Beyond Candidate Coverage: 3D Point-Profile Re-scoring for Individual Tree Detection in Agroforestry

Murilo Salem[1,2], Luísa Böhm[1,2], Henrique dos Reis[1], Lúcio Vagner Carvalho[1], Theo Ferrugem[1], Diniz Arruda Carvalho[2], Anderson Ferrugem[1]

[1] Universidade Federal de Pelotas, Centro de Desenvolvimento Tecnológico (UFPel)
Pelotas – RS – Brasil

{mcsalem, lcbohm, hdreis, lvcarvalho, tdferrugem, ferrugem}@inf.ufpel.edu.br

[2]Center for Embedded Devices and Research in Digital Agriculture (CEDRA)
São Leopoldo – RS – Brasil

{murilo.salem, luisa.bohm, diniz.arruda}@senairs.org.br

Abstract. *Individual tree detection in agroforestry systems is harder than in natural forests: dense, overlapping canopies generate spurious local maxima that are easily confused with real tree apices. We address this with a multimodal benchmark of 10 agroforestry plots (∼3,600 field-annotated trees) combining aerial RGB, 13-channel structural LiDAR, a canopy height model, and the raw 3D point cloud, evaluated under a leakage-free protocol (fixed validation/test plots, macro-averaged F1). We propose a two-stage decoupled detector: a convolutional network generates high-recall candidates, which an ExtraTrees classifier then re-scores using per-candidate engineered features, the vertical 3D profile of the raw point cloud, and a cross-detector agreement signal. The pipeline reaches* $\mathrm{F}_1 = 0.567 \pm 0.004$*, above an end-to-end CNN ensemble (0.530; below the lower bound of an instance-level bootstrap confidence interval, though inside the wider between-plot band) and a classical watershed pipeline (0.352), raising recall from 0.50 to 0.60 while holding precision near 0.55. Critically, a controlled experiment shows that expanding candidate coverage to 99.8% of trees leaves F1 flat, and that none of the ∼25 re-scoring families we tested surpasses the same ceiling: randomized tree ensembles converge on it, while higher-capacity learners fall short of it. We conclude that the performance ceiling is largely a discriminability limit imposed by the available sensors and resolution—rather than by candidate coverage or model capacity—reframing future work toward richer spectral/full-waveform data and label-position review instead of larger models.*

1. Introduction

Mapping individual trees from remote sensing underpins forest inventory, carbon accounting, and—increasingly—the monitoring of agroforestry systems, where trees are deliberately intercropped with agricultural production. Accurate per-tree localization in these systems enables yield estimation, biomass tracking, and management at the level of the individual plant.

Individual tree detection (ITD) is comparatively well studied in natural and plantation forests, where crowns are often large, well separated, and amenable to local-maxima

or watershed segmentation on a canopy height model (CHM). Agroforestry breaks these assumptions. Crowns are dense and mutually overlapping, heights are heterogeneous, and the canopy surface is rough: a single physical tree can produce several local maxima, while several neighbouring trees can merge into one. The result is a regime in which the dominant error mode is not *missing* trees but *failing to discriminate* real apices from spurious peaks that are geometrically indistinguishable on a 2D height surface.

This work makes that difficulty explicit and confronts it directly. We assemble a multimodal benchmark of ten agroforestry plots ($\sim$3,600 field-annotated trees) that pairs aerial RGB, a 13-channel structural LiDAR raster, a CHM, and the raw 3D point cloud, and we evaluate under a leakage-free, plot-held-out protocol with macro-averaged F_1—a deliberately conservative setup that prevents a model from exploiting within-plot spatial correlation.

Our central methodological contribution is a two-stage decoupled detector. A convolutional network is decoded at a tight non-maximum-suppression (NMS) radius to emit high-recall candidates, decoupling the recall problem from the precision problem. A second stage re-scores each candidate with an ExtraTrees classifier that consumes per-candidate engineered features, the vertical 3D profile sampled from the raw point cloud around the candidate, and a cross-detector agreement signal. This pipeline reaches $\mathrm{F}_1 = 0.567 \pm 0.004$, above an end-to-end CNN ensemble (0.530) and a classical watershed pipeline (0.352), and it raises recall from 0.50 to 0.60 without collapsing precision.

