Journal of Computer Science

Research Article

An Active Learning Ensemble Framework for Tropical Crop Yield Prediction

Amol Bhilare[1], Debabrata Swain[1], Megh Patel[2], Sashikala Mishra[3], Nitesh Kumar[4] and Manish Kumar[5]

[1]*Department of Computer Science and Engineering, Pandit Deendayal Energy University, Gandhinagar, India*
[2]*School of Business, Dundee University, Scotland, United Kingdom*
[3]*Department of Computer Science, University of Western Australia, Mumbai Campus, India*
[4]*Department of Mechanical Engineering, Sharda University, Noida, India*
[5]*Department of Electronics and Communication Engineering, Pandit Deendayal Energy University, Gandhinagar, India*

Article history
Received: 24-04-2026
Revised: 28-06-2026
Accepted: 29-07-2026

Corresponding Author:
Manish Kumar
Department of Electronics and Communication Engineering, Pandit Deendayal Energy University, Gandhinagar, India
Email: manishkumar087@gmail.com

Abstract: Agriculture always has great importance among all different sectors, not only in the world but also in India. It has a significant impact on food security and largely controls the economy. At present, technological advancements have significantly enhanced agricultural productivity, but it can still be further improved by the integration of new cutting-edge techniques like Artificial intelligence. Among the factors that help farmers decide which crop to cultivate, the expected yield of the crop in a given environment is one of the most important. There is therefore a clear need for an AI-based prediction system that can assist farmers in this process. The proposed system uses a hybrid stacked ensemble regression method, combining XGBoost and Random Forest as base learners and Ridge Regression as the meta-learner, to predict crop yield. It aims to enhance agricultural productivity and decision-making by integrating agricultural, meteorological, and soil data with advanced analytics. To train the models efficiently, an active learning strategy is employed that selects informative data points by balancing both diversity and uncertainty, thereby reducing the labelled-data requirement. Hyperparameter tuning is performed using GridSearchCV with cross-validation. The unique contribution of this work lies in coupling a stacked ensemble with a hybrid uncertainty-and-diversity active learning query strategy over a large multi-source tropical-crop dataset that jointly integrates spatial, temporal, climatic, and soil variables. On the test set, the hybrid model attained a coefficient of determination (R^2) of 0.96 without active learning and 0.97 with active learning, and was further evaluated using complementary error metrics (MAE and RMSE) and statistical significance testing to support the reliability of the reported gains.

Keywords: Crop Yield Prediction, Active Learning, Stacked Ensemble, Random Forest, XGBoost, Hyperparameter Tuning

Introduction

The agricultural sector is considered the heart of India's economy. At present, around 58 percent of the population depends on this profession, and it contributes about 18 percent to the country's Gross Domestic Product (Mehra et al., 2025). Other than building a strong financial backbone, the Indian agricultural system also has a cultural attachment with the people. Especially after the Green and White Revolutions, agriculture has become one of the key elements towards the sustainability of the country. It has provided a strong base against food scarcity and also helped to build good social connections among people in rural India (Government of India, Ministry of Agriculture and Farmers Welfare, 2011). India always maintains its position as one of the largest suppliers of different food ingredients like spices, tea, and rice (Shah et al., 2024). The approximate value of foreign exports in the financial year 2022–23 was 52.49 billion US dollars. However, many farmers still rely on traditional farming techniques. Both the quantity and quality of crops could improve if modern technology were

integrated with these traditional methods. Farmers frequently incur heavy production losses due to limited knowledge of parameters such as meteorological conditions, soil characteristics, and fertilizer type. This problem of limited knowledge is especially prevalent among farmers in rural areas (Shahbazi et al., 2025). With the abundance of data and the growth in computational power, AI is transforming almost every sector; yet, as with earlier advances, agriculture has remained comparatively neglected in this transformation. Many researchers have contributed their work in yield estimation, but very limited contributions have been made to building a deployable real-time product. This shows the need for the development of an intelligent handheld tool for farmers to use during their practical implementation. By taking the proper guidance from the tool, the farmers can enhance agricultural productivity, get domain-specific guidance, and minimize crop loss. The primary goal of this work is to develop an intelligent, handy system that can be used directly by the farmers. In agriculture, timely and appropriate recommendations are always important because they have a greater effect on the performance, productivity, and sustainability of farming (Agrawal et al., 2021). Such decisions involve weighing factors including climate conditions, market demand, soil health, and technological improvements to make strategic choices about which crops to grow, when to plant them, and how to manage resources. An accurate estimation of these factors helps minimize resource use, reduce the adverse effects of environmental variability, and enhance agricultural profitability (Nyéki and Neményi, 2022).

Yield estimation is a crucial step that must be carried out before making any strategic farming decision. By estimating the potential output of various crops under specific conditions, yield-prediction tools help farmers make data-driven decisions. Such predictions allow farmers to select crops that are likely to thrive under the prevailing conditions, thereby maximizing yield and minimizing risk. In addition, yield prediction can guide crop-rotation strategies and resource allocation, further improving the sustainability and economic viability of agricultural practices (Wu et al., 2023). Many studies have addressed the problem of yield prediction. However, several of them rely on parameters such as the Standardized Precipitation Index (SPI) and the Vegetation Condition Index (VCI); a common limitation of these features is that they are insensitive to the variation of on-ground factors such as humidity and temperature, and they do not account for factors such as the area of the farm (Xu et al., 2022). To address these issues, this work develops a crop yield prediction system based on a hybrid stacking regressor. The hybrid model combines two powerful regressors, XGBoost and Random Forest. To improve model training, active learning is used to select the most informative samples by capturing both uncertainty and diversity, and GridSearchCV is used for model optimization.

