A Machine Learning Framework for Crop Productivity Classification and Risk Assessment

João Pedro de Moraes Xavier[1], Kelyn Schenatto[1], Glauco Vieira Miranda[2], Claudio Leones Bazzi[1], Ricardo Sobjak[1] and Marlon Rodrigues[1]

[1]Federal University of Technology – Paraná (UTFPR), Medianeira, Paraná, Brazil

[2]Federal University of Technology – Paraná (UTFPR), Santa Helena, Paraná, Brazil

Corresponding authors: joaopedroxavier@alunos.utfpr.edu.br; kschenatto@utfpr.edu.br

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. Field-scale crop productivity classification from satellite data was evaluated for corn (Zea mays), soybean (Glycine max), and wheat (Triticum aestivum) in western Paraná, Brazil. Four machine learning (ML) algorithms — Random Forest (RF), Gradient Boosting (GB), K-Nearest Neighbors (KNN), and Logistic Regression (LR) — were compared on a harmonized Landsat/Sentinel-2 time-series integrating HLS v2.0 satellite reflectance with NASA POWER climate data across 29,835 observations from 2,560 unique 10×10 m grid cells. A dominant temporal feature (harvest_day_of_year) accounted for 47–64% of model importance and was removed to prevent near-deterministic data leakage. Without it, three-class accuracy ranged from 43.6–49.8% (corn), 45.7–48.4% (soy), and 58.0% (wheat), all above a 33.3% dummy baseline. The wheat binary model reached 84.5% accuracy (GB, LR, KNN), 17.8 percentage points (pp) above the majority-class baseline, and was confirmed as a genuine pre-harvest signal by three structurally different algorithms. Bootstrap 95% confidence intervals and McNemar's test confirm that differences among the top algorithms are not statistically significant within each crop.

Keywords. remote sensing; feature engineering; Random Forest; Gradient Boosting; precision agriculture

Introduction

Early prediction of poor crop yields allows farmers to redirect inputs or adjust logistics. While satellite-based classification offers a scalable solution, agronomic literature is inconsistent regarding its practical success. Two gaps motivated this work: limitations in single-mission studies due to cloud cover, and reliance on snapshot spectral indices rather than full-crop-cycle features.

The authors are solely responsible for the content of this paper, which is not a refereed publication. Citation of this work should state that it is from the Proceedings of the 17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture. EXAMPLE: Last Name, A. B. & Coauthor, C. D. (2026). Title of paper. In Proceedings of the 17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture (unpaginated, online). Monticello, IL: International Society of Precision Agriculture and Brazilian Congress on Precision and Digital Agriculture.

We address these by using the Harmonized Landsat and Sentinel-2 (HLS) dataset and by conducting systematic field-scale feature evaluation of corn, soybean, and wheat in western Paraná, Brazil.

A key practical advantage is that the entire pipeline operates without any field visits. All inputs (HLS surface reflectance and NASA POWER climate data) are freely available through public archives and APIs, allowing pre-harvest risk assessments to be generated for any field with a known location and harvest calendar. This no-visit design makes the framework especially relevant for large-scale precision agriculture applications where ground surveys are costly or impractical.

An important methodological contribution is the explicit identification of harvest_day_of_year as a near-deterministic leakage feature. Before ablation, it accounted for 47–64% of Gini importance across all five tested algorithms, and the wheat binary task reached 98.7% accuracy identically across five algorithm families — a diagnostic sign of task degeneracy rather than genuine model learning. The remainder of this paper reports exclusively on ablated results using only pre-harvest predictors.

Materials and Methods

Ground-truth yield monitor data (t/ha) spanning diverse climatic seasons (2013–2017, 2023–2026) were cleaned with a per-crop, per-year interquartile range (IQR) criterion to remove harvester artifacts, and aggregated to a 10×10 m grid. The final dataset comprises 29,835 observations from 2,560 unique grid cells covering approximately 25.6 ha, representing about 39% of the 64.9 ha bounding box. The complete data processing and machine learning workflow is systematically illustrated in Fig. 1, spanning from spatial aggregation to final model evaluation. Three crops are represented: corn in five seasons (13,905 obs.), soybean in five seasons (12,258 obs.), and wheat in two seasons (3,672 obs.). The non-contiguous multi-year design intentionally spans both drought years and favorable seasons to give the models exposure to climatic variability.

