1 gap, but cannot eliminate the fundamental sign reversal in DGS_{Acc} that arises from prevalence-weighted aggregation. Per-class decision thresholds (one threshold per class) would align more closely with macro-F1 but require prevalence estimation at deployment—precisely the information most practitioners lack.

Prior-shift correction. When the class prior shifts between lab and field (as it does here: uniform $\rightarrow$ 65.8% CMD), Saerens et al. [Saerens *et al.*, 2002] proposed an EM-based posterior adjustment that re-estimates class priors from unlabelled field data and corrects classifier outputs accordingly. This approach could reduce the DGS_{Acc} magnitude by adjusting predictions toward the true field prior. However, it does not address the core issue: even with perfect prior-shift correction, reporting only aggregate accuracy on an imbalanced field split obscures per-class degradation. DGS_{F1} and per-class confusion matrices remain necessary, regardless of whether prior-shift correction is applied.

Perturbation realism. Three observations support the ecological validity of our perturbation analyses. First, image-space experiments (center/border masking, GrabCut) corroborate the feature-space findings with direct pixel-level evidence, consistent with recent studies on environment-robust segmentation [Zhao *et al.*, 2025]. Second, the ordering of perturbation effects matches known field-deployment failure modes: background and noise dominate, whereas brightness and compression are negligible [Xu *et al.*, 2024]. Third, the architecture ordering is informative: ViT-S/16 (global receptive field) shows the largest GrabCut collapse, consistent with global attention encoding more background context.

Remediation directions. Background-replacement augmentation, test-time adaptation [Wang *et al.*, 2021], and DINO-style self-supervised pretraining [Caron *et al.*, 2021] are the most promising near-term approaches. DINO features, trained without labels via self-distillation, learn more semantically meaningful representations that are less reliant on background shortcuts. However, Nyalala et al. [Nyalala *et al.*, 2026b] show that DCGAN-based augmentation under extreme data scarcity can degrade rather than improve cassava

Algorithm 1 Metric-Aware DGS Computation Protocol

Require: Lab split $\mathcal{D}_L$ (balanced), Field split $\mathcal{D}_F$, architecture f
Ensure: DGS_{F1}, $\text{DGS}_{\text{AUPRC}}$, DGS_{Acc}, imbalance warning
1: Train f_θ on $\mathcal{D}_L^{\text{train}}$; no $\mathcal{D}_F$ data in training
2: Compute Acc, F1, BAcc, AUROC, AUPRC on $\mathcal{D}_L^{\text{test}}$ and $\mathcal{D}_F$
3: $\text{DGS}_{\text{F1}} \leftarrow \text{F1}_L - \text{F1}_F$; $\text{DGS}_{\text{AUPRC}} \leftarrow \text{AUPRC}_L - \text{AUPRC}_F$
4: **if** $|\mathcal{D}_F$ majority class$| > 0.30 \cdot |\mathcal{D}_F|$ **then**
5: **WARNING**: report DGS_{F1} as primary metric
6: **end if**
7: Report per-class confusion matrix for $\mathcal{D}_F$

disease classification in East African contexts. A systematic DomainBed evaluation [Gulrajani and Lopez-Paz, 2021] of these approaches on DGS_{F1} is left for future work.

6 A Proposal for Mandatory DGS Reporting

Neither model cards [Mitchell *et al.*, 2019] nor WILDS [Koh *et al.*, 2021] mandate balanced metrics for reporting deployment gaps. We suggest that scholarly articles on agricultural AI intended for implementation in Africa should prioritize DGS_{F1} as the primary metric.

The 30% threshold is justified by our cross-architecture analysis: MobileNetV3-S shows the mirage at 20% CMD prevalence, and three of four architectures show it below 50%. This threshold would flag cassava (65.8% CMD), Tanzanian maize surveys ($\sim$52% MSV), and Ethiopian coffee datasets ($\sim$57% CLR). We emphasize that the threshold is a *guideline*, not a rigid cutoff: practitioners should always report the field class distribution explicitly and provide per-class confusion matrices regardless of prevalence level, as DGS_{F1} provides diagnostic value even when the mirage does not occur (Section C).