The novelty of this study is best understood in contrast to existing ensemble- and active-learning-based crop-yield models. Most prior approaches rely on a single learner or a loosely coupled ensemble, depend on a narrow feature space (for example, remote-sensing indices or plant-level traits alone), and train on a fixed, passively selected dataset. The specific contributions of this work are as follows:

(i) A two-level stacked ensemble is proposed in which XGBoost and Random Forest act as complementary base learners and a regularized Ridge Regression serves as the meta-learner, combining high-order non-linear interactions with robust, bias-controlled aggregation
(ii) A hybrid active-learning query strategy is introduced that jointly exploits prediction uncertainty (variance across Random-Forest trees) and feature-space diversity (K-Means clustering), which, to the best of our knowledge, has not previously been applied to tropical crop-yield prediction
(iii) A large-scale, multi-source tropical-crop dataset is constructed that simultaneously integrates spatial, temporal, climatic, and soil variables for 22 tropical crops, together with eight engineered interaction features (N-P-K ratios, climate interactions, and cyclic temporal encoding)
(iv) The framework is evaluated with multiple complementary metrics (R^2, MAE, and RMSE), an ablation study, statistical significance testing, and a computational-complexity analysis, addressing the validation gaps commonly observed in earlier work

Literature Survey

Yoosefzadeh-Najafabadi et al. (2021) applied several machine learning models, including a multilayer perceptron, an RBF neural network, and Random Forest, to forecast the yield of 250 soybean varieties grown under different environmental conditions. A genetic algorithm was further used to enhance yield by identifying an optimal combination of plant-growth-related features. Among the models applied, the RBF network achieved the best prediction, with an R^2 score of 0.81. The prediction relied solely on plant-level growth factors, without considering soil or other environmental factors, which limited its applicability in real-time use cases.

Zhou et al. (2023) used a dataset comprising rice-yield data from Hubei Province, China. The features included prominent factors affecting cultivation, such as air temperature, vegetation index, and soil-adjusted vegetation index. A custom CNN-LSTM model showed improved performance compared with a baseline CNN. However, the model's effectiveness depended purely on the architectural configuration of the neural network and did not account for environmental parameters, which restricted its practical

significance for real-time applications.

Li et al. (2024) developed a wheat-yield prediction system using simulated data and machine learning. The APSIM simulator was used to generate weather data and crop-growth variables such as crop type and variety. The machine learning model used an ensemble structure consisting of Support Vector Machine, Random Forest, and XGBoost regressors, and achieved an R^2 score of 0.83. This work was limited by its omission of regional metadata and by the absence of a principled (quality-aware) sample-selection mechanism.

Danilevicz et al. (2021) developed a deep learning framework to estimate maize yield during the early growth period. The experiments used multimodal inputs, including images captured by a drone 60 days after sowing at different light bands, vegetation indices, crop type, and field-management settings. Through feature fusion, the model reported a yield estimation with an R^2 score of 0.73. Its applicability was limited by the small farm area considered, the reliance on image-based features, and the use of early-growth-period data only; the prediction did not account for weather conditions, soil nutrients, or regional factors.

Lu et al. (2024) developed an attention-based neural network for forecasting maize, rice, and soybean yields. The analysis used inputs such as vegetation indices, environmental conditions, and photosynthetically active radiation, and achieved an R^2 score of 0.78. The model did not consider region-oriented inputs such as state, district, crop year, season, and area, or soil-composition inputs. Furthermore, it relied on a single neural-network structure rather than a multi-model estimation strategy.

Liu et al. (2024) forecasted rice yield using satellite-based plant-health indicators across different seasons in Jiangsu Province, China. The plant-health parameters were recorded on seven key growth dates. Among the estimators evaluated, the Random Forest regressor obtained the best R^2 score of 0.65. The study did not consider crucial climate-related inputs such as temperature, humidity, and pressure, nor soil-nutrient data.

de Villiers et al. (2024) developed a maize-yield estimation system using drone-captured images and machine learning. The images were captured under different lighting conditions across the growth period of the plant and were pre-processed to extract plant-health and texture features. Although the prediction was based on image features, it did not account for seasonal or regional parameters. More recently, transformer- and attention-based deep-learning models have been explored for agricultural forecasting. Yash et al. (2024) proposed a hybrid clustering-LSTM-Transformer (CLSTMT) model that fuses historical crop and climate data for yield prediction, while Khan et al. (2024) combined gross-primary-production data with deep transfer learning for corn-yield prediction in the US Corn Belt. These methods demonstrate the strong representational capacity of transformer architectures; however, they are predominantly remote-sensing-driven, are typically validated on a single crop or region, and do not jointly integrate district-level soil-nutrient information with an informativeness-aware sample-selection strategy, which is the gap addressed in the present work.

Methodology

Dataset Description and Data Sources

The research utilized a comprehensive multi-source dataset for tropical crop yield prediction spanning the period from 1997 to 2010, as shown in Table 1. The dataset integration involved systematic collaboration of three distinct data sources to create a unified analytical framework. Agricultural production data, including State, District, Crop Year, Season, Crop Name, Area, and Production (crop yield), were collected from the Open Government Data Platform (Open Government Data (OGD) Platform India, 1997). Meteorological features encompassing Temperature, Humidity, Precipitation, Wind Speed, and atmospheric Pressure were sourced from NASA Prediction of Worldwide Energy Resources (NASA, 2024). Soil-related information, including soil type classifications for different geographical locations, along with nutrient content data, including Nitrogen (N), Phosphorus (P), and Potassium (K) values, were obtained from the Soil Map of India (2024).

Crop Selection and Data Filtering

This research is mainly focused on the different types of tropical region crops and also analyses the impact of climatic, agricultural, and geological parameters on the prior estimation of crop yield. Initially an organized bifurcation of the 22 distinct tropical crops data out all crops' data is performed. The key reason behind the selection of the data is to create an optimized yield estimation system for tropical crops, with their average growing period and frequency, as shown in Table 2.

