Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Small-Sample Tree-Ensemble Models for Climate-Smart Crop Selection in Low-Resource Ethiopian Agriculture

Anna Chen[1], Lumeng Han[2]
[1]*Industrial Engineering and Operations Research, UC Berkeley, Berkeley, CA, USA*
[2] *International Agricultural Development, University of California, Davis, CA, USA*
anna.chen0818@yahoo.com

DOI: 10.69987/JACS.2024.40608

Keywords

climate-smart agriculture; crop recommendation; small-sample learning; tree ensembles; XGBoost; CatBoost; RandomForest; Ethiopia

Abstract

Low-resource agricultural decision support requires models that can learn from small, heterogeneous, and locally collected tabular data. This study evaluates a crop-selection workflow for an Ethiopian crop recommendation table containing soil color, soil pH, macro- and micronutrients, seasonal humidity, maximum and minimum temperature, precipitation, wind, cloud amount, surface pressure, and crop labels. Because the main supervised table contains crop labels but no row-level yield target, the experiment is framed as multiclass crop-label recommendation rather than yield regression. To clarify the data scope, a regional cereal yield table and the EthCT2020 georeferenced crop-type reference are used as contextual datasets, not as merged row-level training targets. CatBoost, RandomForest, and XGBoost were fitted on a fixed stratified 75/25 split with low-resource training sizes of 120, 240, 480, 960, and 2,900 records. On the full training pool, RandomForest achieved the highest top-1 accuracy at 0.501, followed by XGBoost at 0.492 and CatBoost at 0.481. XGBoost produced the strongest top-3 shortlist accuracy at 0.838, while CatBoost reached 0.825 and RandomForest reached 0.810. Macro-F1 remained low for all models, showing that rare crops were not reliably recovered under the observed class imbalance. The feature ablation and RandomForest importance analysis show that soil color, pH, and nutrient fields carry most of the separative signal, while climate variables add context but are weaker when used alone. The results support fast tabular crop recommendation for data-scarce settings, while also showing that rare-crop coverage and row-level yield prediction require additional field-level data.

1. Introduction

Climate-smart agriculture combines productivity, resilience, and mitigation goals under changing weather, soil degradation, and food-security pressure [1]. In sub-Saharan Africa, crop-selection decisions are often made with sparse local data, seasonal rainfall uncertainty, and uneven access to agronomic advisory services [2], [3]. Ethiopia is a representative setting because smallholder production includes diverse cereals and legumes, while field-level soil measurements, climate summaries, and historical crop labels are not always available at large scale.

Crop recommendation from tabular soil and weather features is a practical decision-support task. Soil pH, nutrient availability, and soil color can indicate suitability constraints, while seasonal temperature, precipitation, humidity, wind, cloud cover, and surface pressure provide broad agro-climatic context. Machine-learning models can combine these variables nonlinearly, but small samples and class imbalance make evaluation sensitive to the validation design. A

model that performs well on common cereal crops may still under-recommend rare but locally important crops.

Tree ensembles are strong baselines for such tables. RandomForest reduces variance through bagging and random feature selection [4], [5]. Gradient boosting fits additive trees to residual structure [6], and XGBoost and CatBoost are widely used implementations for structured data [7], [8]. These models are practical for local analysts because they train quickly, work with mixed tabular variables, and do not require large deep-learning infrastructure.

This study evaluates a low-tuning tree-ensemble workflow for Ethiopian crop recommendation. The analysis is limited to crop-label classification because the row-level table contains crop labels but no yield target. Additional Ethiopia-specific yield and crop-type datasets are reviewed only to describe data coverage and contextual limitations. The paper reports model performance, class imbalance, confusion patterns, feature-group ablation, feature importance, and runtime in a single ordered set of tables and figures.

2. Data and Method

The overall evaluation workflow is shown in Figure 1. The workflow starts with a data audit, uses one fixed stratified test set, repeats low-resource training samples, and reports model performance and interpretation in matching tables and figures.

