FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 21, November, 2026, pp 364 – 372
DOI: https://doi.org/10.33003/fjs-2026-1021-6242

Comparative Evaluation of Machine Learning Models for Predicting Crop Yields Across Sixteen Crops in Nigeria

Chima Samson Oleforo and *Terna Godfrey Ieren

Department of Mathematics and Computer Science, Faculty of Science, Rev. Fr. Moses Orshio Adasu University Makurdi

*Corresponding authors' email: ternagodfrey@gmail.com Phone: +2348134508572
https://orcid.org/0000-0003-3878-1773

Received: 16 Aug. 2026
Accepted: 21 Oct. 2026
Published: 09 Nov. 2026

ABSTRACT
Accurate crop yield prediction supports agricultural planning, food security and the scheduling of harvest-related resources. This study compared machine learning models for predicting the yields of sixteen major crops in Nigeria, using secondary agricultural and climatic data for 2015–2019 from the Agricultural Performance Survey of Nigeria (National Agricultural Extension and Research Liaison Services and Federal Department of Agricultural Extension) and the Nigerian Meteorological Agency. After preprocessing and the removal of 25 extreme yield observations, 1,962 crop–state–year records were analysed. Five approaches were evaluated: Random Forest; XGBoost; CatBoost; Ridge Regression; and a stacked ensemble combining the three tree-based models with a Ridge Regression meta-learner. XGBoost and CatBoost were run with their original libraries. Performance was assessed with crop-specific five-fold cross-validation, repeated three times, using R^2, RMSE, MSE and MAE. Under the random, record-level With-State design, in which state was an input and almost all test states also appeared in the training data, CatBoost achieved the highest mean R^2 (0.357; RMSE = 1.031 t/ha; MAE = 0.754 t/ha), followed by XGBoost (0.331) and Random Forest (0.315); its advantage over Random Forest (0.042; 95% CI 0.018 to 0.067) and XGBoost (0.026; 95% CI 0.003 to 0.049) was small relative to differences among crops. Crop-level R^2 was highest for yam (0.720), soybean (0.549), onion (0.542) and tomato (0.524) and low for groundnut, ginger, benniseed and okra. The stacked ensemble (mean R^2 = 0.281) did not outperform the strongest individual models and was below the best individual model in 14 of 16 crops. When states were withheld from training, mean R^2 was negative for every model (−0.054 to −0.249), indicating no useful transfer to unseen states. The reported accuracy therefore applies only to settings with comparable state-level data and should not be extrapolated nationwide. The study predicts yield, not post-harvest loss. Validation across years and regions, and richer agronomic predictors, are needed before operational use.

Keywords: Crop yield prediction, Machine learning, Ensemble learning, Predictive performance, Nigeria

INTRODUCTION

Agriculture is essential to food security, livelihoods and economic development, particularly where households depend heavily on agricultural production. Losses can occur throughout the post-harvest chain, and inadequate handling, storage, packaging and transport can worsen food wastage (Nath et al., 2024; Ogundele, 2022). Climate variability and extreme weather also affect crop growth and expected production, increasing uncertainty for farmers and agricultural planners (Araya et al., 2022; Kuradusenge et al., 2023).

Accurate crop yield prediction is therefore an important component of agricultural planning. Reliable estimates of expected production can inform harvest scheduling, storage capacity, transportation, market supply and food-security interventions (Asamoah, 2022; Benos et al., 2021). Conventional statistical and process-based approaches may have difficulty capturing nonlinear crop–climate interactions and localised yield variability (Morales & Villalobos, 2023; Nandikolla et al., 2025). Machine learning offers a data-driven alternative that learns relationships from historical agricultural and environmental observations, and hybrid machine learning and deep learning approaches have been applied to yield prediction (Agarwal & Tarar, 2021; Benos et al., 2021). Tree-based and boosting methods, including Random Forest and boosting-based approaches, have produced promising results in several settings (Ahmed et al., 2023; Islam et al., 2024; Nagesh et al., 2024; Patil et al., 2023), and comparative studies have shown that the best-performing algorithm differs among datasets and crops (Cedric et al., 2022; Sharma et al., 2024).

Ensemble learning combines the predictions of several learners and can improve accuracy when the learners contribute complementary information (Wolpert, 1992; Zhou, 2012). Gains from ensembling have been reported in other prediction domains (Yagyanath & Sharma, 2023) and in stacked crop-yield models (Ramesh & Kumaresan, 2025). The benefit is not guaranteed, however: it depends on the diversity of the base learners and on the amount of data available to fit the combining model (Zhou, 2012). Soil and agronomic information may also improve yield prediction (Vhatkar et al., 2025), but such information is often absent from routinely collected survey data.

Three gaps motivate the present study. First, several Nigerian yield studies address a single crop (Yecho et al., 2025) or a limited region (Ahmed et al., 2023), so a common benchmark across many crops is lacking. Second, in the Nigerian yield literature reviewed here, stacked ensembles were not

compared with their own constituent learners under identical conditions, so it is unclear whether stacking is worthwhile for survey-based data of modest size. Third, results from random cross-validation, in which records from the same state appear in both training and testing data, are not always contrasted with results for states withheld from training, although the two answer different questions. The independent contribution of this paper is a common-dataset comparison of five approaches across sixteen crops that reports both validation designs and separates predicted yield from measured post-harvest loss. A yield model does not demonstrate that losses are reduced; at most, forecasts can inform planning decisions that may influence losses.