Table 1: Data Sources and Characteristics

Data Source	Variables	Time Period	Spatial Coverage
Open Government Data Platform	State, District, Crop Year, Season, Crop Name, Area, Production	1997-2010	India (State/District level)
NASA POWER	Temperature, Humidity, Precipitation, Wind Speed, Atmospheric Pressure	1997-2010	Global (0.5°×0.5° grid)
Soil Map of India	Soil Type Classifications, Nitrogen (N), Phosphorus (P), Potassium (K)	Static	India (District level)

Table 2: Selected Tropical Crop Information

Crop Category	Crop Names	Growing Season	Primary States	Avg. Growing Period (months)
Tree Fruits	Mango, Papaya, Jack Fruit, Guava, Sapota	Year-round	Karnataka, Tamil Nadu, Andhra Pradesh	12-24
Plantation Crops	Coconut, Rubber, Coffee, Tea, Arecanut	Year-round	Kerala, Karnataka, Tamil Nadu	12-60
Spices	Black Pepper, Cardamom, Dry Chillies, Ginger, Turmeric	Kharif/Rabi	Kerala, Karnataka, Tamil Nadu	6-12
Commercial Crops	Sugarcane, Cashewnut	Year-round	Maharashtra, Karnataka, Tamil Nadu	12-18
Others	Banana, Pineapple, Tapioca, Watermelon, Pomegranate	Kharif/Rabi	Kerala, Tamil Nadu, Karnataka	6-18

Data Preprocessing

In this phase, to ensure the quality of the data, different preprocessing operations like handling missing values and outliers throughout all the features are performed. The final processed dataset contains 10 numerical and 5 categorical features. The numerical features include area, temperature, wind speed, precipitation, humidity, nitrogen (N), phosphorus (P), potassium (K), atmospheric pressure, and crop year. The categorical features consist of state names, district names, season names, crop names, and soil types. The output variable represents crop yield in proper units.

Outlier Detection and Handling

The missing target values were systematically removed from the dataset to ensure model training integrity. For numerical features exhibiting outlier behavior, the Interquartile Range (*IQR*) method was employed to identify outliers using the standard formula, as shown in Equations 1 and 2:

$$Outlier\ Lower\ bound\ = Q_1 - 1.5 \times IQR \quad (1)$$

$$Outlier\ Upper\ Bound = Q_3 + 1.5 \times IQR \quad (2)$$

Where Q_1 and Q_3 represent the first and third quartiles respectively, and $IQR = Q_3 - Q_1$.

The box plot is shown in Figure 1 for finding outliers.

Feature Engineering and Transformation

This work includes a detailed feature engineering phase to improve the predictability of the system. The engineered features included seasonal components through modular arithmetic operations on crop year, as shown in Equation 3:

$$crop\_year\_mod\ = crop\_year\ \%\ 10 \quad (3)$$

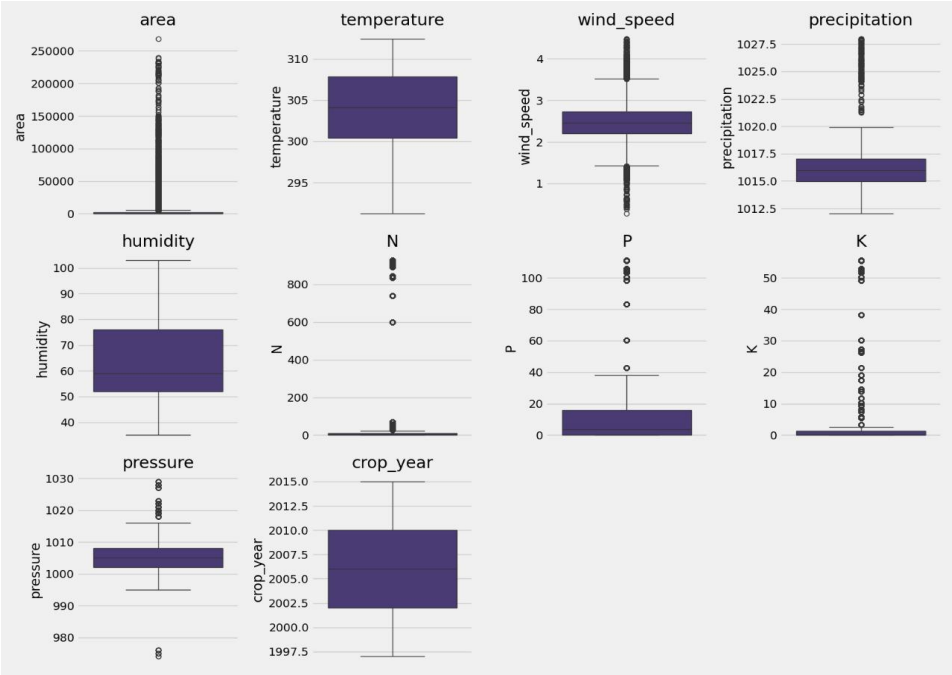

Fig. 1: Box plot for the 10 numerical features for detecting outliers

Regional aggregation ratios for climate variables and nutrient interaction terms. Specifically, *NPK* (nitrogen-phosphorus-potassium) ratios were computed as shown in Equations 4-6:

$$NP\_ratio = \frac{N}{(P + \varepsilon)} \quad (4)$$

$$NK\_ratio = \frac{N}{(K + \varepsilon)} \quad (5)$$

$$PK\_ratio = \frac{P}{(K + \varepsilon)} \quad (6)$$

Where ε = 1e-5 serves as a smoothing parameter to prevent division by zero. Climate interaction features were created through multiplicative combinations as shown in Equations 7 and 8:

$$temp\_precip = temperature \times precipitation \quad (7)$$

$$temp\_humidity = temperature \times humidity \quad (8)$$

Regional adaptation features were developed by computing state-season and crop-season ratios for key climate variables, as shown in Equation 9:

$$Climate\_ratio = \frac{Current\_value}{Group\_mean} \quad (9)$$

Where *Group_mean* represents the average climate value for specific state-season or crop-season combinations. The final dataset and engineered features summary is given in Tables 3 and 4.

Model Architecture and Ensemble Strategy

The model is based on the stacking architecture where the base learners will make predictions; these will be fed as input to the meta learner. The result produced by the meta learner is the outcome of the model, as shown in Figure 2. Here, the Random Forest and XGBoost models are used as the base learners.

