Operational Satellite Weed Detection Across 23,000 Sugarcane Fields: Lessons in Temporal Feature Design

C. Ferraz-Pueyo[1], M. Cabrera Dengra[1], R. Fortes Gallego[1], J. L. Poli[2], E. J. Bernardes Júnior[2]

[1] HEMAV Technology SL, Fontsanta 46, 08970 Sant Joan Despí, Barcelona, Spain

[2] HEMAV Brasil Tecnologia EIRELI, Av. T-63, 1296, Ed. New World, sala 1706, Setor Bueno, CEP 74.230-105, Goiânia (GO), Brazil

Corresponding author: cferraz@hemav.com

A paper from the Proceedings of the

17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture

13-15 July 2026

Porto Alegre, Brazil

Abstract. *Weed competition reduces sugarcane (Saccharum spp. hybrids) yields, and managing it at estate scale remains an unsolved problem. Most remote-sensing studies cover tens of fields; scaling to tens of thousands introduces challenges in processing throughput, ground-truth scarcity, and feature design that smaller studies do not encounter. We describe an operational satellite system that now covers 22,969 sugarcane fields (273,493 ha) across eight grower groups in Brazil — anchored on the four mills of the anchor grower group in São Paulo and extended to seven further estates — and delivers two products: weed-presence maps and tree masks. The system ingests Sentinel-2 (10 m), Sentinel-1 SAR, and GammaEarth S2DR 1 m super-resolution imagery through a Bronze–Silver–Gold–Platinum pipeline with daily and monthly refresh. We report results and limitations honestly. For weeds, a temporal Random Forest on per-pixel NDRE statistics provides strong field-level discrimination (cross-validated F1 = 0.778 on the cleaned-label set; 0.893 on the original, pre-cleaning set — see the label-quality finding below), and a training-free earth-observation heuristic reaches F1 = 0.641 but with recall ≈ 1.0 (it flags most of the estate, so it is a coarse pre-filter, not a localiser). The clearest lesson is that* ***label quality, not model capacity, governs the ceiling****: cleaning ambiguous annotations moves F1 from 0.893 to 0.778, a larger swing than any difference between model architectures. Tree/weed spectral confusion is resolved by a dedicated tree detector (field-grouped CV F1 = 0.90). Crucially, the 1 m ML model also localises weeds within fields: on agronomist transect data (fields carrying both weed and clean labels along a transect), its leakage-free within-field AUC is ≈ 0.69, versus ≈ 0.52 (chance) for the EO heuristic — so EO serves as a field-level pre-filter while the 1 m ML model resolves sub-field weed pattern. Dedicated drone validation is the planned next step to tighten the within-field estimates. The pipeline runs operationally with a daily refresh and near-complete processing success on the* *reprocessed anchor-group catalogue (14,605/14,606 fields).*

Keywords. *sugarcane; weed management; temporal features; Sentinel-2; Sentinel-1; S2DR super-resolution; texture analysis; operational deployment; ground-truth curation; precision agriculture*

1. Introduction

Sugarcane underpins Brazil's sugar and bioethanol sectors, and weed competition is a persistent constraint on its productivity. Uncontrolled weed competition during the critical period (approximately the first four months, until canopy closure) can reduce cane yield by roughly 15–40%, exceeding 50% under heavy infestations of aggressive grasses; the magnitude depends on weed species, density, and competition duration (Kuva et al. 2001, 2003). Losses concentrate before canopy closure, when fast-growing grasses and vines exploit the open inter-row.

Weed pressure is observable from space in principle, and a large literature demonstrates weed and crop-stress detection from satellite and UAV imagery. Yet most such studies operate on tens of fields under controlled conditions. Scaling to tens of thousands of operational fields across multiple growers introduces qualitatively new difficulties: acquisition and feature extraction must run reliably at high throughput; ground-truth labels are scarce, noisy, and unevenly distributed; and feature-design choices that look sound at small scale can fail in ways visible only across a large, heterogeneous field population.