The algorithms were chosen to represent complementary modelling strategies. Random Forest is a bagging ensemble that is relatively robust to correlated predictors (Breiman, 2001). XGBoost builds trees sequentially to correct earlier errors, using a regularised objective and second-order gradient information (Friedman, 2001; Chen & Guestrin, 2016), and CatBoost is a gradient boosting library that uses ordered boosting and symmetric trees to reduce prediction shift (Prokhorenkova et al., 2018). Ridge Regression is a regularised linear benchmark suited to the severe multicollinearity found in the climatic predictors and serves as the meta-learner in stacking. Including a linear model indicates how much of the predictive signal requires nonlinear learners.

The aim of this study was to compare Random Forest, XGBoost, CatBoost, Ridge Regression and a Ridge-based stacked ensemble for predicting the yields of sixteen crops in Nigeria, and to assess how well their accuracy transfers to states not used in training.

MATERIALS AND METHODS

Analytical Framework

The analysis followed the workflow shown in Figure 1: agricultural and climatic data were collected, merged, cleaned and explored; predictors were prepared; and models were trained and validated separately for each crop using five-fold cross-validation. The out-of-fold predictions of the three tree-based learners were then combined by Ridge-based stacking, and all models were evaluated and interpreted as yield forecasting tools only.

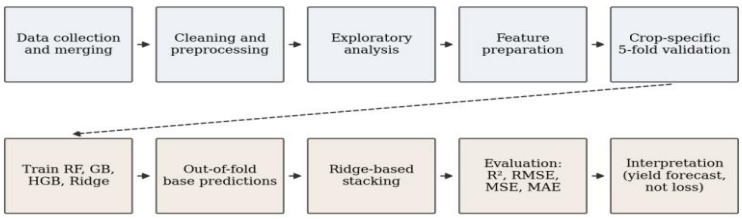


Figure 1: Workflow for Crop-yield Prediction and Model Comparison. RF = Random Forest; XGBoost = extreme Gradient Boosting; CatBoost = Categorical Boosting

Data, Design and Validation

A quantitative, experimental design based on supervised machine learning was adopted. The analysis was predictive and comparative: individual models were evaluated alongside a stacked ensemble. The principal validation design was the With-State design, in which state was retained as an explanatory variable.

Data Sources and Matching: Crop yield, land area, state and year were taken from the Agricultural Performance Survey of Nigeria associated with the National Agricultural Extension and Research Liaison Services (NAERLS) and the Federal Department of Agricultural Extension (FDAE) for 2015–2019 (2,450 entries). Climatic and soil-moisture variables were obtained from the Nigerian Meteorological Agency (NiMet) for the same period (228 entries). The two files were merged on state and year to give the analysis file used here, which contains 1,987 crop–state–year records in 14 columns (state, crop, land area, yield and ten climatic and soil-moisture variables) and carries no year label. Climate values are therefore specific to a state and year, not to a crop or a farm: the data contain 185 distinct state–year weather combinations (37 states × 5 years), and every crop recorded in the same state and year shares the same climate values. This design feature is important for interpreting the results.

Missing Values and Cleaning: The analysis file contained no missing values and no duplicate records. Extreme yields were removed for each crop when a record exceeded both the third quartile plus five interquartile ranges and eight times the crop median. This removed 23 soybean observations (yields of 20–202 t/ha against a median of 0.8 t/ha) and two cotton observations, 25 records in all, which look like unit or entry errors; the remaining 1,962 records were used for modelling. Any missing predictor values would be imputed with the training-fold median, as described below. The dataset covered 36 states and the Federal Capital Territory and 16 crops: maize, okra, rice, cassava, tomato, groundnut, cowpea, yam, sorghum, cocoyam, millet, benniseed, soybean, cotton, onion and ginger.

Variables: The dependent variable was crop yield in tonnes per hectare. Predictors represented spatial, agronomic, climatic and soil-moisture information (Table 1). All predictors were retained on agronomic grounds before any modelling, so no data-driven feature selection was performed and none could leak test information.

Unit of splitting and leakage. The unit of analysis was the crop–state–year record. For each crop, records were randomly assigned to five folds, with about 80% of the crop's records used for training and 20% for testing in each fold. The five-fold procedure was repeated three times with different random seeds (42–44); metrics were computed on the pooled out-of-fold predictions of each repeat and averaged over the three repeats. Because a state contributes about four records per crop on average (range one to five), records from the same state and from neighbouring years usually appeared in both training and testing folds. This random design therefore permits spatial and temporal leakage: a model can recognise a state it has already seen. About 99% of test records had their state in the training data. To measure how far accuracy depends on this, a complementary State-wise design withheld

whole states from training (grouped five-fold cross-validation, also repeated three times) and tested only on states not used for fitting. A forward-year validation, which trains on earlier years and tests on later years, was not performed.

Preprocessing and Tuning within Folds: Median imputation, one-hot encoding of state and standardisation (for Ridge Regression only) were fitted on the training portion of each fold and then applied to the test portion. Hyperparameters were fixed in advance, were identical for every crop (Table 2), and were not tuned on test data. Tree-based models were trained on unscaled inputs; state was one-hot encoded for every model, so CatBoost's native categorical encoding was not used. Models were implemented in Python with scikit-learn 1.9.1, xgboost 3.4.1 and catboost 1.2.10.

Stacking. The stacked ensemble used out-of-fold predictions from the three base learners as inputs to the Ridge meta-learner. Within each training portion, these inputs were produced by an internal three-fold procedure, so test records were never used to fit the meta-learner.

