

Leakage-Aware Ablation of Data Augmentation for Sugarcane Leaf Disease Classification: Operator- and Severity-Specific Effects

Jarot Budiasto¹, Hasanudin Jayawardana², Tri Kustanti Rahayu³, Yosefina Mangera⁴

^{1,2,3}Department of Information Systems, Faculty of Engineering, Universitas Musamus, Merauke, Indonesia

⁴Department of Agricultural Engineering, Faculty of Agriculture, Universitas Musamus, Merauke, Indonesia

Received:

May 2, 2026

Revised:

September 5, 2026

Accepted:

September 21, 2026

Published:

October 24, 2026

Corresponding Author:

Author Name*:

Jarot Budiasto

Email*:

jarot@unmus.ac.id

DOI:

10.63158/journalisi.v8i5.1896

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)



Abstract. Conventional image augmentation is widely used to improve model generalization; however, its benefit in fine-grained plant disease classification can be overstated when duplicated or near-duplicated images cross the train-test boundary. This study tests whether conventional augmentation remains beneficial for sugarcane leaf disease classification after controlling duplicate leakage. The original Sugarcane Leaf Dataset contained 6,748 images and was audited using MD5, perceptual hash (pHash), and structural similarity index measure (SSIM), yielding 5,389 curated images. A leakage-aware stratified 70/15/15 train-validation-test split evaluated 13 augmentation configurations using ImageNet-pretrained MobileNetV2 across three random seeds, totaling 39 runs. Macro-F1 performance was assessed through paired stratified bootstrap analysis and an exact Page test. The baseline attained the highest mean Macro-F1 (0.9646 ± 0.0068), while photometric augmentation produced comparable performance (0.9643 ± 0.0079). Geometric policies generally reduced performance, and combined augmentation achieved the lowest Macro-F1 (0.8029). The Page test indicated an overall negative ordered trend, although the medium recipe outperformed the light recipe. These results demonstrate that augmentation policies must be validated by operator and severity level, rather than enabled by default in plant disease recognition.

Keywords: Sugarcane leaf disease classification, data augmentation, dataset leakage, image deduplication, MobileNetV2.

1. INTRODUCTION

Sugarcane is an economically important crop for sugar, ethanol, and agricultural by-products. Its productivity and quality can be reduced by foliar diseases with overlapping visual manifestations, including discoloration, streaks, spots, pustules, and broader changes in leaf appearance. Manual field diagnosis remains dependent on expert availability and may be inconsistent when symptoms are subtle, mixed, or observed under variable illumination. Image-based deep learning has consequently become an increasingly investigated route for rapid and scalable disease screening [1], [2], [3], [4], [5], [6].

Recent sugarcane studies have largely emphasized architecture selection, transfer learning, attention mechanisms, ensemble models, and mobile deployment. High-capacity convolutional networks have reported strong classification performance on the 11-class Sugarcane Leaf Dataset (SLD) [2], while MobileNet variants and attention-based systems have demonstrated computational efficiency and deployment potential [3], [4]. MobileNetV2 is particularly attractive for resource-constrained applications because its inverted residual blocks and linear bottlenecks balance representational capacity with computational cost [7].

Data augmentation is commonly incorporated to increase apparent sample diversity and reduce overfitting. Standard operations include flips, rotation, zoom, translation, and photometric jitter [8], [9], [10]. However, an augmentation is valid only when the transformed image remains compatible with its assigned label and approximates variation that can plausibly occur at inference time. Strong or semantically inappropriate transformations can disrupt salient regions, create label noise, or impose invariances unsupported by the task [11], [12], [13], [14]. This concern is especially relevant to fine-grained plant disease recognition, where class boundaries may depend on localized lesions, spatial arrangement, orientation, leaf shape, edge structure, or subtle color and texture differences [15], [16]. Recent plant-disease augmentation research increasingly preserves or targets discriminative regions rather than assuming that random diversity is always beneficial [17], [18], [19].

A second concern is dataset quality. Duplicate and near-duplicate images can violate training-test independence, inflate evaluation estimates, and obscure the effect of experimental interventions [20], [21], [22], [23]. Similar problems have motivated curated benchmark recommendations in other image domains [24]. Because an augmentation study often interprets relatively small performance differences between controlled conditions, unrecognized leakage can compromise both the estimates and the scientific conclusion.

Table 1 positions the present study against prior work. Srinivasan et al. [2] benchmarked deep architectures on the 6,748-image SLD without reporting an exact/near-duplicate audit, while Daphal and Koli [3], [4] focused on transfer learning, ensembles, and deployment. LeafGAN and Enhanced-RICAP [17], [18] examined advanced augmentation in other plant-disease settings but did not combine duplicate control with operator-specific repeated-seed ablation on the SLD. Thus, Table 1 highlights methodological scope rather than providing a systematic review.

Prior work has largely treated architecture selection, transfer learning, deployment, or augmentation as separate concerns. Few studies have examined duplicate leakage, and no prior sugarcane study identified here jointly evaluates leakage control, operator-specific augmentation ablation, and empirical severity under one fixed repeated-seed protocol. Therefore, this study evaluates augmentation only after explicit dataset-integrity control, with objectives to quantify contamination, compare augmentation policies with an unaugmented baseline, test the ordered severity hypothesis, and identify the operators and classes associated with the largest changes.

Table 1. Methodological Positioning of Related Studies

Study	Dataset and primary aim	Integrity/leakage control	Augmentation evaluation	Uncertainty and error analysis
Daphal and Koli [3]	2,569 images; 5 classes; architecture/ensemble comparison	Not a stated focus	Augmentation supported training; no	Model-level performance comparison

Study	Dataset and primary aim	Integrity/leakage control	Augmentation evaluation	Uncertainty and error analysis
			operator/severity ablation	
Srinivasan et al. [2]	6,748 images; 11 classes; architecture benchmark	No exact/near-duplicate audit reported	Not positioned as a controlled augmentation study	70/15/15 evaluation and cross-validation
Cap et al. (LeafGAN) [17]	Cucumber; 5 classes; generative augmentation	Latent similarity discussed; dataset curation not central	Attention-guided GAN versus vanilla CycleGAN	External-environment evaluation
Diallo et al. (Enhanced-RICAP) [18]	Cassava and tomato; saliency-guided augmentation	Not a stated focus	Multiple augmentation methods across architectures	Comparative accuracy, precision, recall, and F1
This study	Curated SLD; 5,389 images; 11 classes	MD5 + pHash/SSIM; conflict removal; split-safe groups	13 controlled configurations with operator-specific ablations	3 seeds; exact Page test; paired bootstrap; class-wise errors

The central premise of this study is that augmentation performance cannot be interpreted independently of dataset integrity: leakage can obscure genuine treatment effects, whereas transformations that violate task-relevant spatial regularities can reduce generalization even when they increase visual diversity.

The study addresses four research questions:

- 1) RQ1: What types of exact-duplicate, cross-label, and near-duplicate contamination are present in the public SLD, and to what extent?