This paper reports on an operational system spanning **22,969 fields (273,493 ha) across eight grower groups**, anchored on the four mills of the anchor grower group in São Paulo and extended to seven further estates, delivering two products (weeds and tree masks) from a shared pipeline. Our contributions are:

- A **multi-grower operational architecture** delivering two products at daily-to-monthly cadence, with the engineering and data choices that make it run at scale.
- A direct comparison of earth-observation heuristics and several machine-learning architectures for weed detection under a common ground-truth regime, with **honest negative results** (deep sequence models, SAR fusion) reported rather than suppressed.
- The finding that **label quality, not model capacity, sets the performance ceiling** — a larger F1 swing comes from cleaning ambiguous annotations than from changing architecture.
- Evidence that the **1 m ML model localises weeds within fields**, validated leakage-free against agronomist transects (within-field AUC ≈ 0.69), where the 10 m EO heuristic operates only at field level (≈ chance within fields).

2. Study Area and Data

2.1 Study area

The system covers the sugarcane supply areas of eight grower groups across Brazil, totalling **22,969 georeferenced fields and 273,493 ha** (Table 1; Fig. 6). The anchor estate (Group A) comprises four mills in São Paulo (14,272 fields) on which the methods were developed; the system was then extended to seven further grower groups (Groups B–H in Table 1), one of which is organised into six sub-brands. The field population includes both plant cane and multiple ratoon cycles, with planting and harvest dates tracked in the operational database. Geometry validation is applied before processing because a minority of fields are stored as bounding boxes rather than polygons.

Table 1. Operational scope by grower group (June 2026).

Grower group	Fields
Group A (anchor; 4 mills)	14,272
Group B	3,355
Group C	1,958
Group D	1,033
Group E (6 sub-brands)	1,003

Group F	610
Group G	583
Group H	155
Total	**22,969 (273,493 ha)**

2.2 Satellite data sources

Three sources feed the pipeline. **Sentinel-2** surface reflectance at 10 m (Drusch et al. 2012; bands B02, B03, B04, B05, B08, B8A, B11) supports NDVI (Rouse et al. 1974), NDRE (Gitelson and Merzlyak 1994), EVI2 (Jiang et al. 2008), GNDVI, and NDWI over a 365-day lookback; cloud/shadow are screened with the Scene Classification Layer (SCL; Sinergise n.d.), only acquisitions with ≤ 5% field cloud cover are kept, and temporal aggregation requires ≥ 3 valid observations per pixel. **Sentinel-1** C-band SAR (VV/VH, descending, 10 m; Torres et al. 2012) is evaluated for cloud-penetrating monitoring. **GammaEarth S2DR** provides a 1 m product (indices NDVI, NDRE, NDMI, LWS) used for the within-field weed analysis in §5.2.

S2DR is an external GammaEarth product: its super-resolution algorithm, point-spread function, and effective resolution are not available to us, so we treat it as a black-box 1 m input (for the general super-resolution approach, see Lanaras et al. 2018) and are careful (§5.2) not to attribute structure in its outputs to recovered detail.

Table 2. Data acquisition and quality configuration.

Parameter	Value
Sentinel-2	L2A, 10 m; B02–B11; NDVI/NDRE/EVI2/GNDVI/NDWI
Sentinel-1	C-band SAR, VV/VH, descending, 10 m
S2DR (GammaEarth)	external 1 m super-resolution; NDVI/NDRE/NDMI/LWS
Lookback window	365 days
Cloud screening	SCL, ≤ 5% field cloud cover
Min. valid obs. per pixel	3

3. System Architecture

The system is an operational pipeline run continuously over the multi-grower estate, organised as a four-layer medallion architecture with all I/O mediated by a storage abstraction (Fig. 1).

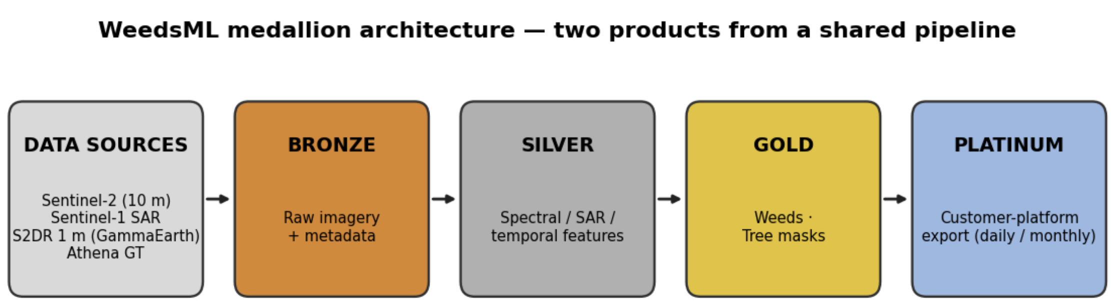

Fig. 1 WeedsML medallion architecture. Three sources (Sentinel-2, Sentinel-1, S2DR 1 m) feed Bronze; Silver computes spectral/SAR/temporal features; Gold produces two products (weeds, tree masks); Platinum exports to the customer agronomy platform.