Our second contribution is diagnostic and, we argue, the more transferable insight. Through a controlled experiment we show that the residual error is not a coverage problem: expanding the candidate pool until 99.8% of reference trees have a nearby candidate leaves F_1 essentially flat, and none of the roughly twenty-five re-scoring model families we tested surpasses the same ceiling—randomized tree ensembles converge on it, while higher-capacity learners fall short. We conclude that performance is bounded by *discriminability*—the available sensors and resolution simply do not carry enough signal to separate the remaining ambiguous cases—rather than by candidate coverage or model capacity. This reframes the path forward toward richer data (additional spectral bands, full-waveform LiDAR) and label-position review, rather than larger models.

2. Related Work

The classical ITD family operates directly on the canopy height model, combining local-maxima filtering with fixed or variable windows and marker-controlled watershed or region-growing segmentation [Chen et al. 2006, Popescu and Wynne 2004, Koch et al. 2006, Dalponte and Coomes 2016]. These methods are fast, interpretable, and remain strong baselines in coniferous and open-canopy settings, but they break down wherever a single crown is multi-peaked—the agroforestry regime. The tuned watershed pipeline we adopt as our classical baseline [Dalponte and Coomes 2016] ($\mathrm{F}_1 = 0.352$) illustrates the failure mode concretely: its recall saturates early, well short of the coverage the learned proposers reach.

Learning-based detectors have largely displaced this hand-crafted segmentation. Center- and heatmap-regression networks, together with density-estimation approaches, learn crown appearance directly from RGB or rasterized LiDAR [Weinstein et al. 2019, Weinstein et al. 2020] and improve over watershed given sufficient training data. On

small, high-variance datasets, however, they tend to trade recall against precision and are sensitive to architecture scale—a regime our own end-to-end CNN ensemble ($\mathrm{F}_1 = 0.530$) sits within, and one in which we find that enlarging the encoder hurts rather than helps, through overfitting.

A complementary line of work fuses optical and LiDAR signals, which has repeatedly improved canopy tasks by combining RGB with CHM and structural LiDAR rasters [Balestra et al. 2024]. Most of this work fuses *rasterized* LiDAR; comparatively little exploits the raw 3D point distribution at the level of an individual candidate. This is the gap our re-scorer targets: we show that the vertical point profile yields a small but consistent gain because it encodes the apex-versus-spurious-peak structure that any 2D raster flattens away.

The proposal-then-classify pattern we build on is standard in general object detection, and its ITD analogue—"detect then refine"—is usually realized as a second convolutional network. We depart from that convention by pairing a high-recall CNN proposer with a tree-level *tabular* re-scorer, and we report a negative result that motivates the choice: a CNN refiner operating on local patches failed badly ($\mathrm{F}_1 = 0.23$), whereas a randomized tree ensemble over engineered and 3D features succeeded. This is consistent with the broader observation that small-sample, label-noisy regimes favour randomized tree ensembles over high-capacity learners.

Finally, transfer from large external tree datasets is an attractive route when target labels are scarce, and pre-training on resources such as NEON [Weinstein et al. 2021] is a natural candidate. We tested it and found that it did not transfer to agroforestry (fine-tuned $\mathrm{F}_1 = 0.473$, below the 0.499 from-scratch baseline), pointing to a natural-forest versus plantation domain gap. We treat this result as preliminary, however, because the run is confounded by a reduced training batch size forced by hardware contention, as we detail in Section 5.

3. Methodology

3.1. Dataset and evaluation protocol

The benchmark comprises ten agroforestry plots, each exhaustively annotated in the field with the position and height of every tree (208–518 trees per plot; $\sim$3,600 in total). All modalities are co-registered on a common grid at 0.1 m/pixel: aerial RGB (3 channels), a structural LiDAR raster (13 channels: the height statistics $z_{\max}$, z_{mean}, z_{std}, and z_{p95}; point densities in six vertical strata (total, 0.5–2, 2–5, 5–10, 10–20, and $>$20 m); canopy-cover fractions above 2 m and 5 m; and a vertical-roughness measure), a canopy height model, and the raw LiDAR point cloud. The point cloud is retained per plot so that vertical structure can be sampled directly at inference time, rather than only through its rasterized projection. Table 1 summarizes the four co-registered modalities.