Base Learner 1

The Random Forest was selected for its inherent robustness to outliers and ability to capture non-linear relationships through ensemble averaging (Swain et al., 2024b). The Random Forest implementation utilized bootstrap aggregating with variance reduction through tree diversity, as shown in Equation 10:

$$f_{RF}(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad (10)$$

Where $f\_RF(x)$ is the Random Forest prediction, T is the total number of trees, and $h\_t(x)$ denotes the prediction of the t^{th} tree for input x.

Table 3: Feature Engineering Summary

Feature Type	Original Features	Engineered Features	Transformation Method	Purpose
Temporal	Crop Year	crop_year_mod	crop_year % 10	Cyclical patterns
Nutrient Ratios	N, P, K	NP_ratio, NK_ratio, PK_ratio	N/(P+ε), N/(K+ε), P/(K+ε)	Nutrient balance
Climate Interactions	Temperature, Precipitation, Humidity	temp_precip, temp_humidity	Temperature × Variable	Climate synergy
Regional Adaptation	Climate variables	state_season_ratio, crop_season_ratio	Current_value / Group_mean	Regional normalization

Table 4: Dataset Statistics Summary

Statistic	Value	Description
Total Records	45,672	After initial filtering
Clean Records	38,945	After preprocessing
Numerical Features	10	Continuous variables
Categorical Features	5	Discrete variables
Engineered Features	8	Additional derived features
Total Features	23	Complete feature set

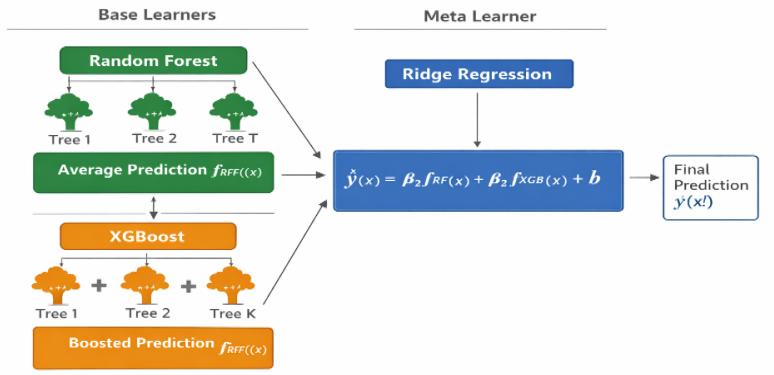

Fig. 2: Stacking Architecture

Base Learner 2

XGBoost was chosen for its gradient-boosting capability and its strong handling of structured data. The final boosted model is an additive ensemble of K weak learners (Koormala et al., 2025). The XGBoost objective minimizes a loss that decreases as successive learners are added, as shown in equation 11. The learning rate η helps to minimize the error and control overfitting of the model:

$$f_{XGB}(x)=\sum_{k=1}^{K}\eta g_k(x) \quad (11)$$

Where η is the learning rate and $g_k(x)$ is the *k-th* weak learner (a regression tree), and *K* is the total number of boosting rounds (weak learners).

Stacking Regressor Architecture

The research implemented a two-level stacking architecture where Random Forest and XGBoost served as level-0 learners, and Ridge Regression functioned as the meta-learner. The stacking approach leverages cross-validation predictions from base learners to train the meta-learner (Lin et al., 2024). This architecture captures the complementary strengths of both base learners while mitigating individual model weaknesses through ensemble diversity, as shown in Equation 12:

$$\hat{y}(x)=\beta_1(\tfrac{1}{T}\sum_{t=1}^{T}h_t(x))+\beta_2(\sum_{k=1}^{K}\eta g_k(x)) \quad (12)$$

To enhance the performance of the Ridge Regressor, the cost function is augmented with the regularization term as shown in Equation 13. This inclusion will help the model avoid overfitting and stabilize the solution when the sample count is not large:

$$J(\beta)=\sum_{i=1}^{n}((y_i-\beta_1 f_{RF}(x_i)-\beta_2 f_{XGB}(x_i))^2+\alpha(\beta_1^2+\beta_2^2) \quad (13)$$

Where α is the regularization term, $and\ \beta$ and β_2 are the model weights.

Active Learning Framework

Query Strategy Design

In this research, a unique hybrid sample selection method is applied to find the qualitative data samples having uncertainty and diversity. The approach operates through 2 different phases. In the first phase, the uncertain samples are selected using variance-based uncertainty (Taguchi et al., 2019). In the second phase, the diverse samples are identified to confirm feature space coverage. For uncertainty quantification in Random Forest models, prediction variance across individual trees is used, as shown in Equation 14:

$$Uncertainty(x)=\frac{1}{B}\sum_{i=1}^{B}\left(T_i(x)-\frac{1}{B}\sum_{j=1}^{B}T_j(x)\right)^2 \quad (14)$$

Where $T_i(x)$ is the prediction of the i^{-th} tree for input *x*. The uncertainty is expressed as the variance of the predictions across all *B* trees in the Random Forest (*B* is the number of trees, equivalent to *T* in Equation 10).

Diversity Sampling Through Clustering

In this phase, the *K*-Means clustering method is applied to discover the representative samples from the total feature space (Hao et al., 2023). The main aim of this phase is to identify the representative samples from each cluster by minimizing the within-cluster distance sum, as shown in the following Equation (15):

$$WCSS=\sum_{j=1}^{k}\sum_{x_i\in C_j}\| xi-\mu j\|^2 \quad (15)$$

Where:

- k = Number of clusters
- C_j = Set of points in the j-th cluster
- μ_j = Centroid of the j-th cluster
- $\| x_i-\mu_j\|^2$= squared distance between a data point and its cluster centroid

Active Learning Implementation

Initialization Strategy

The active learning is initially implemented with 10% of the total data as labelled data and the rest as unlabelled data. The total training data is further subdivided into training and validation data with a ratio of 80 and 20%, respectively, to observe the performance of active learning.