- **Bronze (raw):** Sentinel-2, Sentinel-1, and S2DR scenes, plus field metadata and ground truth. Immutable.
- **Silver (features):** per-date spectral/SAR indices and per-pixel temporal statistics.
- **Gold (products):** weed-probability maps, tree masks, per-field reports.

- **Platinum (delivery):** exports in the customer platform's format, consumed by the agronomy teams.

Products are combined at delivery through a soft mask-reduction rule that prevents trees from being reported as weeds (Eq. 1):

```
weed_adj = weed_prob × (1 − tree_prob)(1)
```

Cadence differs by product. The **optical weed** pipeline (earth-observation) refreshes **daily** across all growers using parallel workers; the machine-learning weed and tree models refresh **monthly**. A full estate reprocessing of the ~14,606-field anchor-group catalogue (which predates later geometry pruning to the 14,272 anchor-group fields in Table 1) completed **14,605 / 14,606 fields (99.99%)**; that success rate refers to that reprocessing population, not the full 22,969-field multi-grower scope.

4. Methods

4.1 Ground-truth curation

Ground truth derives from agronomist field reports. The report's *category* field records the campaign an inspector was dispatched under — almost always "Weeds" — whereas the free-text *comment* records what was observed; we therefore let comment evidence override the category ("daninha"/"capim" → weed; "cana limpa" → crop; "árvore" → tree; "falha" → gap; "solo exposto" → exposed soil), with uncertain notes down-weighted and non-observations excluded. The curated set has grown over successive iterations to thousands of samples.

4.2 Earth-observation heuristic (weeds)

The training-free baseline exploits temporal instability rather than absolute greenness. A fixed NDVI threshold does not generalise (F1 = 0.374). A single interaction feature, `variation_index = ndvi_mean × ndvi_cv` thresholded at 0.15, raises F1 to 0.641 — but with recall ≈ 1.0 and precision ≈ 0.47 (near the base rate), i.e. it predicts "weed" almost everywhere. Its F1 is therefore largely prevalence-driven, and the heuristic is useful only as a coarse estate-wide pre-filter, not a localiser.

4.3 Temporal feature engineering (weeds)

The machine-learning approach operates on per-pixel time series. For each field the 365-day Sentinel-2 stack is reduced to per-pixel NDRE — computed from the red-edge band (B05) and narrow-NIR band (B8A), both 20 m-native and resampled to 10 m — and 19 temporal statistics are computed per pixel (mean, std, min, max, range, CV, trend, R^2, peak value and timing, percentiles, IQR, skewness, kurtosis, early/late growth, stability, valid-observation count), summarising the shape of the phenological trajectory.

4.4 Machine-learning weed classifiers (10 m)

Under a common ground-truth regime, with cross-validation grouped by field, we compared a temporal Random Forest (Breiman 2001) on the 19 NDRE statistics against KNN-DTW (dynamic time warping; Sakoe and Chiba 1978), an LSTM with attention (Hochreiter and Schmidhuber 1997), and a Transformer (Vaswani et al. 2017); fold counts and sample sizes differed across architectures (see §5.1). The temporal Random Forest was carried forward as the operational 10 m model.

4.5 S2DR 1 m features

From S2DR's four 1 m indices we derive per-pixel texture (local std/CV/range of NDRE at 5×5 and 11×11 windows, NDVI local std; cf. Haralick et al. 1973), spectral, and ratio features, used for the within-field weed analysis in §5.2.

4.6 Tree detector

Trees inside and around fields mimic dense weed patches. The production tree detector is a Random Forest over 40 features (spectral statistics, SAR summaries, and **within-field rank features** — a pixel's percentile rank of CV/min within its own field), trained on 3,560 samples (1,300 tree / 2,260 non-tree) across 68 fields and evaluated with **field-grouped** five-fold cross-validation. The within-field rank features make the model spatially aware while remaining leakage-free. Its output is applied as a soft-reduction mask.