- 2) RQ2: How do photometric and progressively configured geometric augmentation recipes affect repeated-seed classification performance relative to unaugmented training?
- 3) RQ3: Does empirically measured geometric distortion exhibit an overall ordered relationship with model performance?
- 4) RQ4: Which spatial operators and disease classes account for the largest augmentation-associated performance changes?

The contributions are fourfold. First, the study provides a leakage-aware audit of the 11-class SLD and documents exact duplication, conflicting labels, and verified perceptual similarity. Second, it presents a controlled 39-run augmentation ablation under fixed data, architecture, and optimization settings. Third, it distinguishes individual and combined transformation policies from multi-operator recipes and evaluates an ordered severity hypothesis using an exact repeated-measures Page test [25]. Fourth, it quantifies test-sample uncertainty through paired bootstrap analysis and performs class-level error decomposition to separate observed performance changes from proposed mechanisms.

2. METHODS

2.1. Experimental Overview

Figure 1 summarizes the six-stage experimental workflow: integrity audit and conservative cleaning of 6,748 readable images; leakage-aware splitting of the resulting 5,389-image curated set into 3,771 training, 809 validation, and 809 test images; evaluation of 13 augmentation configurations; MobileNetV2 training across seeds 42, 7, and 123; and analysis using class-sensitive metrics, paired differences, the exact Page test, paired bootstrap intervals, and class-level errors. Augmentation was applied only to training images; validation and test images received deterministic resizing and MobileNetV2 preprocessing. Figure 2 presents the higher-level research flow from problem formulation through interpretation.

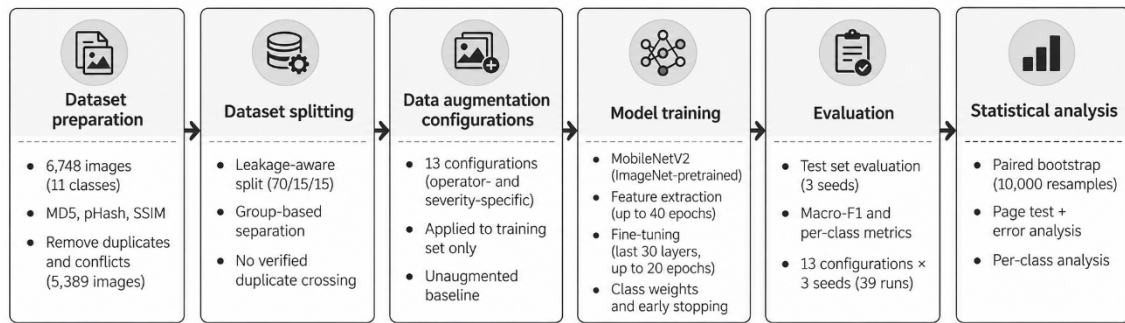


Figure 1. Leakage-Aware Experimental Workflow

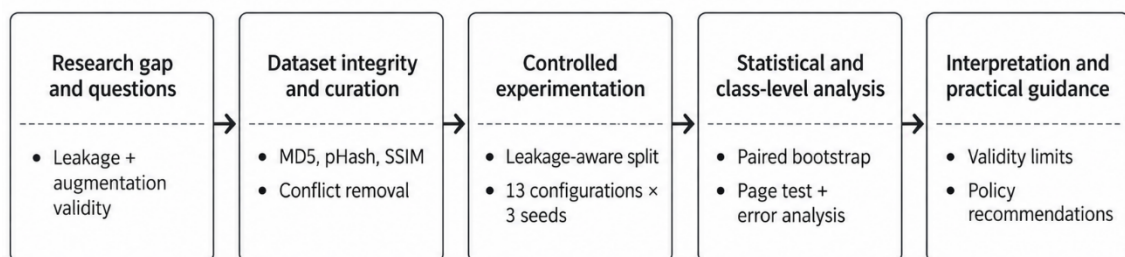


Figure 2. Research Flowchart for Leakage-Aware Augmentation Evaluation

2.2. Dataset and Integrity Audit

The experiments used the open SLD introduced by Thite et al. [1]. The audited copy contained 6,748 readable JPEG images with a uniform width of 768 pixels and varying heights, predominantly 1,024 pixels. Eleven classes were represented: nine disease categories, Healthy, and Dried. The original class distribution ranged from 297 Pokkah Boeng images to 1,722 Brown Spot images. Exact duplicates were identified from MD5 digests computed directly from file bytes in 1-MiB chunks. If byte-identical images occurred under more than one class label, the entire MD5 group was treated as an irreconcilable annotation conflict and removed. Within same-label exact-duplicate groups, one canonical image was retained using deterministic relative-path ordering.

Near-duplicate screening followed exact-duplicate cleaning. A 64-bit pHash was computed after EXIF correction, grayscale conversion, Lanczos resizing to 32 × 32 pixels, orthonormal DCT, and median thresholding of an 8 × 8 low-frequency block with the DC coefficient excluded. A BK-tree retrieved pairs with Hamming distance $d_H \leq 6$ for broad audit screening; pHash alone never triggered deletion. Only candidates with $d_H \leq 2$ underwent SSIM verification after grayscale fitting to 256 × 256 pixels (Lanczos,

data_range = 255). Pairs with SSIM ≥ 0.985 were operationally treated as high-confidence near-duplicates. The threshold roles were intentionally asymmetric and conservative for this dataset: the broader pHash threshold broadened candidate retrieval, the stricter pHash threshold restricted SSIM verification to only the closest pHash matches, and the 0.985 SSIM threshold required very high structural similarity before a pair was treated as a high-confidence near-duplicate. Empirically, this staged procedure reduced 5,200 pHash candidate pairs to three SSIM-verified pairs, all with the same label. These values are dataset-specific operational criteria rather than universal cut-offs. Verified same-label pairs were constrained to one subset, while verified cross-label pairs, if present, would have been removed. The resulting audit outcomes are summarized in Table 2.

Table 2. Dataset Integrity Audit and Cleaning Outcomes

Audit outcome	Count
Original readable images	6,748
Unique MD5 hashes	5,695
Exact-duplicate groups	1,051
Cross-label exact-duplicate groups	306
Images removed from conflicting groups	614
Redundant same-label copies removed	745
pHash candidate pairs	5,200
SSIM-verified near-duplicate pairs	3
Final curated dataset	5,389
Overall reduction	20.1%
Verified duplicate groups crossing final split	0

2.3. Leakage-Aware Data Splitting

After cleaning, the 70/15/15 partition was selected through a two-stage group-preserving GroupShuffleSplit search. Stage 1 generated 500 candidates for an approximately 70/30 training/temporary-holdout split (random_state = 42), and Stage 2 generated 500 candidates to divide the holdout equally into validation and test sets (random_state = 43). Candidates were eligible only when all 11 classes appeared on both sides. The candidate-selection objective was defined as

$$J = |h - q| + 2 \times \frac{1}{C} \sum_{c=1}^C |p_c - p_c(H)| \quad (1)$$

Where h is the observed holdout fraction, q the target fraction, $C = 11$, and p_c and $p_c(H)$ are the class- c proportions in the source and candidate holdout. The minimum- J candidate was retained at each stage. Verified near-duplicate groups were indivisible, final assignments were fixed for all policies and seeds, and a post-split audit confirmed zero verified duplicate groups crossing subsets. The final class distribution across the training, validation, and test subsets is reported in Table 3.