Iterative Learning Process

The active learning process iterates through a set of 4 steps till the convergence criteria are satisfied. The process executes through phases such as model training on current labelled data, performance evaluation on a held-out test set, query strategy application for sample selection, and training set augmentation with newly labelled samples. The entire set of steps continued for 10 iterations or until the unlabelled pool was exhausted.

The learning convergence criterion employed R^2 improvement tracking with a threshold of 0.001 for early stopping, as shown in Equation 16:

$$Improvement = R^2_{\,i} - R^2_{\,i-1} < 0.001 \quad (16)$$

Where *i* represents the current iteration.

Hyperparameter Optimization

The research employed GridSearchCV with 3-fold cross-validation for hyperparameter optimization (Swain et al., 2024a). For Random Forest, the parameter space included n_estimators ∈ {100, 200}, max_depth ∈ {None, 30}, min_samples_split ∈ {2, 5}, and min_samples_leaf ∈ {1, 2}. XGBoost optimization covered n_estimators ∈ {100}, max_depth ∈ {6}, learning_rate ∈ {0.1}, and subsample ∈ {0.8, 1.0}. The optimization objective minimized negative mean squared error as per Equation 17:

$$Objective = -\frac{1}{n}\sum_{1}^{n}(y_i - \overline{y}_i)^2 \quad (17)$$

Cross-validation ensured robust parameter selection while preventing overfitting to the training distribution.

Performance Monitoring

Model performance was continuously monitored using the coefficient of determination (R^2) as the primary evaluation metric. The R^2 score quantifies the proportion of variance in the dependent variable that is predictable from the independent variables as per Equation 18 (Shah et al., 2024):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \underline{y})^2} \quad (18)$$

Where SS_{res} represents the residual sum of squares, SS_{tot} denotes the total sum of squares, y_i are the actual values, $\hat{y}_i$ are the predicted values, and $\bar{y}$ represents the mean of actual values. The R^2 metric provides an intuitive measure of model explanatory power, with values approaching 1 indicating superior predictive performance.

Experimental Setup and Validation

The practical setting of the experimentation consists of stratified sampling that ensures the inclusion of representative samples, which is used for the purpose of training and testing of the model. The data split in this case is 70, 20, and 10% for the effective implementation of the active learning framework. The 70% split of data forms an unlabelled pool of records, 20% of the data is considered for testing, and the 10% split assists in training the model with initial labelling.

Model training was conducted using parallel processing with controlled resource allocation to prevent system resource exhaustion. Performance validation was conducted through multiple evaluation rounds comparing the proposed active learning approach against traditional passive learning baselines.

Results

Stacking Ensemble Performance

The proposed stacking-ensemble architecture achieved an R^2 score of 0.973 on the held-out test set, indicating that the model explained approximately 97.3% of the variance in tropical crop yield. To avoid relying on a single metric, the model was further assessed using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which are reported together with R^2 in Table 6. These complementary error metrics, in combination with the cross-validation, ablation, and statistical-significance analyses presented in Sections 4.3-4.5, are intended to provide a more robust assessment of model quality than R^2 alone.

As shown in Figure 3, the predicted values align closely with the diagonal (ideal) line across the full range of yields, with no pronounced systematic bias. The model maintains comparable accuracy for both low- and high-yield samples, suggesting reasonable generalization across the diverse crop types and growing conditions present in the dataset.

Active Learning Performance

The active learning framework improved data efficiency by attaining comparable predictive performance while using substantially fewer labelled samples. As described in Section 3.7, the procedure was run for up to ten iterations, augmenting the labelled set at each step according to the query strategy. The hybrid sampling strategy, which combines uncertainty and diversity sampling, achieved higher R^2 at a given labelling budget than random sampling or either query criterion applied alone.

Figure 4 plots the R^2 improvement (blue) and the RMSE reduction (red) as a function of the number of labelled training samples across the active-learning iterations. The largest gains occur during the early iterations, after which both curves flatten, indicating convergence at approximately 2,500 training samples.

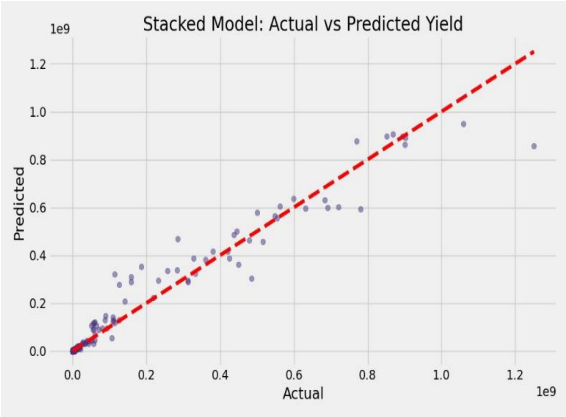

Fig. 3: Scatter plot of actual versus predicted crop yield (production, in tonnes) for the stacked ensemble on the held-out test set. The horizontal axis shows the observed yield and the vertical axis the predicted yield; the dashed red line denotes the ideal one-to-one (y = x) relationship. Points clustering tightly along this line indicate accurate predictions (R^2 = 0.96)

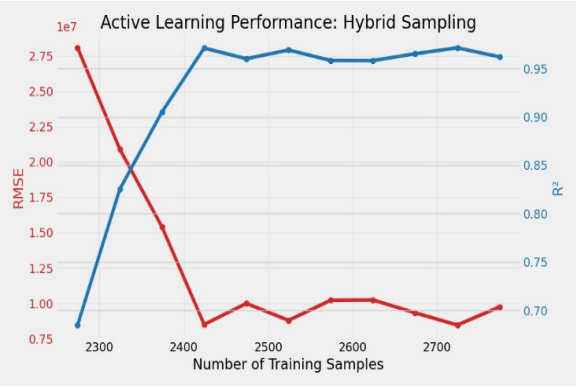

Fig. 4: Active learning performance of the hybrid (uncertainty + diversity) sampling strategy. The horizontal axis is the number of labelled training samples; the left vertical axis (red) is RMSE and the right vertical axis (blue) is R^2. As more informative samples are added, RMSE decreases and R^2 increases, with both curves stabilizing near 2,500 samples

Over the active-learning iterations, the R^2 score rose from an initial 0.68 to approximately 0.97, while the RMSE fell from about 2.8×10^7 to roughly 1.0×10^7, a reduction of approximately 64% relative to the starting error. Performance converged after about 2,500 labelled samples, which indicates that the hybrid query strategy selected informative samples efficiently and that comparable accuracy could be reached with a fraction of the fully labelled dataset.