5. Results

5.1 Comparison of weed-detection approaches

On a common time-series dataset (Fig. 2a), KNN-DTW achieved the highest cross-validated F1 (0.882), ahead of the Transformer and temporal Random Forest (both 0.838) and the LSTM (0.792); fold counts and sample sizes differed across architectures (KNN-DTW: 3-fold/819 samples; others: 5-fold), so this is an indicative rather than strict head-to-head, and KNN-DTW's edge does not survive its inference cost at estate scale. The temporal Random Forest was carried forward. The training-free EO heuristic improved from 0.374 to 0.641, subject to the recall ≈ 1.0 caveat of §4.2 (Fig. 2b). All reported F1 are at each model's F1-maximising threshold and are therefore optimistic; we prefer threshold-independent metrics where available.

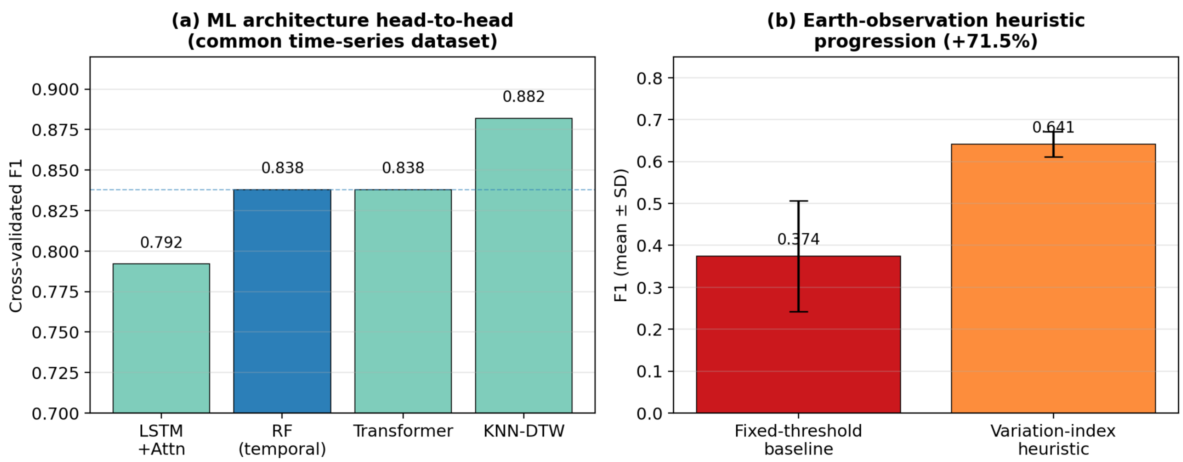

Fig. 2 (a) ML architecture comparison on a common time-series dataset (indicative; see fold-count caveat). (b) EO heuristic progression from a fixed-threshold baseline (0.374) to the variation-index rule (0.641).

Table 3. Weed-detection F1 by approach. EO, 10 m, and S2DR 1 m families use different evaluation sets and are compared within family only; we make no cross-family superiority claim.

Approach	Resolution	Cross-validated F1	Notes
Fixed-threshold NDVI (EO baseline)	10 m	0.374 ± 0.132	Does not generalise
Variation-index heuristic (EO)	10 m	0.641 ± 0.030	recall ≈ 1.0; coarse pre-filter only
LSTM + attention	10 m	0.792	no cost advantage
Transformer	10 m	0.838	no gain over RF
Temporal RF (common set)	10 m	0.838	carried forward
Temporal RF v14 (NDRE, original curated set)	10 m	0.893	pre-cleaning; see §5.3
Temporal RF (cleaned-label set, v16–v33)	10 m	0.73–0.79	operational range
S2DR 1 m weed models (texture/spectral/combined)	1 m	0.76–0.84	separate set; see §5.2

5.2 Within-field weed localisation, validated on transects

A spatially explicit weed product must resolve *where within a field* weed pressure sits, not just whether a field is weedy. The agronomist ground truth lets us test this directly: many fields are sampled along **transects**, and 41 fields carry both weed and clean labels within the same field (≈ 1,160 samples). Using leakage-free out-of-fold predictions, we ask whether each model ranks the weedy locations above the clean ones within a field (within-field AUC).

