GMR
The Original

CROSS-DOMAIN ROBUSTNESS OF COTTON LEAF CURL DISEASE CLASSIFICATION

Renu[1]*, Dr. Neha Gaba[2]

[1]*Research Scholar, Department of Computer Science and Applications, Baba Mastnath University, Rohtak, India, rkbaghel09876@gmail.com
[2]Associate Professor, Department of computer science and application, Baba Mastnath University, Rohtak, India, Nehagaba05@gmail.com
Corresponding author: rkbaghel09876@gmail.com

ABSTRACT
High internal performance does not guarantee that a plant-disease classifier will behave similarly under new backgrounds, lighting, cameras or visually overlapping conditions. This paper evaluates the frozen cross-domain robustness of eight Cotton Leaf Curl Disease (CLCuD) classifier configurations developed under a common ResNet34 training protocol. External evaluation used 257 Healthy Leaf and 400 Curl Virus original field images from SAR-CLD-2024 after exact and perceptual overlap auditing against the 3,462-image internal cohort. Twenty-three available checkpoints were evaluated without retraining, external threshold fitting or calibration under raw and automatically leaf-isolated conditions. Adaptive disease-preserving enhancement (A1) achieved the highest seed-ensemble balanced accuracy in both settings: 0.802 raw (95% source-group bootstrap CI 0.770–0.834) and 0.817 isolated (0.787–0.847). A prespecified stress extension added all 1,449 images from five other repository conditions, producing 2,106 images and 96,876 checkpoint predictions. Best full-cohort balanced accuracy was 0.597 raw and 0.620 isolated. Herbicide Growth Damage was the strongest category-specific confounder. Quantitative attribution and cue-intervention audits showed that stronger-looking foreground localization did not necessarily imply stronger cross-domain classification. The study provides a frozen, overlap-audited account of acquisition shift and class-boundary stress, identifies bounded adaptive preprocessing as the most consistent component in the two-class field comparison, and establishes explicit multiclass differential-diagnosis supervision as the next system requirement.

KEYWORDS: Cotton Leaf Curl Disease; domain shift; external validation; hard negatives; field images; robustness; explainable artificial intelligence

1. INTRODUCTION

Image classifiers for plant disease are often developed on cleaned or visually controlled collections. Their operating behaviour can change when photographs come from another source with different backgrounds, lighting, devices, framing or disease prevalence [11–12]. For Cotton Leaf Curl Disease (CLCuD), this problem is especially relevant because curling, margin deformation and colour or texture changes can also appear in other diseases, pest damage and abiotic stress.

MorphoReli-CLCuD was developed as a family of controlled ResNet34-based configurations containing adaptive disease-preserving enhancement, pseudo-vein guidance and morphology-reliability fusion. The companion architecture study evaluates these components on a locked internal cohort. The present paper asks a separate question: how do the frozen checkpoints behave when moved unchanged to an external field-image source and then challenged with visually overlapping cotton conditions?

The study addresses three questions:

- **RQ1:** Which frozen model components transfer most consistently to external raw and automatically leaf-isolated Healthy Leaf versus Curl Virus images?
- **RQ2:** How much do preprocessing condition and training seed affect the operating point under acquisition shift?
- **RQ3:** Which other cotton conditions produce CLCuD-positive calls, and what do quantitative XAI and cue interventions reveal about the relationship between localization and prediction?

The contribution is an evaluation protocol rather than a new training claim. It combines exact and perceptual overlap auditing, conservative source grouping, fixed-threshold raw and isolated evaluation, checkpoint-level seed analysis, a complete hard-negative extension and quantitative attribution with input interventions. No external image is used to retrain a checkpoint or select its threshold.

Figure 1 summarizes the frozen external evaluation protocol.

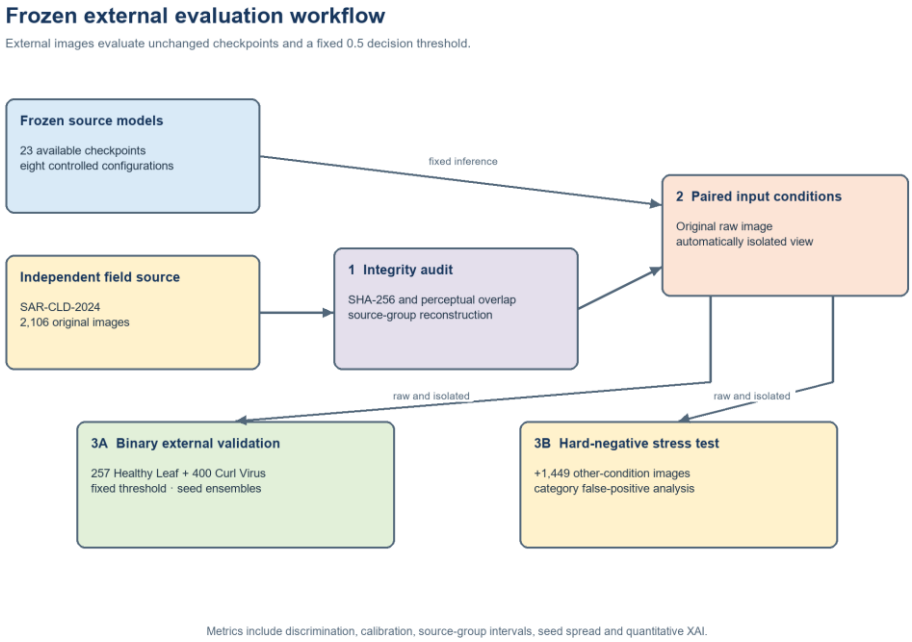


Figure 1. Frozen external evaluation workflow. Twenty-three available checkpoints and the independent SAR-CLD-2024 images enter an overlap-audited protocol without retraining, external threshold selection or calibration. Raw and automatically isolated inputs are evaluated in the 657-image binary comparison and the complete 2,106-image hard-negative stress cohort.

