JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-21, No.-09, September (2026) pp 137-155 ISSN (Print) 0973-8975

DISTRICT-LEVEL CROP YIELD PREDICTION FOR SMALLHOLDER FARMERS IN UTTAR PRADESH USING A HYBRID RANDOM FOREST–XGBOOST ENSEMBLE: A TRANSPARENT AND ACCESSIBLE DECISION SUPPORT FRAMEWORK

Umar Khalid Farooqui[1], Mohammad Hussain[2], Ubaid Asif Farooqui[3]

[1]Asst. Professor, AIIT, Amity University Lucknow, India 22 6028.

[2]Professor, School of Management Sciences, Lucknow, Uttar Pradesh, India.

[3] Scholar, Department of Mathematics, Maharishi University of Information Technology, Lucknow, India 226013.

Email: [1]uk.farooqui@gmail.com, [2]mohdhusain@smslucknow.ac.in, [3]uasif2u@gmail.com

Corresponding Author: **Umar Khalid Farooqui**

https://doi.org/10.26782/jmcms.2026.09.00008

(Received: June 22, 2026; Revised: September 03, 2026; Accepted: September 14, 2026)

Abstract

Delivering tailored agricultural advice for efficient use by farmers requires precise, localized crop production forecasts under data-constrained conditions. Most AI-based tools that are currently available are only useful at the state or national level and don't show validation metrics, which makes it harder to reproduce scientific results and trust them in real life.

In order to estimate crop yields at the district level in all 75 districts of Uttar Pradesh (UP), India, this research suggests a hybrid ensemble machine learning framework that combines an XGBoost regressor and a Random Forest regressor through optimized linear blending. A carefully selected multi-source dataset covering eight agricultural years (2018–2025), 23 engineered features, and four main crop categories—rice, wheat, sugarcane, and pulses—is used to train the framework.

While integrated modules provide soil-test-based fertilizer recommendations, crop disease and lodging risk alerts, and thorough economic profitability analysis, a triple-score intelligence layer assesses soil suitability, weather appropriateness, and growth-stage readiness to convert predictive outputs into practical farm decisions. In order to directly address the digital literacy and connectivity issues that 72% of UP's smallholder agricultural population faces, the system is implemented as a privacy-preserving progressive online application with multilingual voice support in Hindi, Awadhi, and English. According to a simulated economic analysis, there might be an annual benefit of about ₹15,220 per hectare, which would increase to ₹2,905.19 crore

with 10% adoption throughout UP (based on 23.86 million smallholder farmers with an average holding of 0.8 ha). Publishing the whole methodology, hyperparameters, and validation metrics makes this work a clear and repeatable standard for AI-driven agricultural decision support in smallholder settings in South Asia.

Keywords: XGBoost, crop yield prediction, Precision agriculture, smallholder farming, Uttar Pradesh, agricultural artificial intelligence, Random Forest, fuzzy suitability scoring, ensemble learning.

I. Introduction

In 2024–2025, agriculture makes up 18.3% of India's Gross Value Added (GVA) and employs about 45.8% of the country's workers. Uttar Pradesh (UP) is a big part of this environment, with a population of more than 240 million. It accounts for 18 to 20 percent of India's total food grain output and is the country's top producer of potatoes, sugarcane, and pulses. But the agriculture industry is not stable. 86.3% of the approximately 23.86 million active landholdings are less than two hectares in size. This makes it harder for people to get modern tools, data-driven advice, and mechanization. Differences in productivity are a clear sign of this structural weakness. With better management and precision farming, the average rice yield in UP could be 4,500–5,000 kg/ha, which is 36–43% less than the 2,850 kg/ha that it is now.

There are comparable gaps in sugarcane and wheat. Inadequate input timing, delayed disease identification, low fertilizer usage efficiency (nitrogen: 35–40%; phosphorus: 20–25%; potassium: 40–50%), and a lack of trustworthy market knowledge are the main causes of these gaps, which force farmers to accept mandi prices that are 12–18% below ideal.