The **S2DR 1 m ML model localises weeds within fields**: mean per-field within-field AUC = 0.67 and a within-field-centred pooled AUC of 0.69 (Fig. 3), consistently above chance across model variants. The per-field distribution is bimodal — roughly half the fields (often the smaller ones) score a perfect 1.0 while ≈ 40% sit at or below chance — so the centred pooled AUC of 0.69 is the preferred single summary, not the higher 0.88 median. The **EO heuristic does not** localise within fields (within-field AUC ≈ 0.52, indistinguishable from chance), consistent with its recall ≈ 1.0 behaviour — it is a field-level pre-filter. EO and ML are each evaluated on their own out-of-fold predictions over different transect subsets (ML: 41 fields / 1,162 samples; EO: 64 fields / 269 samples), so each AUC should be read against its own chance line of 0.5 rather than as a matched head-to-head; the asymmetric claim (ML above chance, EO at chance) holds on each set independently.

This is the operationally important claim — the 1 m model sees inside the field — and it is established on real agronomist data, not synthetic variance. We keep two honest bounds. First, per-field AUC is variable: a minority of fields sit at chance, and the transect set is modest (41 fields). Second, within-field *variance* alone is not evidence of localisation — a cubic-upsampling control of the 10 m map reproduces the 1 m variance (control σ ≈ 0.17 vs S2DR 1 m σ ≈ 0.15 vs 10 m σ ≈ 0.03) — so it is the *transect agreement*, not the map texture, that supports the claim. Dedicated drone/orthomosaic reference at sub-metre scale is the planned next step to tighten this estimate.

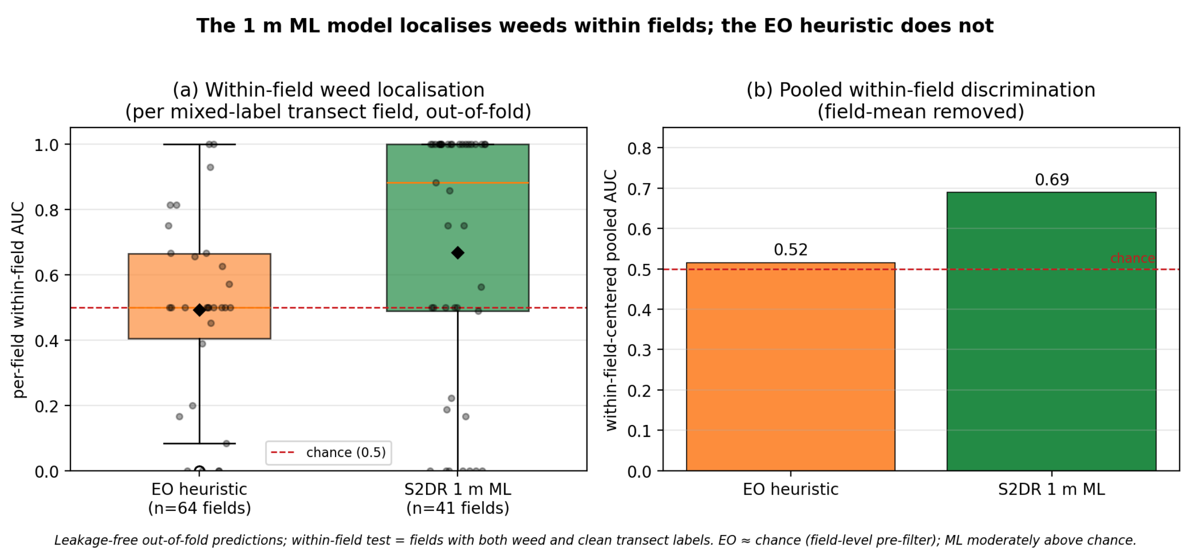

Fig. 3 Within-field weed localisation on agronomist transect fields (those with both weed and clean labels), using leakage-free out-of-fold predictions. (a) Per-field within-field AUC: the S2DR 1 m ML model sits well above the chance line; the EO heuristic at chance. (b) Within-field-centred pooled AUC (field means removed): ML 0.69 vs EO 0.52.

5.3 The label-quality ceiling

The temporal Random Forest reaches F1 = 0.893 on the original 1,282-sample NDRE set. After ambiguous annotations are removed and labels cleaned, the same approach settles near 0.778 on the smaller cleaned set (≈960 samples), with feature variants spanning ≈ 0.73–0.79 (Fig. 4). The 11–12 point swing attributable to annotation quality alone exceeds the gap between any two architectures we tested (§5.1) — establishing label quality, not model capacity, as the binding constraint. The 0.893 figure should be read as a pre-cleaning artifact, not the operational number.