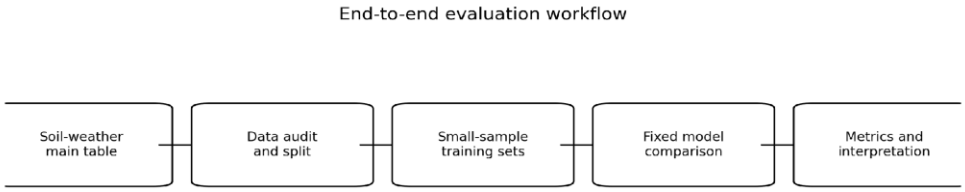

Figure 1. End-to-end experimental workflow.

2.1 Data sources and audit

The main supervised dataset was loaded as a crop recommendation table. The column Ph was renamed pH for readability. After this renaming, the table contained 3,867 rows, 28 predictors, and one crop-label outcome. Soilcolor was the only categorical predictor; the other predictors were numeric soil, nutrient, seasonal climate, and surface-weather fields. The feature groups used in the experiment are summarized in Table 1.

Table 1. Dataset feature groups and operational roles.

Group	Columns	Type	Use in experiment
Soil color	Soilcolor	Categorical field	One-hot encoded; represents visually described soil condition.
Acidity	pH	Numeric soil property	Used directly after renaming Ph to pH.
Macronutrients	K, P, N, S	Numeric nutrient fields	Represent potassium, phosphorus, nitrogen, and sulfur availability.
Micronutrient	Zn	Numeric nutrient field	Represents zinc availability.

Humidity	QV2M-W, QV2M-Sp, QV2M-Su, QV2M-Au	Seasonal climate fields	Specific humidity by season.
Temperature	T2M_MAX-*, T2M_MIN-*	Seasonal climate fields	Seasonal maximum and minimum 2 m air temperature.
Rainfall	PRECTOTCORR-*	Seasonal climate fields	Corrected seasonal precipitation.
Other weather	WD10M, GWETTOP, CLOUD_AMT, WS2M_RANGE, PS	Climate and surface fields	Wind direction, surface wetness, cloud amount, wind-speed range, and pressure.
Outcome	label	Multiclass target	Twelve crop classes for crop-selection classification.

The target label contains twelve crop classes. As shown in Table 2 and Figure 2, the table is strongly imbalanced: Teff, Maize, Wheat, and Barley dominate the observations, while Fallow, Red Pepper, Potato, Niger seed, Dagussa, Sorghum, and Pea have fewer than 100 records each. This imbalance is central to the interpretation of macro-F1 and per-class recall.

Table 2. Crop-label distribution in the main supervised table.

Crop	Records	Percent
Teff	1260	32.58
Maize	732	18.93
Wheat	715	18.49
Barley	503	13.01
Bean	253	6.54
Pea	94	2.43
Sorghum	72	1.86
Dagussa	71	1.84
Niger seed	64	1.66
Potato	48	1.24
Red Pepper	29	0.75

Fallow	26	0.67

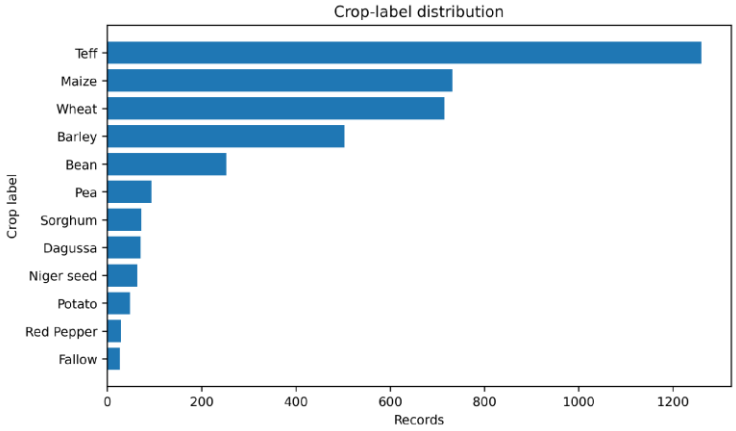

Figure 2. Crop-label imbalance in the evaluated crop recommendation table.