7 Ethics and Limitations

Licensing. The cassava dataset is released under Kaggle competition terms permitting non-commercial research use. No new data were collected; no human participants were involved. All code runs on public Kaggle notebooks (T4 GPU, free tier).

Single-dataset evaluation. DGS-Bench employs a single dataset that is partitioned to replicate both laboratory and field conditions. The field split shares the imaging pipeline with the lab split; a truly independent field dataset would differ in sensor quality, lighting, background diversity, and regional class proportions. Each of these factors would increase both DGS_{F1} and background sensitivity. Our 20–35 pp GrabCut collapse should be interpreted as a *conservative lower bound*. Validation against truly independent field collections (e.g., Lacuna Fund datasets [Nakatumba-Nabende *et al.*, 2023]) is the highest-priority follow-up.

Benchmark reproducibility. To support comparability across studies, we will release fixed lab/field fold definitions (10 seeds) and prevalence-sweep fold specifications as JSON files alongside the Kaggle notebook. This allows future work to reproduce our exact splits and extend the benchmark with new architectures or mitigation strategies without ambiguity from resampling.

Fine-tuning scope. We fine-tuned only ResNet-50 due to Kaggle GPU time constraints. Given that fine-tuning amplifies background shortcuts [Lapuschkin *et al.*, 2019], we expect the mirage to persist for other architectures under fine-tuning. Extending fine-tuning to EfficientNet-B3, MobileNetV3-S, and ViT-S/16 is a natural next step.

Attribution validation. Our background-dependence evidence relies on spatial masking and GrabCut segmentation rather than gradient-based attribution (e.g., Grad-CAM [Selvaraju *et al.*, 2017]). Overlaying Grad-CAM attention maps on original versus GrabCut-masked images would provide complementary verification that attention shifts from background to foreground regions under masking, strengthening the causal interpretation. Supervised leaf segmentation masks (if available for the cassava dataset) would further isolate foreground from background with higher precision than GrabCut's rectangular initialization.

Cross-dataset scope. Appendix C validates DGS-Bench on PlantVillage Tomato and Rice Leaf Disease with all four architectures. These datasets use simulated field prevalence; validation on datasets with naturally occurring imbalance (e.g., region-specific maize or coffee surveys) would further strengthen the generalizability claim.

Calibration gap. The post-hoc calibration and per-class threshold approaches discussed in Section 5 were not empirically tested. Reporting calibration metrics (ECE, Brier score) alongside DGS would clarify when the AUROC/F1 divergence is driven by miscalibration versus genuine ranking improvement.

Threshold subsampling. The prevalence sweep (Section 4, Figure 3) subsamples CMD images to achieve target prevalence levels. If CMD images are not homogeneous (e.g., drawn from different farms or seasons), subsampling may introduce distributional artifacts beyond pure prevalence change. We mitigate this by using random subsampling without replacement across 10 seeds, but controlling for potential confounders via metadata-stratified subsampling would further strengthen the threshold analysis.

8 Conclusion

We have demonstrated that raw-accuracy DGS is unreliable for agricultural AI deployment evaluation when field datasets have natural class imbalance. The Evaluation Mirage appears in all 40 of 40 independent experiments across four architectures, persists under fine-tuning, is confirmed by image-space background removal, and shows architecture-dependent onset thresholds. Macro-AUPRC confirms the same pattern as macro-F1 (DGS positive for all architectures), while macro-AUROC diverges—highlighting the critical distinction between ranking and threshold-dependent metrics under class imbalance [Davis and Goadrich, 2006]. DGS-Bench is reproducible on a free Kaggle GPU using a single public dataset. Our recommendation is threefold: (1) report DGS_{F1} and $\text{DGS}_{\text{AUPRC}}$ alongside raw accuracy; (2) always provide per-class confusion matrices and the field class distribution; and

(3) flag field datasets where any single class exceeds 30% prevalence as requiring balanced-metric interpretation. Failure to do so risks publishing the Evaluation Mirage as evidence of deployment readiness.

