RESEARCH ARTICLE

A Big Data-Driven Hybrid Ensemble Learning Framework for Crop Yield Prediction in Low-Productivity Districts

Somya Singh[1,*] **and Yogesh Kumar Gupta**[1]

[1] Department of Computer Science, Banasthali Vidyapith, Rajasthan 304022, India

Abstract

Low-productivity agricultural belts in India face the compounding realities of land fragmentation, erratic weather, and heterogeneous soils, which together limit harvest outcomes. Precision agriculture and big data analytics together offer an innovative solution that turns large volumes of heterogeneous data into actionable agronomic intelligence. This study proposes a big-data-driven crop yield prediction (CYP) framework tailored to low-productivity districts in India. The architecture combines soil and climatic analytics with data warehousing, machine learning (four base regressors and voting-based hybrid ensembles), and a preprocessing pipeline designed for compatibility with distributed computing frameworks such as Apache Spark, together with SHAP-based explainability for interpreting the model's predictions. The framework provides accurate, district-level yield forecasting to support context-sensitive agronomic decision-making. The paper reports the predictive performance achieved, discusses implementation challenges, and outlines a future research agenda for data-driven agricultural reform in India.

Keywords: crop yield prediction (CYP), machine learning (ML), ensemble learning, big data, precision agriculture, explainable AI (SHAP).

1 Introduction

The Indian economy is heavily agricultural, and a significant share of the population depends on agriculture for its livelihood. Because technological adoption remains incomplete, several Indian districts suffer from low agricultural productivity due to soil erosion, unpredictable rainfall, nutrient deficiency, and limited use of modern agricultural practices. These issues directly affect farmers' income and national food security.

Crop yield prediction is a critical component of precision agriculture because it enables better planning, risk management, and policy-making. The complex, non-linear relationships present in agricultural data are difficult to capture using traditional prediction approaches that rely on statistical models and historical experience alone.

Submitted: 30 March 2026
Revised: 04 April 2026
Accepted: 15 September 2026
Published: 22 September 2026

Vol. 2, **No.** 3, 2026.
10.62762/DIA.2026.295683

***Corresponding author:**
Somya Singh
singhsomya160@gmail.com

Citation
Singh, S., & Gupta, Y. K. (2026). A Big Data-Driven Hybrid Ensemble Learning Framework for Crop Yield Prediction in Low-Productivity Districts. *Digital Intelligence in Agriculture*, 2(3), 158–167.

However, existing machine learning models often lack scalability and do not generalize well to field- and farm-level scenarios. Moreover, little effort has been devoted to hybrid learning models that combine multiple base learners to achieve improved predictive performance.

In this paper, a hybrid machine learning architecture founded on big data is introduced for predicting crop yield in low-productivity districts. The main contributions of this work are:

1. Develop a preprocessing pipeline designed for compatibility with distributed computing frameworks such as Apache Spark.
2. Compare four machine learning base models on district-level agricultural data.
3. Construct hybrid ensemble models via voting regression.
4. Compare model performance across districts and analyze model explainability using SHAP.

2 Related Work

Predicting crop yield has developed beyond conventional statistical methods toward more sophisticated machine learning (ML) and deep learning (DL) models that overcome the shortcomings of early regression and empirical approaches [1, 2]. Combining machine learning models with soil and weather parameters has been shown to achieve better prediction accuracy, as reported in [3, 4].

Random Forest and XGBoost are ensemble learning techniques that have become popular because they perform well on agricultural prediction tasks [5, 6]. Nevertheless, single models can be biased and exhibit high variance, and combined machine learning methods have been shown to enhance stability and generalization [7].

Recent developments in deep learning emphasize the capacity of neural networks to learn complex non-linear relationships. Deep neural networks have been effectively used to model genotype-environment interactions [8], and deep learning combined with dimensionality reduction has been used to improve yield prediction for Indian regional crops [9]. Other architectures, such as convolutional and recurrent neural networks, are also being actively applied in agricultural analytics [10].

Geographically distributed agricultural data depends heavily on spatio-temporal modeling. Fan et al. [11] introduced a GNN-RNN model that captures spatial and temporal relationships, improving generalization in prediction. Likewise, physics-informed neural networks (PINNs) have been proposed to incorporate domain knowledge into learning models, increasing reliability and interpretability [12].

Big data frameworks are needed to support large-scale agricultural data operations. Ngo et al. [13] proposed a scalable data warehouse system for crop yield prediction. Distributed computing technologies and big data frameworks have been shown to enable efficient and scalable crop growth simulations [14], and big data frameworks in smart farming can integrate heterogeneous agricultural data sources and turn large volumes of complex data into actionable agronomic insight [15].