1.1 Abbreviations and operational terms

CLCuD denotes Cotton Leaf Curl Disease. ADPE-Lite is adaptive disease-preserving enhancement; VAAM is the vein-aware attention module; MorphoReli denotes morphology-reliability fusion; CNN denotes convolutional neural network; RGB denotes red-green-blue input; and XAI denotes explainable artificial intelligence. Grad-CAM++ denotes Gradient-weighted Class Activation Mapping++, and IG denotes Integrated Gradients. SHA-256 is Secure Hash Algorithm 256; pHash and dHash denote perceptual hash and difference hash; DOI denotes digital object identifier; U²-Net is a nested U-structured segmentation network and U2NetP is its compact implementation; and AdamW is Adam optimization with decoupled weight decay. BA is balanced accuracy; AUROC is area under the receiver operating characteristic curve; AP is average precision; FPR is false-positive rate; CI is confidence interval; SD is standard deviation; and ECE is expected calibration error. SAR-CLD-2024 is the external cotton-leaf dataset name used by its publisher. A hard negative is a non-CLCuD abnormal leaf that can be mistaken for CLCuD. The F1 configuration label is distinct from the F1 classification metric.

2. Related Work and Evaluation Rationale

Direct CLCuD and cotton-disease studies have established CNN, Transformer, attention, fusion and explainability approaches [1–10,16–19]. SAR-CLD-2024 has also supported seven-class cotton recognition studies [15,17–18]. Results from different datasets should not be treated as a single leaderboard because source images, class definitions, augmentation, splitting and test protocols differ.

Leakage is a central concern when near-duplicate images or repeated captures can cross dataset partitions [11]. Plant-disease models can also lose performance when transferred from controlled imagery to field conditions [12]. External evaluation should therefore verify source separation, preserve the trained model and report acquisition-realistic inputs rather than silently optimizing on the external answers.

Attribution methods such as Grad-CAM++ and Integrated Gradients provide complementary views of model influence [13–14]. A plausible heatmap is not proof of correct diagnosis or biological reasoning. The present study measures foreground and proxy-region attribution and pairs those measurements with foreground-only, background-only and structural-region interventions.

3. MATERIALS AND METHODS

The evaluated checkpoints came from a separate 3,462-image internal healthy-versus-CLCuD development study using a locked reconstructed-group split and identical optimization budgets. That study assessed

architecture development and internal validation for the same model family. The present manuscript uses the resulting frozen checkpoints only for external and hard-negative evaluation and makes no new training claim.

3.1 Frozen source models and checkpoints

Table 1 defines the eight frozen model configurations and the controlled purpose of each comparison.

Table 1. Controlled model matrix.

ID	Configuration	Controlled purpose
B0	ResNet34 classifier	Standard diagnosis baseline
A1	B0 + ADPE-Lite	Adaptive preprocessing contribution
V1	B0 + VAAM and pseudo-vein loss	Pseudo-venation contribution
C0	Four-branch gated architecture with diagnosis loss only	Parameter-matched architecture control
C1	Four morphology branches with concatenation and structural losses	Fusion reference
C2	C1 + learned quality-gated fusion	Correctly paired pseudo-morphology supervision
P1	C2 architecture with cross-source-group mismatched structural targets	Supervision-specificity control
F1	ADPE-Lite + VAAM + MorphoReli fusion	Integrated proposed configuration

Note. Every identifier is evaluated with the same locked partitions, optimization budget and checkpoint-selection rule.

All eight configurations use the same locked groups, data loading, optimizer, training-transformation policy and validation-based checkpoint rule. Inputs are resized to 224 x 224 pixels. ADPE-Lite operates in [0,1] before ImageNet normalization, and each geometric transformation is applied identically to RGB images and structural pseudo-labels. AdamW optimization uses an initial learning rate of 3 x 10^-4, weight decay of 10^-4, batch size 32 and ImageNet-pretrained encoders. Each configuration is trained for 12 epochs at seeds 20260908, 20260909 and 20260910; the epoch with maximum validation macro-F1 supplies the test prediction for that run. The diagnosis loss uses inverse-frequency class weights $N/(2N_c)$, where N_c is the number of training images in class c. A cosine-annealing schedule spans the 12 epochs.