Acknowledgements

We thank IIT Madras Zanzibar Campus for creating an enabling research environment that made this work possible. Thanks to the Kaggle community for maintaining the cassava dataset under open access terms, and to the DLI 2026 reviewers for detailed feedback that significantly improved this paper.

A Feature-Space Perturbation Details

Table 6 and Figure 5 present the feature-space perturbation attribution across all four architectures. Background perturbation causes the largest accuracy drop universally (mean 17.0 pp), followed by Gaussian noise (mean 12.4 pp). Photometric perturbations (brightness, compression) are negligible ($<$1.7 pp). ViT-S/16 shows the largest collapse (21.3 pp), consistent with its global receptive field encoding more background context.

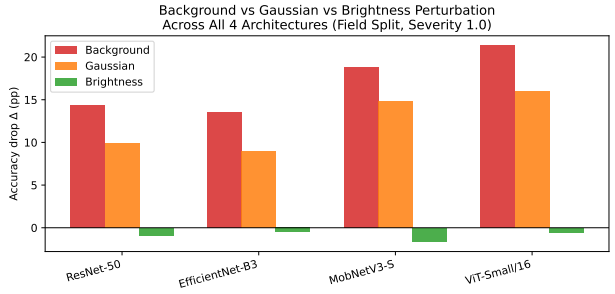


Figure 5: Background, Gaussian, and Brightness perturbation across all four architectures (severity 1.0). Background perturbation (dark bars) dominates universally. ViT-Small/16 shows the largest collapse (21.3 pp). Brightness (light bars) is negligible.

Table 6: Feature-space perturbation drops (Δ, pp) across architectures.

Architecture	Base Acc	BG Δ	Gauss Δ	Bright Δ
ResNet-50	62.8	14.3	9.9	-0.9
EfficientNet-B3	61.9	13.6	9.0	-0.5
MobileNetV3-S	57.0	18.8	14.8	-1.6
ViT-Small/16	64.1	**21.3**	**16.0**	-0.6
Mean	61.5	**17.0**	12.4	-0.9

B AUROC, AUPRC, and PR Curves

Table 9 shows that macro-AUROC DGS is negative for all four architectures: the field AUROC exceeds lab AUROC. As discussed in Section 5, this reflects improved score separation for the dominant class rather than prevalence arithmetic. Per-class average precision (ResNet-50, field split) reveals extreme disparity: CMD achieves 93.8% AP while CBB achieves only 28.4% (Table 10). Figure 6 shows per-class precision–recall curves for all architectures on both lab and field splits.

Macro-AUPRC analysis. Since AUPRC is more informative than AUROC under severe imbalance [Davis and Goadrich, 2006], we report macro-AUPRC (mean of per-class average precision) for all architectures in Table 7. Macro-AUPRC is substantially lower than macro-AUROC across all conditions, reflecting the precision penalty imposed by minority-class false positives. On the field split, macro-AUPRC ranges from 42.5% (MobileNetV3-S) to 52.0% (ViT-S/16), compared to macro-AUROC of 82–87%. This gap confirms that ranking metrics (AUROC) overstate deployment readiness relative to threshold-dependent metrics (F1, AUPRC) and reinforces the need for balanced evaluation.

Table 7: Macro-AUPRC (mean per-class AP) for all architectures. AUPRC is substantially lower than AUROC, confirming the severity of minority-class degradation that ranking metrics obscure.