Satellite-based analytics and remote sensing have also become important. Machine learning models have been applied to estimate soil parameters from satellite and other environmental data [16], and IoT-based systems can forecast yields in real time using sensor and climate information [17]. Region-specific machine learning methods further improve prediction accuracy by accounting for local conditions [18, 19].

Despite these developments, current research mainly emphasizes improving model accuracy but pays little attention to variability among low-productivity districts. Moreover, very little research has examined the effectiveness of hybrid machine learning frameworks in such areas. This paper addresses these gaps by proposing a scalable and robust framework that combines distributed preprocessing, ensemble learning, hybrid modeling, and district-level analysis for crop yield prediction.

3 Methodology

This section describes the overall workflow of the proposed hybrid, big-data-driven machine learning framework for crop yield prediction. The approach includes distributed data preprocessing, ensemble-based predictive modeling, hybrid learning, and system deployment considerations. The framework is designed to be scalable, precise, and viable for low-productivity agricultural areas.

3.1 Overall System Architecture

The proposed system follows a layered architecture consisting of interconnected processing stages, as illustrated in Figure 1.

A. Data Acquisition Layer: Collects and compiles data from agricultural departments, remote sensing

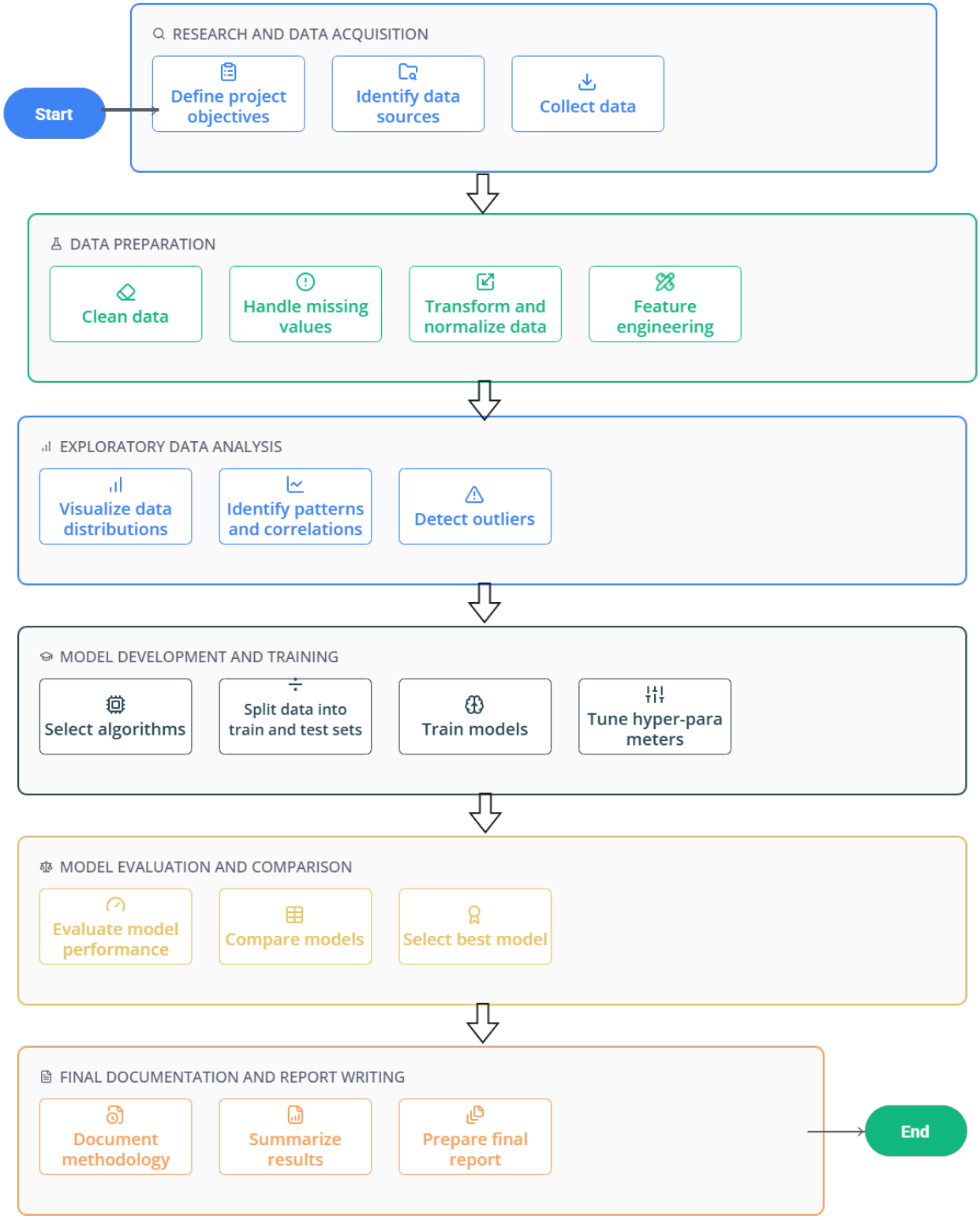

Figure 1. General system architecture of the proposed hybrid crop-yield-prediction framework.

facilities, and meteorological stations into a unified, district-level dataset.

As shown in Table 1, the acquired district-wise dataset integrates administrative, agronomic, soil, and vegetation-index fields into a single schema suitable for downstream processing.

B. Big Data Processing Layer: Aggregates, cleans, and transforms the data using a pipeline designed for compatibility with Apache Spark, as described in Algorithm 1.

In preprocessing, both temporal consistency and feature dependency were addressed to improve data

Table 1. Schema of the acquired district-wise agricultural dataset.

Field	Description
State, District	Administrative identifiers
Year, Season	Cropping year and season (Kharif/Rabi)
Crop	Crop type
Area_ha	Sown area (hectares)
Production_tonnes	Total production (tonnes)
Yield_kg_per_ha	Target variable: yield (kg/ha)
Kharif/Rabi_Rainfall_mm	Seasonal rainfall (mm)
pH	Soil pH
Organic_Carbon	Soil organic carbon
N, P, K	Soil nitrogen, phosphorus, potassium
Avg_NDVI	Mean vegetation index

Algorithm 1: Data Preprocessing

Data: Raw agricultural dataset D
Result: Aggregated dataset A