Evaluation follows **Protocol A**, which is designed to suppress spatial leakage. We partition the plots into three fixed folds; each fold holds out one validation plot and one disjoint test plot, so that no plot ever contributes to both training and evaluation within the same fold. A prediction is scored as a true positive when it falls within 1.5 m of a reference tree, with assignments resolved by greedy one-to-one matching so that a single detection cannot claim multiple references. We report macro-averaged F_1 together

Tabela 1. Input modalities of the benchmark, all co-registered at 0.1 m/pixel.

Modality	Channels	Role / contents
Aerial RGB	3	Optical crown appearance
Structural LiDAR raster	13	$z_{\max}$, z_{mean}, z_{std}, z_{p95}; densities in six height strata; cover above 2/5 m; roughness
Canopy height model (CHM)	1	Rasterized surface height
Raw LiDAR point cloud	—	Per-plot 3D structure for vertical-profile features

with its precision and recall components. Because the held-out test set spans only three plots, every figure we report is seed-averaged and accompanied by bootstrap confidence intervals [Efron and Tibshirani 1993].

3.2. Stage 1 — high-recall candidate generation

The first stage is a center-based convolutional detector [Zhou et al. 2019] that regresses a tree-center heatmap from the fused multimodal input. We train three complementary variants—a warm-up schedule with stochastic weight averaging [Izmailov et al. 2018] (V5), a decoder augmented with CBAM attention [Woo et al. 2018] (V8), and a strong-augmentation variant (V6)—and decode each at a deliberately tight NMS radius of 0.5 m, well below the radius that would maximize any single model's F_1. The intent is not to produce a clean detection but to over-generate proposals at high recall: decoding tightly decouples recall from precision and lifts candidate coverage far above any individual model's operating point. A single detector recovers roughly 0.68 of the reference trees at this stage (macro recall; up to 0.72 on the easiest plot), whereas the union of the three reaches 0.889—pushing the recall ceiling that the second stage inherits as high as possible before any filtering is applied. Additional higher-recall variants (V2, V11) are trained but enter only the coverage experiment of Section 4.3; they are not part of the main pipeline.

3.3. Stage 2 — tree-level re-scoring

Each surviving candidate is encoded by a compact, heterogeneous feature vector that combines three complementary views of the proposal: cheap 2D summaries from the rasters, the vertical structure of the raw point cloud, and the consensus among the independent detectors. Table 2 lists the three groups and their dimensionality.

Tabela 2. Per-candidate feature groups consumed by the re-scorer (19 features in total).

Feature group	Dim.	Contents
Engineered tabular	8	Heatmap confidence, predicted height, CHM value, peakness, candidate density, source flags
3D vertical profile	10	Point count; mean, std., and 25/50/75/95 percentiles of relative height; RMS lateral offsets (x, y, radial) within 2 m of the candidate
Cross-detector agreement (n_{agree})	1	Number of detectors firing within 0.6 m

The first group captures cheap appearance and context cues—the proposer's heatmap confidence and predicted height, the local CHM value, a local peakness measure, the local candidate density, and binary flags recording which detector(s) produced the candidate. The second group is the one that 2D rasters cannot supply: ten statistics computed from the raw points within a 2 m radius of the candidate—the point count; the mean, standard deviation, and 25th/50th/75th/95th percentiles of height relative to the local minimum; and the root-mean-square lateral offsets along x, along y, and radially—which encode the apex geometry that any height-surface projection necessarily flattens. The third is a single but unusually informative scalar, the cross-detector agreement — the number of independent detectors that fired within 0.6 m of the candidate, a cheap yet strongly discriminative true-positive prior.

An ExtraTrees classifier [Geurts et al. 2006] ($n = 1000$ trees, $\texttt{max_depth} = 16$, $\texttt{min_samples_leaf} = 4$) maps this representation to a calibrated tree/non-tree score. The choice of an extremely-randomized ensemble is empirical rather than incidental (Section 4.3): under the small-sample, label-noisy conditions of this benchmark it consistently outperforms both gradient-boosting and CNN-based re-scorers.

Figure 1 makes the intuition behind the vertical-profile features concrete: two candidates that look equally plausible on a 2D height surface can differ sharply in the distribution of raw LiDAR points around the proposed apex, and it is this difference the re-scorer learns to exploit.