Table 3. Final Class Distribution and Data Split

Class	Final n	Train	Validation	Test
Brown Spot	1,411	991	209	211
Yellow Leaf	885	618	133	134
Sett Rot	652	458	96	98
Mosaic (Viral)	545	373	90	82
Healthy	417	292	61	64
Grassy Shoot	346	243	53	50
Banded Chlorosis	313	219	50	44
Dried	250	171	38	41
Brown Rust	207	153	26	28
Pokkah Boeng	187	129	26	32
Smut	176	124	27	25
Total	5,389	3,771	809	809

2.4. Augmentation Configurations

Thirteen configurations were evaluated: six core strategies (baseline, photometric, three progressively configured geometric recipes, and a combined geometric-photometric recipe) plus seven operator-specific configurations. Transformations were generated on the fly. No custom probability was supplied to Keras RandomFlip; each enabled direction used the implementation's 0.50 stochastic decision during training, and horizontal-and-vertical mode sampled the two directions separately, allowing unchanged, single-direction, or combined-flip realizations. Table 4 lists the complete configurations.

Table 4. Augmentation Strategies and Operator-Specific Configurations

Configuration	Operations	Purpose
Baseline	None	Reference condition
Photometric	Brightness ± 0.20 ; contrast ± 0.20	Non-spatial intensity variation
Geometric light	Rotation 0.05; zoom 0.05; stochastic H+V flip policy	Light multi-operator recipe
Geometric medium	Rotation 0.10; zoom 0.10; translation 0.05; stochastic H+V flip policy	Intermediate multi-operator recipe
Geometric heavy	Rotation 0.15; zoom 0.15; translation 0.10; stochastic H+V flip policy	Strong multi-operator recipe
Combined	Geometric heavy + photometric	Strongest composite recipe
Horizontal flip policy	Stochastic horizontal flip	Orientation-focused ablation
Vertical flip policy	Stochastic vertical flip	Orientation-focused ablation
H+V flip policy	Stochastic horizontal-and-vertical flip	Combined flip policy
Rotation only (0.05)	Rotation 0.05	Single-operator ablation
Zoom only (0.05)	Zoom 0.05	Single-operator ablation
Translation only (0.05)	Translation 0.05	Single-operator ablation
Rotation + zoom (0.05 each)	Rotation 0.05; zoom 0.05; no flip	Two-operator combination

2.5. Empirical Augmentation Severity

Configured parameter magnitudes do not guarantee an empirical ordering of image distortion. Severity was therefore audited on 240 training images with three stochastic realizations per policy. Mean absolute pixel difference (MAD) on the 0-255 intensity scale was defined as

$$\text{MAD}(x, \tilde{x}) = \frac{1}{HWC} \sum_{h=1}^H \sum_{w=1}^W \sum_{c=1}^C |x_{hwc} - \tilde{x}_{hwc}| \quad (2)$$

Structural similarity was computed as

$$\text{SSIM}(x, \tilde{x}) = \frac{(2\mu_x\mu_{\tilde{x}} + C_1)(2\sigma_{x\tilde{x}} + C_2)}{(\mu_x^2 + \mu_{\tilde{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\tilde{x}}^2 + C_2)} \quad (3)$$

For the light-medium-heavy sequence, increasing median MAD together with decreasing median SSIM was required to support the intended empirical severity ordering.

2.6. MobileNetV2 Transfer Learning

Images were resized to 224×224 pixels and processed with MobileNetV2 preprocessing. The ImageNet-pretrained MobileNetV2 backbone [7] was followed by global average pooling, dropout 0.30, and an 11-unit softmax layer; the backbone was called with `training=False` so Batch Normalization remained in inference mode. Feature extraction froze the backbone for up to 40 epochs with Adam at 1×10^{-3} . Fine-tuning retained only the final 30 backbone layers as trainable for up to 20 epochs at 1×10^{-5} . Inverse-frequency class weights addressed imbalance [26], [27]. Early stopping monitored `val_loss` (patience = 10, `restore_best_weights` = True) after a 12-epoch minimum guard; `ReduceLROnPlateau` used factor 0.3, patience 4, and $\min_{lr} = 1 \times 10^{-6}$.

Seeds 42, 7, and 123 were pre-specified as matched experimental blocks. For each seed, Python, NumPy, and TensorFlow random states were reset, one set of initial weights was generated, and those weights were reused across all augmentation policies; dataset assignments also remained fixed. Full GPU-operation determinism was not enabled, so the three seeds characterize residual training variability rather than population-level superiority. Wall-clock training time covered feature extraction through completion of fine-tuning, including validation and callbacks, but excluded model construction, test inference, and post-hoc metrics. The training and optimization settings held constant across augmentation configurations are summarized in Table 5.

Table 5. Training and Optimization Settings

Setting	Value
Input size	$224 \times 224 \times 3$
Batch size	32
Backbone	MobileNetV2, ImageNet pretrained

Setting	Value
Classification head	Global average pooling; dropout 0.30; dense softmax (11 classes)
Optimizer	Adam
Loss	Categorical cross-entropy with balanced class weights
Feature extraction	40 epochs maximum; learning rate 1×10^{-3}
Fine-tuning	Last 30 backbone layers; 20 epochs maximum; learning rate 1×10^{-5}
Early stopping	Validation loss; minimum epoch requirement; restore best weights
Learning-rate reduction	Factor 0.3; patience 4
Seeds	42, 7, and 123
Software and hardware	TensorFlow 2.20; NVIDIA Tesla T4 GPU

2.7. Evaluation Metrics and Seed-Paired Comparisons

Model performance was evaluated using complementary class-sensitive metrics, with Macro-F1 designated as the primary measure because the curated dataset remained imbalanced. To isolate augmentation-associated changes from seed-specific variability, each configuration was also compared with the unaugmented baseline trained using the same seed.

For class c , precision, recall, and F1 were defined as

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c}, \quad F1_c = \frac{2P_cR_c}{P_c + R_c}. \quad (4)$$

With $C = 11$ classes, the primary metric was

$$\text{MacroF1} = \frac{1}{C} \sum_{c=1}^C F1_c, \quad (5)$$

and multiclass balanced accuracy was

$$\text{BalancedAccuracy} = \frac{1}{C} \sum_{c=1}^C R_c. \quad (6)$$

Balanced accuracy is therefore numerically identical to macro-recall in this single-label multiclass setting. Results are reported as mean $\pm$ sample standard deviation across seeds.

Because each policy was paired with the baseline by seed, the seed-level effect was

$$\Delta F1_{a,s} = F1_{a,s} - F1_{\text{baseline},s}, \quad (7)$$

where a denotes an augmentation policy and s a seed. Directional consistency was summarized by

$$D_a = \sum_{s=1}^3 \mathbb{I}(\Delta F1_{a,s} < 0). \quad (8)$$

A value of $D_a = 3$ indicates that a policy was below baseline for all three seeds.