Remove missing/invalid values from D to obtain D_{clean};
Group D_{clean} by District, Year, and Season;
for *each group* **do**
  Aggregate rainfall;
  Aggregate soil parameters;
  Aggregate nutrient values (N, P, K);
  Aggregate NDVI;
  Aggregate yield;
end
Store all aggregated results in A;
return A

quality. NDVI and soil characteristics were temporally aligned by aggregating data on a seasonal basis. Mean NDVI was calculated across the crop growth period and matched with the corresponding soil and weather variables using common keys (district, year, season) to ensure consistency across heterogeneous data sources.

Moreover, tree-based ensemble models were used to address multicollinearity between environmental features (e.g., soil nutrients: nitrogen, phosphorus, potassium) and climatic variables (e.g., rainfall, NDVI), since these models are naturally robust to correlated inputs through their hierarchical feature-selection mechanism. Domain-knowledge-based feature selection further reduced redundant features to ensure a meaningful input representation.

C. Modeling Layer: Both individual and hybrid machine learning models were used to predict yields during training, following the procedure in Algorithm 2.

D. Evaluation Layer: Performance is assessed using standard regression measures (Section 3.6).

Algorithm 2: Base Model Training

Data: Training data (X_{train}, y_{train}), testing data (X_{test}, y_{test})
Result: Model performance metrics
Initialize models: Linear Regression, Random Forest, XGBoost, Extra Trees;
for *each model* **do**
  Train the model using X_{train}, y_{train};
  Predict y_{pred} from X_{test};
end
Compute R^2 and Mean Absolute Error using y_{test} and y_{pred};
Store the results;
return *comparison of model performance*

This modular architecture enables large-scale data processing to be combined with predictive analytics.

3.2 Base Model Development

Four regression models were employed as base learners.

1) Linear Regression (LR): provides a baseline, interpretable model.

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \epsilon \tag{1}$$

where β_i are the model coefficients and ϵ is the error term.

2) Random Forest (RF): captures non-linear relationships using an ensemble of decision trees.

$$Y = \frac{1}{T} \sum_{t=1}^{T} f_t(X) \tag{2}$$

where T is the number of trees and f_t is the prediction from each tree.

3) XGBoost (XGB): utilizes gradient boosting for optimized sequential tree learning.

$$Y = \sum_{k=1}^{K} f_k(X) \tag{3}$$

where f_k denotes sequential trees, each correcting the errors of the previous ones.

4) Extra Trees (ET): introduces extreme randomization in split selection to enhance generalization.

$$Y = \frac{1}{T} \sum_{t=1}^{T} f_t(X, \theta_t) \tag{4}$$

where θ_t denotes the random split parameters of each tree.