Exploratory analysis: Across the 1,962 modelling records, pooled yield showed only weak linear association with the predictors: land area (r = 0.15), humidity (r = 0.14), radiation (r = 0.12), soil moisture (r = 0.12), soil-surface moisture (r = 0.12), root-zone moisture (r = 0.10), rainfall (r = 0.09) and minimum temperature (r = 0.07); maximum temperature showed a weak negative association (r = −0.12). Pooled correlations mix crops with very different yield levels, so they understate relationships within a crop. Variance Inflation Factors indicated severe multicollinearity among soil-surface moisture (136.4), root-zone moisture (98.9), humidity (63.4) and soil moisture (50.7), with values above 10 for radiation (11.1), which supported the inclusion of tree-based ensembles and a regularised linear model.

Variables and Data Coverage

The final modelling dataset contained 1,962 observations across 16 crops. Each observation represented a crop–state–year combination, with yield in tonnes per hectare. Table 1 lists the variables and their observed ranges, which refer to the 1,987 records in the analysis file before extreme yields were removed.

Table 1: Variables and Coding Used in the Crop-Yield Dataset

Variable	Role	Scale	Description (unit)	Observed range
state	Spatial identifier	Nominal	Nigerian state or FCT of the observation	37 categories
crop	Crop identifier	Nominal	Crop cultivated (16 crops)	16 categories
land_area	Predictor	Ratio	Area cultivated for the crop (ha)	0.21–1,489.45
humidity	Climatic	Ratio	Average relative humidity (%)	36.44–89.12
max_temp	Climatic	Interval	Maximum temperature (°C)	31.13–44.19
min_temp	Climatic	Interval	Minimum temperature (°C)	5.73–21.40
rainfall	Climatic	Ratio	Total rainfall (mm)	316.4–3,243.2
radiation	Climatic	Ratio	Solar radiation intensity (MJ/m²/day)	34.99–35.82
max_speed	Climatic	Ratio	Maximum wind speed (m/s)	2.63–10.02
min_speed	Climatic	Ratio	Minimum wind speed (m/s)	0.00–0.09
soil_surface	Soil	Ratio	Soil-surface moisture (m³/m³)(Ratio)	0.22–0.89
soil_moisture	Soil	Ratio	Soil moisture content (m³/m³) (Ratio)	0.24–0.96
root_zone	Soil	Ratio	Root-zone moisture (m³/m³) (Ratio)	0.23–0.91
yield	Target	Ratio	Crop yield per unit land area (t/ha)	0.02–202.42

Machine Learning Models

Random Forest. Random Forest builds many decision trees on bootstrap samples and averages their predictions. For regression the forest prediction is

$$\hat{y}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \quad (1)$$

where B is the number of trees and $f_b(x)$ is the prediction of the b-th tree (Breiman, 2001). The model captures nonlinear relationships and interactions and is relatively robust to correlated predictors.

XGBoost. XGBoost is a gradient boosting implementation that builds a sequence of regularised regression trees, each trained to reduce the errors of the preceding ensemble (Friedman, 2001; Chen & Guestrin, 2016):

$$F_m(x) = F_{m-1}(x) + \nu h_m(x) \quad (2)$$

where $F_m(x)$ is the updated ensemble, $h_m(x)$ the regression tree fitted at iteration m, and ν the learning rate. At each iteration the tree is chosen to minimise a regularised objective that uses first- and second-order gradient information (Chen & Guestrin, 2016):

$$\text{Obj}^{(m)} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(m-1)} + h_m(x_i)\right) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} {w_j}^2 \quad (3)$$

where L is the squared-error loss, T the number of leaves of the tree, w_j the weight of leaf j, γ a penalty per leaf and λ the L2 penalty on the leaf weights. The original xgboost library (XGBRegressor, version 3.4.1) was used.

CatBoost. CatBoost follows the additive update in Equation (2) but builds symmetric (oblivious) trees and uses ordered boosting, in which the gradient for each training record is computed from a model fitted only to the records that precede it in a random permutation, to reduce prediction shift (Prokhorenkova et al., 2018). In principle, the pseudo-residual of record i at iteration m is:

$$r_i^{(m)} = -\left[\frac{\partial L(y_i, s)}{\partial s}\right]_{s=M_{i-1}^{(m-1)}(x_i)} \quad (4)$$

where s is the model prediction, L the loss and $M_{i-1}^{(m-1)}$ the ensemble after $m-1$ iterations fitted to the first $i-1$ records of the permutation. The original catboost library (CatBoostRegressor, version 1.2.10) was used.

Ridge Regression. Ridge Regression is a regularised linear model that penalises the squared size of the coefficients but not the intercept β_0:

$$\hat{\beta} = \text{argmin}_{\beta}\{\|y - \beta_0\mathbf{1} - X\beta\|^2 + \lambda\|\beta\|^2\} \quad (5)$$

where λ controls the strength of regularisation. Ridge Regression was included as a simple linear benchmark and as the meta-learner of the stacked ensemble.

Stacked ensemble. The stacked ensemble combined Random Forest, XGBoost and CatBoost as base learners. Their out-of-fold predictions were passed to Ridge Regression, which learned a linear combination:

$$\hat{y}_{\text{stack}} = \beta_0 + \beta_1 \hat{y}_{\text{RF}} + \beta_2 \hat{y}_{\text{XGB}} + \beta_3 \hat{y}_{\text{CB}} \quad (6)$$

The purpose of stacking was to test whether a learned combination of the three tree-based predictors could outperform the individual models.