2.8. Statistical and Uncertainty Analysis

The hypothesized monotonic ordering from the baseline to light, medium, and heavy geometric augmentation was evaluated using the Page test for ordered alternatives [25]. Seeds were repeated blocks. If r_{ij} denotes the within-seed rank for treatment position j , the Page statistic is

$$L = \sum_{j=1}^k j R_j, \quad R_j = \sum_{i=1}^n r_{ij}, \quad (9)$$

with $n = 3$ blocks and $k = 4$ treatments. The exact one-sided null distribution was obtained by enumerating all $24^3 = 13,824$ within-block rank permutations. A significant result indicates an overall ordered trend and does not require every adjacent pair to decrease strictly.

Test-sample uncertainty was evaluated with a paired class-stratified bootstrap [28]. For each seed and policy, 10,000 resamples used identical within-class indices for the baseline and comparison policy to obtain percentile 95% CIs for paired Macro-F1 differences. These intervals are conditional on the fitted model pair and fixed test sample and do not replace between-seed variability.

Class-specific effects were examined through mean per-class F1, seed-aggregated row-normalized confusion matrices, and transitions between baseline-correct and policy-incorrect decisions. Because the same 809 test images were evaluated by three trained seed models, seed-aggregated confusion matrices are descriptive summaries of predictions, not 2,427 independent samples.

3. RESULTS AND DISCUSSION

3.1. RQ1: Dataset Contamination and Leakage Control

The audit found 5,695 unique MD5 hashes among 6,748 files, corresponding to 1,053 redundant files in 1,051 exact-duplicate groups. Of particular concern, 306 groups contained conflicting labels. Conservative removal excluded 614 images from those groups and removed 745 redundant same-label copies. The curated dataset contained 5,389 images, a 20.1% reduction.

Perceptual hashing generated 5,200 audit candidate pairs with Hamming distance $d_H < = 6$. Only the higher-similarity subset with $d_H \leq 2$ underwent SSIM verification; three pairs met the predefined SSIM threshold of 0.985. None was cross-label. These same-label groups were assigned as indivisible units, and the final audit confirmed no verified duplicate group crossing training, validation, or test subsets. Thus, RQ1 revealed substantial exact duplication and label conflict while the final split prevented verified duplicate groups from crossing data subsets.

3.2. RQ2 and the Operator-Level Component of RQ4: Repeated-Seed Performance

Table 6 reports accuracy, balanced accuracy, Macro-F1, and seed-paired Macro-F1 differences for all 13 configurations; Figure 3 visualizes the paired differences. Baseline had the highest mean Macro-F1 (0.9646 ± 0.0068), while photometric augmentation was nearly identical on average (0.9643 ± 0.0079) but changed direction across seeds. Horizontal flipping produced the next-smallest mean reduction, followed by translation and zoom; vertical flip, H+V flip, rotation, and rotation plus zoom showed larger reductions. All multi-operator geometric recipes were below baseline in every seed, with combined augmentation showing the largest mean reduction and variability. Exact seed-wise paired 95% bootstrap CIs are reported in Table 7.

Table 6. Test Performance and Paired Differences Across Augmentation Configurations; seed-wise paired 95% bootstrap CIs are reported in Table 7

Configuration	Accuracy	Balanced Accuracy	Macro-F1	Paired Macro-F1 difference vs. baseline
Baseline	0.9670 ± 0.0076	0.9754 ± 0.0014	0.9646 ± 0.0068	-
Photometric	0.9683 ± 0.0089	0.9747 ± 0.0024	0.9643 ± 0.0079	-0.0003
Horizontal flip policy	0.9600 ± 0.0014	0.9729 ± 0.0026	0.9601 ± 0.0014	-0.0045
Translation only (0.05)	0.9518 ± 0.0064	0.9720 ± 0.0038	0.9535 ± 0.0049	-0.0111
Zoom only (0.05)	0.9522 ± 0.0147	0.9685 ± 0.0071	0.9530 ± 0.0134	-0.0116
Vertical flip policy	0.9341 ± 0.0056	0.9606 ± 0.0026	0.9341 ± 0.0040	-0.0305
H+V flip policy	0.9328 ± 0.0103	0.9606 ± 0.0065	0.9327 ± 0.0091	-0.0319
Rotation only (0.05)	0.9201 ± 0.0144	0.9506 ± 0.0097	0.9270 ± 0.0139	-0.0376
Rotation + zoom (0.05 each)	0.9007 ± 0.0029	0.9337 ± 0.0013	0.9080 ± 0.0020	-0.0566
Geometric medium	0.8842 ± 0.0019	0.9164 ± 0.0025	0.8919 ± 0.0004	-0.0727
Geometric light	0.8677 ± 0.0057	0.9084 ± 0.0057	0.8774 ± 0.0075	-0.0872
Geometric heavy	0.8307 ± 0.0097	0.8681 ± 0.0094	0.8343 ± 0.0109	-0.1303
Combined	0.7969 ± 0.0395	0.8499 ± 0.0296	0.8029 ± 0.0373	-0.1617

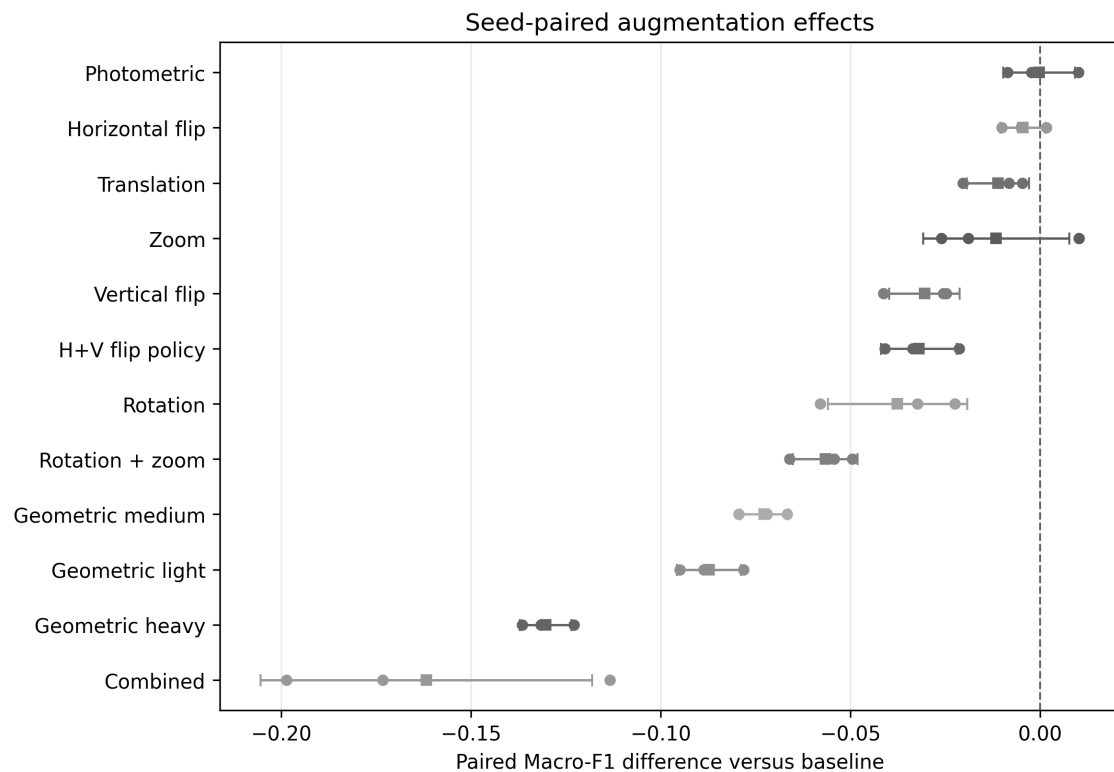


Figure 3. Seed-Paired Macro-F1 Differences from Baseline

Across the operator-specific configurations, photometric augmentation and horizontal flipping remained closest to baseline, translation and zoom showed smaller average reductions than rotation-based policies, and the largest reductions occurred for the multi-operator geometric recipes. These processed performance comparisons answer RQ2 and provide the operator-level component of RQ4; their interpretation is developed in Section 3.2.