5) Hybrid Ensemble Models: the hybrid model combines the predictions of the constituent base models by simple averaging,

$$Y_{ensemble} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{5}$$

where Y_i is the prediction from each constituent model and n is the number of models combined.

Each model was trained independently using identical training data, ensuring a fair comparison.

3.3 Model Configuration

Table 2. Hyperparameters of the base models.

Model	Parameters
Random Forest	n_estimators=300, max_depth=15, min_samples_split=2
XGBoost	n_estimators=300, learning_rate=0.05, max_depth=6, subsample=0.8
Extra Trees	n_estimators=300, random_state=42
Linear Regression	Default

As summarized in Table 2, all models were trained on an 80:20 train–test split. Model parameters were learned using the training dataset, and performance was measured on the held-out testing dataset; the same split was applied to all models for consistency. Hyperparameters were selected empirically based on validation performance: the number of estimators, tree depth, and learning rate were tuned to balance bias and variance without overfitting. Tree-based ensemble models increase model complexity by integrating multiple decision trees, enabling them to capture non-linear associations between environmental factors and crop yield; this added complexity is mitigated through randomness and parameter tuning. The chosen models mix a simple baseline (Linear Regression) with complex ensemble-based methods (Random Forest, XGBoost, Extra Trees), enabling a detailed comparison between linear and non-linear, and single and ensemble, learning methods.

3.4 Hybrid Construction of Ensembles

Hybrid ensemble models based on voting regression were constructed to improve robustness and predictive stability. Base learners were combined through averaging to exploit differences among models and reduce variance. Hybrids were formed both in pairs and as a combination of all base learners. This grouping strategy mitigates overfitting and is therefore applicable to heterogeneous agricultural locations.

3.5 Optimization and Training of Models

All models were trained and tested on the dataset using the same 80:20 split. Hyperparameters (tree depth, learning rate, number of estimators) were tuned empirically against validation performance to balance bias and variance.

3.6 Performance Evaluation

Two standard regression measures were used to evaluate model performance.

Coefficient of Determination (R^2): measures how well the model explains the variance in the target variable.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{6}$$

where y_i is the actual yield value, $\hat{y}_i$ is the predicted yield value, and $\bar{y}$ is the mean yield value.

Mean Absolute Error (MAE): measures the average magnitude of the prediction error.

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i| \tag{7}$$

where n is the number of observations. All base models and hybrids were compared using these two measures to determine the best configuration.

3.7 Explainability and Interpretability Analysis

Post-hoc interpretation and transparency of model behavior were obtained using SHAP (SHapley Additive exPlanations) [20]. The Extra Trees model was analyzed using TreeExplainer, yielding (i) global feature importance, (ii) local, instance-level explanations, and (iii) district-level attribution patterns, showing the relative effect of soil, climatic, and vegetation variables on yield predictions.

3.8 District-wise Performance Assessment

Spatial heterogeneity at the district level was analyzed by computing district-level predictions and individual error values, enabling the detection of region-specific constraints, such as nutrient deficits or environmental vulnerability, that can inform regional agricultural planning.

4 Experimental Results and Discussion

4.1 Experimental Setup

The crop yield prediction framework was implemented in Python using Google Colab. PySpark and Pandas were used for data preprocessing and aggregation, and Scikit-learn and XGBoost were used to implement the algorithms. The processed data was split 80:20 into training and testing sets to evaluate performance on unseen data and avoid overfitting. Prediction quality was assessed using two standard regression measures: the Coefficient of Determination (R^2), which measures the explained variance, and the Mean Absolute Error (MAE), which measures the average absolute difference between actual and predicted yield.

4.2 Performance Comparison of Base and Hybrid Models

Table 3 and Figure 2 show the performance comparison between individual machine learning models and hybrid ensemble models in crop yield prediction. The results reveal clear differences in predictive ability across the compared methods.

Table 3. R^2 and MAE of all base and hybrid models, sorted by R^2.