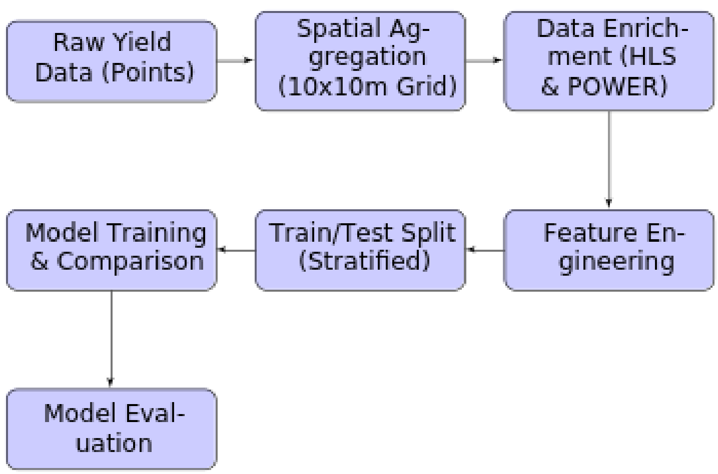

Fig. 1. Data processing and machine learning workflow.

Satellite reflectance came from HLS v2.0, which harmonizes Sentinel-2 MSI and Landsat-8 OLI observations by correcting for BRDF effects. Clouds and cloud shadows were masked using the Fmask quality band; observations with NDVI ≥ 0.99 (physically unrealistic saturation) were additionally excluded. A Whittaker smoother ($\lambda = 100$) was applied to reconstruct a continuous daily time-series, fully resolving all cloud-induced gaps. Climate variables (precipitation, temperature, solar radiation, VPD) were obtained from NASA POWER.

All features were computed from observations available up to a fixed prediction cutoff of 30 days before the expected harvest date (`prediction_lag_days = 30`). Two feature categories were engineered: (1) vegetation phenology indices from the smoothed HLS time-series (peak NDVI, NDVI AUC, NDVI standard deviation, green-up slope, senescence slope, EVI at cutoff); and (2) climate features aggregated over phenological windows (growing degree days, solar radiation, VPD, precipitation, dry-day counts, heat×drought interaction per stage). The top 10 features were selected via Mutual Information (MI) from approximately 57–64 candidates.

Productivity was divided into terciles (Low, Medium, High). We compared four classifiers—RF (class_weight=‘balanced’), GB (stochastic regularization via subsample, no explicit class balancing), LR (class_weight=‘balanced’, StandardScaler), and KNN (weights=‘distance’, StandardScaler)—using hyperparameter tuning via randomized search (30 iterations, 5-fold stratified CV, macro-F1 scoring). Data were split 75/25 with stratification. A majority-class dummy classifier provided baselines: 33.3% (three-class) and 66.7% (binary). Performance was assessed via accuracy, macro-F1, bootstrap 95% CIs (1,000 resamples), and McNemar pairwise significance tests with Edwards’ correction.

Results

Overall classification performances for both three-class and binary tasks across the four evaluated algorithms are summarized in Table 1.

Table 1. Classification accuracy (%) for three-class and binary tasks across four algorithms. Best result per row in bold.

Crop	Task	RF (%)	GB (%)	KNN (%)	LR (%)
Corn	3-class	**49.8**	49.7	48.8	43.6
Soy	3-class	47.4	47.5	**48.4**	45.7
Wheat	3-class	**58.0**	**58.0**	52.3	53.4
Corn	Binary	72.5	73.4	**73.5**	61.7
Soy	Binary	72.7	**74.9**	**74.9**	71.5
Wheat	Binary	72.8	**84.5**	**84.5**	**84.5**

Note: Dummy baselines: 33.3% (three-class); 66.7% (binary). RF = Random Forest; GB = Gradient Boosting; KNN = K-Nearest Neighbors; LR = Logistic Regression.

Productivity Classes and Three-Class Performance

Productivity was divided into perfectly balanced terciles (Low, Medium, High), resulting in total field counts of 13,905 for corn, 12,258 for soy, and 3,672 for wheat. Because these were terciles, the boundaries between classes were extremely narrow. For instance, the soy Low–Medium boundary (3.67 t/ha) and Medium–High boundary (4.31 t/ha) were separated by only 0.64 t/ha.

This narrow separation made three-class classification difficult. Compared to a 33.3% baseline, accuracies across all four algorithms clustered tightly: 43.6–49.8% for corn (max macro-F1: 0.479) and 45.7–48.4% for soy (max macro-F1: 0.474). Wheat was more tractable, with both RF and GB reaching 58.0% (F1: 0.590). In every crop, the Medium class had the worst F1 score, reflecting that Medium fields are spectrally indistinguishable from their neighbors by definition. Because model accuracies are clustered within a 6 percentage-point margin, the path forward for these crops is better features, not better models.