3.2.1. Paired Bootstrap Evidence

The seed-wise paired stratified bootstrap results are shown in Table 7. For photometric augmentation, the interval was below zero for seed 42, above zero for seed 7, and crossed zero for seed 123. Horizontal flip intervals crossed zero in all three seeds. Translation was below zero only for seed 123. For zoom, the interval was above zero for seed 7 and below zero for seeds 42 and 123. In contrast, vertical flip, H+V flip, rotation, rotation plus zoom, all three geometric recipes, and combined augmentation had intervals below zero in all three seeds.

Table 7. Seed-Wise Paired Bootstrap Differences in Macro-F1

Configuration	Seed 42 delta [95% CI]	Seed 7 delta [95% CI]	Seed 123 delta [95% CI]	CIs below zero
Photometric	-0.0086 [-0.0157, -0.0024]	+0.0101 [+0.0021, +0.0188]	-0.0023 [-0.0087, +0.0033]	1/3
Horizontal flip policy	-0.0050 [-0.0131, +0.0023]	+0.0016 [-0.0047, +0.0086]	-0.0101 [-0.0205, +0.0001]	0/3
Translation only (0.05)	-0.0047 [-0.0140, +0.0053]	-0.0082 [-0.0190, +0.0032]	-0.0203 [-0.0320, -0.0082]	1/3
Zoom only (0.05)	-0.0190 [-0.0294, -0.0091]	+0.0103 [+0.0019, +0.0191]	-0.0260 [-0.0403, -0.0115]	2/3
Vertical flip policy	-0.0255 [-0.0379, -0.0132]	-0.0247 [-0.0376, -0.0120]	-0.0412 [-0.0567, -0.0259]	3/3
H+V flip policy	-0.0213 [-0.0332, -0.0093]	-0.0336 [-0.0482, -0.0195]	-0.0409 [-0.0557, -0.0259]	3/3
Rotation only (0.05)	-0.0224 [-0.0334, -0.0122]	-0.0324 [-0.0464, -0.0191]	-0.0580 [-0.0757, -0.0409]	3/3
Rotation + zoom (0.05 each)	-0.0543 [-0.0715, -0.0380]	-0.0495 [-0.0677, -0.0325]	-0.0661 [-0.0850, -0.0478]	3/3
Geometric medium	-0.0720 [-0.0941, -0.0514]	-0.0667 [-0.0872, -0.0468]	-0.0794 [-0.1003, -0.0591]	3/3
Geometric light	-0.0949 [-0.1167, -0.0744]	-0.0781 [-0.0995, -0.0582]	-0.0886 [-0.1099, -0.0683]	3/3
Geometric heavy	-0.1229 [-0.1483, - 0.0992]	-0.1364 [-0.1642, - 0.1104]	-0.1316 [-0.1576, - 0.1066]	3/3
Combined	-0.1732 [-0.1987, - 0.1479]	-0.1134 [-0.1391, - 0.0896]	-0.1986 [-0.2270, -0.1711]	3/3

The bootstrap does not establish population-level superiority or equivalence across training procedures, but it distinguishes policies with small, seed-dependent differences from baseline from policies whose paired differences were below zero across all three seeds on the observed test set.

3.2.2. RQ3: Empirical Severity and Ordered Trend

Table 8 reports empirical distortion. Median MAD increased from 36.20 for geometric light to 38.54 for medium and 39.31 for heavy, while median SSIM decreased from 0.5372

to 0.5305 and 0.5209. The intended severity order was therefore supported under both metrics. Stochastic flip-only policies had zero median MAD and unit median SSIM because at least half of their realizations were no-ops or symmetry-equivalent at the pixel-comparison stage. This pattern shows that pixel-level severity can remain zero for stochastic no-op or pixel-equivalent realizations.

Table 8. Empirical Augmentation Severity

Configuration	MAD median [Q1, Q3]	SSIM median [Q1, Q3]
Photometric	25.91 [13.91, 37.88]	0.9729 [0.9394, 0.9890]
Geometric light	36.20 [18.97, 47.99]	0.5372 [0.0946, 0.6791]
Geometric medium	38.54 [23.52, 47.94]	0.5305 [0.0963, 0.6582]
Geometric heavy	39.31 [25.85, 48.78]	0.5209 [0.0936, 0.6446]
Combined	49.27 [39.13, 57.36]	0.4901 [0.0825, 0.6242]
Horizontal flip policy	0.00 [0.00, 38.07]	1.0000 [0.5305, 1.0000]
Vertical flip policy	0.00 [0.00, 40.10]	1.0000 [0.5299, 1.0000]
H+V flip policy	0.00 [0.00, 46.23]	1.0000 [0.5046, 1.0000]
Rotation only (0.05)	17.89 [11.86, 37.76]	0.6355 [0.1109, 0.7074]
Zoom only (0.05)	4.97 [3.19, 19.68]	0.8459 [0.4523, 0.9077]
Translation only (0.05)	16.26 [9.65, 35.83]	0.6270 [0.0976, 0.7146]
Rotation + zoom (0.05 each)	18.13 [12.14, 37.22]	0.6303 [0.1187, 0.7104]