The numeric audit in Table 3 confirms that the main table has complete values for the evaluated predictors. Figure 3 provides the numeric correlation matrix used to inspect redundant or strongly associated features before model fitting. These summaries were used for interpretation, not for feature selection.

Table 3. Descriptive data audit by feature group.

Feature group	Raw features	Mean of means	Mean SD	Minimum	Maximum
soil_and_nutrients	6	59.1261	40.6897	0.0000	2119.00
seasonal_climate_weather	21	19.8558	6.5462	0.0000	354.88

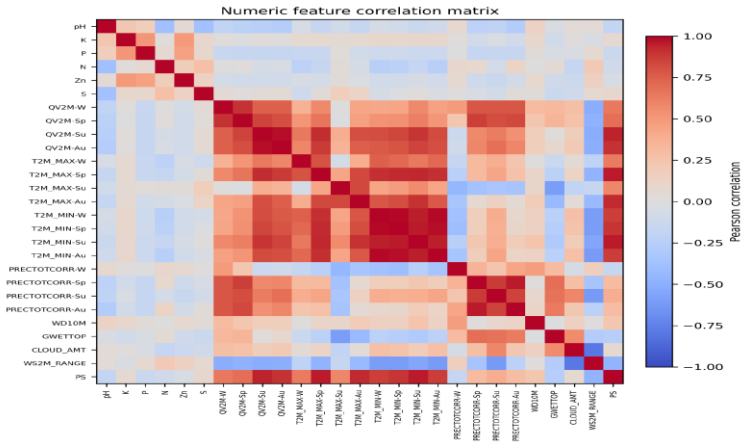

Figure 3. Numeric feature correlation matrix used for data inspection.

Two Ethiopia-specific contextual datasets were also reviewed. The regional cereal yield table covers region-year-crop records from 1996 to 2022. It is useful for describing historical yield variability, but it cannot be joined directly to the row-level soil-weather records without coordinates, field identifiers, and year-specific management information. EthCT2020 provides georeferenced crop-type samples and is useful for understanding external crop-type coverage, but its remote-sensing structure differs from the soil-weather recommendation table. Table 4 states how these contextual datasets are used in the study.

Table 4. Supplementary Ethiopia data context.

Dataset	Observed scope	Role in study
Regional cereal yield table	1,820 rows; 10 regions; 7 crop types; 1996-2022	Describes historical Ethiopia-wide yield context; not merged with the row-level classifier.
EthCT2020 crop-type reference	2,793 georeferenced samples; 7 crop groups; 22 crop classes	Provides external crop-type context; not merged because its remote-sensing structure differs from the soil-weather table.

Table 5 summarizes the overlapping cereal yield context. The regional cereal table shows median yields of 585.5 kg/ha for Teff, 639.0 kg/ha for Barley, 851.0 kg/ha for Wheat, 1,507.0 kg/ha for Sorghum, and 2,014.5 kg/ha for Maize. Because the table is aggregate and includes extreme entries, it is treated as descriptive background rather than a supervised yield target.

Table 5. Regional yield context for crop types overlapping the main label set.

Crop	Records	Mean yield	Median yield	Min yield	Max yield
Maize	260	6528.2	2014.5	0.0	1149000.0
Sorghum	260	5584.6	1507.0	0.0	1063000.0
Wheat	260	929.5	851.0	0.0	3364.0
Barley	260	726.2	639.0	0.0	2856.0
Teff	260	598.7	585.5	0.0	2769.0

2.2 Experimental protocol

The main table was split once into a 2,900-record training pool and a 967-record held-out test set using a stratified 75/25 split with random state 2026. Stratified low-resource training samples of 120, 240, 480, and 960 records were drawn from the training pool with seeds 101, 202, and 303. The full 2,900-record training pool was also evaluated under the same three seeds. Table 6 summarizes the protocol.

Table 6. Experimental protocol.