3.4. Operating-point selection

The decision threshold and a final NMS radius are the pipeline's only post-hoc hyperparameters. Both are fixed on the pooled validation plots and never tuned on test data; the selected values (threshold ≈ 0.30, NMS ≈ 1.0 m) are then applied unchanged to every held-out plot. Reporting a single, validation-selected operating point—rather than an oracle tuned per test plot—keeps the protocol honest and ensures that the headline scores reflect deployment-realistic behaviour.

4. Results

4.1. Main comparison

The proposed two-stage detector is the strongest method on the benchmark (Table 3). It improves macro F_1 by $+0.037$ over the end-to-end CNN ensemble and by $+0.215$ over the watershed baseline. The gain decomposes cleanly. Applied to the same V5+V8 pool used by the baseline ensemble, the re-scorer alone lifts macro F_1 from 0.530 to 0.555 ($+0.026$); adding the third detector (V6) to the candidate pool accounts for the remainder. Most of the improvement is therefore attributable to re-scoring rather than to the enlarged proposal set. The nature of the gain is as important as its size: where the high-recall CNN variants buy recall only by sacrificing precision, our pipeline lifts recall from 0.50 to 0.60 while holding precision near 0.55, shifting the entire operating point outward rather than trading one axis for the other. The improvement is also consistent across folds—per test plot (canonical seed) $\texttt{plot_09}$ reaches 0.645, and the hardest fold, $\texttt{plot_03}$, gains the most relative to the ensemble, indicating that the re-scorer helps precisely where the base detectors struggle most.

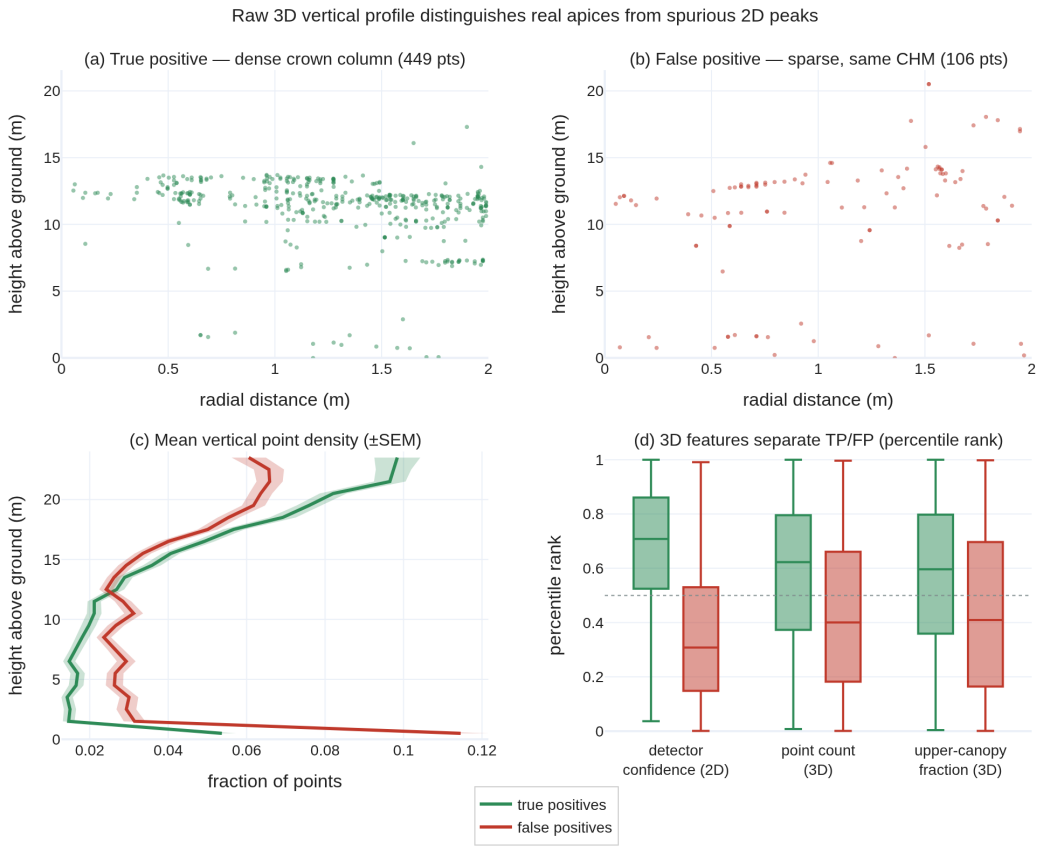


Figura 1. Candidate-level 3D vertical point profiles extracted from the raw LiDAR point cloud within a 2 m radius of each proposal. A true tree candidate typically exhibits a compact vertical structure around the canopy apex, whereas a false candidate may produce a similar CHM response but a less distinctive vertical distribution. The averaged profiles summarize the additional discriminative signal the raw 3D point cloud provides beyond rasterized height features.