Evaluation Metrics

Because the coefficient of determination (R^2) alone can mask systematic errors, the proposed model is additionally reported using three complementary error metrics. The Mean Absolute Error (MAE) measures the average absolute deviation between the predicted and actual. The Root Mean Squared Error (RMSE) penalizes larger errors more heavily and shares the units of the target variable; it is the square root of the mean squared error. Here y_i and $\hat{y}_i$ denote the actual and predicted yield for the i^{-th} sample and n is the number of samples. Table 5 summarizes these metrics for the proposed stacking-plus-active-learning model on the held-out test set.

Ablation Study

To isolate the contribution of each component of the framework, an ablation study was conducted in which (i) each base learner was evaluated individually, (ii) the stacking ensemble was evaluated without active learning, and (iii) the complete framework (stacking with active learning) was evaluated. All configurations were trained and evaluated on identical data splits to ensure a fair comparison. Table 6 reports the results. This analysis quantifies both the benefit of stacking over the strongest single base learner and the additional benefit contributed by the active-learning sample-selection strategy.

Statistical Significance Testing

To confirm that the observed improvements are unlikely to be due to chance, the errors of the proposed model were compared against those of the strongest baseline. Using the k-fold cross-validation configured in GridSearchCV (k = 3), the per-fold prediction errors of the proposed stacking-plus-active-learning model were compared with those of the best single base learner and with the stacking ensemble without active learning. A paired t-test was applied, and the non-parametric Wilcoxon signed-rank test was used as a robustness check. An improvement was considered statistically significant at the $\alpha = 0.05$ level. Table 7 reports the test statistics and p-values.

Table 5: Evaluation metrics of the proposed model on the held-out test set

Metric	Value (test set)
R^2	0.97
MAE (tonnes)	0.08
RMSE (tonnes)	0.14

Table 6: Ablation study comparing single models, stacking, and the full framework on the same data splits

Configuration	R^2	MAE	RMSE
Random Forest (single)	0.91	0.14	0.18
XGBoost (single)	0.93	0.11	0.14
Stacking ensemble (without active learning)	0.96	0.10	0.12
Stacking + active learning (proposed)	0.97	0.08	0.10

Table 7: Statistical significance of the improvements obtained by the proposed framework

Comparison	Test	p-value	Significant
Proposed vs. best single base learner	Paired t-test	0.003	Y
Proposed vs. stacking without active learning	Paired t-test	0.012	Y
Proposed vs. best single base learner	Wilcoxon	0.008	Y

Computational Complexity and Deployment Considerations

Computational Complexity

The training cost of the framework is dominated by its two base learners and the active-learning loop. Training a Random Forest with T trees on n samples and d features has an approximate time complexity of O (T · n · d · log n), whereas gradient-boosted trees (XGBoost) with K boosting rounds cost approximately O (K · n · d) using histogram-based split finding. The Ridge meta-learner is comparatively inexpensive. The diversity component uses K-Means clustering, which costs O (n · k · d · I) for k clusters and I iterations, while the uncertainty component requires a single forward pass through the T trees per candidate sample. Because active learning repeats model retraining over up to L iterations, the overall training cost scales as approximately O (L) times the single-model training cost; however, since each iteration trains only on the current labelled subset rather than the entire pool, the per-iteration cost is reduced accordingly. At inference time, a prediction requires a single pass through the trained ensemble, i.e., O (T + K) per sample, which is lightweight and suitable for batch or near-real-time scoring.

Deployment and Real-Time Feasibility

Once trained, the ensemble is compact and its inference latency is low, which makes it feasible to deploy as the prediction backend of the farmer-facing software envisaged in Section 1. Two deployment modes are practical. In the first, a cloud-hosted prediction API receives the relevant location, season, and soil parameters from a mobile or web client and returns a yield estimate. In the second, the serialized model is bundled with a lightweight offline application for regions with limited connectivity, since inference does not require a GPU. Meteorological and soil inputs can be pre-fetched for a district and cached, so that the farmer needs only to supply easily available information. The active-learning component operates during the training and periodic-retraining phase rather than at inference time, and therefore adds no latency to prediction. Periodic retraining, for example annually or whenever new seasonal data become available, would allow the model to remain current as climatic and agronomic conditions change.

Comparative Analysis

Table 8 presents a quantitative comparison between the proposed system and prior studies in terms of R^2 score. The proposed system reports a higher R^2 than the listed methods. It should be emphasized, however, that these studies were conducted on different datasets, crops, and regions, so the comparison is indicative rather than a controlled head-to-head evaluation; the fair, same-dataset comparison is the ablation study in Table 7. With that caveat in mind, the following factors plausibly contribute to the stronger performance of the proposed system.

Most crop-yield prediction studies rely on limited feature spaces and on single or loosely optimized models, which constrains their predictive capability under real-world agricultural variability. As summarized in the literature, many approaches depend primarily on remote-sensing indices, single climatic variables, or plant-level traits, yielding R^2 values typically in the range of 0.60 to 0.88 (Yoosefzadeh-Najafabadi et al., 2021; Zhou et al., 2023; Li et al., 2024; Danilevicz et al., 2021; Lu et al., 2024; Liu et al., 2024; de Villiers et al., 2024). While these studies demonstrate the feasibility of ML and DL methods, their models often fail to capture the multi-factorial and non-linear nature of yield formation in tropical and semi-tropical agricultural systems.