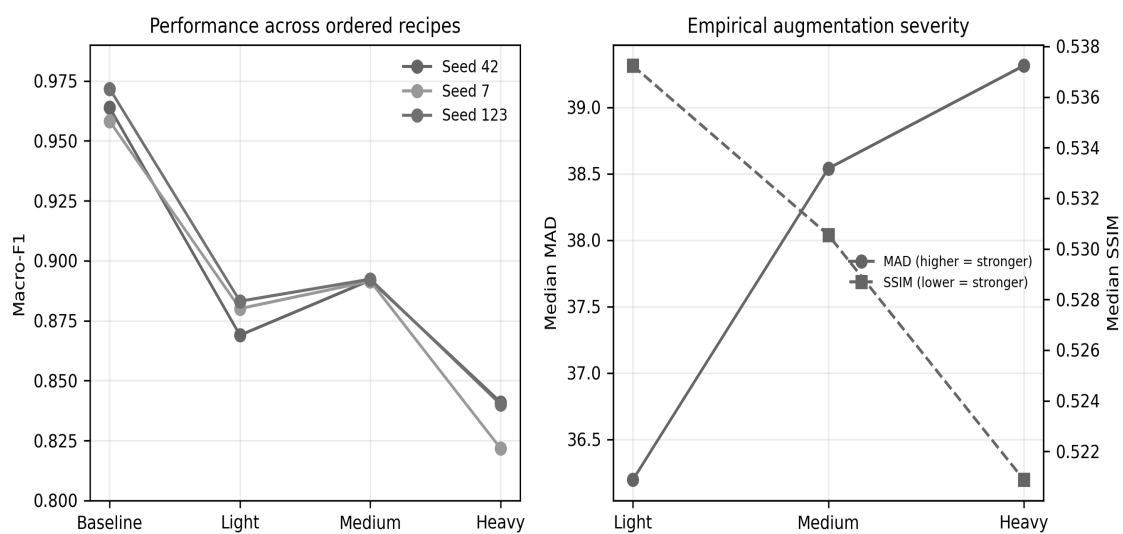

Figure 4. Performance and Severity of Geometric Augmentation

Figure 4 combines repeated-seed performance with the severity summary. The exact Page test yielded $L = 87$ (maximum 90) with $p = 0.0070$, supporting an overall negative ordered trend. However, geometric medium outperformed geometric light in all three seeds. RQ3 is therefore characterized by an overall severity-associated ordered trend with a consistent local inversion, not by a strictly monotonic dose-response law.

3.2.3. Class-Specific Sensitivity and Error Decomposition

Figure 5 shows mean per-class F1 across all policies. The strongest combined-recipe declines in mean seed-level F1 occurred for Brown Rust, Banded Chlorosis, Brown Spot, Healthy, and Yellow Leaf. Sett Rot and Grassy Shoot were comparatively stable. In Figure 5, H-flip and V-flip denote the stochastic horizontal and vertical flip policies, while Rot+Zoom denotes rotation plus zoom without flipping.

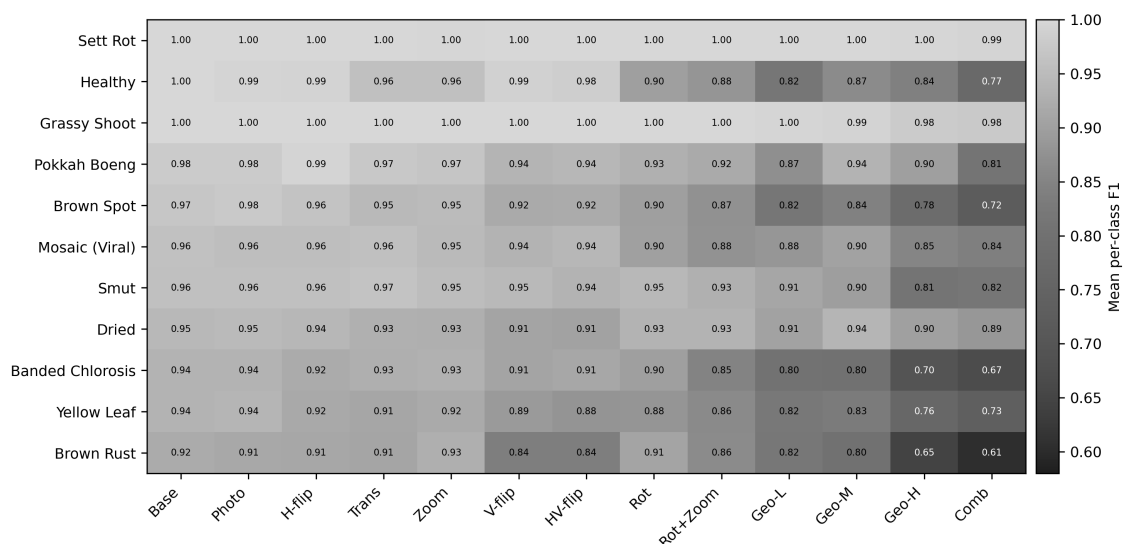


Figure 5. Class-Wise F1 Across Augmentation Policies

Figure 6 and Table 9 show distinct class-level error profiles. Under combined augmentation, Brown Spot recall decreased from 0.949 to 0.592 while precision remained 0.938. Healthy retained recall of 0.984 but precision fell to 0.634, indicating increased misclassification of diseased images as Healthy. Brown Rust showed a similar precision-dominant pattern, whereas Smut and Yellow Leaf had stronger recall losses. Confusion-matrix abbreviations are BC (Banded Chlorosis), BR (Brown Rust), BS (Brown Spot), Dr (Dried), GS (Grassy Shoot), He (Healthy), MV (Mosaic), PB (Pokkah Boeng), SR (Sett Rot), Sm

(Smut), and YL (Yellow Leaf). Seed-aggregated values are descriptive and do not treat repeated predictions per image as independent observations.

Table 9. Class-Level Precision, Recall, and F1 Profiles

Class	Baseline P / R / F1	Combined P / R / F1	Aggregated delta F1	Dominant error-profile change
Brown Rust	0.889 / 0.952 / 0.920	0.452 / 0.905 / 0.603	-0.316	Precision collapse with relatively preserved recall
Banded Chlorosis	0.885 / 0.992 / 0.936	0.577 / 0.795 / 0.669	-0.267	Both precision and recall declined
Brown Spot	0.997 / 0.949 / 0.972	0.938 / 0.592 / 0.726	-0.246	Recall-dominant degradation
Healthy	1.000 / 1.000 / 1.000	0.634 / 0.984 / 0.771	-0.229	Precision-dominant degradation; recall remained high
Yellow Leaf	0.951 / 0.920 / 0.936	0.785 / 0.689 / 0.734	-0.202	Joint precision and recall degradation
Sett Rot	1.000 / 1.000 / 1.000	0.990 / 1.000 / 0.995	-0.005	Stable reference class

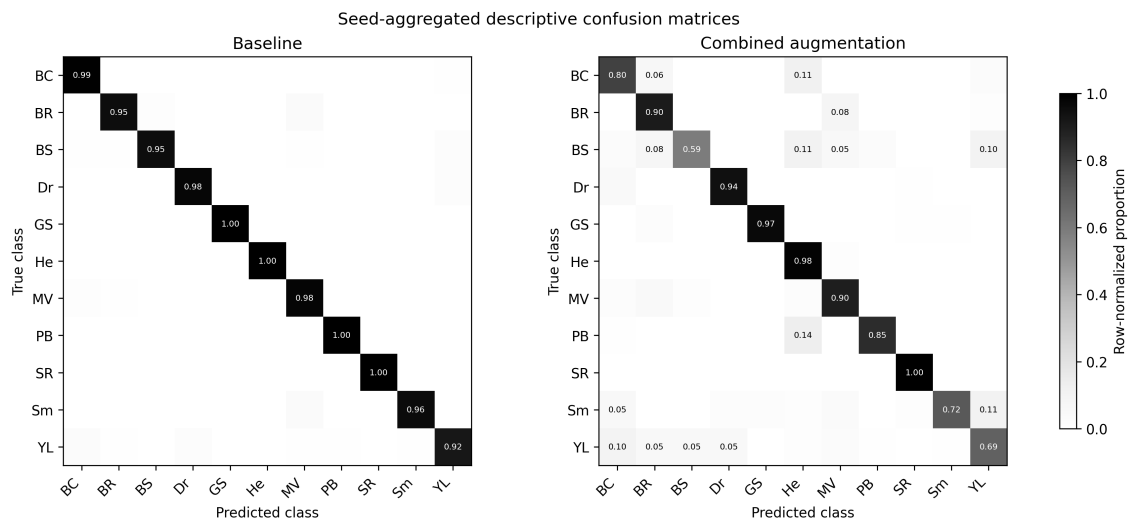


Figure 6. Seed-Aggregated Baseline and Combined Augmentation Confusion Matrices

Seed-by-seed transition analysis further showed that combined augmentation changed an average of 77.0 Brown Spot decisions per seed from baseline-correct to combined-