3.2 External field-image validation

External evaluation used original images from SAR-CLD-2024, a seven-condition cotton-leaf dataset collected in the field [15]. The binary cohort comprised all 257 available Healthy Leaf originals and a deterministic natural-filename-order sample of 400 of the 431 Curl Virus originals. Repository categories are treated as external dataset labels rather than diagnoses independently adjudicated by this study. All 657 images decoded at 800 x 800 pixels and passed SHA-256 verification. No exact SHA-256 or joint perceptual-hash (pHash)/difference-hash (dHash) near-overlap candidate at Hamming distance at most four was found against the 3,462-image internal cohort. Because the repository provides no leaf, plant or capture-session identifier, perceptually similar external images were joined conservatively, yielding 650 source groups for resampling. This grouping reduces, but cannot prove elimination of, repeated-subject dependence.

The training images have black backgrounds, whereas the external originals are field photographs. Two conditions were therefore locked before inference. The raw condition retained each original RGB image and used an aligned estimated foreground mask only for morphology inputs and analysis. The isolated condition placed the estimated leaf foreground on black, square-padded it and resized it to 180 x 180 before the model's standard 224 x 224 resize. Foregrounds were estimated with the compact U²-Net model U2NetP. Masks below 5% image area triggered a centred GrabCut and green-region fallback; the final automated audit contained no mask below 5%. These masks and derived boundary, distance and pseudo-vein maps are unreviewed preprocessing proxies, not expert segmentations.

The frozen evaluation collection comprised three checkpoints for B0, A1, V1, C0, C1, C2 and P1. F1 was evaluated at seeds 20260908 and 20260910. The completed F1 seed-20260909 run retains its training history, validation/test predictions and internal metrics, but its selected best_model.pt file is absent from the transferred archive; it is therefore excluded from external re-inference and F1 external summaries are explicitly two-checkpoint descriptive ensembles. All were evaluated at the unchanged 0.5 threshold without external fitting.

The primary comparison comprised B0, C0, C2 and P1; the other configurations were prespecified secondary ablations. Probabilities were averaged across available seeds to form each ensemble. Balanced accuracy, macro-F1, sensitivity, specificity, AUROC, average precision, Brier score and 10-bin expected calibration error were calculated. Ninety-five percent intervals resampled the 650 external source groups. Paired group bootstraps compared C2 with B0, C0 and P1 and A1 with B0. A deterministic 12-image subset, balanced by repository class and sampled evenly in filename order, received Grad-CAM++, eight-step Integrated Gradients and cue-removal analysis for B0 and C2 at seed 20260908 in both conditions. This small XAI audit is descriptive and was not used for model selection.

3.3 Cross-disease hard-negative stress test

The hard-negative stress test retained the exact 257 Healthy Leaf and 400 Curl Virus anchors used in Section 3.2 and added every original image from the five remaining SAR-CLD-2024 categories: 250 Bacterial Blight, 280 Herbicide Growth Damage, 225 Leaf Hopper Jassids, 578 Leaf Redding and 116 Leaf Variegation images. “Leaf Redding” preserves the repository label. The resulting frozen cohort contained 2,106 images: 400 binary positives and 1,706 binary negatives, of which 1,449 were hard negatives. Repository categories are not treated as new expert diagnoses. The independently transferred source archive passed its cyclic-redundancy integrity check, and all 657 previously verified anchor files matched byte-for-byte by SHA-256. The original SAR-CLD-2024 digital object identifier (DOI) and Creative Commons Attribution 4.0 provenance remain the source record. The overlap and grouping protocol from Section 3.2 was rerun on the extended cohort. No exact SHA-256 or joint pHash/dHash candidate at Hamming distance at most four was found against the internal images. Conservative perceptual grouping produced 2,095 external source groups. The previously prepared raw and leaf-isolated derivatives and predictions were reused for the 657 anchors; the same locked preprocessing generated corresponding derivatives for the 1,449 new images. No image was removed after model output was examined.

All 23 available frozen checkpoints were evaluated without retraining, calibration or threshold selection. Seed probabilities were averaged by model, and the threshold remained 0.5. The main endpoint was false-positive rate on the 1,449 hard negatives, reported together with sensitivity, specificity and balanced accuracy on the full binary cohort. Category-specific positive rates identified confounders. Two thousand bootstrap replicates resampled the 2,095 source groups, and paired group bootstraps compared A1 and C2 with B0. A focused 11-image subset per preprocessing condition received Grad-CAM++, eight-step Integrated Gradients and cue-removal probes for B0, A1 and C2 at seed 20260908. The subset deliberately included hard-negative false positives, low-confidence Curl Virus cases and correctly rejected hard negatives. It is a descriptive targeted stress audit and was not used to choose a model or threshold.

4. RESULTS

4.1 External field-image validation

The external cohort contained 257 healthy and 400 Curl Virus images in 650 reconstructed source groups. No exact or joint perceptual overlap candidate was found against the internal cohort. Table 2 and Figure 2 report frozen seed-ensemble performance. A1 achieved the highest balanced accuracy in both conditions: 0.802 on raw field images (95% source-group bootstrap CI 0.770–0.834) and 0.817 after leaf isolation (0.787–0.847). B0 ranked second in the isolated condition (0.814, 0.785–0.845) and third on raw images (0.759, 0.724–0.792). C1 ranked second on raw images at 0.766 (0.735–0.798).

Table 2. External SAR-CLD-2024 seed-ensemble operating-point performance in Panel A (raw) and Panel B (isolated) at the unchanged 0.5 threshold.