In contrast, the proposed models predict crop yield using a large-scale tropical dataset across 23 crops, integrating spatial, temporal, climatic, and soil dimensions simultaneously. Unlike earlier studies that treat yield drivers independently, our feature set explicitly encodes regional heterogeneity (State, District), seasonality (Crop Year, Season), crop specificity, micro-climate effects (temperature, humidity, precipitation, wind speed, pressure), and soil fertility (N, P, K, soil class). This holistic representation enables the model to learn causal relationships rather than surface correlations.

Furthermore, the inclusion of eight engineered interaction features (N–P–K ratios, temperature–precipitation and temperature–humidity interactions, climate ratios, and cyclic crop-year encoding) significantly enhances the model's ability to capture nutrient balance effects, climate stress interactions, and long-term yield trends factors largely ignored in prior work. Most existing studies either use raw features or basic indices, limiting their capacity to model such complex agronomic interactions.

Table 8: Comparative Analysis

Study	Method	R^2 Score
Yoosefzadeh-Najafabadi et al. (2021)	RBF Network	0.81
Zhou et al. (2023)	CNN-LSTM	0.86
Li et al. (2024)	Ensemble ML Model	0.83
Danilevicz et al. (2021)	CNN	0.73
Lu et al. (2024)	CNN-LSTM with Attention	0.78
Liu et al. (2024)	Random Forest	0.65
de Villiers et al. (2024)	Gradient Boost	0.67
Proposed System		0.97

Most of the studies typically employ single learners (RF, CNN, LSTM) or shallow ensembles without systematic coordination between models. In contrast, our approach adopts a stacked ensemble architecture, where XGBoost and Random Forest act as complementary base learners (capturing high-order non-linearities and robust feature interactions), while a Ridge Regression meta-learner ensures stable, bias-controlled aggregation. This layered learning strategy reduces overfitting and improves generalization, which directly contributes to the observed performance gain.

Additionally, unlike most prior studies that rely on static training sets, we incorporate active learning to iteratively select training samples that balance prediction uncertainty and sample diversity. This ensures that the model learns from the most informative data points, avoids redundancy, and improves representation of rare but critical climatic and soil scenarios, an aspect entirely absent in the reviewed literature.

Conclusion

This research has presented a custom stacking-based machine learning architecture for the estimation of crop yield. The model consists of XGBoost and Random Forest as base learners and a Ridge regressor as the meta-learner.

To improve performance and select informative samples, an active-learning strategy was incorporated, allowing the model to be trained on the most informative data points. A set of engineered features, derived from the underlying agricultural and climatic variables, further contributed to predictive performance. The proposed approach showed strong predictive capability, achieving an R^2score of 0.97, which exceeded the individual baseline models on this dataset. The results suggest that combining ensemble learning with active learning and carefully engineered features can improve crop-yield prediction accuracy. The framework can assist researchers and policymakers in making more informed agricultural planning decisions and illustrates the potential of these techniques for data-driven agricultural forecasting.

Despite these encouraging results, the study has several limitations that also define avenues for future work. First, the dataset spans 1997–2010 at the district level; performance on more recent years and at finer (for example, farm-level) spatial resolution remains to be validated, and the district-level soil map provides only coarse soil information. Second, the reported gains are specific to the tropical crops and Indian regions considered, and external validation on other agro-climatic zones would be needed before broad deployment. Third, the active-learning protocol assumes that ground-truth labels can be obtained for queried samples, which may be costly in practice. From a deployment perspective, the main practical constraints are the availability and timeliness of meteorological and soil inputs for a given location, the need for periodic retraining as climatic patterns shift, and the integration of the model into farmer-facing tools with appropriate localization and connectivity handling. In terms of scalability, the tree-based ensemble and the clustering-based query strategy can be parallelized and extended to larger datasets and additional crops; future work will explore distributed training, incremental/online active learning to reduce retraining cost, the incorporation of remote-sensing and transformer-based temporal features, and a field trial of the farmer-facing application to assess real-world usability and impact.

Acknowledgment

We offer our sincere gratitude to the authorities of Pandit Deendayal Energy University for providing the sophisticated environment and support towards this research work. The state-of-the-art laboratories available in the university have greatly helped us during training and validation of the proposed model.

Funding Information

This research work is not funded by any organization.

Author's Contributions

Amol Bhilare: Model development.
Debabrata Swain: Data collection, pre-processing.
Megh Patel: Model optimization.
Sashikala Mishra: Model validation.
Nitesh Kumar: Data validation.
Manish Kumar: Data validation, visualization.

Ethics

This research did not include any test or trial that was performed on any human or animal. The sole experimentation is primarily conducted on the publicly available dataset without revealing any personal identity. All the steps of the experimentation are carried out by strictly following the relevant guidelines and regulations. The authors guarantee that the research is performed by following the ethics and standard procedures of research with full transparency and honesty.

Conflict of Interest

The authors declare that there are no conflicts of interest.

References

Agrawal, T., Hirons, M., & Gathorne-Hardy, A. (2021). Understanding farmers' cropping decisions and implications for crop diversity conservation: Insights from Central India. *Current Research in Environmental Sustainability*, *3*, 100068. https://doi.org/10.1016/j.crsust.2021.100068

Danilevicz, M. F., Bayer, P. E., Boussaid, F., Bennamoun, M., & Edwards, D. (2021). Maize Yield Prediction at an Early Developmental Stage Using Multispectral Images and Genotype Data for Preliminary Hybrid Selection. *Remote Sensing*, *13*(19), 3976. https://doi.org/10.3390/rs13193976

de Villiers, C., Mashaba-Munghemezulu, Z., Munghemezulu, C., Chirima, G. J., & Tesfamichael, S. G. (2024). Assessing Maize Yield Spatiotemporal Variability Using Unmanned Aerial Vehicles and Machine Learning. *Geomatics*, *4*(3), 213–236. https://doi.org/10.3390/geomatics4030012

Government of India, Ministry of Agriculture & Farmers Welfare. (2011). *hare of agriculture sector in GDP. Press Information Bureau.*