Binary Classification Performance

Collapsing Medium and High into a “Not Low” category (66.7% baseline) substantially improved performance. GB, LR, and KNN all reached 84.5% accuracy (F1: 0.801) for wheat, 17.8 pp above baseline, proving that the remaining pre-harvest feature set is genuinely sufficient for strong binary separation. The convergence of three structurally different classifiers — a boosting ensemble, a linear model, and a distance-based method — to the same result confirms this is a genuine signal, not algorithm-specific overfitting.

For GB and KNN, accuracy (F1: 0.663) was tied at 74.9%, an 8.2-point gain over the baseline. For corn, three algorithms beat the baseline, peaking with KNN at 73.5% (F1: 0.647). This revealed vegetation-based discriminative signals previously hidden by data leakage. However, Low-class recall for corn remained poor across all models (0.33–0.43), meaning most genuinely poor corn fields were still missed.

Statistical Validation

Bootstrap 95% confidence intervals confirm that the wheat binary result (GB: [82.1, 86.7]%) is clearly separated from corn ([72.0, 74.8]%) and soy ([73.3, 76.4]%). To evaluate the robustness of the top-performing models, bootstrap 95% confidence intervals were computed for each crop-task combination, as detailed in Table 2. McNemar pairwise tests (Edwards' correction) confirm that differences between the top models are not statistically significant within each crop: corn RF vs. GB $p = 0.78$; soy RF vs. GB $p = 0.92$. For wheat three-class and wheat binary, RF and GB produced identical predictions ($b = c = 0$), reported as complete agreement rather than a formal p-value. For the binary tasks, GB vs. KNN was statistically indistinguishable for soy ($p = 0.86$) and corn ($p = 0.79$), while both were significantly better than LR ($p < 0.001$). These results confirm that algorithm choice is essentially irrelevant once the feature set is fixed.

Table 2. Bootstrap 95% confidence intervals (1,000 resamples) for the best model per crop-task. Wheat binary [82.1, 86.7]% is clearly separated from all other binary results.

Crop	Task	Model	Accuracy (%)	Acc. 95% CI (%)	F1 95% CI
Corn	3-class	RF	49.8	[48.2, 51.5]	[0.463, 0.496]
Soybean	3-class	KNN	48.4	[46.8, 50.1]	[0.447, 0.482]
Wheat	3-class	RF	58.0	[55.0, 61.4]	[0.560, 0.623]
Corn	Binary	KNN	73.5	[72.0, 74.8]	[0.629, 0.664]
Soybean	Binary	GB	74.9	[73.3, 76.4]	[0.638, 0.678]
Wheat	Binary	GB	84.5	[82.1, 86.7]	[0.771, 0.829]

Note. CI = Confidence Interval; Acc. = Accuracy. RF = Random Forest; GB = Gradient Boosting; KNN = K-Nearest Neighbors.

Feature Importance

After removing `harvest_day_of_year`—which previously accounted for 47–64% of model importance and caused near-deterministic leakage—importance profiles became more evenly distributed and agronomically interpretable. SHAP (SHapley Additive exPlanations) values, Gini impurity reduction, and Mutual Information were computed independently and agreed on the same dominant feature groups across all six crop-task combinations.

For corn, peak_ndvi_to_date leads all three metrics (SHAP: 0.039, Gini: 26.1%, MI: 0.126), but the top seven NDVI-derived features are nearly indistinguishable (MI range: 0.118–0.128), confirming no single spectral metric dominates the corn yield signal. For soybean, SHAP places inter_heat_dry_Flowering (mean |SHAP| = 0.044) at the top—a heat×drought interaction during the flowering window—which is the most notable divergence from Gini (rank 4, 14.5%). For wheat, all three methods agree: peak_ndvi_to_date (SHAP: 0.045) and dry_days_Tillering (0.043) rank first and second, consistent with the decisive role of peak canopy development and water stress at tillering in winter cereals. The wheat binary model's top features shift toward recent climate: t2m_mean_30d (SHAP: 0.068), precip_7d (0.061), and total_precip_Ripening (0.060), reflecting that temperature and precipitation close to harvest are the most informative pre-harvest signals.

Discussion

The clearest finding is that for corn and soy, algorithm choice is essentially irrelevant for three-class accuracy. All four algorithms clustered within 6 pp, and McNemar tests confirm no statistically significant differences between the top models within any crop. When the seven most informative features have nearly indistinguishable MI scores, no classifier, however expressive, can extract a signal that is not present. For these crops, the path forward is better features, not better models.