Component	Value	Protocol choice
Main supervised table	3,867 records; 28 predictors; 12 crop classes	Crop-label classification
Data quality	0 missing cells; no row-level yield column	Yield regression is not used
Split	2,900 train-pool records and 967 held-out test records	Stratified 75/25 split, random state 2026
Low-resource sizes	120, 240, 480, 960, and full 2,900 training records	Stratified subsampling from the train pool
Repeated seeds	101, 202, and 303	Metrics are reported as mean +/- sample SD
Metrics	Top-1 accuracy, macro-F1, weighted-F1, balanced accuracy, top-3 accuracy, fit time, prediction time	Top-3 represents a short advisory list

All models received the same processed representation. Soilcolor was one-hot encoded with unknown-category handling, and numeric features were passed through unchanged. No imputation was required because the evaluated table had no missing cells. The evaluation used top-1 accuracy, macro-F1, weighted-F1, balanced accuracy, top-3 accuracy, fit time, and prediction time. Top-3 accuracy is included because an extension officer may prefer a short crop shortlist instead of a single automatic recommendation.

The model configurations were fixed before reporting. Table 7 lists the three classifiers used in the empirical comparison. These settings deliberately avoid grid search so that the reported results reflect a low-tuning workflow suitable for small local datasets.

Table 7. Fixed model configurations used in the empirical evaluation.

Model label	Estimator used	Fixed configuration	Reason
CatBoost	CatBoostClassifier	50 iterations, depth 5, learning rate 0.08, multiclass loss	Ordered-boosting tree baseline.
RandomForest	RandomForestClassifier	100 trees, sqrt feature sampling, balanced subsample class weights	Bagging ensemble baseline.
XGBoost	XGBClassifier	20 trees, depth 3, learning rate 0.20, histogram tree method	Gradient-boosted tree baseline.

Two interpretive analyses were run after the main comparison. First, a RandomForest ablation compared soil color plus nutrient features, climate-weather features alone, and all features together. Second, RandomForest feature importances from the full training pool were aggregated back to original feature names. These analyses were used to explain model behavior and were not used to tune the classifiers.

3. Results and Discussion

3.1 Low-resource learning curves

The low-resource learning curves show that all three models improve as more training records are added, but the amount of improvement differs by metric. Table 8 and Figure 4 report top-1 and top-3 accuracy. CatBoost is strongest at 120 records, with 0.424 +/- 0.005 top-1 accuracy and 0.770 +/- 0.012 top-3 accuracy. With the full 2,900-record training pool, RandomForest has the highest top-1 accuracy, while XGBoost has the highest top-3 accuracy.

Table 8. Low-resource top-1 and top-3 accuracy learning curves.

Model	n	Top-1 accuracy	Top-3 accuracy
CatBoost	120	0.424 +/- 0.005	0.770 +/- 0.012
RandomForest	120	0.405 +/- 0.008	0.715 +/- 0.017
XGBoost	120	0.367 +/- 0.017	0.707 +/- 0.022
CatBoost	240	0.443 +/- 0.004	0.800 +/- 0.015
RandomForest	240	0.426 +/- 0.019	0.767 +/- 0.005
XGBoost	240	0.420 +/- 0.025	0.769 +/- 0.012
CatBoost	480	0.468 +/- 0.005	0.803 +/- 0.007
RandomForest	480	0.446 +/- 0.004	0.772 +/- 0.023
XGBoost	480	0.458 +/- 0.007	0.788 +/- 0.004
CatBoost	960	0.469 +/- 0.006	0.823 +/- 0.002
RandomForest	960	0.468 +/- 0.006	0.799 +/- 0.010
XGBoost	960	0.470 +/- 0.017	0.815 +/- 0.006
CatBoost	2900	0.481 +/- 0.004	0.825 +/- 0.002
RandomForest	2900	0.501 +/- 0.004	0.810 +/- 0.003
XGBoost	2900	0.492 +/- 0.000	0.838 +/- 0.000