Table 2: Models and Principal Implementation Settings

Model	Role	Implementation and settings
Random Forest	Individual model and base learner	scikit-learn RandomForestRegressor; 200 trees; minimum leaf size 2; 60% of inputs tried per split; random seeds 42–44
XGBoost	Individual model and base learner	xgboost.XGBRegressor (version 3.4.1); 200 trees; learning rate 0.03; maximum depth 4; row and column subsampling 0.8; L2 penalty 1; random seeds 42–44
CatBoost	Individual model and base learner	catboost.CatBoostRegressor (version 1.2.10); 200 iterations; learning rate 0.05; depth 5; L2 leaf regularisation 3; random seeds 42–44
Ridge Regression	Individual model and meta-learner	Regularisation strength 10 (standalone) and 1 (stacking meta-learner)
Stacked Ensemble	Combined model	RF + XGBoost + CatBoost → Ridge meta-learner; internal three-fold out-of-fold predictions

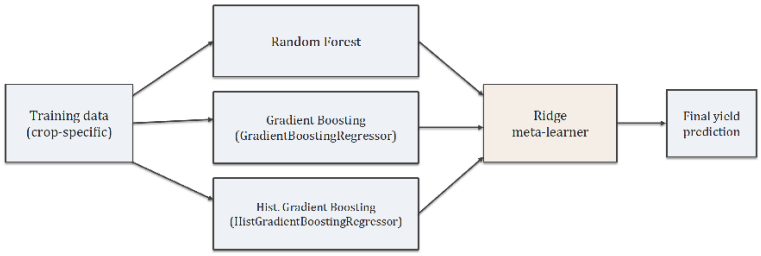


Figure 2: Stacked Ensemble Architecture with Random Forest, XGBoost and CatBoost base Learners and a Ridge Regression Meta-learner

Evaluation Metrics

Model performance was assessed with the coefficient of determination (R^2), root mean squared error (RMSE), mean squared error (MSE) and mean absolute error (MAE):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (8)$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (9)$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (10)$$

Higher R^2 indicates that a greater share of observed yield variation is explained, whereas lower RMSE, MSE and MAE indicate smaller errors. R^2 is negative when a model predicts worse than the training mean. Metrics were computed for each crop on its pooled out-of-fold predictions, separately for each of the three repeats, and averaged over the repeats. The overall figures in Table 4 are unweighted means of the 16 crop-specific values, so each crop counts equally regardless of sample size

RESULTS AND DISCUSSION

Dataset Characteristics and Description

The final modelling dataset contained 1,962 observations across 16 crops. Records per crop ranged from 25 for ginger to 190 each for okra and rice (Table 3). With five-fold validation, the test portion of each fold contained only about 5 ginger and 10 onion records. R^2 estimated from so few points is unstable, so results for ginger, onion and cotton (12 test records) should be treated as indicative only.

Table 3: Dataset Characteristics and Sample Sizes by Crop

Crop	Rows in file	Rows removed	Rows used	States	Train rows (~80%)	Test rows (~20%)	Mean yield (t/ha)
Maize	184	0	184	37	147	37	1.94
Okra	190	0	190	37	152	38	1.71
Rice	190	0	190	37	152	38	2.15
Cassava	175	0	175	37	140	35	6.66
Tomato	163	0	163	33	130	33	4.39
Groundnut	151	0	151	33	121	30	1.25
Cowpea	152	0	152	37	122	30	0.82

Crop	Rows in file	Rows removed	Rows used	States	Train rows (~80%)	Test rows (~20%)	Mean yield (t/ha)
Yam	135	0	135	28	108	27	8.37
Sorghum	110	0	110	24	88	22	1.20
Cocoyam	101	0	101	34	81	20	7.56
Millet	98	0	98	20	78	20	0.97
Benniseed	94	0	94	29	75	19	0.62
Soybean	111	23	88	25	70	18	0.88
Cotton	60	2	58	16	46	12	0.42
Onion	48	0	48	16	38	10	2.67
Ginger	25	0	25	6	20	5	6.51
Total	1,987	25	1,962				

Note. Train and test rows are average numbers per fold.

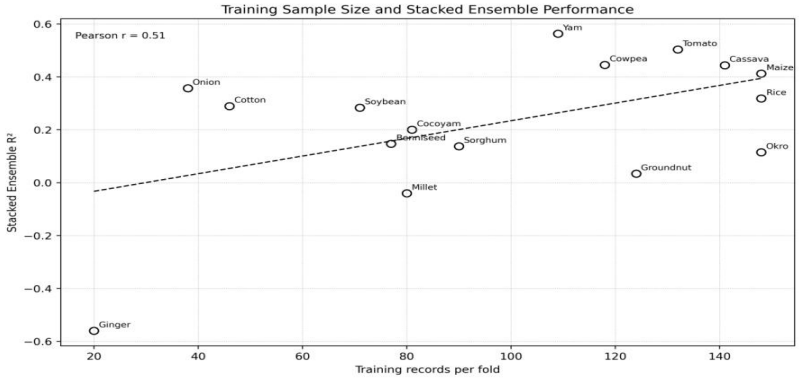


Figure 3: Training Sample Size and STACKED Ensemble R² under the With-State Design

Training sample size was only weakly to moderately related to the Stacked Ensemble's R² (Pearson r = 0.36; Spearman ρ = 0.12; Figure 3). Cotton and onion, with 46 and 38 training records per fold, reached R² of 0.44 and 0.54 under CatBoost, whereas okra, with 152 training records, reached 0.19 under its best model. Sample size therefore explains only part of the differences between crops.