Qualitative examples in Figure 2 show that most remaining errors occur in visually ambiguous regions, where overlapping crowns produce multiple plausible apex candidates or hide individual tree centers.

Significance. With only three held-out plots, uncertainty must be reported at two levels. A 5,000-iteration instance-level bootstrap (resampling trees within each test plot) yields a 95% confidence interval of $[0.547, 0.593]$ for macro F_1; the end-to-end ensemble baseline (0.530) lies below its lower bound, so the margin is robust to within-plot sampling noise. This interval is optimistic about between-plot variation, however: per-plot F_1 spans 0.526–0.645, leave-one-plot-out estimates span $[0.533, 0.592]$, and a conservative plot-level bootstrap gives $[0.526, 0.645]$, which contains the baseline score. We therefore read the improvement as consistent across all three folds rather than as statistically established against between-plot variability—a question only a larger panel of test plots can settle.

Tabela 3. Main comparison on the held-out test plots (macro-averaged). Proposed-pipeline metrics are means over five seeds (± s.d. across seeds). The *V5+V8 pool* row isolates the re-scorer's contribution from the enlarged candidate set.

Method	F_1 (macro)	P	R
Classical watershed (Stage 1+2)	0.352	0.394	0.348
End-to-end CNN, best single (V5)	0.499	—	—
End-to-end CNN ensemble (V5+V8)	0.530	0.576	0.497
Two-stage + ExtraTrees (V5+V8 pool)	0.555	—	—
Two-stage + ExtraTrees (V5+V8+V6, ours)	$\mathbf{0.567 \pm 0.004}$	0.552 ± 0.011	0.598 ± 0.009