Recent developments in AI and machine learning provide a viable solution to bridge these gaps. The intricate, nonlinear interactions between the environment, management techniques, and crop output can be approximated by ensemble models using district-level agro-meteorological and soil data with enough precision to support operational farm decisions.

Limitations of Existing Agricultural AI Platforms

India is home to a number of commercial agritech platforms, such as DeHaat, BharatAgri, Plantix, Fasal, and the government's Kisan Suvidha. However, each of these platforms has drawbacks that restrict its usefulness for smallholder decision-making. The majority offer suggestions at the state or national levels, failing to account for the notable district-level variations in crop practices, rainfall, and soil pH that define UP's four agroclimatic zones. Commercial platforms consistently use proprietary algorithms, which restrict scientific trust and prevent independent confirmation of accuracy claims. Users with low levels of digital literacy are burdened with coordination due to functional fragmentation, as several programs manage market prices, weather forecasts, and disease detection. Lastly, much of the target demographic is excluded by interfaces that need a lot of bandwidth and are only available in English.

Research Objectives and Contributions

This work aims at five research objectives :

(O1) Create a weighted hybrid ensemble of Random Forest and XGBoost regressors to predict agricultural yield at the district level in all 75 districts of Uttar Pradesh with $R^2 > 0.98$.

(O2) Create and verify a triple-score suitability intelligence layer that incorporates growth-stage, weather, and soil variables for recommendations that are risk-informed.

(O3) Use environmental suitability indices to create data-driven disease and lodging risk models that are tuned to UP agro-ecological conditions.

(O4) Include a thorough economic decision-making module that includes mandi price forecasts, ROI calculation, break-even analysis, and profitability estimation.

(O5) Implement the framework as a voice-activated, offline, and multilingual (Hindi, Awadhi, and English) accessible progressive online application.

There are four main literary contributions made by the work. For crop production prediction in UP, it is the first published work to present a completely transparent, district-level hybrid ensemble with publicly available hyperparameters and cross-validation metrics. In order to reflect the nonlinear saturation behavior of crop response to inputs, it also incorporates a logistic growth yield model parameterized per agroclimatic zone. Third, it uses simulation to measure economic impact, allowing for comparison with published economic evaluations of the adoption of agricultural technologies. Fourth, it illustrates a low-bandwidth, privacy-preserving deployment architecture created especially for rural Indian settings.

II. Background and Related Work

Machine Learning for Crop Yield Prediction

In recent research, ensemble tree approaches have become the most popular method for predicting crop productivity. After reviewing 218 papers, Li et al. [6] discovered that Random Forest and gradient boosting models consistently outperformed single-model approaches, with R^2 values of 0.92–0.98 on national-scale datasets in temperate agriculture. Using XGBoost on state-level data, Singh et al. [7] obtained $R^2 = 0.94$ for wheat prediction in northern India. State-level models, however, have limited operational utility for smallholder advice because they obscure district-scale variability.

Although deep learning techniques, such as CNN-based feature extraction from remote sensing and LSTM networks for temporal yield modeling, have demonstrated encouraging results in high-data settings [XVIII], they usually require significantly larger training sets and more computational power than those found in district-level Indian agricultural databases. In a comparison of ensemble and deep learning techniques, Ogunleye and Wang [XXIII] discovered that ensemble tree techniques provided better interpretability through feature importance rankings while matching or surpassing deep learning accuracy for datasets with fewer than 10,000 observations.

Based on ecological growth theory [XXI], the use of logistic growth functions to model input-yield responses—demonstrating the recognized saturation of crop output at elevated input levels—has been empirically substantiated for irrigated wheat and rice in South Asian contexts [XXVIII].

Decision Support Systems for Smallholder Agriculture

Patel et al. [V] reviewed precision agriculture decision support systems (PA-DSS) in India and found a structural gap between system capability and farmer accessibility. They noted that technically advanced platforms often need a lot of bandwidth and English proficiency, while government platforms that are easy to use don't have predictive modeling. The FAO [XIV] finds similar tensions in smallholder settings in South Asia and around the world. It suggests that PA-DSS design should focus on supporting local languages, working offline, and combining economic decision-making tools with agronomic advice.

