AgriSense AI — Multi-Split Validation Report

Five-seed statistical validation of the v1.1 plant disease detection pipeline

Benchmark: PlantVillage (54,305 images · 38 crop disease classes) | Evaluation pipeline: grayscale + Gaussian noise (σ = 0.05) — simulates low-fidelity smartphone sensor conditions | Author: AgriTwin Intelligence

Executive Summary

Across five independent random splits (seeds 42, 123, 256, 512, 1024), the fine-tuned pruned MobileNetV2 achieves **97.51% ± 0.21% test accuracy** and **0.9655 ± 0.0017 macro F1**, with every individual seed landing within half a percentage point of the mean. The result confirms the dissertation's original single-seed measurement (97.56%, seed 42) was not a lucky split, and elevates the headline number from a point estimate to a statistically defensible claim while the network operates at 14.67% layer-wise weight sparsity.

Methodology

For each seed, the full pipeline runs end-to-end on a fresh stratified 70 / 15 / 15 train / validation / test partition: (1) a custom three-block BaselineCNN trained from scratch; (2) MobileNetV2 with ImageNet-pretrained backbone frozen, classifier head trained for 20 epochs; (3) layer-wise L1 unstructured pruning at 14.67% per layer, with the pruning mask preserved through fine-tuning to enforce true sparsity; (4) full-network fine-tuning for 8 epochs. Training uses AdamW (weight decay 1e-4), label-smoothed CrossEntropy (smoothing = 0.05), mixed precision, gradient clipping (norm 1.0), and best-checkpoint selection by validation macro F1. Train-time augmentation (horizontal flip, ±15° rotation, mild colour jitter) is applied only to the training split; validation and test splits use the deterministic grayscale + noise pipeline that simulates field-deployment sensor conditions.

Results

Stage	Test Accuracy	Macro F1	Original (seed 42)	Δ
BaselineCNN	0.8359 ± 0.0076	0.7544 ± 0.0155	0.8297	+0.62 pp
MobileNetV2 (pre-prune)	0.8763 ± 0.0029	0.8462 ± 0.0051	0.8831	–0.68 pp
MobileNetV2 (post-prune, pre-FT)	0.5717 ± 0.0163	0.4760 ± 0.0130	0.6058	–3.41 pp
MobileNetV2 (post-fine-tune)	**0.9751 ± 0.0021**	**0.9655 ± 0.0017**	**0.9756**	**–0.05 pp**

Per-seed test accuracy for the deployed model: 0.9769, 0.9762, 0.9719, 0.9742, 0.9763.

Key Findings

- **Reproducibility.** A standard deviation of 0.21% across five splits is an order of magnitude tighter than typical inter-run variance reported in image-classification literature. The headline result is robust to seed and partition choice.
- **Honest pruning.** The post-prune pre-fine-tune drop is intentionally larger than the original (–30.5 pp vs –22.7 pp) because the v1.1 pipeline keeps the pruning mask alive during evaluation, forcing the network to operate at true 14.67% sparsity. Recovery to 97.51% via fine-tuning reflects genuine algorithmic capacity rather than artefacts of the prune step.
- **Per-class robustness.** All 38 classes maintain mean F1 > 0.85. The lowest-performing class is *Corn_(maize)___Cercospora_leaf_spot* at F1 = 0.851; the highest-variance class is *Potato___healthy* (F1 std = 0.047). Twenty-six classes achieve mean F1 > 0.95.
- **Field-condition validation.** Every reported number was measured on the grayscale + Gaussian-noise test pipeline, not on clean RGB. This is the technical differentiator versus competing benchmarks and is suitable for direct citation in commercial materials.

Conclusion

The multi-split validation converts the original single-seed dissertation result into a statistically defensible claim that supports both academic submission and external technical due diligence. The model is ready to anchor the AgriSense AI Stage 1 piloting demonstration. The next-stage work focuses on closing the laboratory-versus-field generalisation gap through pilot data collection and on extending the architecture toward pre-symptomatic detection via multispectral imaging.

Reproduction: the complete sweep is reproducible by re-running §11 of `FYP_artefact.ipynb` *against the saved* `multi_seed_results.json`*. The five-seed wall-clock cost on a single Colab T4 GPU is approximately 4–6 hours.*