Overall Predictive Performance

Under the With-State design, CatBoost achieved the highest mean R² (0.357), followed by XGBoost (0.331), Random Forest (0.315), the Stacked Ensemble (0.281) and Ridge Regression (0.223). These values describe random, record-level cross-validation in which almost all test states were also present in training. They are not estimates of accuracy for a new state or a new year.

Table 4: Overall Predictive Performance Under the With-State Design (Unweighted Means of 16 Crop-Specific Values)

Model	Mean R² (SD across crops)	RMSE (t/ha)	MSE (t/ha)²	MAE (t/ha)
Random Forest	0.315 (0.159)	1.082	2.246	0.806
XGBoost	0.331 (0.190)	1.049	2.093	0.753
CatBoost	0.357 (0.184)	1.031	2.031	0.754
Ridge Regression	0.223 (0.153)	1.159	2.581	0.891
Stacked Ensemble	0.281 (0.251)	1.086	2.292	0.807

Note. Each metric was computed for each crop and then averaged without weighting. The mean RMSE squared does not equal the mean MSE (for CatBoost, $1.031^2 = 1.063$ against 2.031) because averaging does not commute with squaring: crops with large yields (cassava, yam, cocoyam, ginger) have large squared errors that dominate the mean MSE. Error values in t/ha are driven by these high-yielding crops and are not comparable across crops of different scale. SD = standard deviation of crop-level R² across the 16 crops; it describes heterogeneity among crops, not resampling uncertainty.

The differences among the three tree-based models are small relative to the variation among crops. Across the 16 crops, the mean paired difference in R² between CatBoost and XGBoost was 0.026 (95% CI 0.003 to 0.049); CatBoost was higher in 13 of 16 crops (Wilcoxon signed-rank p = 0.021). CatBoost also exceeded Random Forest by 0.042 (95% CI 0.018 to 0.067) and was higher in 11 of 16 crops (p = 0.005), whereas XGBoost and Random Forest were not distinguishable (difference 0.016, 95% CI −0.017 to 0.050; p = 0.14). CatBoost is therefore treated as the best single model under this design, but its advantage is modest. These intervals treat crops as exchangeable units, do not capture fold-to-fold variability within crops, and the p-values are not adjusted for multiple comparisons. By contrast, all three tree-based models exceeded Ridge Regression (mean differences of 0.092 for Random Forest, 0.109 for XGBoost and 0.135 for CatBoost; higher in 15, 14 and 15 of 16 crops), and CatBoost exceeded the Stacked Ensemble in 14 of 16 crops (mean difference 0.076, 95% CI 0.019 to 0.134), whereas Random Forest (0.034, 95% CI −0.031 to 0.099) and XGBoost (0.050, 95% CI −0.020 to 0.121) did not differ detectably from it.

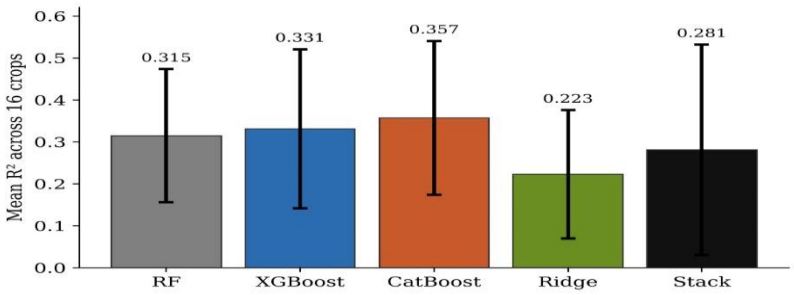


Figure 4: Mean R^2 of the Five Approaches under the With-State Design. Error Bars show the Standard Deviation Across the 16 Crops

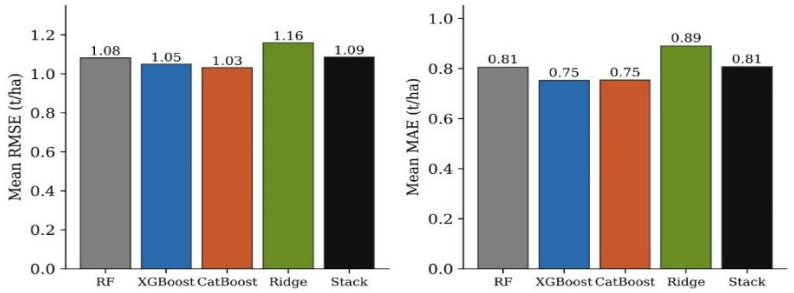


Figure 5: Mean RMSE and MAE of the Models Across all 16 Crops under the With-State Design

Crop-Specific Performance

Predictive performance varied substantially across crops (Table 5; Figure 6). The strongest fit was obtained for yam using the Stacked Ensemble ($R^2 = 0.7199$, a negligible 0.0001 above XGBoost), followed by soybean (0.5489), onion (0.5422), tomato (0.5238), cassava (0.4613), cowpea (0.4463) and cotton (0.4382), all using CatBoost, and maize using the Stacked Ensemble (0.4263). Groundnut, ginger, benniseed and okra had best-model R^2 of approximately 0.19 or below. CatBoost was the best model for 9 of 16 crops, Random Forest and XGBoost for two each, and Ridge Regression for groundnut. The Stacked Ensemble was the best model only for yam and maize, by margins of 0.0001 and 0.0069 over the best individual model; its R^2 was below the best individual model in the other 14 crops (mean shortfall across all 16 crops 0.09). Ginger shows the risk of a learned combiner on a very small sample: its stacked R^2 was −0.356 against 0.129 for XGBoost.