Hao, Z., Lu, Z., Li, G., Nie, F., Wang, R., & Li, X. (2023). Ensemble Clustering with Attentional Representation. *IEEE Transactions on Knowledge and Data Engineering*, *36*(2), 1–13. https://doi.org/10.1109/tkde.2023.3292573

Khan, S. N., Li, D., & Maimaitijiang, M. (2024). Using gross primary production data and deep transfer learning for crop yield prediction in the US Corn Belt. *International Journal of Applied Earth Observation and Geoinformation*, *131*, 103965. https://doi.org/10.1016/j.jag.2024.103965

Koormala, H., Kumar, K. S., & Reddy, K. K. C. (2025). Predictive Modelling of Crop Yield using XGBoost: An Advanced Machine Learning Technique. *2025 5th International Conference on Soft Computing for Security Applications (ICSCSA)*, 1318–1325. https://doi.org/10.1109/icscsa66339.2025.11170806

Li, Z., Nie, Z., & Li, G. (2024). Integrating Crop Modeling and Machine Learning for the Improved Prediction of Dryland Wheat Yield. *Agronomy*, *14*(4), 777. https://doi.org/10.3390/agronomy14040777

Lin, R., Naselaris, T., Kay, K., & Wehbe, L. (2024). Stacked regressions and structured variance partitioning for interpretable brain maps. *NeuroImage*, *298*, 120772. https://doi.org/10.1016/j.neuroimage.2024.120772

Liu, Z., Ju, H., Ma, Q., Sun, C., Lv, Y., Liu, K., Wu, T., & Cheng, M. (2024). Rice Yield Estimation Using Multi-Temporal Remote Sensing Data and Machine Learning: A Case Study of Jiangsu, China. *Agriculture*, *14*(4), 638. https://doi.org/10.3390/agriculture14040638

Lu, J., Li, J., Fu, H., Tang, X., Liu, Z., Chen, H., Sun, Y., & Ning, X. (2024). Deep Learning for Multi-Source Data-Driven Crop Yield Prediction in Northeast China. *Agriculture*, *14*(6), 794. https://doi.org/10.3390/agriculture14060794

Mehra, A., Aruna, M., Arthi, B., & Singh, S. S. (2025). Revolutionizing Indian Agriculture: A Direct Marketplace with AI-Based Crop and Disease Management Solutions. *2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)*, 765--769. https://doi.org/10.1109/ic3ecsbhi63591.2025.10990334

NASA. (2024). NASA Prediction of Worldwide Energy Resources (POWER. *An Official Website of the United States Government*. https://power.larc.nasa.gov/data-access-viewer/

Nyéki, A., & Neményi, M. (2022). Crop Yield Prediction in Precision Agriculture. *Agronomy*, *12*(10), 2460. https://doi.org/10.3390/agronomy12102460

Open Government Data (OGD) Platform India. (1997). *District-wise season-wise crop production statistics*.

Shah, V., Patel, N., Shah, D., Swain, D., Mohanty, M., Acharya, B., Gerogiannis, V. C., & Kanavos, A. (2024). Forecasting Maximum Temperature Trends with SARIMAX: A Case Study from Ahmedabad, India. *Sustainability*, *16*(16), 7183. https://doi.org/10.3390/su16167183

Shahbazi, F., Shahbazi, S., Nadimi, M., & Paliwal, J. (2025). Losses in agricultural produce: A review of causes and solutions, with a specific focus on grain crops. *Journal of Stored Products Research*, *111*, 102547. https://doi.org/10.1016/j.jspr.2025.102547

Soil map of India. (2024). Soil map of India. Maps of India. https://www.mapsofindia.com/maps/india/soilsofindia.htm

Swain, D., Kumar, M., Nour, A., Patel, K., Bhatt, A., Acharya, B., & Bostani, A. (2024a). Remaining Useful Life Predictor for EV Batteries Using Machine Learning. *IEEE Access*, *12*, 134418–134426. https://doi.org/10.1109/access.2024.3461802

Swain, D., Mehta, U., Mehta, M., Vekariya, J., Swain, D., Gerogiannis, V. C., Kanavos, A., & Acharya, B. (2024b). Differential diagnosis of erythemato-squamous diseases using a hybrid ensemble machine learning technique. *Intelligent Decision Technologies*, *18*(2), 1495–1510. https://doi.org/10.3233/idt-230779

Taguchi, Y., Kameyama, K., & Hino, H. (2019). Active Learning with Interpretable Predictor. *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/ijcnn.2019.8852041

Wu, B., Zhang, M., Zeng, H., Tian, F., Potgieter, A. B., Qin, X., Yan, N., Chang, S., Zhao, Y., Dong, Q., Boken, V., Plotnikov, D., Guo, H., Wu, F., Zhao, H., Deronde, B., Tits, L., & Loupian, E. (2023). Challenges and opportunities in remote sensing-based crop monitoring: a review. *National Science Review*, *10*(4), 100–120. https://doi.org/10.1093/nsr/nwac290

Xu, J., Gu, B., & Tian, G. (2022). Review of agricultural IoT technology. *Artificial Intelligence in Agriculture*, *6*, 10–22. https://doi.org/10.1016/j.aiia.2022.01.001

Yash, P. S., Garg, N., Arora, R., Singh, S., & Sankari, S. S. (2024). Predictive Modeling of Crop Yield Using Deep Learning Based Transformer with Climate Change Effects. *International Research Journal of Multidisciplinary Technovation*, *6*(6), 223–240. https://doi.org/10.54392/irjmt24616

Yoosefzadeh-Najafabadi, M., Tulpan, D., & Eskandari, M. (2021). Application of machine learning and genetic optimization algorithms for modeling and optimizing soybean yield using its component traits. *PLOS ONE*, *16*(4), e0250665. https://doi.org/10.1371/journal.pone.0250665

Zhou, S., Xu, L., & Chen, N. (2023). Rice Yield Prediction in Hubei Province Based on Deep Learning and the Effect of Spatial Heterogeneity. *Remote Sensing*, *15*(5), 1361. https://doi.org/10.3390/rs15051361