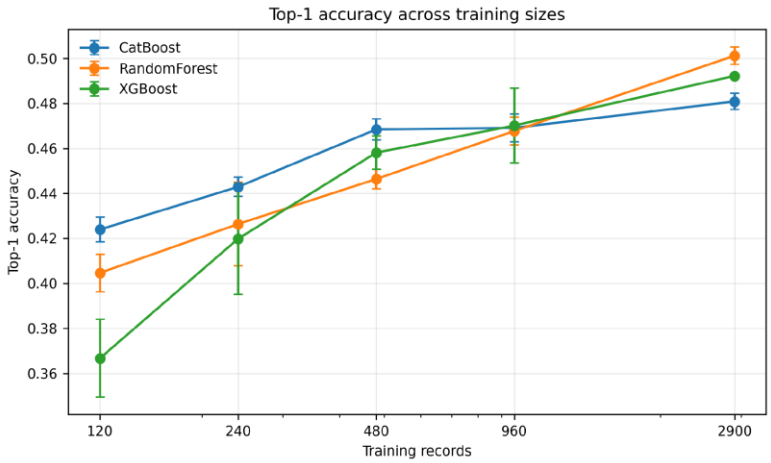

Figure 4. Top-1 accuracy learning curves across low-resource training sizes.

Macro-F1 and balanced accuracy tell a different story, as shown in Table 9 and Figure 5. The values remain much lower than top-1 and top-3 accuracy because rare classes have few examples. This is an important operational result: a model can be useful for common-crop shortlisting while still being weak for minority crops.

Table 9. Macro-F1 and balanced accuracy under low-resource sampling.

Model	n	Macro-F1	Balanced accuracy
CatBoost	120	0.138 +/- 0.006	0.148 +/- 0.004
RandomForest	120	0.188 +/- 0.014	0.194 +/- 0.012
XGBoost	120	0.170 +/- 0.007	0.169 +/- 0.010
CatBoost	240	0.149 +/- 0.009	0.159 +/- 0.005
RandomForest	240	0.200 +/- 0.017	0.199 +/- 0.023
XGBoost	240	0.211 +/- 0.018	0.205 +/- 0.017
CatBoost	480	0.151 +/- 0.007	0.163 +/- 0.006
RandomForest	480	0.179 +/- 0.008	0.179 +/- 0.006
XGBoost	480	0.186 +/- 0.005	0.184 +/- 0.004
CatBoost	960	0.151 +/- 0.002	0.164 +/- 0.002
RandomForest	960	0.192 +/- 0.017	0.191 +/- 0.011
XGBoost	960	0.198 +/- 0.020	0.197 +/- 0.019
CatBoost	2900	0.158 +/- 0.003	0.170 +/- 0.002
RandomForest	2900	0.210 +/- 0.014	0.206 +/- 0.010

XGBoost	2900	0.209 +/- 0.000	0.204 +/- 0.000

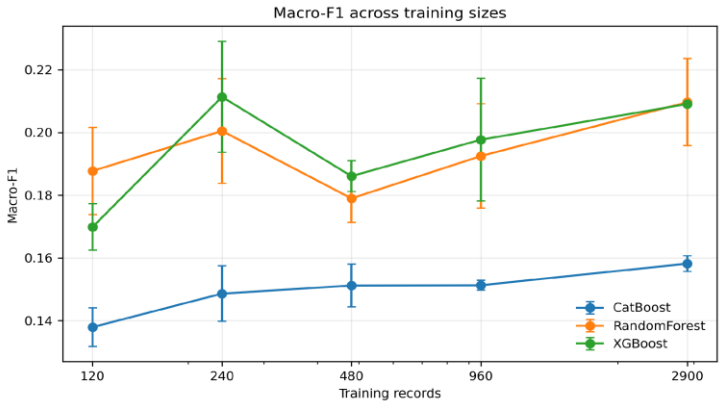

Figure 5. Macro-F1 learning curves across low-resource training sizes.

3.2 Full training-pool comparison

Table 10 compares the models on the full training pool. RandomForest achieved the highest top-1 accuracy at 0.5012 and the highest weighted-F1 at 0.4641. XGBoost achieved the highest top-3 accuracy at 0.8376 and a macro-F1 very close to RandomForest. CatBoost had competitive top-3 accuracy but lower macro-F1, indicating weaker minority-class recovery under the fixed configuration.

Table 10. Full training-pool comparison on the fixed test set.

