RESEARCH SUMMARY

Carly Chery

A preregistered equivalence test of phone-deployable vision transformers for ordinal coffee leaf disease severity grading

PG 26038 · Licenciatura thesis, defended December 2026 · EARTH University · Advised by Víctor Hugo Morales Peña, Ph.D.

OSF DOI 10.17605/OSF.IO/9VX3G · registered 2026-04-26 · CC BY 4.0 · open data, code, weights

THE QUESTION

Coffee leaf rust (*Hemileia vastatrix*) and American leaf spot (*Mycena citricolor*) collectively cost Latin American producers up to half a crop's yield at epidemic peak. Severity is graded by eye on coarse ordinal scales whose inter-rater reliability is fragile (ICC 0.80 to 0.96 for trained raters; meaningfully lower without standard-area-diagram support). Manual grading is the operational bottleneck and the noise floor against which any computational method must be judged.

This thesis asks a different question than the one most deep-learning agricultural-pathology papers ask. It asks whether **a phone-sized model can replace a research-grade one**. This is not a superiority test. It is an **equivalence** test: failing to find a difference between two models is not the same as affirmatively claiming they are equivalent. Only a Two One-Sided Tests (TOST) procedure (Schuirmann, 1987; Lakens, 2017) supports that claim.

DESIGN

A preregistered comparative deep-learning framework for automated severity grading of coffee leaf rust and American leaf spot on *Coffea arabica* var. Esperanza L4A5. Methodology, hypotheses, sample size, and analysis plan were locked on OSF *before any leaf was photographed*.

Two paradigms, one comparison

Method A. Semantic segmentation. SegFormer-B3 (Xie et al., 2021): 45 M parameters, 79 G FLOPs, MiT-B3 hierarchical encoder with All-MLP decoder. Compound loss: 0.5 FocalTversky + 0.3 Boundary + 0.2 WeightedCE. Severity grade derived from diseased-area ratio via preregistered ordinal thresholds.

Method B. Ordinal classification. FastViT-T8 (Vasu et al., 2023) with DINOv2-ViT-S/14 LoRA r=8 teacher (Oquab et al., 2024) and CORN ordinal head (Shi et al., 2021). 4 M parameters, 0.7 G FLOPs. KL distillation τ=4.0. Post-hoc temperature scaling (Guo et al., 2017). Phone-deployable; sub-millisecond inference; approximately 113$\times$ more efficient than the segmenter at 11$\times$ fewer parameters.

Three baselines (ResNet-18, U-Net, DINOv2 linear probe) are reported alongside both primary methods.

Five preregistered hypotheses

Family-wise $\alpha = 0.05$ controlled by Holm-Bonferroni over $\{H_1, H_3\}$ only (H_2 is a decomposed component of H_3 and shares its α budget; H_4 is conditional on H_2 or H_3; H_5 is descriptive).

H	Test	Operator	Margin / power
H_1	Difference (BCa paired bootstrap)	$\kappa_{\text{Seg}} \neq \kappa_{\text{FastViT}}$	power ≈ 0.99 at $\lvert\Delta\kappa\rvert \geq 0.05$; MDE ≈ 0.034 at power 0.80
H_2	Non-inferiority (decomposed)	$\kappa_{\text{FastViT}} > \kappa_{\text{Seg}} - 0.05$	shared α with H_3-upper
H_3	**Equivalence (PRIMARY)** TOST	$\lvert\kappa_{\text{Seg}} - \kappa_{\text{FastViT}}\rvert \leq 0.05$	joint $\alpha = 0.05$; power ≈ 0.97 at $\Delta\kappa = 0$
H_4	Efficiency (conditional)	$\text{FLOPs}_{\text{FastViT}} \leq \text{FLOPs}_{\text{Seg}}/10$ AND Memory ratio ≥ 10	expected 113× FLOPs, 11× params
H_5	Ablation (descriptive)	$\Delta\kappa \geq 0.02$ per component	leave-one-out vs. full pipeline

A Bayesian companion (BEST; Kruschke, 2013) is reported alongside every frequentist test. Vanbelle and Albert's (2008) BCa paired bootstrap is the inference engine for κ comparisons.

Materials & Sample

1,400 unique leaves stratified into seven categories of 200 each: Healthy (shared), *H. vastatrix* G1/G2/G3, *M. citricolor* G1/G2/G3. **4,200 calibrated images**: three orientations per leaf (adaxial 0°, 45° oblique, abaxial). The 1,400 adaxial images are annotated for training; the 2,800 oblique and abaxial images are reserved for augmentation and viewpoint-robustness testing. **15 % held-out test** (210 leaves) plus **3-fold grouped CV** (1,190 out-of-fold) gives $N = 1{,}400$ paired predictions for confirmatory tests.

Field site. EARTH Forestal Farm, Caribbean lowlands of Costa Rica (10°13′ N, 35 m elevation). Q2 2026 collection.