The ablation methodology is the most transferable contribution of this study. The diagnostic pattern—five structurally diverse algorithms converging to identical accuracy—provides a reusable template for detecting dominant-feature artifacts in agricultural ML pipelines. Any temporal feature correlated with calendar-based management decisions (planting date, harvest timing) should be treated as a leakage candidate, particularly when accuracies across structurally diverse algorithms converge.

Wheat shows the strongest operational case. GB, LR, and KNN reached 84.5% binary accuracy, 17.8 pp above baseline, using only pre-harvest vegetation and climate features. In practice, an agronomist can run this model 3–4 weeks before harvest by querying the HLS archive and NASA POWER API, generating a field-level risk map without any field visits. The RF configuration (72.8% accuracy, perfect Low-class recall) may be preferable in insurance contexts where false negatives are more costly than false alarms.

Soybean is practically usable at 74.9% binary accuracy, provided users understand that approximately one in three at-risk fields will be missed (Low-class recall: 0.35). Three-class performance remains limited by canopy saturation: once the canopy closes, NDVI and EVI lose sensitivity to yield variation. Corn binary classification is not recommended for standalone operational use. Low-class recall of 0.33–0.43 means the model misses more than half of genuinely poor fields; inter-annual climate variability not resolvable from the current feature set appears to be the dominant source of ambiguity.

Three limitations warrant emphasis. First, all data come from a single spatially contiguous study area (~25.6 ha) in western Paraná; the 75/25 split was performed at the grid-cell level without spatial blocking, so results should be interpreted as an *upper bound* on performance within the same study area. Second, temporal generalization has not been validated: the models were tested on held-out grid cells from the same seasons as training, not on a withheld growing year. Third, the wheat dataset covers only two seasons, limiting inter-annual generalizability. The highest-priority extensions are: expanding the season count, adopting year-blocked validation, and replacing scalar feature summaries with sequence-aware architectures (LSTM, Temporal CNN).

Conclusion

Removing `harvest_day_of_year`—which previously caused massive data leakage—revealed the true predictive limits of satellite and climate data. For corn and soy, feature quality limits 3-class performance more than model choice. The binary wheat model (84.5% accuracy, three algorithms) is operationally promising for pre-harvest risk detection. Soybeans are practically usable as a triage tool. Corn binary classification remains insufficient for standalone deployment without additional seasons and improved feature engineering. The ablation methodology is transferable to other agricultural ML pipelines where temporal management-correlated features may dominate model importance.

Acknowledgements

The authors thank AgriLab at UTFPR (Medianeira) and LabSolos at UTFPR (Pato Branco) for their indispensable technical assistance. The specialized equipment and facilities provided by these institutions were fundamental to the successful acquisition of experimental data for this study.

References

Benos, L., Tagarakis, A.C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. Sensors, 21, 3758. https://doi.org/10.3390/s21113758.

Claverie, M., Ju, J., Masek, J.G., Dungan, J.L., Vermote, E.F., Roger, J.C., et al. (2018). The harmonized Landsat and Sentinel-2 surface reflectance data set. Remote Sensing of Environment, 219, 145–161.

Helfer, G.A., Barbosa, J.L.V., Santos, R.d., & da Costa, A.B. (2020). A computational model for soil fertility prediction in ubiquitous agriculture. Computers and Electronics in Agriculture, 175, 105602.

Maimaitijiang, M., Sagan, V., Sidike, P., Daloye, A.M., Erkbol, H., & Fritschi, F.B. (2020). Soybean yield prediction from UAV using multimodal data fusion and deep learning. Remote Sensing of Environment, 237, 111599.

NASA Langley Research Center. POWER Project. Available online: https://power.larc.nasa.gov/ (last accessed 1 October 2025).

Pelletier, C., Webb, G.I., & Petitjean, F. (2019). Temporal convolutional neural network for the classification of satellite image time series. Remote Sensing, 11, 523. https://doi.org/10.3390/rs11050523.

van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709.

Wulder, M.A., Roy, D.P., Radeloff, V.C., Loveland, T.R., Anderson, M.C., Dungan, J.L., et al. (2022). Fifty years of Landsat science and impacts. Remote Sensing of Environment, 280, 113195.

Xue, J., & Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. Journal of Sensors, 2017, 1353691.

Zhong, L., Hu, L., & Zhou, H. (2019). Deep learning based multi-temporal crop classification. Remote Sensing of Environment, 221, 430–443. https://doi.org/10.1016/j.rse.2018.11.032.