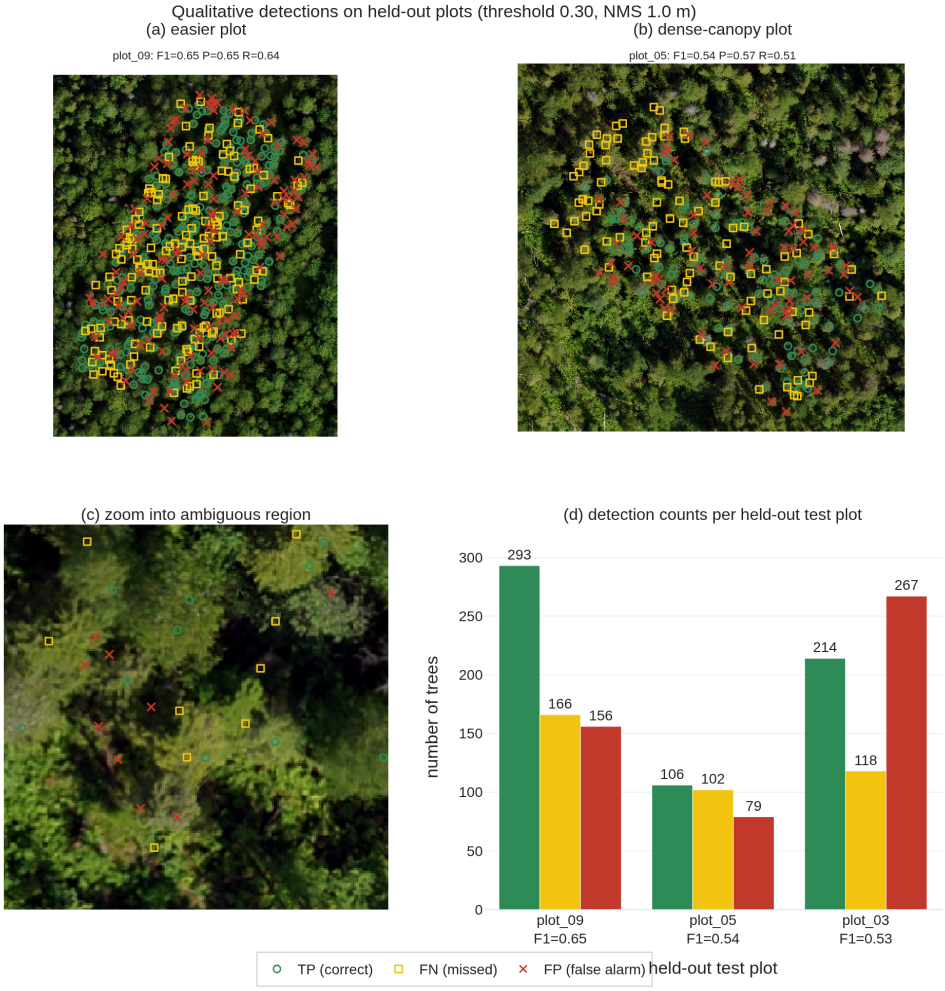


Figura 2. Qualitative detection examples on held-out test plots. Ground-truth tree positions are shown together with true positives, false positives, and false negatives produced by the proposed two-stage detector. The examples highlight the main difficulty of agroforestry ITD: in dense and overlapping canopies, real apices and spurious local maxima can be visually and geometrically similar, making the task substantially harder than detection in more regularly spaced or open-canopy forests.

4.2. Feature ablation

The re-scoring features contribute additively, each layer adding signal the previous one lacked (multi-source pool, seed-averaged; Table 4). Starting from heatmap confidence alone, the engineered 2D tabular features lift F_1 from 0.513 to 0.535, the 3D vertical profile adds a further $+0.012$, and the cross-detector agreement feature completes the model at 0.567. This 3D gain is consistent with the profile's intended role—separating a genuine apex from a spurious 2D peak, a discrimination that vertical structure affords but a flattened height raster cannot. The agreement feature n_{agree} is the single most influential signal of all—it produces the largest single jump in the table. Without it, the validation-selected threshold collapses to roughly 0.2, yielding a high-recall, low-precision regime; adding it moves the operating point to ≈ 0.30 and restores a balanced precision/recall trade-off.

Tabela 4. Feature ablation for the re-scorer (cumulative; rows add one feature group at a time).

Feature set	F_1
Heatmap confidence only	0.513
+ engineered 2D tabular	0.535
+ 3D vertical profile	0.547
+ cross-detector agreement	$\mathbf{0.567}$

4.3. Re-scorer model family

Holding the candidate pool and feature set fixed, we sweep roughly twenty-five classifier families as the re-scorer. ExtraTrees wins clearly and the margin is stable across seeds (Table 5). The runners-up—QDA, kNN, and a Nyström-approximated SVC—trail closely at 0.562–0.564, while the dedicated gradient-boosting libraries (XGBoost [Chen and Guestrin 2016], LightGBM [Ke et al. 2017], CatBoost [Prokhorenkova et al. 2018], HistGBM) settle slightly lower despite their capacity. The two high-capacity or ranking-oriented framings fail outright: a CNN patch refiner collapses to 0.23 and a LambdaRank [Burges 2010] learning-to-rank formulation reaches only 0.40.

The pattern is coherent with the operating regime. With only $\sim$15k candidates and $\sim$20 features, and with label noise concentrated at the decision boundary, bagged, extremely-randomized trees generalize better than either boosting (which chases that boundary noise) or high-capacity learners (which overfit it). We therefore regard this axis as saturated: model choice within the tabular family is essentially settled, and further gains will not come from a better classifier.

4.4. Coverage is not the bottleneck (controlled experiment)

To test whether the remaining error is a coverage problem, we progressively enlarge the candidate pool by unioning in additional (including high-recall) detectors and re-running the full re-scorer end to end (Table 6).

Figure 3 makes the precision–recall redistribution explicit: adding more proposals increases recall, but the corresponding precision loss prevents any meaningful improvement in final F_1.

Tabela 5. Re-scorer model family comparison.

Family	F_1
ExtraTrees (ours)	**0.567**
QDA / kNN / SVC-Nyström	0.562–0.564
HistGBM	~ 0.55
XGBoost / LightGBM / CatBoost	0.548–0.553
CNN patch refiner	0.23
Learning-to-rank (LambdaRank)	0.40

Tabela 6. Effect of candidate-pool coverage on final performance. Values come from the frozen-operating-point verification run, which accounts for the minor offset from the seed-averaged figures of Table 3.