Model	Accuracy	Macro-F1	Weighted-F1	Balanced acc.	Top-3 acc.	Fit time (s)	Predict time (s)
CatBoost	0.4809	0.1582	0.4201	0.1704	0.8249	1.048	0.150
RandomForest	0.5012	0.2096	0.4641	0.2059	0.8101	0.858	0.066
XGBoost	0.4922	0.2091	0.4421	0.2040	0.8376	1.406	0.010

The per-class RandomForest report in Table 11 and the confusion matrix in Figure 6 show why macro-F1 remains low. Teff, Maize, Wheat, and Barley receive usable recall, while the rarest crops are often absorbed into common labels. The result suggests that additional sampling for minority crops would be more valuable than small algorithmic changes alone.

Table 11. Per-class RandomForest report for the full train-pool run with seed 101.

Crop	Precision	Recall	F1-score	Support
Barley	0.5111	0.3651	0.4259	126
Bean	0.5385	0.1111	0.1842	63

Dagussa	0.0000	0.0000	0.0000	18
Fallow	0.0000	0.0000	0.0000	6
Maize	0.5161	0.6120	0.5600	183
Niger seed	0.0000	0.0000	0.0000	16
Pea	0.4000	0.0833	0.1379	24
Potato	0.0000	0.0000	0.0000	12
Red Pepper	0.0000	0.0000	0.0000	7
Sorghum	0.0000	0.0000	0.0000	18
Teff	0.5532	0.7587	0.6399	315
Wheat	0.3918	0.4246	0.4075	179

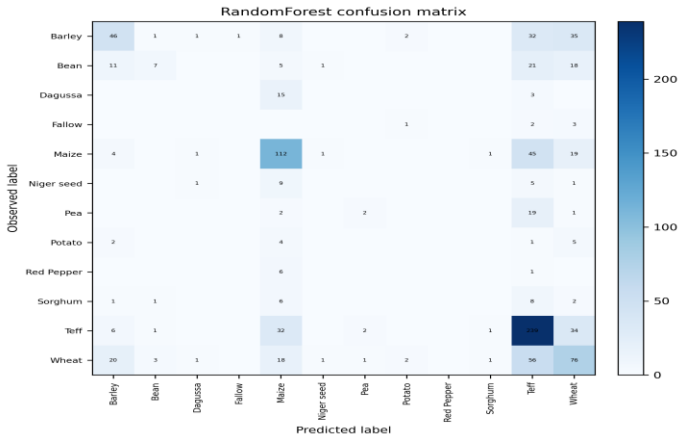

Figure 6. Confusion matrix for RandomForest with the full training pool and seed 101.

3.3 Feature groups, importance, and runtime

The RandomForest feature-group ablation in Table 12 shows that soil color plus nutrients carry most of the predictive signal. Soil color plus pH and nutrients achieved 0.4281 top-1 accuracy and 0.7766 top-3 accuracy, while climate-weather fields alone reached only 0.1624 top-1 accuracy and 0.4881 top-3 accuracy. Using all features increased top-1 accuracy to 0.4984 and top-3 accuracy to 0.8066 in the seed-101 RandomForest run.

Table 12. RandomForest feature-group ablation on the fixed test set.

Feature set	Raw features	Accuracy	Macro-F1	Balanced accuracy	Top-3 accuracy	Fit time (s)
soil_color_plus_nutrients	7	0.4281	0.1760	0.1731	0.7766	1.8754

climate_weather_only	21	0.1624	0.1278	0.2831	0.4881	0.7128
all_features	28	0.4984	0.1963	0.1962	0.8066	0.8747

The feature-importance results in Table 13 and Figure 7 are consistent with the ablation. Potassium, sulfur, nitrogen, phosphorus, pH, zinc, and soil color are the leading predictors. Seasonal humidity and weather variables appear after the soil and nutrient block, which indicates that climate context contributes but is less discriminative in the current table.

Table 13. Top RandomForest feature importances aggregated to original feature names.