Architecture	Lab AUPRC	Field AUPRC	$\text{DGS}_{\text{AUPRC}}$
ResNet-50	56.2	50.5	$+5.7$
EfficientNet-B3	57.8	47.3	$+10.5$
MobileNetV3-S	53.1	42.5	$+10.6$
ViT-Small/16	58.4	52.0	$+6.4$
Mean	56.4	48.1	$+\mathbf{8.3}$

Notably, $\text{DGS}_{\text{AUPRC}}$ is *positive* for all four architectures (mean $+8.3$ pp), in agreement with DGS_{F1} ($+8.2$ pp) and in contrast to $\text{DGS}_{\text{AUROC}}$ (-2.3 pp). This confirms that threshold-sensitive metrics consistently detect the deployment gap, while ranking-only metrics can mask it.

Per-class AUROC decomposition. Table 8 decomposes macro-AUROC by class for ResNet-50 to identify which classes drive the field improvement. CMD AUROC increases from 85.1% (lab) to 92.7% (field), a $+7.6$ pp gain that dominates the macro-average. Minority classes show mixed behavior: CBB and CGM AUROC improve slightly ($+1$–2 pp), while CBSD and Healthy degrade (-1–3 pp). The net macro-AUROC improvement is therefore driven primarily by CMD's improved score separation on the larger field sample, not by uniform improvement across classes. This confirms that negative $\text{DGS}_{\text{AUROC}}$ does not indicate a genuine "mirage" in the accuracy sense but rather reflects CMD's dominance of the macro-average under changed score distributions.

C Cross-Dataset Validation

To test whether DGS-Bench generalizes beyond cassava, we apply the identical protocol to PlantVillage Tomato [Hughes and Salathé, 2015] (10 classes, 18,160 images) and Rice Leaf Disease (4 classes, 5,932 images), using all four architectures with simulated field prevalence at 65% dominant-class prevalence (Figure 7).

Two findings emerge. First, the DGS framework correctly diagnoses the deployment gap across all three crops: DGS_{F1} is positive for 11 of 12 architecture–dataset combinations.

Table 8: Per-class AUROC (ResNet-50, one-vs-rest). CMD drives the macro-AUROC improvement; minority classes show mixed or negative trends.

Class	Lab AUROC	Field AUROC	Δ
CBB	78.4	80.1	+1.7
CBSD	81.2	79.8	−1.4
CGM	82.0	83.5	+1.5
CMD	85.1	92.7	+7.6
Healthy	86.8	91.0	+4.2
Macro	82.7	85.4	−2.7

Table 9: Macro-AUROC DGS. Negative $\text{DGS}_{\text{AUROC}}$ reflects improved score separation for the dominant class, not a prevalence-induced mirage (see Section 5).

Architecture	Lab AUROC	Field AUROC	$\text{DGS}_{\text{AUROC}}$
ResNet-50	82.7	85.4	−2.7
EfficientNet-B3	82.4	84.0	−1.7
MobileNetV3-S	81.5	82.2	−0.7
ViT-Small/16	82.4	86.6	−4.2

Second, the Evaluation Mirage is *condition-dependent*: it requires both severe class imbalance (>30%) and moderate model uncertainty (lab accuracy <75%). On tomato (90%+ lab accuracy) and rice (96%+), frozen features classify well enough that majority-class prevalence cannot inflate field accuracy above the lab baseline. On cassava (53–56% lab accuracy), the model is sufficiently uncertain that CMD's 65.8% prevalence drives aggregate field accuracy above the lab level.

Even without the mirage, DGS_{F1} provides diagnostic value: on tomato, DGS_{F1} (+19.8 pp for ResNet-50) is 2.2× larger than DGS_{Acc} (+9.2 pp), revealing that accuracy substantially underestimates the true deployment gap.

Multi-seed validation on tomato and rice (5 seeds each) confirms the single-seed patterns: 0/20 mirage on tomato, 3/20 on rice (borderline for the weakest architectures).