Model	R^2	MAE (kg/ha)
Extra Trees	0.989	30.44
RF + ET	0.986	32.07
Random Forest	0.973	45.96
LR + RF + XGB + ET	0.966	56.33
XGB + ET	0.963	46.84
LR + ET	0.958	58.05
RF + XGB	0.950	61.12
LR + RF	0.933	71.33
LR + XGB	0.910	86.73
XGBoost	0.870	76.61
Linear Regression	0.838	121.03

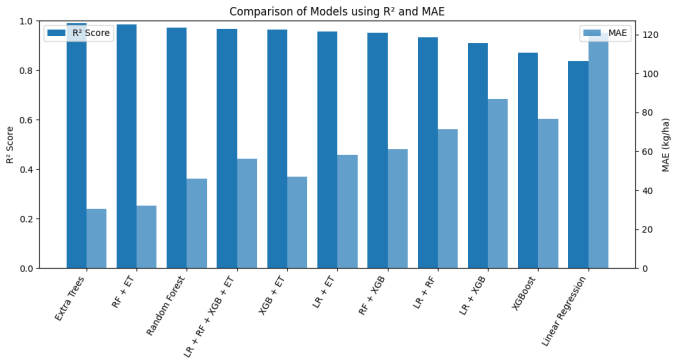

Figure 2. Comparison of prediction models using R^2 and MAE metrics.

The Extra Trees regression model was the most precise, with a Mean Absolute Error of 30.44 kg/ha and R^2 of 0.989, implying that the model explains 98.9% of the variance in crop yield – a strong predictive result. The RF + ET hybrid, with R^2 and MAE of 0.986 and 32.07 kg/ha respectively, also performed well, showing that combining ensemble models can improve prediction robustness by reducing model variance. Random Forest alone achieved a high R^2 of 0.973, indicating that tree-based ensemble models are well suited to modeling complex agricultural data. It is noted that R^2 and MAE do not always rank models identically, since R^2 measures explained variance while MAE measures absolute prediction error; minor inversions in ranking between the two metrics (e.g., XGB + ET versus LR + RF + XGB + ET, and XGBoost versus LR + XGB) therefore reflect differences in error distribution rather than inconsistency in the results.

4.3 Computational and Scalability Analysis

Training time was used as a proxy for scalability, since it reflects computational efficiency. Random Forest and Extra Trees are parallelizable tree-based ensemble techniques and can therefore be applied to large-scale datasets. Although the framework was not deployed in a fully distributed setting in this study, the modular pipeline and its compatibility with Apache Spark allow it to be scaled in future work.

4.4 Analysis of Hybrid Ensemble Models

To increase prediction reliability, hybrid ensemble models were developed by combining base learners through voting regression to obtain superior generalization across different algorithms. Results indicate that hybrids of similar ensemble methods (e.g., RF + ET) perform well, presumably because both rely on non-linear, tree-based modeling. However, adding a weaker model such as Linear Regression has a mildly negative effect on performance: the LR + RF + XGB + ET hybrid achieves an average score of $R^2 = 0.966$, lower than Extra Trees alone, since the accuracy of the weaker model can pull down the ensemble's overall performance.

Although hybrid ensembling was explored, the Extra Trees model proved most effective on its own. Hybrid models average the predictions of multiple learners, and this averaging can introduce a small reduction in peak accuracy even as it improves stability and reduces variance across different data splits.

4.5 Error Analysis

The Extra Trees model shows the lowest Mean Absolute Error (30.44 kg/ha), corresponding to a deviation of roughly 1–2% relative to the overall mean crop yield of approximately 2000 kg/ha, indicating high reliability. In contrast, Linear Regression recorded the largest MAE (121.03 kg/ha), likely because its linear assumption cannot capture the non-linear relationships between environmental factors and crop yield.

4.6 Cross-Validation Analysis

A 5-fold cross-validation was performed to assess robustness. The Extra Trees model obtained a mean R^2 of 0.993 ± 0.003 and a mean MAE of 28.91 ± 2.34 kg/ha, indicating that the model remains stable across different data splits and generalizes well despite the limited size of the test data.

4.7 Visualization of Model Performance

As shown in Figure 2, the ensemble-based methods consistently outperform the traditional regression baseline across both evaluation measures, with the Extra Trees model achieving the best performance overall and confirming its suitability for crop yield prediction on data-intensive agricultural tasks.

4.8 Confusion-Matrix-Based Evaluation

Figure 3 shows the normalized confusion matrix of the best-performing Extra Trees model when crop yield is discretized into three categories (low, medium, high) using percentile-based thresholds.