Panel A Raw condition

Model	Raw balanced accuracy	Raw sensitivity	Raw specificity
B0	0.759	0.860	0.658
A1	**0.802**	0.795	0.809
V1	0.693	0.460	0.926
C0	0.713	0.478	0.949
C1	0.766	0.625	0.907
C2	0.696	0.443	0.949
P1	0.731	0.513	0.949
F1	0.727	0.505	0.949

Panel B Isolated condition

Model	Isolated balanced accuracy	Isolated sensitivity	Isolated specificity

B0	0.814	0.810	0.817
A1	**0.817**	0.813	0.821
V1	0.747	0.545	0.949
C0	0.773	0.608	0.938
C1	0.790	0.685	0.895
C2	0.756	0.590	0.922
P1	0.787	0.640	0.934
F1	0.702	0.450	0.953

Note. F1 averages seeds 20260908 and 20260910; every other row averages three. Bold marks the largest balanced accuracy in each condition. Complete discrimination, calibration and interval estimates are in the accompanying machine-readable results.

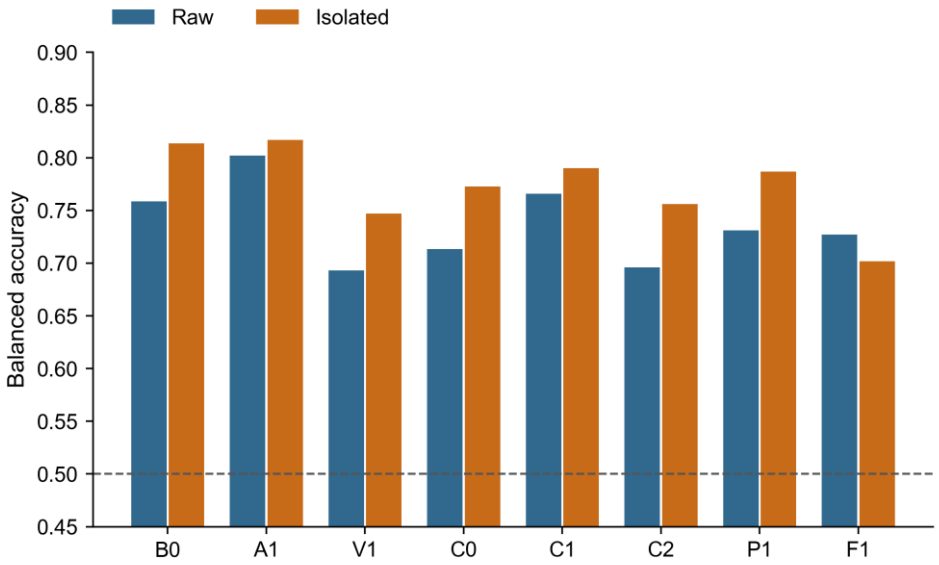


Figure 2. Balanced accuracy for all frozen seed ensembles on original raw field photographs and estimated leaf-isolated photographs. The dashed line marks chance balanced accuracy. Isolation is a controlled preprocessing analysis rather than an independently acquired cohort.

External C2 operating behavior differed from the internal comparison. Relative to B0, its balanced accuracy was lower by 0.0627 on raw images (paired group-bootstrap 95% CI -0.0996 to -0.0256) and by 0.0573 after isolation (-0.0869 to -0.0283). It was also lower than P1 in both conditions: -0.0349 (-0.0570 to -0.0132) raw and -0.0310 (-0.0527 to -0.0104) isolated. The raw C2-C0 difference was -0.0175 (-0.0340 to -0.0013), while its isolated difference of -0.0166 (-0.0359 to 0.0022) was inconclusive. The difference was driven chiefly by disease sensitivity: C2 detected 44.3% of Curl Virus images raw and 59.0% isolated while retaining specificity above 92%. In contrast, A1 exceeded B0 on raw images by 0.0431 (0.0170–0.0704) and was indistinguishable after isolation, 0.0032 (-0.0185 to 0.0253).

Checkpoint-level variation was substantial (Figure 3). Raw C2 balanced accuracy ranged from 0.612 to 0.816 across seeds, and isolated C2 ranged from 0.698 to 0.789. P1 showed similarly wide ranges. A1 was more stable: 0.740–0.778 raw and 0.778–0.829 isolated. Depending on model, leaf isolation changed 85–139 of 657 ensemble decisions, so its effect was neither small nor uniformly beneficial. The results therefore support reporting both acquisition-realistic and harmonized conditions rather than selecting the condition that favours one architecture.