Candidate pool	Coverage	F_1	P	R
V5+V8+V6	0.889	0.569	0.565	0.581
+V11	0.976	0.568	0.510	0.653
+V2	0.993	0.558	0.510	0.634
+V2+V11	0.998	0.560	0.500	0.642

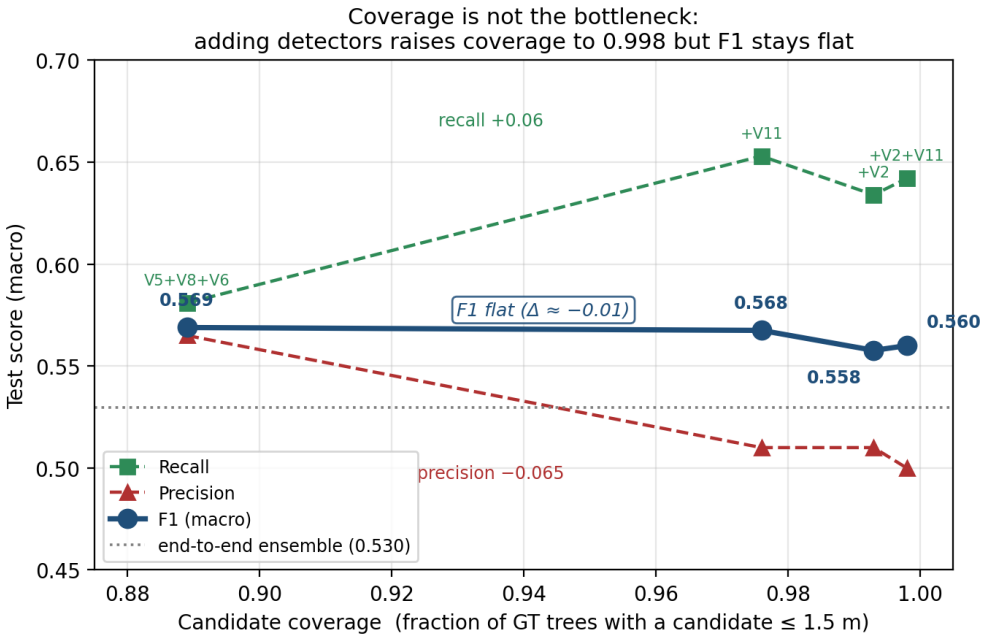


Figura 3. Effect of increasing candidate-pool coverage on final detection performance. Although candidate coverage rises from 0.889 to 0.998, final F_1 remains essentially flat. The additional candidates increase recall but also introduce enough false positives to reduce precision by a comparable amount. This trade-off supports the conclusion that the residual error is not primarily caused by missing proposals, but by limited discriminability among candidate apices under the available sensing modalities and resolution.

As coverage climbs from 0.889 to 0.998—to the point where nearly every reference tree acquires at least one nearby candidate—F_1 does not move. The added proposals are simply redistributed across the precision/recall ledger: recall rises from 0.58 to 0.65 while precision falls by an almost identical margin, from 0.565 to 0.50. Read alongside the model-family plateau of Section 4.3 and the fact that a larger CNN encoder (31.5 M parameters) overfits and degrades to 0.472, this experiment isolates the true bottleneck. The $\sim$40% of trees that go undetected are not absent from the candidate pool; they are present but *indistinguishable* from false positives given RGB, CHM, structural LiDAR, and the 3D profile at the current resolution. The ceiling is therefore one of discriminability—a limit of the available data and features—rather than of candidate coverage or model capacity.

5. Limitations

Several caveats bound the strength of our conclusions. The first is the **small test base**: macro F_1 is averaged over only three held-out plots, and although our bootstrap confidence interval resamples instances within those plots, it is therefore optimistic about between-plot variance; the per-plot, leave-one-plot-out, and plot-bootstrap intervals of Section 4.1 make this gap explicit. The reported figures should be read as benchmark-relative rather than as cross-dataset state-of-the-art. Related to this is a **validation-plot density bias**—two of the three folds use the densest plot for validation, which skews the pooled-validation threshold; a fixed global threshold actually reached slightly higher test F_1, but because it cannot be selected cleanly on held-out data, we report the more honest pooled-validation operating point instead.