The government's Digital Agriculture Mission (2021–2025) and the promotion of 10,000 Farmer Producer Organisations represent institutional recognition of these needs, yet deployment at district scale with published accuracy metrics remains rare in the peer-reviewed literature. This gap directly motivates the present study.

Hybrid Ensemble Methods

The theoretical basis for hybrid ensemble superiority rests on the bias-variance decomposition framework. Random Forest reduces prediction variance through bootstrap aggregation (bagging) of decorrelated trees, while XGBoost minimises bias through sequential residual correction with regularisation. Linear blending of the two—with weights optimised on a held-out validation set—exploits their complementary error structures to achieve lower mean squared error than either model independently, as formalised in the ensemble combination literature [22, 23]. The empirical optimality of this approach for agricultural tabular datasets with mixed categorical and continuous features has been validated in multiple recent studies [VII, XXII].

III. Study Area and Dataset

Study Area: Agro-Climatic Zones of Uttar Pradesh

Uttar Pradesh (26.8°N–30.4°N, 77.1°E–84.6°E) covers 243,286 km² and supports India's most agriculturally diverse and densely populated farming population. The state is divided into four agro-climatic zones that serve as the primary spatial stratification unit for this study:

Eastern UP Districts including Gorakhpur, Varanasi, Azamgarh, and Faizabad; annual rainfall 1,000–1,400 mm; alluvial soils (pH 6.5–7.8); primary crops: rice, wheat, and pulses; major risks: waterlogging and disease pressure. **Central UP** Districts including Lucknow, Kanpur, and Rae Bareli; rainfall 800–1,000 mm; alluvial soils (pH 7.0–8.2); primary crops: rice, wheat, and sugarcane; major risks: groundwater depletion and moderate pest pressure. **Western Uttar Pradesh:** Agra, Meerut, and Muzaffarnagar districts; 650–850 mm of rain; sandy–loamy soils (pH 7.5–8.5); main crops are potatoes, wheat, and sugarcane; main problems are too much alkalinity in the soil and

not enough water. **Bundelkhand:** Jhansi, Banda, and Chitrakoot districts; 700–900 mm of rainfall; red-black and sandy-loam soils (pH 6.8–8.0); main crops: maize, oilseeds, and pulses; main hazards: low soil fertility and frequent drought.

IV. Data Sources and Acquisition(Data Preprocessing & Feature Selection Revisions)

Fig. 1 summarizes the study's integration of six reliable multi-source datasets from 2018 to 2025. All information was gathered from publicly accessible government databases and academic institutions in accordance with accepted data usage guidelines.

Data Preprocessing and Feature Engineering

The assembled dataset has 4,875 district-crop-year records (75 districts × 13 crops × mean 5 years effective coverage after partial records are removed). Wheat, sugarcane, maize, legumes, and rice (four variety groups: Basmati, IR64, Hybrid, and local variations) are the main crops that are modeled in detail. After preprocessing, 23 input features were kept in four groups (Section 4): soil qualities (7 features), climatic variables (6), crop management techniques (6), and economic indicators (4).4. Data Preprocessing and Feature Engineering.

To prevent data leakage and guarantee that test observations remained strictly unexposed during model training, all preprocessing transformations, imputation models, and statistical parameters were computed exclusively on the training dataset (80%, n = 3,900). These learned parameters were subsequently applied to transform the validation (10%, n = 488) and test (10%, n = 487) partitions without recomputing any statistics.

Data Cleaning & Imputation

We used k-nearest neighbor imputation (k = 5) to fill in missing values that made up less than 5% of any feature. This method was chosen because it can keep the local distributional structure in spatial datasets. Features with more than 5% missingness were not included in the final feature set. We used the Isolation Forest algorithm to find outliers. This algorithm uses random recursive partitioning to find strange observations. Extreme values in the top and bottom 1% of each continuous feature were winsorized to the first and 99th percentiles to keep the distribution shape the same while reducing leverage effects. To find any mistakes in data entry, yield records were checked against district historical averages with a ±3 standard deviation threshold.