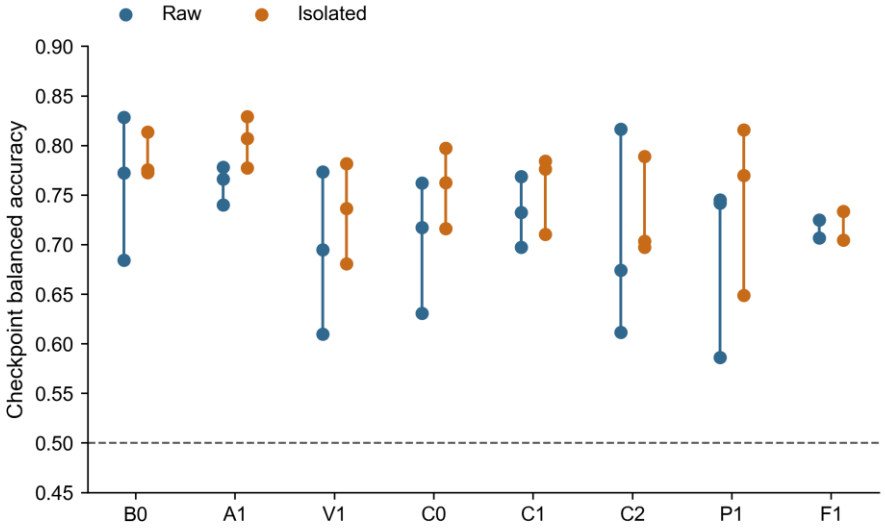


Figure 3. Balanced accuracy for every available frozen checkpoint in each external condition. Vertical segments show the minimum-to-maximum seed range; points are individual checkpoints. The F1 external summary contains seeds 20260908 and 20260910, as specified in Section 3.2.

On the deterministic 12-image XAI subset, C2 placed more Grad-CAM++ attribution inside estimated leaf masks than B0 both raw (0.566 versus 0.375) and isolated (0.667 versus 0.378). Integrated-Gradients leaf fractions were similar on raw images (0.559 versus 0.545) and nearly saturated for both isolated inputs (0.983 versus 0.983). Despite the stronger-looking C2 Grad-CAM++ localization, mean C2 disease probability was lower than B0 in both conditions (0.337 versus 0.620 raw; 0.385 versus 0.558 isolated). Background-only and region-replacement probes also produced condition-dependent shifts, including increased C2 probability for isolated background-only inputs. The small subset, estimated masks and unrealistic interventions make this evidence descriptive. It shows that stronger foreground localization can coexist with different external classification behavior and does not validate causal anatomical reasoning.

4.2 Cross-disease hard-negative stress test

The extended cohort contained 2,106 images in 2,095 reconstructed source groups. All 96,876 expected checkpoint predictions and 33,696 seed-ensemble predictions were present. Table 3 quantifies the binary operating boundary after the 1,449 hard negatives were added. C1 had the largest raw balanced accuracy, 0.597 (95% source-group bootstrap CI 0.571–0.623), while P1 had the largest isolated balanced accuracy, 0.620 (0.595–0.647). Their shift relative to the Healthy Leaf versus Curl Virus results in Table 2 specifies the need for multiclass differential-diagnosis supervision.

Table 3. Frozen seed-ensemble performance in Panel A (raw) and Panel B (isolated) on the full SAR-CLD-2024 cohort at threshold 0.5.

Panel A Raw condition

Model	Raw BA	Raw sensitivity	Raw specificity
B0	0.567	0.860	0.275
A1	0.563	0.795	0.331
V1	0.542	0.460	0.624
C0	0.562	0.478	**0.646**
C1	**0.597**	0.625	0.569
C2	0.529	0.443	0.615
P1	0.560	0.513	0.607
F1	0.557	0.505	0.609

Panel B Isolated condition

Model	Raw hard-negative FPR	Isolated BA	Isolated sensitivity	Isolated specificity	Isolated hard-negative FPR
B0	0.793	0.609	0.810	0.408	0.665
A1	0.754	0.606	0.813	0.400	0.675

V1	0.430	0.605	0.545	**0.666**	**0.384**
C0	**0.408**	0.595	0.608	0.583	0.480
C1	0.491	0.620	0.685	0.555	0.506
C2	0.444	0.585	0.590	0.579	0.482
P1	0.453	**0.620**	0.640	0.601	0.458
F1	0.451	0.558	0.450	0.665	0.386

Note. BA is balanced accuracy. Hard-negative FPR is the fraction of the 1,449 non-healthy, non-Curl Virus images called CLCuD. F1 uses seeds 20260908 and 20260910; all other ensembles use three. Bold marks each condition's largest balanced accuracy, specificity or lowest hard-negative FPR.

Figure 4 shows the category-specific positive-call rates that identify the strongest hard-negative confounders.