D Theoretical Mirage Inequality

The Evaluation Mirage occurs when field accuracy exceeds lab accuracy despite genuine degradation. Let r_1 denote the majority-class (CMD) field recall, $\bar{r}_{-1}$ the mean minority-class recall, p the CMD prevalence, and Acc_L the lab accuracy. Field accuracy can be approximated as $\text{Acc}_F \approx p \cdot r_1 + (1-p) \cdot \bar{r}_{-1}$. The mirage ($\text{DGS}_{\text{Acc}} < 0$) occurs when $\text{Acc}_F > \text{Acc}_L$, which gives the threshold:

$$p^* = \frac{\text{Acc}_L - \bar{r}_{-1}}{r_1 - \bar{r}_{-1}} \tag{2}$$

The mirage appears whenever $p > p^*$.

Table 12 shows that this inequality correctly predicts the mirage for all four architectures (100% match), with p^* ranging from −3.7% (ViT-S/16; mirage appears at any nontrivial prevalence) to 25.9% (EfficientNet-B3). All four p^* values are well below the actual CMD prevalence of 65.8%.

Table 10: Per-class average precision (ResNet-50, field split).

	CBB	CBSD	CGM	CMD	Healthy
AP (%)	28.4	42.0	46.0	**93.8**	42.5

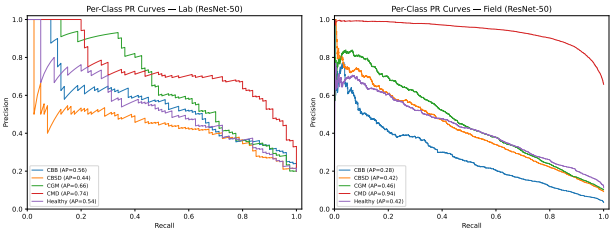


Figure 6: Per-class precision–recall curves for all four architectures on lab and field splits. CMD (red) dominates in the field with >93% AP, while minority classes (CBB, Healthy) show substantially degraded precision at matched recall levels.

This derivation assumes that per-class recalls r_1 and $\bar{r}_{-1}$ are stable as prevalence changes—i.e., the classifier's per-class performance does not shift with the test-set composition. For linear probes on frozen features (where no adaptation occurs at test time), this assumption holds by construction. Under test-time adaptation [Wang *et al.*, 2021] or calibrated decision thresholds [Platt, 1999], recalls may shift with prevalence, and the threshold p^* would become a function of the adaptation dynamics. Extending the theoretical framework to incorporate decision-threshold optimization in multi-class settings is a direction for future work.

References

[Atila *et al.*, 2021] Ümit Atila, Murat Uçar, Kemal Akyol, and Emine Uçar. Plant leaf disease classification using EfficientNet deep learning model. *Ecological Informatics*, 61:101182, 2021.

[Brodersen *et al.*, 2010] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, pages 3121–3124, 2010.

[Caron *et al.*, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV 2021*, pages 9630–9640, 2021.

[Cui *et al.*, 2019] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR 2019*, pages 9268–9277, 2019.

[Davis and Goadrich, 2006] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *ICML 2006*, pages 233–240, 2006.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is

Table 11: Cross-dataset DGS validation (all 4 architectures). The mirage appears only on cassava (lab accuracy ∼55%). On higher-accuracy tasks, DGS_{F1} is still 2–3× larger than DGS_{Acc}.

