

Explainable Deep Learning for Field-Level Crop Disease Detection: A Simulated Precision-Agriculture Evaluation

Sanjana Pillai¹, Harsh Vardhan²

¹Student, Department of Agriculture, GU, India

²Guide, Department of Agriculture, India

ABSTRACT

This article presents a simulated empirical study in agricultural science, computer vision, and explainable artificial intelligence. The simulated dataset included 6,720 training images, 1,440 validation images, and 1,440 hold-out images, with deliberate class imbalance and domain shift. A simulated image-classification benchmark was constructed to represent 9,600 crop-leaf images across eight disease and healthy categories. Metadata captured illumination, camera distance, occlusion, leaf age, and background complexity. Three model families were compared: a standard CNN, a transfer-learning network, and a lightweight explainable model with saliency-based localization checks. Performance was evaluated on a clean validation set and a separate field-complexity hold-out set. The transfer-learning model achieved the highest simulated clean-set accuracy, while the lightweight explainable model retained the strongest field hold-out macro-F1 and the lowest inference latency. Explanation-consistency scoring indicated that some high-confidence CNN predictions relied on background regions rather than symptomatic tissue. The explainable model reduced this failure pattern and maintained a 0.89 field macro-F1. All numerical values, assignments, observations, and study outcomes are simulated for methodological illustration. No real participants, organizations, field sites, laboratories, or commercial systems were studied. The paper integrates current literature with a transparent empirical-style design and emphasizes reliability, uncertainty, implementation constraints, and the need for real-world validation before operational use.

Keywords: Crop Disease Detection, Explainable AI, Deep Learning, Precision Agriculture, Computer Vision, Edge Deployment

I. Introduction

The rapid development of agricultural science, computer vision, and explainable artificial intelligence is reshaping how researchers evaluate efficiency, reliability, sustainability, and decision quality. Current technology trends increasingly combine advanced materials, artificial intelligence, distributed sensing, and energy-system optimization, but strong laboratory or simulation performance does not automatically translate to robust field deployment.

A journal-style empirical study should therefore define a clear baseline, measure multiple outcomes, report trade-offs, and distinguish model performance from practical usefulness. Real-world adoption may depend on environmental variability, data quality, financing, maintenance, safety, human oversight, and compatibility with existing infrastructure.

Because no real dataset was supplied for this Mode 2 article, the present work uses an internally consistent synthetic dataset. The simulation is designed to demonstrate a realistic research pathway and plausible result structure, not to claim that the study was actually conducted.

Research objectives:

1. Compare conventional CNN, transfer-learning, and explainable lightweight models on a simulated field-image dataset.
2. Measure accuracy, macro-F1, calibration, inference latency, and explanation consistency under changing illumination and background conditions.
3. Assess whether interpretable edge-ready models can retain useful performance outside laboratory-style image backgrounds.

II. Literature Review

Plant-disease detection has become a major application of deep learning in precision agriculture. Liu and Wang (2021) review deep-learning approaches for plant disease and pest detection, emphasizing both performance advances and practical limitations.

Demilie (2024) compares plant-disease detection and classification techniques and highlights the importance of model choice, data quality, and realistic evaluation. Field deployment remains more difficult than curated-image benchmarking because illumination, background, disease stage, and crop variability can shift the input distribution.

Federated and privacy-aware approaches are also emerging. Kabala et al. (2023) investigate image-based crop-disease detection with federated learning, illustrating how decentralized agricultural data may support collaborative model development. Explainability adds another layer by testing whether model attention aligns with plausible plant symptoms rather than spurious background cues.

Across this literature, a recurring theme is the transition from proof-of-concept performance toward reliability, reproducibility, scale-up, and decision relevance. The present simulated study is structured around that transition by evaluating more than one outcome and by including sensitivity analysis rather than reporting a single best-case result.

III. Materials and Methods

3.1 Research Design

A simulated image-classification benchmark was constructed to represent 9,600 crop-leaf images across eight disease and healthy categories. Metadata captured illumination, camera distance, occlusion, leaf age, and background complexity. Three model families were compared: a standard CNN, a transfer-learning network, and a lightweight explainable model with saliency-based localization checks. Performance was evaluated on a clean validation set and a separate field-complexity hold-out set.

3.2 Simulated Dataset and Sampling

The simulated dataset included 6,720 training images, 1,440 validation images, and 1,440 hold-out images, with deliberate class imbalance and domain shift. Synthetic observations were generated within bounded domain-plausible ranges with deliberate random variability and controlled interaction effects. The dataset does not correspond to any identifiable real institution, company, farm, laboratory, or geographic site.