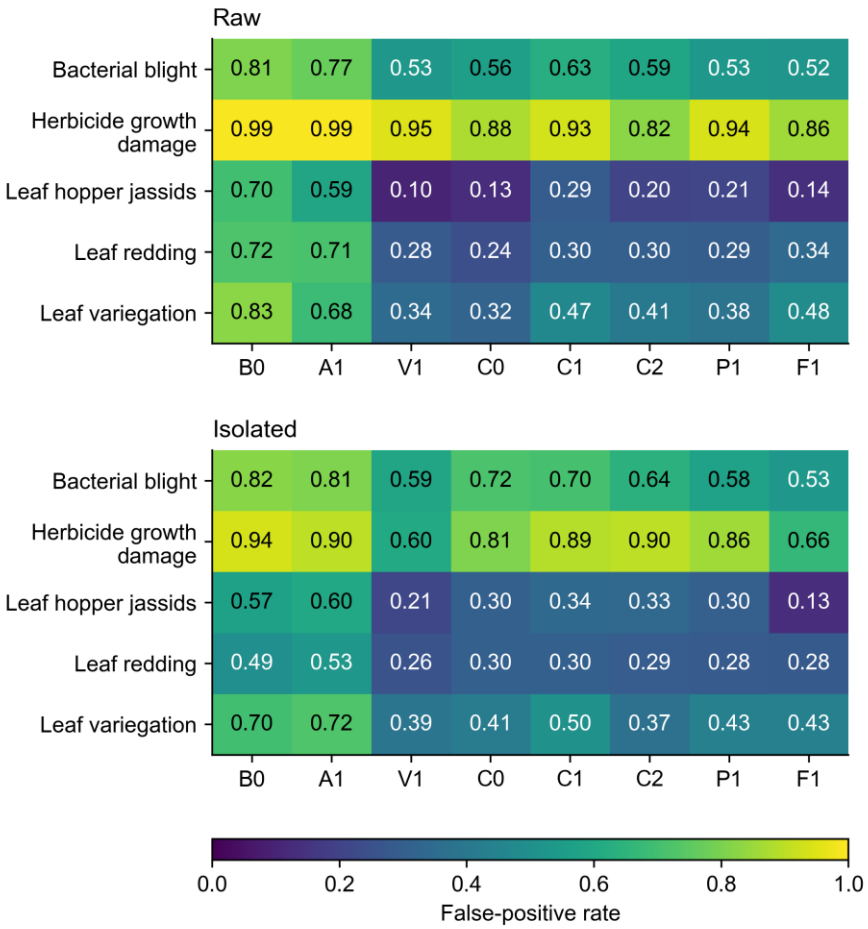


Figure 4. Seed-ensemble positive-call rate by repository category and preprocessing condition. For the five hard-negative rows, a positive call is a false positive. Repository labels are reproduced as supplied and are not independent diagnoses by this study.

Herbicide Growth Damage was the dominant confounder (Figure 4). Raw false-positive rates ranged from 0.825 for C2 to 0.989 for B0 and A1; isolated rates ranged from 0.604 for V1 to 0.936 for B0. Bacterial Blight, Leaf Hopper Jassids, Leaf Redding and Leaf Variegation also produced substantial model-dependent positive-call rates. A1's hard-negative false-positive rate was 0.754 raw and 0.675 isolated, compared with 0.793 and 0.665 for B0.

Figure 5 compares full-cohort balanced accuracy after inclusion of all five hard-negative categories.

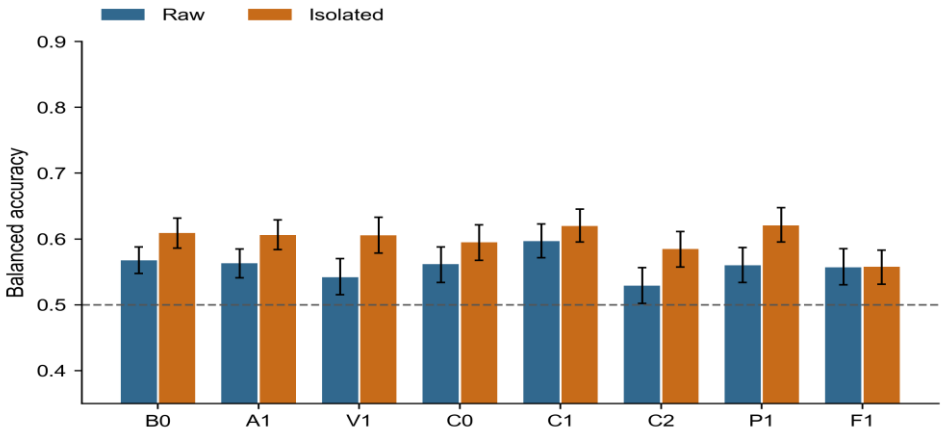


Figure 5. Full-cohort balanced accuracy with 95% source-group bootstrap intervals. All models use the fixed 0.5 threshold and the same 2,106 images.

C2 rejected more raw hard negatives than B0, reducing false-positive rate by 0.349 (paired group-bootstrap 95% CI 0.324–0.374), but its lower Curl Virus sensitivity produced a balanced-accuracy deficit of 0.0385 (-0.0674 to -0.0099). After isolation, C2 reduced hard-negative false positives by 0.183 (0.163–0.204) while remaining lower in balanced accuracy by 0.0244 (-0.0488 to -0.0002). A1 reduced raw false positives relative to B0 by 0.0393 (0.0206–0.0568) without a clear balanced-accuracy difference; neither A1 difference was conclusive after isolation. These operating-point trade-offs motivate the multiclass differential-diagnosis extension specified in Section 6.

Figure 6 summarizes foreground attribution for the focused hard-negative audit.