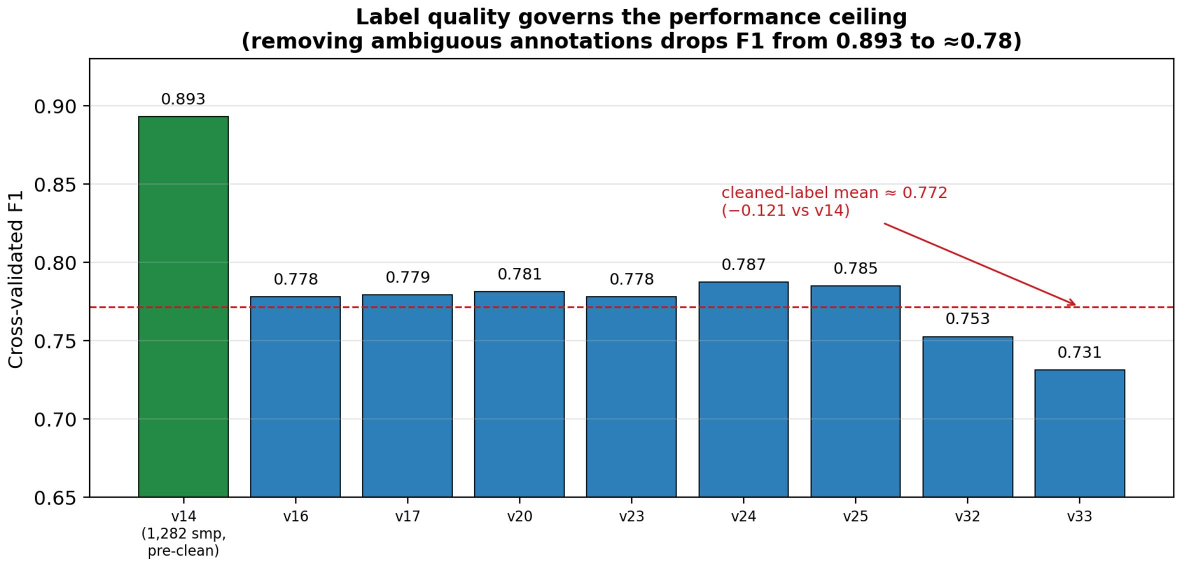

Fig. 4 Label quality governs the performance ceiling. The v14 model scores F1 = 0.893 on the original 1,282-sample set; after ambiguous annotations are cleaned, the same approach settles near 0.778 (cohort v16–v33).

5.4 Tree detector

The production tree detector achieves a **field-grouped cross-validated F1 of 0.900 ± 0.057** (precision 0.954, recall 0.857), resolving the tree/weed confusion that otherwise inflates weed estimates (Table 4). Its most important features are within-field rank features — a pixel's NDVI/EVI2 CV-rank and min-rank relative to its own field (Fig. 5) — which encode spatial context without leaking field identity. (An earlier prototype's apparent F1 of 0.976 was an artifact of a single-field test split and field-constant features; the field-grouped 0.90 is the credible estimate.)

Table 4. Tree-detector performance (Random Forest, 40 features incl. within-field rank features).

Metric	Field-grouped 5-fold CV
F1	0.900 ± 0.057
Precision	0.954
Recall	0.857
Per-fold F1 range	0.815 – 0.970
Samples / fields	3,560 / 68