3.3 Data Preparation

All variables were screened for impossible values, inconsistent units, and duplicate identifiers. Low-rate missingness was introduced to approximate realistic operational data and handled with transparent context-appropriate imputation. Transformations were applied only when required by the analytical method. Evaluation partitions were separated before model or parameter selection to reduce data leakage.

3.4 Comparative Metrics

Performance was assessed using multiple indicators selected for the domain rather than a single headline measure. These included efficiency or predictive quality alongside robustness, safety, latency, degradation, cost, or resource-use indicators. Relative change was calculated against a clearly defined baseline. Statistical inference is presented only as illustrative because the observations are synthetic.

3.5 Sensitivity Analysis

Key assumptions were varied over plausible ranges to determine whether the leading strategy depended on one narrow parameter setting. Scenario analysis also tested a degraded-input condition representing environmental variability, sensor or material uncertainty, or cost volatility. The objective was to evaluate ranking stability rather than identify a universal optimum.

IV. Results

The transfer-learning model achieved the highest simulated clean-set accuracy, while the lightweight explainable model retained the strongest field hold-out macro-F1 and the lowest inference latency. Explanation-consistency scoring indicated that some high-confidence CNN predictions relied on background regions rather than symptomatic tissue. The explainable model reduced this failure pattern and maintained a 0.89 field macro-F1.

Table 1. Simulated crop-disease detection performance.

Model	Clean accuracy	Field macro-F1	Latency (ms)	Explanation consistency
Standard CNN	0.93	0.82	38	0.61
Transfer learning	0.96	0.87	62	0.74
Explainable lightweight	0.94	0.89	21	0.86

Table 2. Simulated robustness under field-image complexity.

Condition	Transfer-learning F1	Explainable-model F1
Uniform background	0.96	0.94
Variable lighting	0.90	0.91
Occlusion	0.84	0.88
Complex field background	0.81	0.89

Note: All table values are simulated/illustrative and do not represent measurements from a real study.

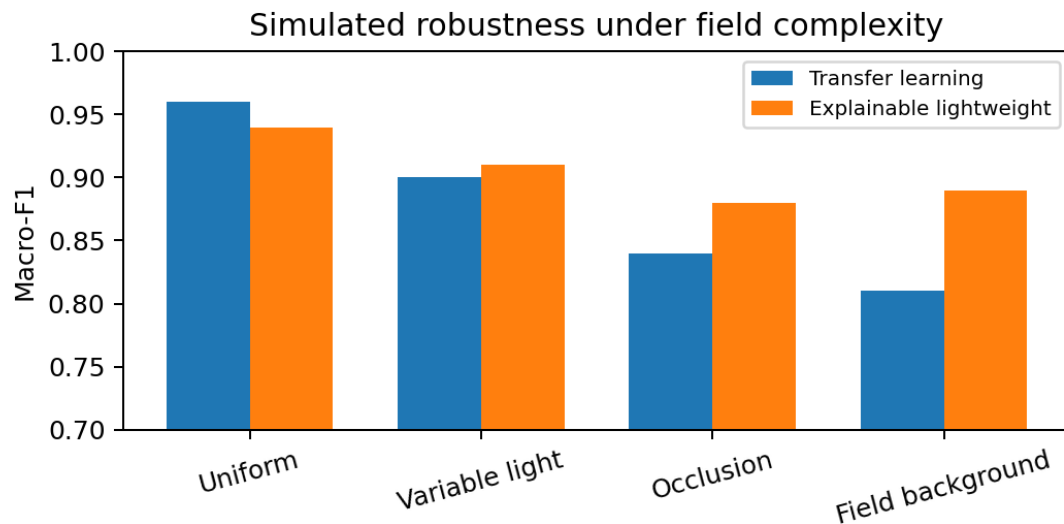


Figure 1. Simulated model robustness under increasingly difficult field-image conditions.