Table 5: Best-Performing Model for Each Crop Under the With-State Design

Crop	Best model	Best R^2	RMSE	MSE	MAE	Stacked R^2
Maize	Stacked Ensemble	0.4263	0.3028	0.0917	0.2172	0.4263
Okra	Random Forest	0.1853	1.3993	1.9614	0.8174	0.1196
Rice	Random Forest	0.2824	0.4392	0.1930	0.2838	0.2636
Cassava	CatBoost	0.4613	2.5648	6.5795	1.8681	0.4524
Tomato	CatBoost	0.5238	2.1451	4.6032	1.4812	0.5217
Groundnut	Ridge Regression	0.1023	0.3258	0.1062	0.2311	0.0539
Cowpea	CatBoost	0.4463	0.1822	0.0333	0.1136	0.4189
Yam	Stacked Ensemble	0.7199	1.9050	3.6393	1.4571	0.7199
Sorghum	CatBoost	0.2369	0.1578	0.0249	0.1162	0.1635
Cocoyam	CatBoost	0.3626	2.8291	8.0121	2.2197	0.2811
Millet	XGBoost	0.3026	0.2813	0.0795	0.1478	0.1407
Benniseed	CatBoost	0.1735	0.2431	0.0592	0.1828	0.0907
Soybean	CatBoost	0.5489	0.3017	0.0914	0.2062	0.4667
Cotton	CatBoost	0.4382	0.1261	0.0159	0.0929	0.2636
Onion	CatBoost	0.5422	0.7186	0.5170	0.5480	0.4716
Ginger	XGBoost	0.1287	2.4315	5.9211	1.9267	−0.3561

Note. RMSE, MSE and MAE refer to the best model for each crop. Crop-level estimates for ginger (25 records), onion (48) and cotton (58) rest on very few test records per fold and are unstable.

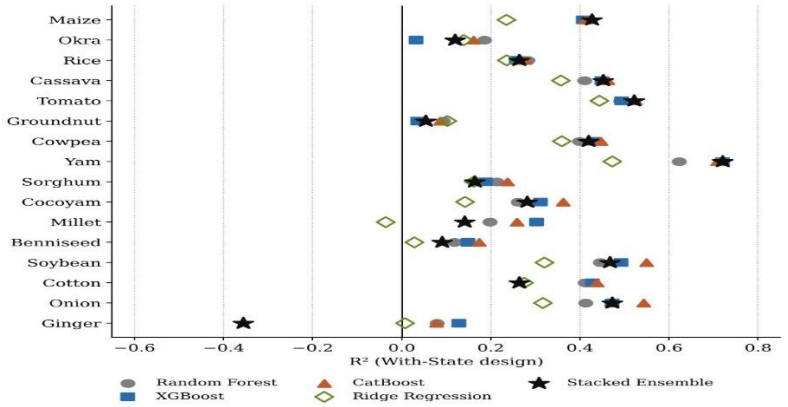


Figure 6: R^2 of Each Model for Each crop under the With-State Design

For several crops the best R^2 was only slightly above that of a state-mean lookup, and for yam, sorghum, benniseed and cotton the state-mean lookup was at least as good as every model. For groundnut, ginger, benniseed and okra every model explained little variance.

State Information and Generalisation

The most important limitation of this study is poor transfer to unseen states. When states were withheld from training, every model produced a negative mean R^2 (Table 6), meaning that all models predicted worse than the training mean.

Table 6: Mean R^2 When States Were Withheld from Training (State-Wise Design)

Model	Mean R^2
Random Forest	−0.068
XGBoost	−0.165
CatBoost	−0.071
Ridge Regression	−0.249
Stacked Ensemble	−0.054

The Stacked Ensemble had the least negative mean R^2 (−0.054) and the narrowest spread across crops (SD = 0.131) in this design, consistent with the variance-reducing effect of ensembling. It did not differ detectably from Random Forest or CatBoost (paired differences 0.014 and 0.018; p = 0.86 and p = 0.67) but exceeded XGBoost by 0.112 (95% CI 0.046 to 0.177; p = 0.004). Its R^2 nevertheless remained below zero, so it offers no useful prediction for new states.

The With-State scores must be read with this in mind. Including state as an input improved mean R^2 by only 0.04 to 0.12 relative to the design without it. This does not mean that state information was absent otherwise: because climate values are specific to each state and year, the 185 distinct weather combinations identify the state of almost every record. About 99% of test records had their state in the training data, and about 58% of yield variance lies between states. The state-mean baseline (R^2 = 0.229) shows that a large share of With-State accuracy reflects recognising states already seen. The With-State results therefore describe performance where comparable state-level data exist and states in testing are largely represented in training. They are not evidence of reliable prediction for new states, and they should not be used to claim nationwide forecasting ability. Further, the models describe predictive association between inputs and yield, not causation; for example, land area was the most influential input of CatBoost in 12 of the 16 crops (fits without state), which does not show that enlarging cultivated area raises yield per hectare.