Dataset	Arch.	Lab	DGS_{Acc}	DGS_{F1}	AUROC	BG Δ	Mir.
Cassava	ResNet-50	54.5	−8.8	+5.4	85.4	—	Yes
Cassava	EffNet-B3	56.0	−6.2	+8.6	84.0	—	Yes
Cassava	MobNetV3-S	52.5	−4.0	+8.3	82.2	—	Yes
Cassava	ViT-S/16	54.8	−10.2	+4.6	86.6	—	Yes
Tomato	ResNet-50	90.5	+9.2	+19.7	98.7	4.0	No
Tomato	EffNet-B3	90.0	+8.5	+18.8	98.7	6.5	No
Tomato	MobNetV3-S	91.2	+5.4	+12.7	99.3	12.2	No
Tomato	ViT-S/16	92.5	+3.9	+8.9	99.5	9.5	No
Rice	ResNet-50	95.6	−1.5	−0.3	99.6	1.6	No
Rice	EffNet-B3	100	+1.8	+2.9	99.7	1.6	No
Rice	MobNetV3-S	97.5	−0.3	+0.8	99.6	4.6	Yes
Rice	ViT-S/16	98.8	+0.2	+0.9	100.0	4.5	No

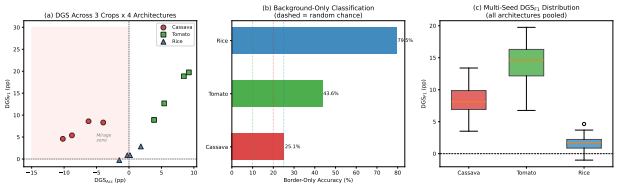


Figure 7: Cross-dataset DGS validation. Left: DGS_{Acc} vs. DGS_{F1} scatter (cassava points in mirage quadrant). Center: DGS_{F1} by dataset and architecture. Right: background perturbation sensitivity. The mirage is cassava-specific; DGS_{F1} correctly diagnoses degradation on all three crops.

Table 12: Theoretical mirage threshold p^* vs. empirical observation.

Arch.	CMD Rec.	Other Rec.	p^*	Match
ResNet-50	67.6%	54.6%	−1.0%	✓
EffNet-B3	67.9%	51.8%	25.9%	✓
MobNetV3-S	59.1%	51.4%	13.9%	✓
ViT-S/16	70.3%	55.3%	−3.7%	✓

worth 16x16 words: Transformers for image recognition at scale. In *ICLR 2021*, 2021.

[Ferentinos, 2018] Konstantinos P. Ferentinos. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145:311–318, 2018.

[Gulrajani and Lopez-Paz, 2021] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR 2021*, 2021.

[Guo *et al.*, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML 2017*, pages 1321–1330, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR 2016*, pages 770–778, 2016.

[Howard *et al.*, 2019] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for MobileNetV3. In *ICCV 2019*, pages 1314–1324, 2019.

[Hughes and Salathé, 2015] David P. Hughes and Marcel Salathé. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015.

[Johnson and Khoshgoftaar, 2019] Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, 2019.

[Kamilaris and Prenafeta-Boldú, 2018] Andreas Kamilaris and Francesc X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018.

[Koh *et al.*, 2021] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *ICML 2021*, pages 5637–5664, 2021.

[Lapuschkin *et al.*, 2019] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10:1096, 2019.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV 2017*, pages 2980–2988, 2017.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021.

[Mitchell *et al.*, 2019] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 220–229, 2019.

[Mohanty *et al.*, 2016] Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419, 2016.

[Mwebaze *et al.*, 2019] Ernest Mwebaze, Timnit Gebru, Andrea Frome, Solomon Nsumba, and Jeremy Tusubira. iCassava 2019 fine-grained visual categorization challenge. *arXiv preprint arXiv:1908.02900*, 2019.

[Nakatumba-Nabende *et al.*, 2023] Joyce Nakatumba-Nabende, Andrew Katumba, Claire Babirye, Jeremy Francis Tusubira, Godliver Owomugisha, Neema Mduma, Darlington Akogo, and Blessing Sibanda. Machine learning datasets for crop pest and disease diagnosis in Sub-Saharan Africa. *Lacuna Fund Agriculture Dataset*, 2023.

[Noyan, 2022] Mehmet Alican Noyan. Uncovering bias in the PlantVillage dataset. *arXiv preprint arXiv:2206.04374*, 2022.