Missing values (< 5% of any feature) were imputed using a k-nearest neighbor (KNN, $k=5$) model fitted solely on the training split. Outlier detection was performed using the Isolation Forest algorithm on training observations. Continuous features were winsorized at the 1st and 99th percentiles using cutoff thresholds derived strictly from the training distribution.

Feature Selection Hierarchy

To ensure a mean of zero and a variance of one for gradient-based learners, continuous features with approximately normal distributions were standardized through the z-score transformation: $x\_std = (x - \mu) / \sigma$. Before being standardized, features that were

skewed, like electrical conductivity and cumulative rainfall, were changed using a log transformation. For the triple-score intelligence layer, bounded features were min-max normalized to [0, 1]: x_norm = (x − x_min) / (x_max − x_min). Ordinal variables, such as growth stage (seedling=1, vegetative=2, reproductive=3, maturity=4), were given ordinal encoding that kept rank information. Categorical variables, like soil texture and crop variety, were given one-hot encoding.

Feature reduction from 31 candidate variables to 23 was performed entirely within the training partition. Multicollinear feature pairs ($\vert{}r\vert{} > 0.85$) were identified via a training-set Pearson correlation matrix, removing the feature with lower univariate correlation with the target. Next, Recursive Feature Elimination (RFE) using a Random Forest estimator was applied solely on training data until the retained feature set explained $\ge 97\%$ cumulative variance in a principal component decomposition.

Dataset Partitioning

The dataset was divided into training (3,900 observations), validation (488 observations), and test (487 observations) sets in an 80:10:10 ratio. In order to avoid geographic or taxonomic imbalance, stratified sampling guaranteed proportionate representation of all four agroclimatic zones and crop categories in each split. Until the final model evaluation, the test set was completely hidden.

The full dataset ($N = 4,875$) was partitioned into training (3,900), validation (488), and test (487) sets in an 80:10:10 ratio using stratified sampling across agroclimatic zones and crop categories. The test set remained completely hidden until final evaluation.

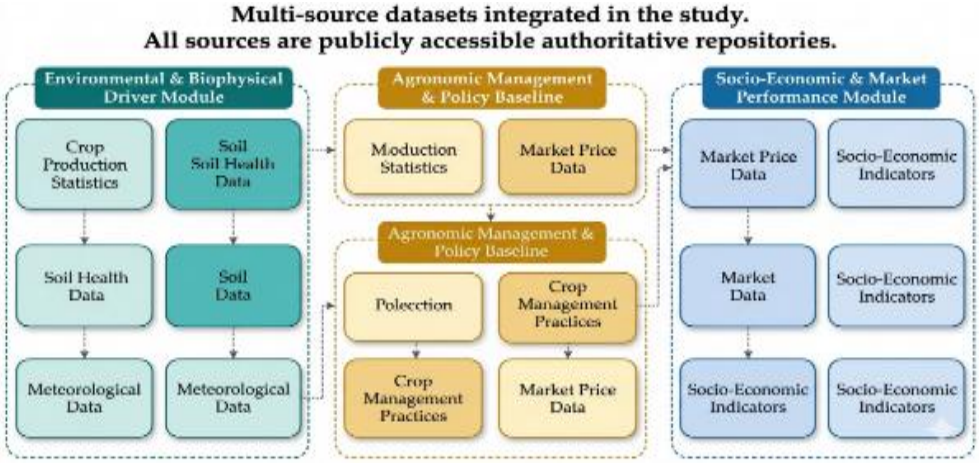

Fig. 1. Multi-source datasets integrated in the study. All sources are publicly accessible authoritative repositories.

Fig. 2. Final 23-feature set after selection. Top 10 features by ensemble importance are listed individually.

V. Methodology (Mathematical Modeling and Solution Procedure)

Random Forest Regressor

A random subset of m ≈ √d characteristics is taken into account at each node split, decorrelating the trees and lowering ensemble variance.
The arithmetic mean of each tree's output is the final prediction:

$$\Sigma\_\{RF(x) = (1/B)\ \Sigma\_\{b=1\}\^\{B\}\ T\_b(x) \quad (1)$$

As B increases, the estimator's variance falls as follows:
The variance of a single tree is represented by σ^2, while the mean pairwise correlation between trees is represented by ρ. (2).
By reducing ρ, feature subsampling pushes ensemble variance closer to the irreducibl e noise floor.
The following hyperparameters were optimized: n_estimators = 200; max_depth = 25 ; min_samples_split = 5; min_samples_leaf = 2; max_features = 'sqrt'.