incorrect, compared with only 1.7 transitions in the opposite direction. The corresponding values were 32.0 versus 1.0 for Yellow Leaf. For Healthy, only one image per seed on average changed from correct to incorrect, confirming that its F1 degradation was driven primarily by an increase in false positives rather than loss of Healthy recall.

3.3. Discussion

3.3.1. Augmentation Is a Task-Dependent Treatment

The central result is not that augmentation is generally ineffective, but that its value depends on whether the induced invariance is plausible for the task. Photometric jitter and horizontal flipping remained near baseline, whereas vertical flipping, rotation, and multi-operator geometric policies were consistently harmful. This is consistent with the principle that augmentation should approximate nuisance variation rather than maximize visual diversity [8], [9], and with information-preserving or saliency-guided approaches designed to protect class-relevant evidence [11], [17], [18].

Evidence, interpretation, and mechanism should remain distinct. Vertical flipping and rotation reduced Macro-F1 in all seeds, with seed-wise bootstrap intervals below zero; this supports interpreting them as unreliable invariances for this dataset-model combination. The mechanism is unresolved, with orientation, leaf geometry, framing, background regularities, and lesion organization remaining plausible contributors.

The findings also differ methodologically from earlier SLD studies. Srinivasan et al. [2] and Daphal and Koli [3], [4] primarily evaluated architectures, transfer learning, ensembles, or deployment on the original dataset rather than augmentation after duplicate auditing. Their accuracy values are therefore not directly comparable with the curated-data baseline. The relevant contrast is methodological: leakage control enables operator-specific augmentation effects to be paired against a fixed repeated-seed baseline. Likewise, LeafGAN and Enhanced-RICAP [17], [18] illustrate the potential of advanced augmentation while reinforcing the need for task-specific policy validation.

3.3.2. Why Medium Consistently Exceeded Light

The medium recipe consistently outperformed the light recipe despite greater measured distortion, showing that severity was not one-dimensional. The recipes also differed compositionally: medium added translation and used larger rotation and zoom ranges

alongside the stochastic flip policy. Because translation and zoom were relatively mild in operator-specific tests, the inversion may reflect transformation composition and co-occurrence rather than nominal magnitude alone. This result explains why the Page test and adjacent-pair behavior are not contradictory. Page's statistic supported an overall ordered trend across the four treatment positions, but it does not assert that each adjacent difference is monotonic. The defensible conclusion is operator-dependent nonlinear degradation, not a scalar dose-response mechanism.

3.3.3. Dataset Leakage Changes the Scientific Question

The dataset audit was integral to the scientific question. More than one thousand redundant files were present, including hundreds of exact-duplicate groups with conflicting labels. An uncurated image-level split could therefore place identical content across training and test sets or assign contradictory targets to byte-identical files, conditions known to distort apparent performance and model rankings [21], [22], [23], [24]. Conservative cleaning produced a smaller but more coherent dataset, so the resulting baseline should not be compared directly with scores from the uncleaned release. The contribution is an auditable estimate of augmentation effects under explicit leakage control rather than another architecture benchmark.

3.3.4. Class-Level Effects Require Error Decomposition

Per-class F1 alone does not reveal whether degradation arises from missed true cases or increased false positives. Under combined augmentation, Brown Spot was primarily recall-limited, while Healthy and Brown Rust retained high recall but lost substantial precision; Banded Chlorosis and Yellow Leaf lost both precision and recall. These patterns support plausible but unverified mechanisms: geometric transformations may disturb combinations of lesion scale, spatial arrangement, or leaf context used for Brown Spot recognition, whereas the Healthy result reflects increased attraction of diseased samples toward the Healthy decision region. Without expert lesion annotations, segmentation masks, or controlled acquisition metadata, these explanations remain hypotheses rather than causal findings.

3.3.5. Practical Implications and Augmentation-Validation Workflow

For fine-grained plant-disease classification, a defensible workflow is to curate duplicate groups before splitting, establish an unaugmented repeated-seed baseline, test plausible

photometric changes independently, isolate spatial operators before combining them, retain only policies without consistent paired degradation, and inspect paired uncertainty and class-specific errors before deployment. For the present SLD-MobileNetV2 setting, baseline or photometric training is preferable to broad geometric augmentation. Horizontal flipping may be retained when domain knowledge supports left-right invariance, whereas vertical flipping and rotation require stronger justification. Automated methods such as RandAugment and TrivialAugment [12], [13] remain promising when their policy spaces are constrained by domain plausibility and evaluated with leakage-safe validation.

Observed training cost also differed: baseline averaged approximately 5.1 minutes per run, whereas geometric recipes required about 19-27 minutes in the same Kaggle NVIDIA Tesla T4 environment. Because wall-clock time included both training stages, validation, callbacks, and configuration-dependent duration, it represents end-to-end procedure cost rather than isolated augmentation overhead.

3.3.6. Threats to Validity

Internal validity. On-the-fly augmentation, variable training duration, and incomplete GPU-operation determinism leave residual stochastic variation despite matched initialization, fixed splits, seed-controlled shuffling, and three repeated seeds. The design was controlled but not fully factorial, so combined-policy differences should not be interpreted as formal interaction coefficients. **Construct validity.** MAD and SSIM measure pixel and structural change rather than preservation of pathological evidence, and pHash/SSIM thresholds cannot identify every semantically related image. Macro-F1 measures class-balanced discrimination but not probability calibration, abstention, or field-level diagnostic utility.

Statistical conclusion validity. Three seeds support descriptive variability and the exact blocked Page test but provide limited power for equivalence or population-level superiority. Paired bootstrap CIs quantify fixed-test-sample uncertainty conditional on fitted models, not the full training process; baseline and photometric performance are therefore described as numerically comparable, not statistically equivalent.

External and model-capacity validity. The study used one public dataset and one lightweight MobileNetV2 backbone without independent field-data validation. Augmentation sensitivity may vary with model capacity and architectural inductive biases, while SLD-specific regularities in orientation, framing, background, scale, or symptom presentation may influence operator rankings. Because lesion masks, expert region annotations, and controlled acquisition metadata were unavailable, mechanistic interpretations remain provisional. The reported rankings should therefore be generalized cautiously, although the principles of leakage control and augmentation-policy validation are broader. Future work should test whether these operator-specific rankings persist across higher-capacity CNN families and transformer-based vision architectures, using additional independent runs and independently acquired field images where feasible.