Open hardware. Custom 3D-printed phone-imaging rig (v5 two-column gantry) in OpenSCAD with all STLs released. Matte-black PLA throughout (for maximum chromatic separation from leaf colours and to align with SAM/EMSAM/DINOv2 pretraining priors). 20 cm working distance at 0° (143 mm at 45°); ColorChecker calibration in NE corner recess; rigid-mount $\sigma \leq 0.5$ mm. Full build under USD $20 in non-printed parts, ∼19 h print, ∼25 min assembly.

Annotation Pipeline

EMSAM → CVAT → STAPLE. EMSAM (Li et al., 2025) provides pre-annotation; human refinement happens in CVAT; STAPLE consensus (Warfield et al., 2004) is computed on a 10 % triplicate subset (140 leaves) annotated independently by three phytopathologists. A 20-leaf calibration block against STAPLE-consensus reference is required of every rater (per-rater Krippendorff $\alpha \geq 0.80$ to proceed). EMSAM anchoring is controlled via a 30-leaf paired blank-vs-seeded subset captured ≥14 days apart, Latin-square within strata; anchoring magnitude > 0.05 IoU triggers an explicit limitation.

Power

At $\text{SE}(\Delta\kappa) \approx 0.012$ (paired $\rho \approx 0.88$, $N = 1{,}400$ out-of-fold), TOST joint power is **0.97 at** $\Delta\kappa = 0$; H_1 achieves 0.99 power at $\lvert\Delta\kappa\rvert \geq 0.05$. Closed-form values are cross-checked by 10,000-iteration Monte Carlo to within 0.01.

Preregistered Target Metrics

QWK ≥ 0.80 · mIoU ≥ 0.75 · Dice ≥ 0.80 · per-class F1 ≥ 0.85 · Krippendorff $\alpha \geq 0.80$. All five reported regardless of outcome.

Reproducibility Stack

Open dataset (CC BY 4.0), open code, open weights, open hardware. Pipeline: Docker, DVC, Weights & Biases, Optuna. Model cards and datasheets ship with each release. Twenty-three preregistered form fields and thirteen supporting files on OSF.

Pre-Specified Limitations

- **Single cultivar.** *C. arabica* var. Esperanza L4A5 only. Generalisation to Caturra, Catimor, Geisha requires separate validation.
- **Single site.** EARTH Forestal Farm. Climate-specific epidemiology constrains external validity beyond similar agro-ecological zones.
- **Single phone.** Primary dataset on iPhone 11 main camera; the rig also supports a Galaxy S24 cradle, but cross-device colour-pipeline generalisation is exploratory only.
- **No co-infected leaves.** Leaves with simultaneously visible *H. vastatrix* and *M. citricolor* are excluded at sampling time; deferred to future work as a multi-label segmentation problem.
- **Single collection year.** Q2 2026 only; inter-annual disease pressure and lighting phenology not captured.

Directions This Opens for Graduate Work

- **Multi-cultivar generalisation** via domain-adversarial training on Caturra, Catimor, Geisha; tests whether the equivalence claim transfers across genotype-driven phenotypes.
- **Mid-altitude validation** (700 to 1,500 m, covering the bulk of Costa Rican coffee) at one or two cooperatives in Tarrazú or Coto Brus; refit colour pipeline; re-test H_3 on an altitude-stratified held-out set.
- **On-farm mobile deployment study.** Port FastViT-T8 to LiteRT (Android) and Core ML (iOS); pair 5–10 extension agents with the offline app over a season; measure model-vs-agronomist agreement.
- **Co-infection multi-label segmentation.** Drop the single-disease scope; train a multi-label head for rust + spot + co-infection.
- **Active-learning loops at scale.** Use calibrated uncertainty to flag the hardest leaves for expert relabelling; close the loop between model uncertainty and annotation budget.

Selected References

Capucho, A. S. et al. (2011) *Plant Pathology* 60:1144 · Esgario, J. G. M. et al. (2020) *Comp. Electron. Agric.* 169 · Eskes, A. B. (1983) Wageningen PhD thesis · Guo, C. et al. (2017) *ICML* · Kruschke, J. K. (2013) *J. Exp. Psychol. Gen.* 142:573 · Lakens, D. (2017) *Soc. Psychol. Personal. Sci.* 8:355 · Li, J. et al. (2025) *Frontiers in Plant Science* 16:1564079 · Oquab, M. et al. (2024) *TMLR* · Schuirmann, D. J. (1987) *J. Pharmacokinet. Biopharm.* 15:657 · Shi, X., Cao, W. & Raschka, S. (2021) arXiv:2111.08851 · Vanbelle, S. & Albert, A. (2008) *J. Stat. Comput. Simul.* 78:1009 · Vasu, P. K. A. et al. (2023) *ICCV* 5785 · Warfield, S. K., Zou, K. H. & Wells, W. M. (2004) *IEEE Trans. Med. Imaging* 23:903 · Xie, E. et al. (2021) *NeurIPS* 34:12077.

Carly Chery · cchery@earth.ac.cr · ORCID 0009-0005-7842-3200 · EARTH University, Geomatics & Remote Sensing Center · Guácimo, Limón, Costa Rica