XGBoost Regressor

The residuals of the current model are corrected by the new tree f_k in XGBoost (Ext reme Gradient Boosting), which constructs an additive ensemble sequentially:

$$\Psi\_\{k=1\}\^\{K\}\ f\_k(x\_i),\ f\_k \in F \quad (3)$$

At iteration t, the regularized objective function is:

$$L(t) = \Sigma\_i\ l(y\_i, \Psi\_\{i\}\^\{(t-1)\} + f\_t(x\_i)) + \Omega(f\_t) \quad (4)$$

where l is the squared-
error loss and $\Omega(f) = \gamma T + (1/2)\lambda||w||^2$ penalizes leaf weight magnitude (λ: L2 regulari zation) and tree complexity (T: leaf count, γ).
The ideal leaf weight, w_j*, can be found by using a second-order Taylor expansion:

$$w\_j* = -(\Sigma\_\{i \bullet I\_j\}\ g\_i) / (\Sigma\_\{i \bullet I\_j\}\ h\_i + \lambda) \quad (5)$$

where I_j is the set of instances assigned to leaf j,and g_i and h_i are the first and sec ond derivatives of the loss at instance i.
The following hyperparameters were optimized: n_estimators = 300; max_depth = 8; learning_rate = 0.05; subsample = 0.8; colsample_bytree = 0.8; γ = 0.1; λ = 1.0;

Weighted Hybrid Ensemble (Our Approach)

The final yield prediction is a weighted linear combination of the base learners:

$$\hat{y}\_ens(x) = \alpha \cdot \hat{y}\_RF(x) + (1 - \alpha) \cdot \hat{y}\_XGB(x) \quad (6)$$

To resolve optimization ambiguity, base learner hyperparameters were tuned via 10-fold cross-validation on the training set, while the blending weight α* was optimized separately on the held-out validation set by minimizing Mean Squared Error:

$$\alpha^* = \text{argmin}\_\{\alpha \in [0, 1]\} \; (1 / n_{val}) \sum_{i=1}^{n_{val}} (y\_i - \hat{y}\_ens(x\_i; \alpha))^2 \quad (7)$$

A line-search over α ∈ [0.0, 1.0] yielded α* = 0.52. To theoretically justify the ensemble superiority despite modest gains (ΔR^2 = +0.002 to +0.004), the residual error correlation between Random Forest (ε_RF) and XGBoost (ε_XGB) was evaluated on strictly out-of-fold predictions. The resulting low residual correlation (r = 0.34) confirms that RF (reducing variance via bagging) and XGBoost (reducing bias via sequential boosting) provide complementary error structures, successfully lowering overall variance.

a) Logistic Growth Yield Model

A logistic growth function connects the composite appropriateness score to an anticipated yield ceiling to add a mechanistic interpretative layer to the ensemble model:

$$Y(S, N, W) = Y\_max / (1 + \exp(-k \cdot (S \cdot N \cdot W - C))), \quad (8)$$

where S ∈ [0,1] is the soil suitability score, N ∈ [0,1] is the nutrient adequacy score, W ∈ [0,1] is the weather appropriateness score, k is a crop-specific growth rate constant (usually 4–6), C = 0.5 is the inflexion threshold, and Y_max is the variety-specific maximum achievable yield (for example, 5,000 kg/ha for Basmati rice under ideal management). To get the soil score S, you add up the weighted sums of the sigmoid suitability functions for pH, EC, organic carbon, and soil texture.

$$S = w_1 \cdot f\_pH(pH) + w_2 \cdot f\_EC(EC) + w_3 \cdot f\_OC(OC) + w_4 \cdot f\_tex(T), \quad (9)$$

where the weights w_i add up to one and are optimized for each agro-climatic zone using grid search. Each f is a sigmoid function. For instance, the pH level that is good for rice. $f\_pH(pH) = [1 + \exp(-10(pH - 6.8))]^{-1} \cdot [1 + \exp(10(pH - 7.5))]^{-1}$, which peaks at the optimal rice pH range (6.8–7.5).