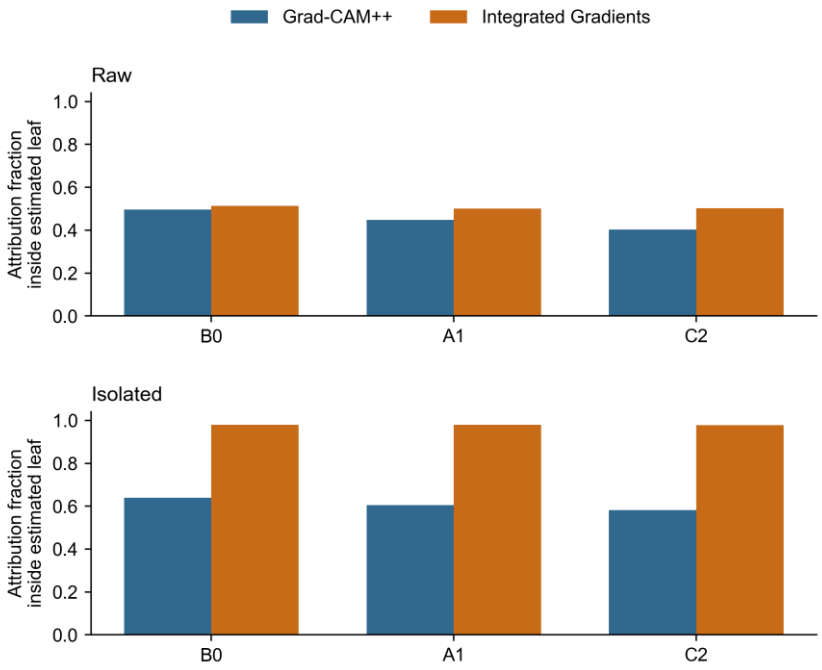


Figure 6. Mean Grad-CAM++ and Integrated-Gradients attribution inside the estimated leaf for the focused 11-image stress subset per condition. The isolated Integrated-Gradients fractions near one mainly reflect removal of the visible background and are not evidence of improved anatomical reasoning.

In the focused XAI audit, raw Grad-CAM++ leaf fractions were 0.496 for B0, 0.448 for A1 and 0.402 for C2; their Integrated-Gradients fractions were about 0.50. Isolation increased Grad-CAM++ leaf fractions and made Integrated Gradients almost entirely foreground by construction. Plausible leaf-centred maps therefore coexisted with high hard-negative positive-call rates. Background-only and foreground-only probes also gave model- and

condition-dependent probabilities, including an isolated background-only mean of 0.850 for C2. These observations identify shortcut sensitivity but do not establish causal anatomy.

5. DISCUSSION

The external binary comparison identifies a clear component-level pattern. A1, which adds bounded adaptive preprocessing to the common ResNet34 baseline, achieved the highest balanced accuracy in both raw and isolated conditions. Its raw advantage over B0 was positive under paired source-group resampling, while the isolated difference was small. This suggests that the adaptive front end was most useful when acquisition style differed visibly from the internal collection.

The morphology-gated configurations occupied more conservative external operating points, typically retaining high specificity while detecting fewer Curl Virus images at the fixed threshold. Checkpoint-level ranges show that seed variation was substantial for several configurations. Reporting seed ensembles and individual checkpoint spread is therefore more informative than presenting one selected external run.

The hard-negative extension answers a different question from the 657-image binary comparison. The frozen models were trained only to separate healthy from CLCuD; they were not trained to name Bacterial Blight, Herbicide Growth Damage, Leaf Hopper Jassids, Leaf Redding or Leaf Variegation. The resulting false-positive patterns identify the alternative conditions that future differential-diagnosis training must represent explicitly. Herbicide Growth Damage was the strongest confounder, consistent with visible deformation overlapping the binary decision boundary.

The XAI findings caution against treating attractive localization as a performance guarantee. In the external subset, stronger leaf-centred Grad-CAM++ attribution could coexist with lower disease probability or incorrect operating behaviour. Integrated Gradients and Grad-CAM++ also differed in within-leaf attribution. These observations support using explanations as measured behavioural evidence together with prediction metrics and interventions.

The paper's main contribution is therefore not a claim of universal external accuracy. It is a frozen, overlap-audited description of how preprocessing, morphology-aware branches, seed variation and alternative cotton conditions affect transfer. The results provide concrete design requirements for the next-stage system: explicit alternative classes, domain-diverse training, expert-reviewed structural targets and independent group-held-out field evaluation.

6. Scope and Future Work

SAR-CLD-2024 supplies useful original field photographs and seven repository categories, but the categories share one repository and lack physical leaf, plant and capture-session identifiers. Conservative perceptual grouping reduces repeated-subject dependence without proving its elimination. Raw and isolated conditions are paired transformations of the same images, and the automatically estimated masks are not expert segmentations.
All analyses retain the original 0.5 threshold and frozen checkpoints. This protects the evaluation from external tuning but also means that the study does not estimate the performance of a model adapted specifically to the external domain. The hard-negative cohort tests a binary operating boundary rather than a trained seven-class classifier. A future multiclass or hierarchical system should model healthy, CLCuD, other disease, pest damage and abiotic deformation explicitly, with development and test sources kept separate.

7. CONCLUSION

Frozen cross-domain evaluation showed that ADPE-Lite produced the most consistent two-class transfer among the evaluated components, reaching balanced accuracy of 0.802 on raw field images and 0.817 after automatic leaf isolation. Preprocessing condition and training seed materially affected operating behaviour. The complete hard-negative audit identified Herbicide Growth Damage and other cotton abnormalities as important CLCuD-like confounders, while quantitative XAI showed that foreground localization alone does not guarantee correct transfer. These findings define the data and supervision requirements for a morphology-aware multiclass field diagnostic system without using external images to retrain or tune the evaluated checkpoints.

8. Data Availability

SAR-CLD-2024 is available under Creative Commons Attribution 4.0 through dataset DOI 10.17632/b3jy2p6k8w.2 [15]. The internal field-image collection and frozen checkpoints are retained by the study team; their final access conditions and repository or archive identifiers will be stated in the journal-specific data-availability declaration before submission.