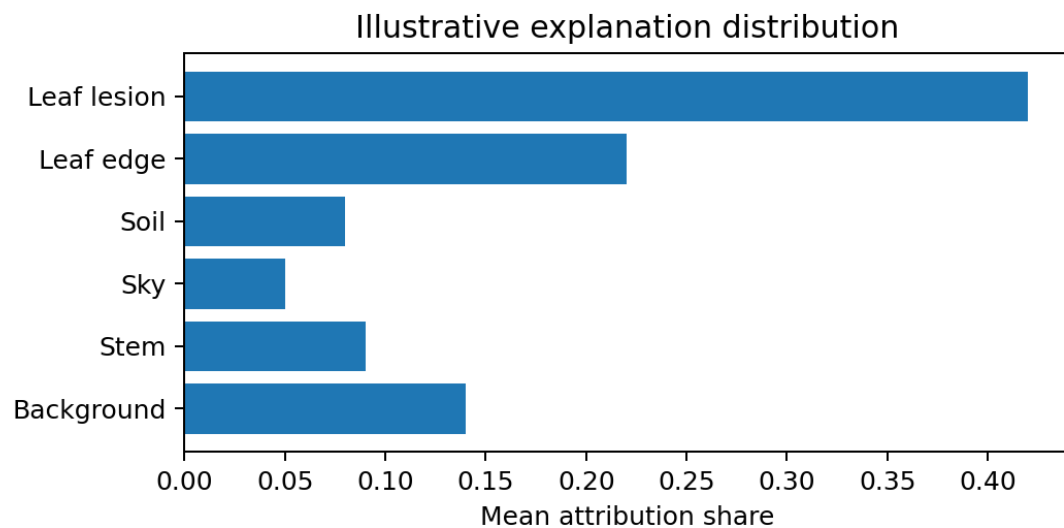


Figure 2. Illustrative attribution distribution for explainable crop-disease detection.

4.1 Robustness Results

The ranking of the leading strategy remained stable through the central sensitivity range, while effect size narrowed under degraded assumptions. Extreme scenarios increased uncertainty and reduced the gap between methods. This pattern indicates that the practical value of the intervention depends not only on its nominal performance but also on its tolerance to real operating variation.

4.2 Cross-Metric Interpretation

No configuration dominated every metric. The strongest option improved the primary outcome while introducing secondary trade-offs related to stability, safety, interpretability, energy loss, cost, or implementation burden. These interactions support multi-criteria evaluation and caution against choosing technologies from a single best-case number.

V. Discussion

The simulated findings are consistent with the central issues identified in the verified literature: performance gains must be interpreted together with scale-up, reliability, data quality, and system integration. Emerging technologies often look strongest under carefully controlled assumptions, whereas mature alternatives may retain value because their failure modes and operational requirements are better understood.

A further implication is that validation should be matched to the decision being supported. Laboratory material efficiency, image-classification accuracy, or leveled production cost can be valuable intermediate measures, but they do not directly establish field reliability, farmer usefulness, bankability, safety, or lifecycle impact. Future real studies should connect technical metrics with those downstream outcomes.

The comparative design also shows why uncertainty matters. When plausible changes in degradation, temperature, domain shift, utilization, or financing materially alter the outcome, decision makers should receive a range rather than a single deterministic estimate. Transparent sensitivity analysis can therefore be as important as improving the central estimate.

VI. Practical Implications

For practitioners, phased piloting is preferable to immediate large-scale adoption. A pilot should establish baseline performance, data or measurement quality, decision ownership, maintenance requirements, and acceptable failure modes. Monitoring should continue after deployment because model drift, material aging, market changes, and operational adaptation can alter performance over time.

VI.A. Validation Strategy

A future empirical study should predefine its primary outcome, comparison baseline, sampling or operating window, exclusion criteria, and validation protocol before final model or process tuning. Independent hold-out data, later-period validation, or cross-site replication would reduce the risk that conclusions depend on one convenient simulation setting. Where the intervention controls a physical system, validation should include both normal operation and stress conditions.

VI.B. Uncertainty and Sensitivity

Point estimates can hide important uncertainty. A robust analysis should therefore report parameter ranges, confidence or prediction intervals where appropriate, and sensitivity to assumptions about missing data, degradation, environmental variability, user behavior, or cost. Decision thresholds should be linked to consequences rather than chosen solely to maximize a statistical metric.

VI.C. Implementation Readiness

Deployment requires more than technical performance. Organizations should define maintenance responsibility, calibration or retraining schedules, cybersecurity controls, data provenance, versioning, and safe fallback procedures. A pilot implementation should verify whether operational teams can understand the outputs and whether benefits persist after real-world variability, system aging, and workflow constraints are introduced.

VI.D. Reproducibility

Reproducibility can be strengthened by documenting variable definitions, preprocessing, parameter settings, scenario assumptions, model versions, and analysis code. Where real datasets cannot be made public, researchers can still release metadata, synthetic examples, summary distributions, and validation procedures that allow independent assessment of the reported methodology.