Two further caveats concern the scope of our central claim. The **discriminability ceiling** is specific to the present sensor suite and to the 0.1 m resolution at which the modalities are co-registered; it does not preclude gains from additional spectral bands, full-waveform LiDAR, or higher-resolution acquisition, and indeed our analysis is precisely what points toward those directions. We also report an **external pre-training** result only as preliminary: pre-training on NEON ($\sim$28k natural-forest trees) did not transfer to agroforestry (fine-tuned $\mathrm{F}_1 = 0.473 < 0.499$), which suggests a natural-forest versus plantation domain gap, but the run is confounded by a smaller training batch size forced by hardware contention and so cannot be read as a clean negative.

Finally, **label-position uncertainty** likely inflates the apparent discriminability gap. Some predictions counted as false positives may in fact correspond to real trees whose field coordinates are imprecise; the matching radius mitigates this effect but does not eliminate it, so a portion of the residual error we attribute to discriminability may instead reflect annotation noise.

6. Conclusion

We presented a multimodal agroforestry ITD benchmark and a two-stage decoupled detector that pairs a high-recall CNN proposer with an ExtraTrees re-scorer over engineered, 3D point-profile, and cross-detector-agreement features. The pipeline sets the strongest result on the benchmark—$\mathrm{F}_1 = 0.567 \pm 0.004$, a consistent $+0.037$ over an end-to-end CNN ensemble and $+0.215$ over watershed—and improves recall without collapsing precision. Beyond the detector, our controlled coverage experiment delivers the more general

message: once nearly every tree has a candidate and no re-scoring family among the $\sim$25 we tested surpasses that score, the residual error is a *discriminability* limit of the available data, not a coverage or capacity limit. For dense agroforestry canopies, the productive next steps are therefore informational—richer spectral or full-waveform sensing, and review of field-label positional accuracy—rather than larger models or more proposals.

Acknowledgements

This work received technical support from the Center for Embedded Devices and Researchin Digital Agriculture (CEDRA) - SENAI-RS.

Referências

Balestra, M., Marselis, S., Sankey, T. T., Cabo, C., Liang, X., Mokroš, M., Peng, X., Singh, A., Stereńczak, K., Vega, C., Vincent, G., and Hollaus, M. (2024). LiDAR data fusion to improve forest attribute estimates: A review. *Current Forestry Reports*, 10:281–297.

Burges, C. J. C. (2010). From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research.

Chen, Q., Baldocchi, D., Gong, P., and Kelly, M. (2006). Isolating individual trees in a savanna woodland using small footprint lidar data. *Photogrammetric Engineering & Remote Sensing*, 72(8):923–932.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794.

Dalponte, M. and Coomes, D. A. (2016). Tree-centric mapping of forest carbon density from airborne laser scanning and hyperspectral data. *Methods in Ecology and Evolution*, 7(10):1236–1245.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 3146–3154.

Koch, B., Heyder, U., and Weinacker, H. (2006). Detection of individual tree crowns in airborne lidar data. *Photogrammetric Engineering & Remote Sensing*, 72(4):357–363.

Popescu, S. C. and Wynne, R. H. (2004). Seeing the trees in the forest: Using lidar and multispectral data fusion with local filtering and variable window size for estimating tree height. *Photogrammetric Engineering & Remote Sensing*, 70(5):589–604.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 6638–6648.

Weinstein, B. G., Graves, S. J., Marconi, S., Singh, A., Zare, A., Stewart, D., Bohlman, S. A., and White, E. P. (2021). A benchmark dataset for canopy crown detection and delineation in co-registered airborne RGB, LiDAR and hyperspectral imagery from the national ecological observatory network. *PLoS Computational Biology*, 17(7):e1009180.

Weinstein, B. G., Marconi, S., Aubry-Kientz, M., Vincent, G., Senyondo, H., and White, E. P. (2020). DeepForest: A Python package for RGB deep learning tree crown delineation. *Methods in Ecology and Evolution*, 11(12):1743–1751.

Weinstein, B. G., Marconi, S., Bohlman, S., Zare, A., and White, E. (2019). Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural networks. *Remote Sensing*, 11(11):1309.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19.

Zhou, X., Wang, D., and Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*.