4. CONCLUSION

This study evaluated conventional data augmentation for sugarcane leaf disease classification under explicit dataset-integrity control. Auditing the 6,748-image SLD identified extensive exact duplication and 306 cross-label conflict groups; conservative curation produced 5,389 images with no verified duplicate group crossing the final subsets. Across 39 MobileNetV2 runs, baseline achieved the highest mean Macro-F1 (0.9646), while photometric augmentation was numerically comparable but seed-dependent. Multi-operator geometric recipes reduced performance, and combined augmentation produced the lowest Macro-F1. MAD and SSIM confirmed the intended light-to-medium-to-heavy distortion order, while the exact Page test supported an overall negative ordered trend; the medium-over-light inversion ruled out a strictly monotonic dose-response interpretation. Operator- and class-level analyses showed heterogeneous effects, supporting individual validation of augmentation policies on leakage-aware data rather than default use. These findings are limited to the curated SLD and MobileNetV2 setting without independent field validation, so operator rankings require confirmation in other datasets and architectures. For real-world agricultural AI deployment, these findings support treating augmentation as a task-specific policy that requires empirical validation rather than as a default preprocessing step.

REFERENCES

- [1] S. Thite, Y. Suryawanshi, K. Patil, and P. Chumchu, "Sugarcane leaf dataset: A dataset for disease detection and classification for machine learning applications," *Data Brief*, vol. 53, p. 110268, 2024, doi: 10.1016/j.dib.2024.110268.
- [2] S. Srinivasan, S. M. Prabin, S. K. Mathivanan, H. Rajadurai, S. Kulandaivelu, and M. A. Shah, "Sugarcane leaf disease classification using deep neural network approach," *BMC Plant Biol*, vol. 25, no. 1, p. 282, 2025, doi: 10.1186/s12870-025-06289-0.
- [3] S. D. Daphal and S. M. Koli, "Enhancing sugarcane disease classification with ensemble deep learning: A comparative study with transfer learning techniques," *Heliyon*, vol. 9, no. 8, p. e18261, 2023, doi: 10.1016/j.heliyon.2023.e18261.
- [4] S. D. Daphal and S. M. Koli, "Enhanced deep learning technique for sugarcane leaf disease classification and mobile application integration," *Heliyon*, vol. 10, no. 8, p. e29438, 2024, doi: 10.1016/j.heliyon.2024.e29438.
- [5] M. Shoaib *et al*, "An advanced deep learning models-based plant disease detection: A review of recent research," *Front. Plant Sci*, vol. 14, Art. no. 1158933, 2023, doi: 10.3389/fpls.2023.1158933.
- [6] J. Liu and X. Wang, "Plant diseases and pests detection based on deep learning: a review," *Plant Methods*, vol. 17, no. 1, p. 22, 2021, doi: 10.1186/s13007-021-00722-9.
- [7] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.
- [8] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019, doi: 10.1186/s40537-019-0197-0.
- [9] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, "A Comprehensive Survey of Image Augmentation Techniques for Deep Learning," *Pattern Recognit*, vol. 137, p. 109347, 2023, doi: 10.1016/j.patcog.2023.109347.
- [10] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2, Art. no. 125, Feb. 2020, doi: 10.3390/info11020125.
- [11] C. Gong, D. Wang, M. Li, V. Chandra, and Q. Liu, "KeepAugment: A Simple Information-Preserving Data Augmentation Approach," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1055–1064. doi: 10.1109/CVPR46437.2021.00111.

- [12] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2020, pp. 3008–3017. doi: 10.1109/CVPRW50498.2020.00359.
- [13] S. G. Muller and F. Hutter, "TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2021, pp. 754–762. doi: 10.1109/ICCV48922.2021.00081.
- [14] Y. Liu, S. Yan, L. Leal-Taixé, J. Hays, and D. Ramanan, "Soft Augmentation for Image Classification," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16241–16250. doi: 10.1109/CVPR52729.2023.01558.
- [15] M. Xu, S. Yoon, A. Fuentes, J. Yang, and D. S. Park, "Style-Consistent Image Translation: A Novel Data Augmentation Paradigm to Improve Plant Disease Recognition," *Front. Plant Sci.*, vol. 12, Art. no. 773142, 2022, doi: 10.3389/fpls.2021.773142.
- [16] M. Xu *et al.*, "Embracing limited and imperfect training datasets: opportunities and challenges in plant disease recognition using deep learning," *Front. Plant Sci.*, vol. 14, Art. no. 1225409, 2023, doi: 10.3389/fpls.2023.1225409.
- [17] Q. H. Cap, H. Uga, S. Kagiwada, and H. Iyatomi, "LeafGAN: An Effective Data Augmentation Method for Practical Plant Disease Diagnosis," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 2, pp. 1258–1267, 2022, doi: 10.1109/TASE.2020.3041499.
- [18] M. B. Diallo *et al.*, "Enhanced-RICAP: a novel data augmentation strategy for improved deep learning-based plant disease identification and mobile diagnosis," *Front. Plant Sci.*, vol. 16, Art. no. 1646611, 2025, doi: 10.3389/fpls.2025.1646611.
- [19] Y. Lu, D. Chen, E. Olaniyi, and Y. Huang, "Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review," *Comput. Electron. Agric.*, vol. 200, p. 107208, 2022, doi: 10.1016/j.compag.2022.107208.
- [20] J. Dong *et al.*, "Data-centric annotation analysis for plant disease detection: Strategy, consistency, and performance," *Front. Plant Sci.*, vol. 13, Art. no. 1037655, 2022, doi: 10.3389/fpls.2022.1037655.
- [21] B. Barz and J. Denzler, "Do We Train on Test Data? Purging CIFAR of Near-Duplicates," *J. Imaging*, vol. 6, no. 6, Art. no. 41, 2020, doi: 10.3390/jimaging6060041.

- [22] I. E. Tampu, A. Eklund, and N. Haj-Hosseini, "Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images," *Sci. Data*, vol. 9, no. 1, p. 580, 2022, doi: 10.1038/s41597-022-01618-6.
- [23] A. Aghabagherloo, A. Abadi, S. Sarkar, V. A. Dasu, and B. Preneel, "Impact of Data Duplication on Deep Neural Network-Based Image Classifiers: Robust vs. Standard Models," in *2025 IEEE Security and Privacy Workshops (SPW)*, 2025, pp. 177–183. doi: 10.1109/SPW67851.2025.00023.
- [24] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," *Med. Image Anal.*, vol. 75, p. 102305, 2022, doi: 10.1016/j.media.2021.102305.
- [25] E. B. Page, "Ordered Hypotheses for Multiple Treatments: A Significance Test for Linear Ranks," *J. Am. Stat. Assoc.*, vol. 58, no. 301, pp. 216–230, 1963, doi: 10.2307/2282965.
- [26] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, p. 27, 2019, doi: 10.1186/s40537-019-0192-5.
- [27] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.
- [28] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, 1st ed. Chapman and Hall/CRC, 1994. doi: 10.1201/9780429246593.