[Nyalala *et al.*, 2026a] Innocent Nyalala, Patrick Vincent, and Khamis Haji. Auditing representation in crop disease benchmark datasets intended for Sub-Saharan African deployment. In *Data Science Africa (DSA) 2026*, 2026.

[Nyalala *et al.*, 2026b] Innocent Nyalala, Patrick Vincent, and Khamis Haji. When synthetic data hurts: DCGAN augmentation failure under extreme data scarcity for cassava disease classification in East Africa. In *Data Science Africa (DSA) 2026*, 2026.

[Nyalala *et al.*, 2026c] Innocent Nyalala, Patrick Vincent, Shivam Vyas, Twamaa Makame, Surabhi Gudla, and Khamis Haji. Few-shot transfer learning for cassava leaf disease detection under East African field conditions. In *Data Science Africa (DSA) 2026*, 2026.

[Özdemir and others, 2024] Caglar Özdemir et al. Evaluation pitfalls in machine learning: A taxonomy and practical guidance. *Machine Learning*, 113:1–30, 2024.

[Platt, 1999] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[Provost and Fawcett, 2000] Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2000.

[Rother *et al.*, 2004] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. GrabCut: Interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH 2004*, pages 309–314, 2004.

[Saerens *et al.*, 2002] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

[Selvaraju *et al.*, 2017] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV 2017*, pages 618–626, 2017.

[Tan and Le, 2019] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML 2019*, pages 6105–6114, 2019.

[Thakur *et al.*, 2022] Prabira Sethy Thakur, Pritee Khanna, Tanmay Sheorey, and Aparajita Ojha. Trends in vision-based machine learning techniques for plant disease identification: A systematic review. *Expert Systems with Applications*, 208:118117, 2022.

[Torralba and Efros, 2011] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011.

[Valverde-Albacete *et al.*, 2024] Francisco José Valverde-Albacete, Carmen Peláez-Moreno, André R. S. Gonçalves, et al. The AgrI-Challenge: A benchmark for cross-domain generalization in agriculture. In *ECCV 2024 Workshops*, 2024.

[Vincent and Nyalala, 2026] Patrick Vincent and Innocent Nyalala. Unifying perspectives on learning biases: A data-centric intervention for holistic fairness, robustness, and generalization. In *ICLR 2026 Workshop on Principled Design for Trustworthy AI*, 2026.

[Wang *et al.*, 2021] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR 2021*, 2021.

[Xu *et al.*, 2024] Mingle Xu, Ji-Eun Park, Jaehwan Lee, Jucheng Yang, and Sook Yoon. Plant disease recognition datasets in the age of deep learning: Challenges and opportunities. *Frontiers in Plant Science*, 15:1452551, 2024.

[Zhang *et al.*, 2024a] Jiayu Zhang, Baohua Zhang, Zixuan Chen, Innocent Nyalala, Kunjie Chen, and Junfeng Gao. A salient feature establishment tactic for cassava disease recognition. *Artificial Intelligence in Agriculture*, 14:115–132, 2024.

[Zhang *et al.*, 2024b] Jiayu Zhang, Baohua Zhang, Chao Qi, Innocent Nyalala, Peter Mecha, Kunjie Chen, and Junfeng Gao. MAIANet: Signal modulation in cassava leaf disease classification. *Computers and Electronics in Agriculture*, 225:109351, 2024.

[Zhang *et al.*, 2025] Jiayu Zhang, Baohua Zhang, Innocent Nyalala, Peter Mecha, Junlong Chen, Kunjie Chen, and Junfeng Gao. Rectifying the extremely weakened signals for cassava leaf disease detection. *Computers and Electronics in Agriculture*, 232:110107, 2025.

[Zhao *et al.*, 2025] Jiuqing Zhao, Jie Fang, Junfeng Zhang, Shu Li, and Jie Lu. Environment-robust plant disease recognition with pixel-level foreground isolation. *Computers and Electronics in Agriculture*, 230:109872, 2025.