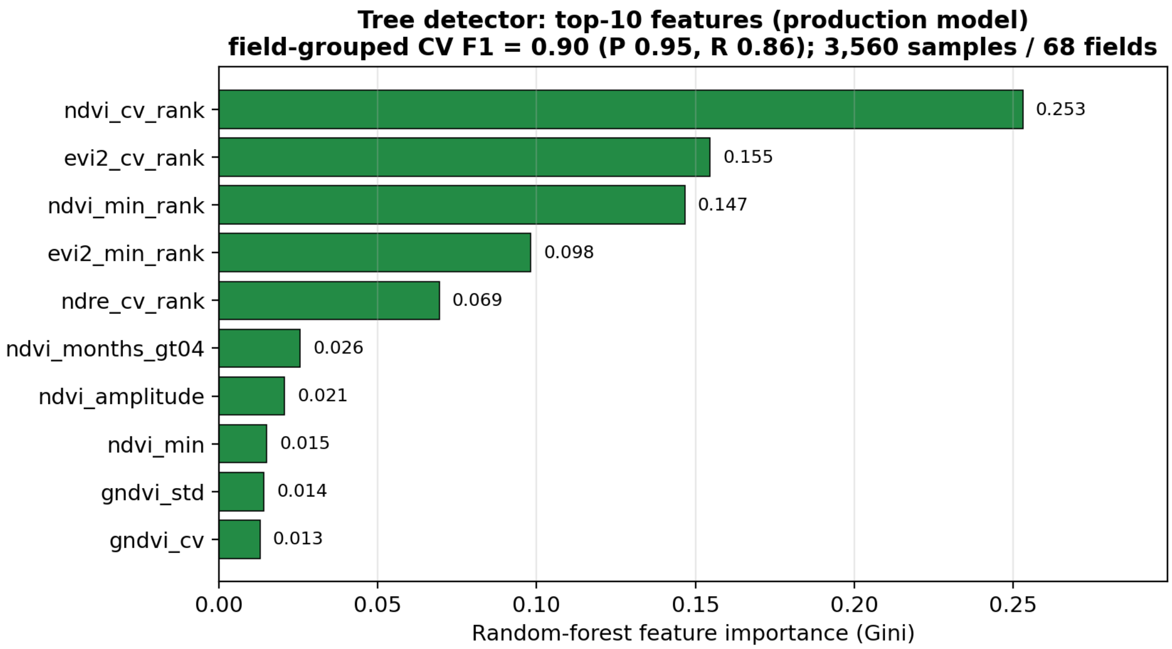

Fig. 5 Tree-detector top-10 feature importances (production model). Within-field rank features dominate — spatially meaningful and leakage-free.

5.5 Operational scale

The system runs across 22,969 fields and eight grower groups (Fig. 6), with a daily EO refresh and monthly ML refresh.

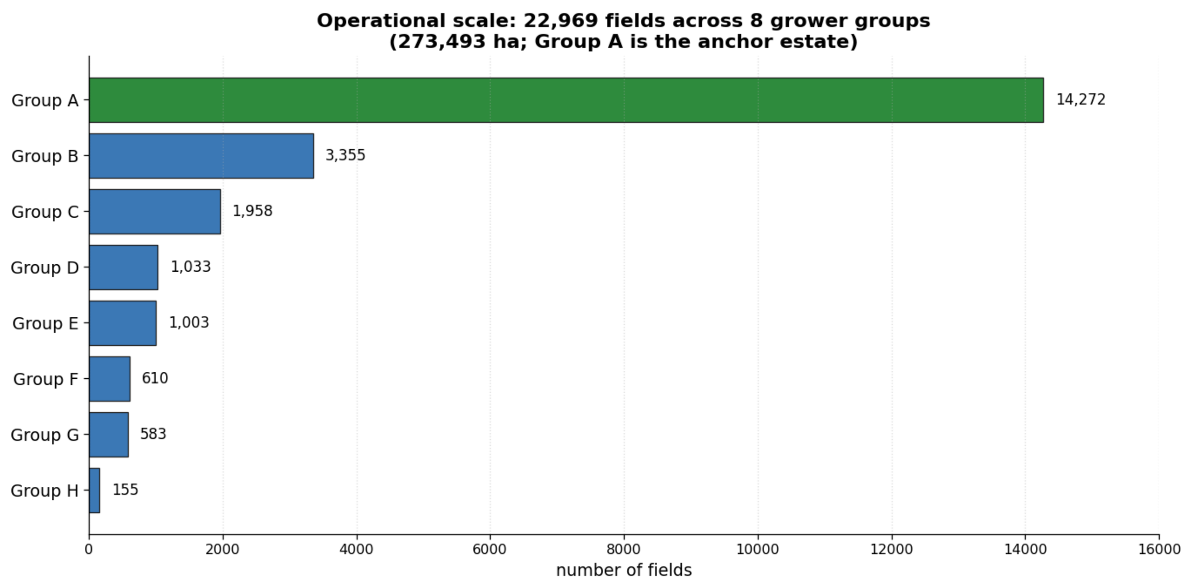

Fig. 6 Operational scale: fields per grower group (22,969 total, 273,493 ha). Group A is the anchor estate; seven further grower groups were onboarded onto the same pipeline.