Feature	**Importance**
K	0.1382
S	0.1295
N	0.1252
P	0.1217
pH	0.1199
Zn	0.1176
Soilcolor	0.0961
QV2M-Au	0.0142
QV2M-Su	0.0110
QV2M-Sp	0.0082
WD10M	0.0081
QV2M-W	0.0079

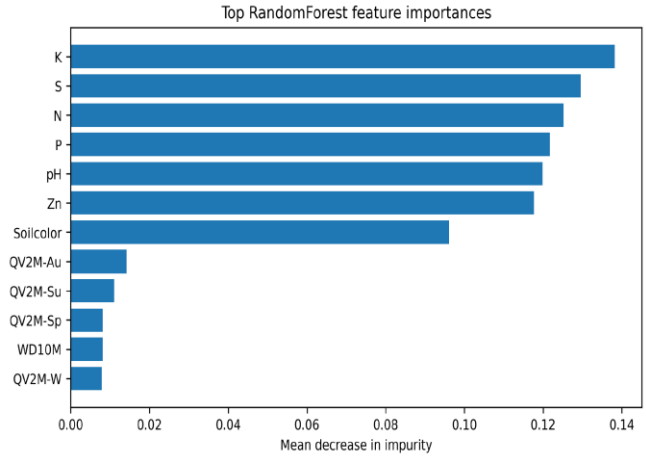

Figure 7. Top RandomForest feature importances.

Runtime results are reported in Table 14 and Figure 8. All fixed configurations are feasible for rapid local retraining. RandomForest fits in under one second on the full training pool, CatBoost takes about one second, and XGBoost takes about 1.4 seconds. Prediction time is also short for all models, so the practical bottleneck is data quality and label coverage rather than computation.

Table 14. Training and prediction time by model and training size.

Model	n	Fit time (s)	Predict time (s)
CatBoost	120	0.428 +/- 0.294	0.068 +/- 0.026
RandomForest	120	0.277 +/- 0.007	0.041 +/- 0.003
XGBoost	120	0.861 +/- 0.161	0.009 +/- 0.000
CatBoost	240	0.732 +/- 0.228	0.146 +/- 0.091
RandomForest	240	0.358 +/- 0.102	0.044 +/- 0.003
XGBoost	240	0.874 +/- 0.062	0.009 +/- 0.000
CatBoost	480	0.771 +/- 0.110	0.098 +/- 0.006
RandomForest	480	0.356 +/- 0.019	0.047 +/- 0.001
XGBoost	480	1.028 +/- 0.172	0.009 +/- 0.000
CatBoost	960	1.101 +/- 0.221	0.128 +/- 0.043
RandomForest	960	0.551 +/- 0.206	0.054 +/- 0.002
XGBoost	960	1.340 +/- 0.047	0.026 +/- 0.032
CatBoost	2900	1.048 +/- 0.225	0.150 +/- 0.123
RandomForest	2900	0.858 +/- 0.008	0.066 +/- 0.002

XGBoost	2900	1.406 +/- 0.122	0.010 +/- 0.002

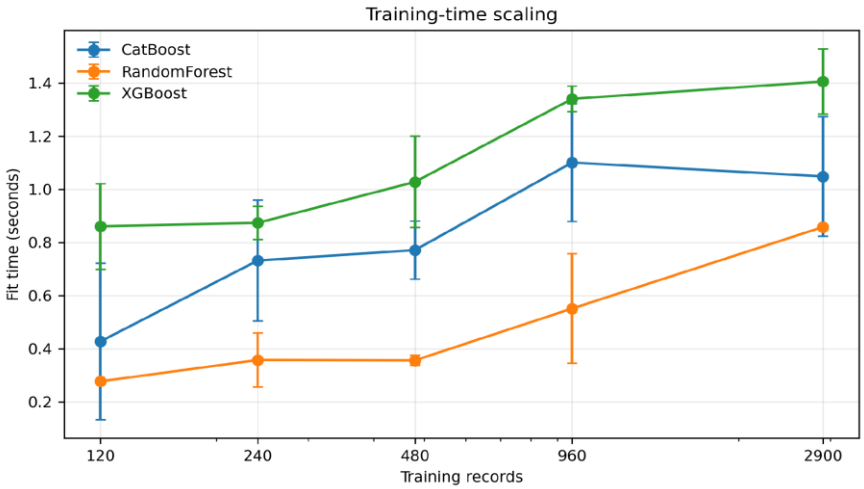

Figure 8. Training-time scaling across models.