9. Code and Reproducibility

The retained evaluation archive contains the external file manifest, SHA-256 digests, overlap-audit outputs, reconstructed source groups, raw and isolated derivatives, checkpoint-level predictions, seed ensembles, bootstrap summaries, XAI records and hard-negative category results. External evaluation uses 23 available checkpoints. The completed F1 seed-20260909 run retains its internal history, predictions and metrics, but its selected checkpoint file is absent; consequently, exact external re-inference for that run is unavailable and the reported F1 external ensembles use the two available checkpoints. Reproduction requires the corresponding source images, available frozen checkpoints, saved manifests and software dependencies.

REFERENCES

[1] A. Magsi et al., "A novel framework for disease severity level identification of cotton plant using machine learning techniques," *International Journal of Scientific & Technology Research*, vol. 10, no. 5, pp. 367–373, 2021.

[2] R. Nazeer et al., "Detection of cotton leaf curl disease's susceptibility scale level based on deep learning," *Journal of Cloud Computing*, vol. 13, 50, 2024. doi:10.1186/s13677-023-00582-9.

[3] N. Ahmad, S. Qadri, and N. Akhtar, "A multimodal fusion framework to diagnose cotton leaf curl virus using machine vision techniques," *Cogent Food & Agriculture*, vol. 10, 2339572, 2024. doi:10.1080/23311932.2024.2339572.

[4] S. K. Noon et al., "Computationally light deep learning framework to recognize cotton leaf diseases," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 6, pp. 12383–12398, 2021. doi:10.3233/JIFS-210516.

[5] S. Otiya, P. Faldu, and P. Goel, "Cotton leaf disease classification using deep convolution neural network with explainable AI," in *ICSCNA*, 2023. doi:10.1109/ICSCNA58489.2023.10370214.

[6] G. Biju et al., "Identification of Cotton Leaf Curl Disease Using CNN and Vision Transformer," in *Intelligent Systems and Applications*, pp. 670–688, Springer, 2024. doi:10.1007/978-3-031-47724-9_44.

[7] N. Sabari et al., "MultiCotNet: a novel multispatial attention-based deep learning architecture for cotton leaf disease classification," *Journal of Cotton Research*, vol. 9, 29, 2026. doi:10.1186/s42397-026-00276-y.

[8] N. Ullah, I. De Falco, and G. Sannino, "Cotton leaf disease detection via dual-backbone CNN–Transformer fusion with quantitative XAI comparison," *Electronics*, vol. 15, no. 16, 3650, 2026. doi:10.3390/electronics15163650.

[9] T. Chen et al., "The intelligent detection method for similar cotton diseases based on high- and low-frequency feature enhancement and attention-guided fusion," *Engineering Applications of Artificial Intelligence*, vol. 168, 114099, 2026. doi:10.1016/j.engappai.2026.114099.

[10] S. Naveena and K. Kavitha, "EnsembleNet model: A hybrid approach for classifying *Gossypium herbaceum* leaf diseases using image feature analysis," *Engineering Applications of Artificial Intelligence*, vol. 164, 113225, 2026. doi:10.1016/j.engappai.2025.113225.

[11] S. Kaufman et al., "Leakage in data mining: formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 4, Article 15, 2012. doi:10.1145/2382577.2382579.

[12] A. Ahmad, A. El Gamal, and D. Saraswat, "Toward generalization of deep learning-based plant disease identification under controlled and field conditions," *IEEE Access*, vol. 11, pp. 9042–9057, 2023. doi:10.1109/ACCESS.2023.3240100.

[13] A. Chattopadhay et al., "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *WACV*, 2018. doi:10.1109/WACV.2018.00097.

[14] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *ICML*, PMLR 70, pp. 3319–3328, 2017.

[15] P. Bishshash, M. A. S. Nirob, M. H. Shikder, A. Sarower, T. Bhuiyan, and S. R. H. Noori, "A comprehensive cotton leaf disease dataset for enhanced detection and classification," *Data in Brief*, vol. 57, 110913, 2024. doi:10.1016/j.dib.2024.110913. Dataset: SAR-CLD-2024, doi:10.17632/b3jy2p6k8w.2.

[16] M. E. Haque et al., "An attention enhanced CNN ensemble for interpretable and accurate cotton leaf disease classification," *Scientific Reports*, vol. 16, 4476, 2026. doi:10.1038/s41598-025-34713-w.

[17] S. Saha, M. J. Israel, and M. Bhattacharya, "Deep Inception: Cotton Leaf Disease Classification with Convolutional Precision Neural Network," in *Data Science and Exploration in Artificial Intelligence*, CCIS 2690, pp. 309–314, Springer, 2026. doi:10.1007/978-3-032-19321-6_28.

[18] S. M. M. R. Swapno et al., "Explainable transformer framework for fast cotton leaf diagnostics and fabric defect detection," *iScience*, vol. 29, no. 2, 114411, 2026. doi:10.1016/j.isci.2025.114411.

[19] M. Naeem et al., "Automatic prediction of cotton leaf's diseases using deep learning techniques," *Scientific Reports*, online ahead of print, 2026. doi:10.1038/s41598-026-60523-9.