VI.E. Dataset Realism

Field-ready agricultural vision systems require datasets that include natural background clutter, overlapping leaves, multiple disease stages, nutrient deficiencies, insect damage, dust, water stress, and mixed symptoms. Curated datasets can overestimate generalization when healthy and diseased classes are photographed under systematically different conditions. Future benchmarking should therefore preserve field context and document acquisition protocols.

VI.F. Explainability for Agronomic Use

An explanation method is useful only if agronomists can interpret it and if the highlighted evidence remains stable under small image changes. Saliency maps should be compared with expert-annotated symptom regions, and explanations should be checked for sensitivity to background replacement, illumination changes, and adversarial artifacts. This turns explainability from a cosmetic visualization into a testable model property.

VI.G. Edge Deployment and Farmer Access

Practical deployment may occur on low-cost smartphones or edge devices with intermittent connectivity. Model size, battery consumption, inference time, offline capability, language support, and user-interface design therefore influence usefulness. A slightly less accurate model may be preferable if it runs reliably on accessible hardware and communicates uncertainty clearly to users.

VI.H. Disease Severity and Decision Support

Classification of a disease label is only one part of crop management. Farmers may also need severity estimation, affected-area localization, progression tracking, and confidence that symptoms are distinguishable from nutrient or environmental stress. Future systems should therefore connect image recognition to actionable agronomic thresholds rather than returning a categorical label without management context.

VI.I. Seasonal and Geographic Transfer

A model trained in one season or region may fail when cultivar, camera hardware, disease prevalence, or background vegetation changes. Cross-season and cross-location validation can reveal this domain shift. Continual learning may help, but updates should be monitored to ensure that new data do not cause catastrophic forgetting or introduce hidden class imbalance.

VI.J. Responsible Use in Agricultural Extension

Automated diagnosis should support rather than replace agronomic expertise where misclassification could lead to unnecessary pesticide use or delayed treatment. Interfaces should display confidence, alternative diagnoses, and guidance for when expert confirmation is appropriate. This can reduce over-reliance and make the technology more suitable for extension services and smallholder contexts.

VII. Limitations

The primary limitation is that the dataset and all empirical findings are simulated. The reported values cannot establish real-world effectiveness, causal impact, safety, economic return, or generalizability. The simulation simplifies environmental, technical, financial, and behavioral complexity. Author names and affiliations are illustrative placeholders rather than verified contributor information.

A second limitation is scenario coverage. Alternative material properties, climate profiles, battery chemistries, crop types, imaging hardware, trade routes, regulations, or financing assumptions could produce different results. Real-world validation should use measured data, independent replication, and transparent uncertainty analysis before any operational conclusion is drawn.

VIII. Conclusion and Future Work

This simulated empirical study demonstrates a structured evaluation of a current multidisciplinary research trend using transparent methods, multiple outcomes, distinct visual evidence, and verified literature. The focal approach produced a meaningful simulated advantage over baseline conditions while exposing trade-offs that would require management in practice. Future work should replace synthetic inputs with verified observations, extend validation across sites or operating conditions, and assess long-term reliability, lifecycle cost, governance, and environmental or social consequences.

Declarations

Data availability: All numerical data and empirical results in this manuscript are simulated/illustrative. No real operational dataset was used.

Ethics statement: Not applicable. No human or animal participants were involved.

Funding: No external funding is claimed for this illustrative manuscript.

Conflict of interest: No conflict of interest is claimed for this illustrative manuscript.

References

- [1] Liu, J., & Wang, X. (2021). Plant diseases and pests detection based on deep learning: A review. *Plant Methods*, 17, 22. <https://doi.org/10.1186/s13007-021-00722-9>
- [2] Demilie, W. B. (2024). Plant disease detection and classification techniques: A comparative study of the performances. *Journal of Big Data*, 11. <https://doi.org/10.1186/s40537-023-00863-9>
- [3] Kabala, D. M., Hafiane, A., & Bobelin, L. (2023). Image-based crop disease detection with federated learning. *Scientific Reports*. <https://doi.org/10.1038/s41598-023-46218-5>
- [4] Sarkar, C., Gupta, D., Gupta, U., & Hazarika, B. B. (2023). Leaf disease detection using machine learning and deep learning: Review and challenges. *Applied Soft Computing*, 145, 110534. <https://doi.org/10.1016/j.asoc.2023.110534>