4. Limitations

This study has several limitations. First, the main supervised table does not contain a row-level yield target; therefore, the experiment evaluates crop-label recommendation rather than yield prediction. The regional yield table is useful for context but is measured at a different observational unit. Second, the class distribution is highly imbalanced, and several rare crops have too few examples for reliable recall. Third, the climate fields are seasonal summaries rather than field-level time series, so the models cannot account for planting date, cultivar, irrigation, fertilizer history, pest pressure, or market constraints. Fourth, the evaluation uses a random stratified split rather than a spatial, temporal, or woreda-level holdout. A deployment study should test spatial generalization and decision impact before operational use. Finally, the present evaluation focuses on fixed, low-tuning classifiers; future work could add calibrated uncertainty, spatial validation, and richer field-management data.

5. Conclusion

This paper presented an empirical evaluation for low-resource climate-smart crop selection using an Ethiopian soil and weather crop recommendation table. The supervised task is crop-label classification because the main table contains crop labels but no row-level yield target. A regional cereal yield table and EthCT2020 are used to clarify external data context without forcing incompatible data units into the row-level classifier. Across fixed low-resource experiments, RandomForest produced the strongest full-pool top-1 accuracy, XGBoost produced the strongest full-pool top-3 shortlist accuracy, and all models remained weak for the rarest crops. The feature analysis shows that soil color, pH, and nutrient measurements dominate the current crop-label signal, while seasonal climate variables add context. The main practical conclusion is that fast tree-ensemble baselines are feasible for local crop recommendation, but stronger rare-crop coverage and yield prediction require additional field-level sampling and compatible row-level targets.

References

[1] Food and Agriculture Organization of the United Nations, Climate-Smart Agriculture Sourcebook. Rome, Italy: FAO, 2013.

[2] P. K. Thornton, P. J. Ericksen, M. Herrero, and A. J. Challinor, “Climate variability and vulnerability to climate change: A review,” Global Change Biology, vol. 20, no. 11, pp. 3313-3328, 2014.

[3] J. W. Jones and P. K. Thornton, “The potential impacts of climate change on maize production in Africa and Latin America in 2055,” Global

Environmental Change, vol. 13, no. 1, pp. 51-59, 2003.
[4] Sihan Zhou, Zeyi Li, and Eric Wang , "Long-Document RAG for Contractual and Insurance Clause Analysis in Receivables RWA Structures", JACS, vol. 4, no. 8, pp. 88–104, Aug. 2024, doi: 10.69987/JACS.2024.40810.
[5] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123-140, 1996.
[6] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.
[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.
[8] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Advances in Neural Information Processing Systems, 2018, pp. 6638-6648.
[9] Ge Liu, Chenyu Li, and Eric Zhang , "OpsLLM for Cloud Incident Triage: Bilingual RAG-Based Root Cause Analysis and Alert Summarization for AI Infrastructure Operations", JACS, vol. 4, no. 4, pp. 97–111, Apr. 2024, doi: 10.69987/JACS.2024.40408.
[10] S. Alemu, "Crop Recommendation using Soil Properties and Weather Prediction Dataset," Mendeley Data, Version 1, 2024, doi: 10.17632/8v757rr4st.1.
[11] Yunhe Li, "Risk-Sensitive Offline Reinforcement Learning for Stable ABR QoE Improvements on Real HSDPA and LTE Traces", JACS, vol. 3, no. 4, pp. 1–11, Apr. 2023, doi: 10.69987/JACS.2023.30401.
[12] A. H. Sparks, "nasapower: A NASA POWER global meteorology, surface solar energy and climatology data client for R," Journal of Open Source Software, vol. 3, no. 30, 2018, Art. no. 1035.
[13] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proceedings of the International Joint Conference on Artificial Intelligence, 1995, pp. 1137-1145.