6. Discussion

Several lessons stand out, all easier to miss on a handful of fields.

First, **the performance ceiling is set by labels, not models.** The 0.893→0.778 swing under label cleaning dwarfs architectural differences; the improvement roadmap is a better ground-truth programme, not a larger model.

Second, **the 1 m ML model localises weeds within fields, and the EO heuristic does not.** On agronomist transects the ML within-field AUC is ≈ 0.69 versus ≈ chance for EO, so the operational division of labour is EO for field-level triage and the 1 m ML model for sub-field pattern. Within-field *variance* is not itself evidence (a non-SR upsampling control reproduces it); the transect agreement is. We argue within-field claims for spatially explicit products should be held to exactly this standard — agreement with within-field reference data, not map texture — and that dedicated drone reference is the right way to strengthen it.

Finally, the **negative results are informative.** Deep sequence models and SAR-optical fusion did not justify their cost in the configurations tested; SAR remains promising specifically for wet-season cloud gaps.

7. Limitations

Weed detection is binary (no species discrimination), limiting direct herbicide selection, and the EO heuristic's recall ≈ 1.0 makes it a coarse field-level pre-filter. Within-field weed localisation (ML) is validated only on a modest transect set (41 fields, ≈ 1,160 samples) at moderate skill (AUC ≈ 0.69) with per-field variability; dedicated drone reference is the planned next step. S2DR is an external black-box product whose effective resolution we cannot independently characterise; its availability is intermittent. Reported F1 are threshold-optimised and the cross-architecture comparison used non-uniform folds. Ground-truth density is uneven across the eight growers, and absolute skill should be read in light of label quality.

8. Conclusion

Operating across 22,969 fields and eight grower groups surfaced lessons that smaller studies cannot: that annotation quality, not model capacity, governs the performance ceiling; that the 1 m ML model localises weeds within fields (transect-validated AUC ≈ 0.69) while the 10 m EO heuristic works at the field level. The pipeline delivers two products — weeds and tree masks — at a daily-to-monthly cadence over 273,493 ha. The clearest paths to improvement are better-curated ground truth and dedicated drone reference to strengthen the within-field validation.

Acknowledgements

This work would not be possible without the agronomists and field teams whose observations provide the foundation for every model evaluated.

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., et al. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120, 25–36.

Gitelson, A. A., & Merzlyak, M. N. (1994). Spectral reflectance changes associated with autumn senescence of *Aesculus hippocastanum* L. and *Acer platanoides* L. leaves: spectral features and relation to chlorophyll estimation. *Journal of Plant Physiology*, 143(3), 286–292.

Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6), 610–621.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

Jiang, Z., Huete, A. R., Didan, K., & Miura, T. (2008). Development of a two-band Enhanced Vegetation Index (EVI2) without a blue band. *Remote Sensing of Environment*, 112(10), 3833–3845.

Kuva, M. A., Gravena, R., Pitelli, R. A., Christoffoleti, P. J., & Alves, P. L. C. A. (2001). Períodos de interferência das plantas daninhas na cultura da cana-de-açúcar. II – *Brachiaria decumbens*. *Planta Daninha*, 19(3), 323–330.

Kuva, M. A., Gravena, R., Pitelli, R. A., Christoffoleti, P. J., & Alves, P. L. C. A. (2003). Períodos de interferência das plantas daninhas na cultura da cana-de-açúcar. III – *Brachiaria decumbens* e *Panicum maximum*. *Planta Daninha*, 21(1), 37–44.

Lanaras, C., Bioucas-Dias, J., Galliani, S., Baltsavias, E., & Schindler, K. (2018). Super-resolution of Sentinel-2 images:

learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146, 305–319.

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. In *Proceedings of the Third Earth Resources Technology Satellite-1 Symposium* (Vol. 1, pp. 309–317). Washington, DC, USA: NASA.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1), 43–49.

Sinergise. (n.d.). Scene classification (SCL). *Sentinel Hub documentation*. Resource document. https://docs.sentinel-hub.com (last accessed 15 June 2026).

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., et al. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment*, 120, 9–24.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Red Hook, NY, USA: Curran Associates.