Discussion

CatBoost achieved the highest mean R^2 (0.357), ahead of XGBoost (0.331), Random Forest (0.315) and Ridge Regression (0.223). The advantage of the tree-based models over Ridge Regression suggests that they captured useful nonlinear structure, but the differences among the three tree-based models are small relative to the variation among crops (Section 3.2). The result is broadly consistent with Ahmed et al. (2023) and Patil et al. (2023), who reported strong Random Forest performance, and with Islam et al. (2024), who found XGBoost best in a Saudi Arabian dataset, in that tree-based ensembles performed well. Taken together, these studies support tree-based ensembles but show that the best algorithm depends on the crop, predictors, geographic coverage and validation design. XGBoost and CatBoost were run with the original libraries and fixed, untuned settings (Table 2), and all three tree-based models fitted the training folds much more closely than the held-out folds (mean train R^2 of 0.82 for Random Forest, 0.97 for XGBoost and 0.92 for CatBoost, against 0.32, 0.33 and 0.36 on held-out records). The ranking therefore describes these settings, and stronger regularisation or tuning might change it.

The Stacked Ensemble did not outperform the individual models. It was below the best individual model in 14 of 16 crops and below CatBoost on average (0.281 against 0.357). This differs from Ramesh and Kumaresan (2025), who reported strong performance for a stacked model with a Decision Tree meta-learner, but the studies differ in datasets, base learners, meta-learners and validation. In the present study the three base learners were all tree-based and used the same predictors, so their errors were probably correlated; Zhou (2012) emphasises that ensemble gains depend on diversity among learners. In addition, crop-level samples were small, so the Ridge meta-learner had to estimate its weights from a few dozen to about a hundred and fifty out-of-fold predictions. The supplementary result that an unweighted average of the three tree-based predictions (R^2 = 0.355) outperformed the Ridge meta-learner (0.281) and matched CatBoost alone (0.357) points to unstable learned weights as a contributing cause. Averaged over crops, the meta-learner gave CatBoost the largest weight (0.44, against 0.23 for

XGBoost and 0.08 for Random Forest), and the Random Forest weight was negative in 6 of the 16 crops. Stacking therefore should not be assumed to help; its value needs to be established against each constituent learner (Wolpert, 1992).

State information materially affects interpretation. Because random folds place records from the same state in both training and testing data, the With-State scores describe prediction where state patterns are already represented. The negative R^2 values under the State-wise design show weak transfer to unseen states. This qualification matters when comparing the present results with studies reporting high accuracy, such as Shuaibu et al. (2024), because results from different split strategies are not directly comparable.

Crop-specific performance also varied. CatBoost was best for 9 crops, including soybean, onion, tomato, cassava, cowpea and cotton; the Stacked Ensemble was best for yam and maize, Random Forest for okra and rice, and XGBoost for millet and ginger. Yecho et al. (2025) reported strong Random Forest performance for cowpea, whereas the best cowpea result here came from CatBoost (R^2 = 0.446); this may reflect data composition, sample size, features and validation choices, not a universal advantage of one algorithm. The weaker results for ginger, groundnut, benniseed, okra and sorghum suggest that the available variables explain yield variation unequally across crops. Yield depends on soil fertility, variety, planting date, fertiliser use, pests and management practices that are absent from the data (Vhatkar et al., 2025).

The practical implication is that forecasts of this kind should be treated as decision-support information, not as measurements of post-harvest loss. Expected production estimates can be combined with storage capacity, transport availability, market demand and historical loss records to support harvest scheduling, aggregation and market planning, but this study predicts yield only and does not estimate or demonstrate reductions in post-harvest loss (Doki et al., 2019; Ogundele, 2022). Three limitations should be stated. Validation was random and record-level, with no forward-year test; climate variables vary only by state and year; and uncertainty is reported across crops, not across folds.

CONCLUSION

Machine learning models provided useful but highly variable information on crop yield when state-level information was available. Under random With-State validation, CatBoost performed best (mean R^2 = 0.357), ahead of XGBoost (0.331) and Random Forest (0.315), although the differences among the tree-based models were small, and the stacked ensemble (0.281) did not improve on the individual models. Accuracy depended strongly on crop and on recognition of states seen in training; mean R^2 was negative for all models when states were withheld. The findings therefore apply to settings with comparable state-level data and do not support nationwide forecasting. Before any operational use, the models should be tested independently across years and geographical areas and extended with agronomic predictors such as soil properties, variety, planting date and fertiliser use.

This study contributes a common-dataset benchmark of five approaches across 16 crops, run with the original XGBoost and CatBoost libraries, evidence that the tested stacking configuration did not outperform its strongest base learners, and a demonstration that validation design changes apparent accuracy.

Data Availability

The yield data come from the Agricultural Performance Survey of Nigeria (NAERLS/FDAE) and the climatic data from NiMet and was supplied from Shuaibu et al. (2024).

REFERENCES

Agarwal, S., & Tarar, S. (2021). A hybrid approach for crop yield prediction using machine learning and deep learning algorithms. *Journal of Physics: Conference Series, 1714*, Article 012012. https://doi.org/10.1088/1742-6596/1714/1/012012

Ahmed, A., Adewumi, S. E., & Yemi-Peters, V. (2023). Crop yield prediction in Nigeria using machine learning techniques: A case study of southern part of Nigeria. *UMYU Scientifica, 2*(4), 31–38. https://doi.org/10.56919/usci.2324.004

Araya, A., Jha, P. K., Zambreski, Z., Faye, A., Ciampitti, I. A., Min, D., Gowda, P. H., Singh, U., & Prasad, P. V. V. (2022). Evaluating crop management options for sorghum, pearl millet and peanut to minimize risk under the projected midcentury climate scenario for different locations in Senegal. *Climate Risk Management, 36*, Article 100436. https://doi.org/10.1016/j.crm.2022.100436