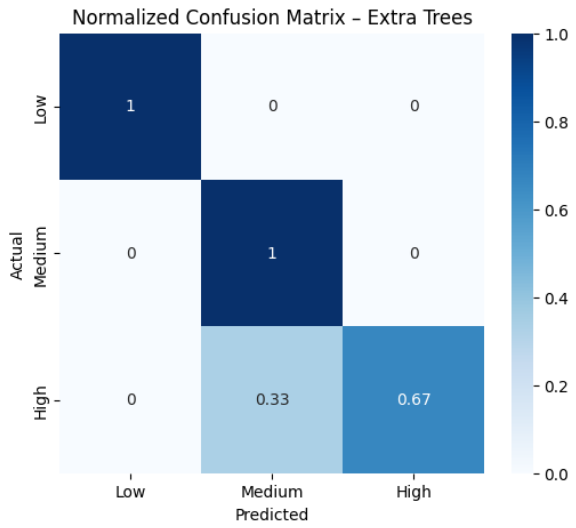

Figure 3. Normalized confusion matrix of the Extra Trees model.

The model perfectly classifies the low- and medium-yield categories, with all samples in these two categories classified correctly. This indicates that the model captures the trends of lower- and moderately-yielding conditions well. In the high-yield category, the model achieves 67% accuracy (2 out of 3 samples), with 33% of high-yield samples classified as medium, because small deviations near the class boundaries can shift feature values between the medium and high ranges.

It is worth noting that the confusion matrix is derived from a regression model with continuous outputs; discretizing continuous predictions introduces sensitivity to class boundaries, so a prediction that is numerically close to the true value may still be misclassified after thresholding. Because the test set is small, the reported percentages (e.g., 67% and 33%) are sensitive to sample variation and should be read as indicative rather than statistically conclusive.

4.9 Regression-Based Uncertainty Analysis

Although classification-based evaluation was used to aid interpretability, the primary model evaluation relies on regression measures such as R^2 and MAE, since classification introduces boundary sensitivity whereby small differences in predicted values can cause misclassification even when the underlying prediction is accurate.

To address this limitation, an uncertainty analysis was conducted using regression-based prediction error bands. A 95% confidence interval was constructed around each predicted value using the residual standard deviation. For a given prediction $\hat{y}$, the confidence interval is

$$\hat{y} \pm 1.96\,\sigma \tag{8}$$

where σ is the standard deviation of the residual errors. Representative sample predictions with their confidence intervals are shown in Table 4.

Table 4. Sample predictions with 95% confidence intervals (kg/ha).

Prediction	Lower bound	Upper bound
2284.63	2216.55	2352.70
1970.58	1902.50	2038.65
2274.00	2205.92	2342.07

These results show that the predicted values fall within a narrow confidence range, indicating stability and reliability. Compared with the discrete outputs of classification, regression-based error bands provide

a continuous and more informative account of prediction uncertainty.

5 Explainability Analysis Using SHAP

The Extra Trees regression model, which achieved the highest R^2 (0.989) and the lowest MAE (30.44 kg/ha), was selected for interpretability analysis. Understanding the drivers of these predictions is important for making reliable agricultural decisions. An explainable AI approach, SHAP (SHapley Additive exPlanations), was therefore used to quantify the contribution of each input feature to the final prediction, providing both global (overall feature importance) and local (individual prediction) explanations. SHAP was applied to the Extra Trees model because it was the best-performing and most stably generalizing model, and tree-based SHAP (TreeExplainer) efficiently supports ensemble tree models.

5.1 Global Feature Importance

Figure 4 shows the global feature importance obtained from the SHAP analysis.

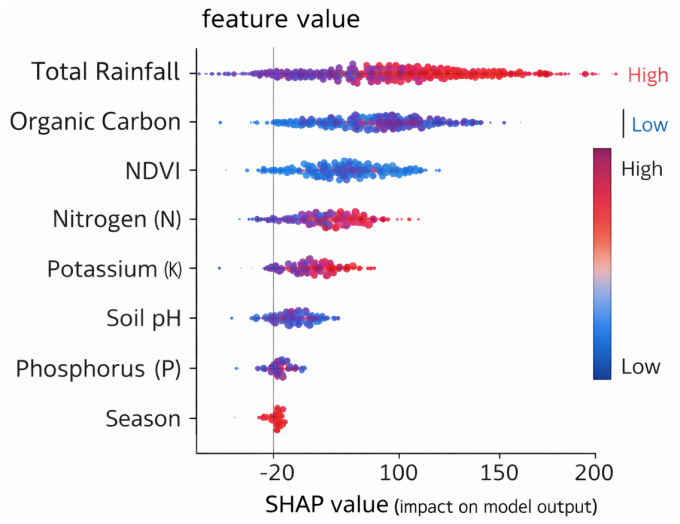

Figure 4. SHAP summary plot showing feature importance for crop yield prediction.