Asamoah, O. Z. (2022). AI-driven crop yield prediction models for smallholder farmers in Sub-Saharan Africa. *Iconic Research and Engineering Journals, 5*(9). https://doi.org/10.64388/IREV5I9-1713512

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors, 21*(11), Article 3758. https://doi.org/10.3390/s21113758

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cedric, L. S., Adoni, W. Y. H., Aworka, R., Zoueu, J. T., Mutombo, F. K., Krichen, M., & Kimpolo, C. L. B. (2022). Crops yield prediction based on machine learning models: Case of West African countries. *Smart Agricultural Technology, 2*, Article 100049. https://doi.org/10.1016/j.atech.2022.100049.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Doki, N. O., Eya, C. I., Tuughgba, M. F., Akahi, O. G., & Ameh, A. (2019). Determinants of post-harvest losses of orange in selected local government areas of Benue State. *International Journal of New Economics and Social Sciences, 2*(10), 295–308. https://doi.org/10.5604/01.3001.0013.8106

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Islam, M. M., Alharthi, M., Alkadi, R. S., Islam, R., & Masum, A. K. M. (2024). Crop yield prediction through machine learning: A path towards sustainable agriculture and climate resilience in Saudi Arabia. *AIMS Agriculture and*

Food, *9*(4), 980–1003. https://doi.org/10.3934/agrfood.2024053

Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A., Uwitonze, C., Ngabonziza, J., & Uwamahoro, A. (2023). Crop yield prediction using machine learning models: Case of Irish potato and maize. *Agriculture*, *13*(1), Article 225. https://doi.org/10.3390/agriculture13010225

Morales, A., & Villalobos, F. J. (2023). Using machine learning for crop yield prediction in the past or the future. *Frontiers in Plant Science*, *14*, Article 1128388. https://doi.org/10.3389/fpls.2023.1128388

Nagesh, O. S., Budaraju, R. R., Kulkarni, S. S., Vinay, M., Ajibade, S. S. M., Chopra, M., Jawarneh, M., & Kaliyaperumal, K. (2024). Boosting enabled efficient machine learning technique for accurate prediction of crop yield towards precision agriculture. *Discover Sustainability*, *5*, Article 78. https://doi.org/10.1007/s43621-024-00254-x

Nandikolla, S., Unhelkar, B., Kshirsagar, P. R., Chakrabarti, P., & Sonekar, S. V. (2025). Survey on crop yield prediction using machine learning algorithm. *Advances in Consumer Research*, *2*(2), 805–812. https://researchgate.net/publication/392705809_Survey_on_Crop_yield_prediction_using_machine_learning_algorithm

Nath, B., Chen, G., O'Sullivan, C. M., & Zare, D. (2024). Research and technologies to reduce grain postharvest losses: A review. *Foods*, *13*(12), Article 1875. https://doi.org/10.3390/foods13121875

Ogundele, F. (2022). Post harvest losses and food security in Nigeria: An empirical review. *African Journal of Agriculture and Food Science*, *5*(3), 77-89.https://doi.org/10.52589/AJAFSC0442Z7J.

Patil, P., Athavale, P., Bothara, M., Tambolkar, S., & More, A. (2023). Crop selection and yield prediction using machine learning approach. *Current Agriculture Research Journal*, *11*(3), 856-864. https://doi.org/10.12944/CARJ.11.3.16.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems* (Vol. 31). https://arxiv.org/abs/1706.09516

Ramesh, V., & Kumaresan, P. (2025). Stacked ensemble model for accurate crop yield prediction using machine learning techniques. *Environmental Research Communications*, *7*(3), Article 035006. https://doi.org/10.1088/2515-7620/adb9c0

Sharma, S., Walia, G. K., Singh, K., Batra, V., Sekhon, A. K., Kumar, A., Rawal, K., & Ghai, D. (2024). Comparative analysis on crop yield forecasting using machine learning techniques. *Rural Sustainability Research*, *52*(347), 63–77. https://doi.org/10.2478/plua-2024-0015

Shuaibu, N., Obunadike, G. N., & Jamilu, B. A. (2024). Crop yield prediction using selected machine learning algorithms. *FUDMA Journal of Sciences*, *8*(1), 61–68. https://doi.org/10.33003/fjs-2024-0801-2220

Vhatkar, K. N., Koparde, S. A., Kothari, S., Sarwade, J., & Sakur, K. (2025). Enhancing prediction of crop yield and soil health assessment for sustainable agriculture using machine learning approach. *MethodsX*, *14*, Article 103418. https://doi.org/10.1016/j.mex.2025.103226

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Yagyanath, R., & Sharma, N. (2023). Ensemble machine learning model improves prediction accuracy for academic performance: A comparative study of default ML vs boosting algorithm. In *Proceedings of the International Conference on Information Management & Machine Intelligence (ICIMMI 2023)*. https://doi.org/10.1145/3647444.3652473

Yecho, T. B., Olanrewaju, O. M., & Echobu, F. O. (2025). Machine learning algorithm for optimal yield prediction of cowpea: An IoT smart farming approach. *Journal of Science Research and Reviews*, *2*(2), 131–139. https://doi.org/10.70882/josrar.2025.v2i2.73

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. CRC Press.

©2026 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license viewed via https://creativecommons.org/licenses/by/4.0/ which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited appropriately.