The SHAP summary plot ranks the input variables by their impact on the crop yield prediction. Each point represents a sample, and its horizontal position (SHAP value) reflects the strength of that feature's impact; features are ranked from most to least influential. Total rainfall is the most important feature, reflecting the strong dependence of crop production on climatic conditions, with higher rainfall associated with higher predicted yield. Soil fertility indicators (organic carbon, nitrogen, phosphorus, potassium) and the vegetation-health indicator (NDVI) are also significant predictors of yield. In summary, the SHAP analysis shows that climatic conditions and soil nutrient status are essential drivers of crop yield prediction in low-productivity districts.

5.2 Local Prediction Explanation

SHAP waterfall plots were generated to further examine individual predictions. Figure 5 shows one such explanation for a single prediction.

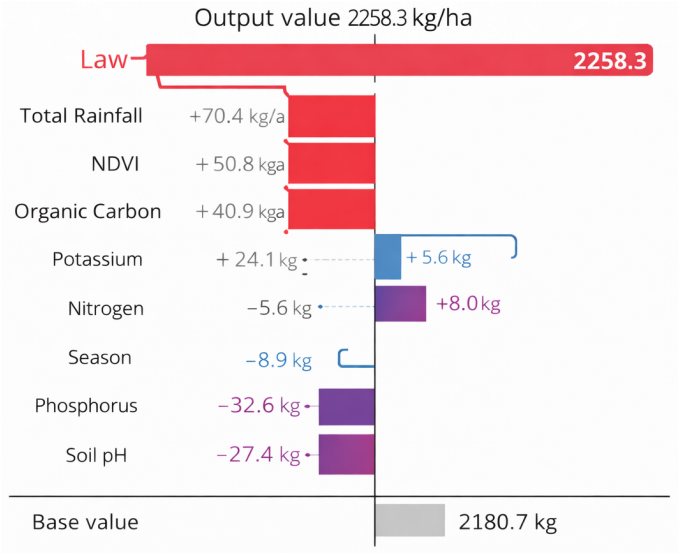

Figure 5. SHAP waterfall plot explaining feature contribution to a sample prediction.

The waterfall plot shows how individual features shift the final predicted crop yield relative to the baseline (the average predicted yield across the dataset), with each feature either increasing or decreasing the forecast. In this example, rainfall, NDVI, and organic carbon positively affect the predicted yield, while soil-nutrient imbalance or unfavorable seasonal conditions reduce it; the cumulative feature contributions determine the final prediction.

This analysis improves interpretability by identifying the agronomic variables that most influence productivity, informing farmers, policymakers, and agricultural planners in managing irrigation, improving soil fertility, and planning crops.

5.3 Discussion of Best-Performing Model

The strong performance of the Extra Trees algorithm stems from its randomized tree construction, which adds extra randomness to feature and split selection compared with classical decision trees, reducing correlation between trees in the ensemble and improving generalization. This randomization allows Extra Trees to effectively capture the complex,

non-linear, noisy, and heterogeneous patterns present in agricultural data (arising from variation in soil, rainfall, and vegetation), leading to improved predictions.

Extra Trees outperformed XGBoost because of its greater randomness in feature and split-threshold selection. This randomness reduces variance and improves generalization, particularly for noisy agricultural data, whereas XGBoost's sequential boosting can be more prone to overfitting on limited or noisy data.

5.4 Key Observations

The experimental findings highlight several key points: ensemble learning techniques outperform conventional regression models for crop yield prediction, with the Extra Trees model performing best overall; hybrid models improve prediction stability but can underperform relative to the strongest individual model when weaker learners are included; Linear Regression performs worst because it cannot capture non-linear relationships; and tree-based ensemble algorithms are particularly effective for heterogeneous, multi-type agricultural datasets. Overall, these findings demonstrate that the proposed hybrid machine learning framework can produce valid and reliable crop yield predictions in low-productivity districts.

6 Conclusion

This paper introduced a hybrid machine learning framework, driven by big data, for predicting crop yield in low-productivity districts. The proposed system combines distributed-capable data preprocessing, ensemble-based predictive modeling, and explainability analysis. A preprocessing pipeline designed for compatibility with Apache Spark was adopted to handle large agricultural datasets, and four base models together with several hybrid ensembles were evaluated. The experiments show that the Extra Trees model achieves the best accuracy and lowest error, while hybrid models further improve prediction stability. SHAP-based explainability provided a clear account of model behavior at both local and global scales, and district-level analysis revealed regional differences in soil and climatic constraints that can inform agricultural interventions. Overall, the proposed framework offers a valid, scalable, and dependable approach to crop yield prediction that can support data-driven decision-making in precision agriculture.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that ChatGPT was used to assist with language editing, structural refinement, and selected drafting of the manuscript. The authors have carefully reviewed, revised, and verified the AI-assisted output and take full responsibility for the content of the manuscript.

Ethical Approval and Consent to Participate

Not applicable. This study is based entirely on secondary, district-level agricultural, soil, and climatic datasets that are publicly available government records accessed in accordance with applicable data use policies, and did not involve human participants, personal data collection, or animal experiments.

References

[1] Hernández Hernández, G. C., Gómez Gómez, J., & Jiménez-Cabas, J. (2025). Predictive models based on artificial intelligence to estimate crop yield: A literature review. *Agriculture, 15*(23), 2438. [CrossRef]

[2] Saha, S., Kucher, O. D., Utkina, A. O., & Rebouh, N. Y. (2025). Precision agriculture for improving crop yield predictions: a literature review. *Frontiers in Agronomy, 7*, 1566201. [CrossRef]

[3] Ahmed, F. U., Das, A., & Zubair, M. (2024, March). A machine learning approach for crop yield and disease prediction integrating soil nutrition and weather factors. In *2024 international conference on advances in computing, communication, electrical, and smart systems (iCACCESS)* (pp. 1-6). IEEE. [CrossRef]

[4] 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), 968. [CrossRef]

[5] Lou, Z., Lu, X., & Li, S. (2024). Yield prediction of winter wheat at different growth stages based on machine learning. *Agronomy, 14*(8), 1834. [CrossRef]

[6] Xu, H., Yin, H., Liu, Y., Wang, B., Song, H., Zheng, Z., ... & Wang, S. (2024). Regional winter wheat yield prediction and variable importance analysis based

on multisource environmental data. *Agronomy, 14*(8), 1623. [CrossRef]

[7] Nosratabadi, S., Imre, F., Szell, K., Ardabili, S., Beszedes, B., & Mosavi, A. (2020). Hybrid machine learning models for crop yield prediction. *arXiv preprint arXiv:2005.04155*. [CrossRef]

[8] Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, 10, 621. [CrossRef]

[9] Subramaniam, L. K., & Marimuthu, R. (2024). Crop yield prediction using effective deep learning and dimensionality reduction approaches for Indian regional crops. *e-Prime-Advances in Electrical Engineering, Electronics and Energy, 8*, 100611. [CrossRef]

[10] Wang, Y., Zhang, Q., Yu, F., Zhang, N., Zhang, X., Li, Y., Wang, M., & Zhang, J. (2024). Progress in research on deep learning-based crop yield prediction. *Agronomy, 14*(10), 2264. [CrossRef]

[11] Fan, J., Bai, J., Li, Z., Ortiz-Bobea, A., & Gomes, C. P. (2022, June). A GNN-RNN approach for harnessing geospatial and temporal information: application to crop yield prediction. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, No. 11, pp. 11873-11881). [CrossRef]

[12] Miranda, M., Charfuelan, M., & Dengel, A. (2024). Exploring physics-informed neural networks for crop yield loss forecasting. *arXiv preprint arXiv:2501.00502*. [CrossRef]

[13] Ngo, V. M., Le-Khac, N. A., & Kechadi, M. (2018). An efficient data warehouse for crop yield prediction. *arXiv preprint arXiv:1807.00035*. [CrossRef]

[14] Knapen, R., de Wit, A., Buyukkaya, E., Petrou, P., Paudel, D., Janssen, S., & Athanasiadis, I. (2025). Efficient and scalable crop growth simulations using standard big data and distributed computing technologies. *Computers and Electronics in Agriculture, 236*, 110392. [CrossRef]

[15] Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017). Big data in smart farming–a review. *Agricultural Systems, 153*, 69-80. [CrossRef]

[16] Kammerlander, C., Kolb, V., Luegmair, M., Scheermann, L., Schmailzl, M., Seufert, M., ... & Schön, T. (2025). Machine learning models for soil parameter prediction based on satellite, weather, clay and yield data. *arXiv preprint arXiv:2503.22276*. [CrossRef]

[17] Talaat, F. M. (2023). Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes. *Neural Computing and Applications, 35*(23), 17281-17292. [CrossRef]

[18] Surana, R., & Khandelwal, R. (2024). Crop yield prediction using machine learning: A pragmatic approach. [CrossRef]

[19] SB, S. H., & Rawat, S. (2026). Crop Yield Prediction Using Different Techniques of Machine Learning for Prayagraj Region. *International Journal of Environment and Climate Change, 16*(2), 386-397. https://hal.science/hal-05511932/

[20] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765-4774.