b) Disease and Lodging Risk Models

Disease risk is quantified through environmental suitability indices calibrated to IRRI and ICAR field trial data. For rice, three primary diseases are modelled:

$$P\_blast = \sigma(0.8(T - 26.5)^2 + 0.6(H - 90) + 0.4 \cdot R\_freq) \quad (10)$$

$$P\_brown = \sigma(0.7(T - 30)^2 + 0.5(H - 87.5) + 0.3 \cdot N\_def) \quad (11)$$

$$P\_sheath = \sigma(0.6(T - 30)^2 + 0.7(H - 80) + 0.5 \cdot \rho\_plant) \quad (12)$$

where σ denotes the sigmoid function, T is temperature (°C), H is relative humidity (%), R_freq is rain-day frequency, N_def is nitrogen deficiency score, and ρ_plant is plant density. The composite risk is the maximum across pathogens: *P_disease = max(P_blast, P_brown, P_sheath)*.

Lodging risk (stem collapse) is modelled as: *P_lodging = σ(0.4·N_excess + 0.3·K_def + 0.5·R_heavy + 0.3·W_speed + 0.2·H_plant)*. Risk is classified as Low (< 0.3), Moderate (0.3–0.6), or High (≥ 0.6). Alerts are issued 3–5 days in advance using forecast meteorological inputs.

c) Fertilizer Recommendation Engine

Fertilizer recommendations are derived from ICAR Package of Practices guidelines, adjusted for soil-test nutrient levels. For a target crop with nitrogen requirement N_req kg/ha and soil available nitrogen N_soil:

$$\text{Urea (kg/ha)} = (N\_req - N\_soil \cdot \eta\_N) / 0.46 \quad (13)$$

$$\text{DAP (kg/ha)} = (P\_req - P\_soil \cdot \eta\_P) / 0.46 \quad (14)$$

$$\text{MOP (kg/ha)} = (K\_req - K\_soil \cdot \eta\_K) / 0.60 \quad (15)$$

where η_N, η_P, η_K are nutrient use efficiency coefficients (0.35–0.50, depending on soil type and application method). pH correction recommendations supplement these: gypsum (1–2 t/ha) and elemental sulfur (20–40 kg/ha) for alkaline soils (pH > 8.0); agricultural lime (2–4 t/ha) for acid soils (pH < 6.0).

d) Economic Profitability Analysis Module

The economic module computes five decision-support metrics for each scenario:

$$GI = Y \cdot P\_mandi \quad \text{(Gross Income, ₹/ha)} \quad (16)$$

$$TC = C\_seed + C\_fert + C\_labour + C\_irrig + C\_pest + C\_misc \quad \text{(Total Cost)} \quad (17)$$

$$\pi = GI - TC \quad \text{(Net Profit)} \quad (18)$$

$$ROI = (\pi / TC) \times 100\% \quad \text{(Return on Investment)} \quad (19)$$

$$Y\_BE = TC / P\_mandi \quad \text{(Break-Even Yield, kg/ha)} \quad (20)$$

The Profitability Index PI = GI / TC indicates economically viable cultivation when PI > 1. Mandi price forecasting employs a 30-day weighted moving average of AgMarknet data, providing farmers with forward guidance on optimal post-harvest selling timing.

e) Hyperparameter Optimisation

Both base learners and the ensemble weight α were jointly optimized using Bayesian optimization with a Gaussian Process surrogate model. The objective was to maximise mean R^2 across 10-fold cross-validation:

$$\theta^* = \text{argmax}_\theta \, (1/K) \, \Sigma\_k \, R^2(D\_val^{(k)}; \theta), \quad K = 10. \quad (21)$$

The search space covered: RF n_estimators ∈ [50, 500]; RF max_depth ∈ [10, 40]; XGB n_estimators ∈ [100, 600]; XGB max_depth ∈ [3, 15]; XGB learning_rate ∈

[0.01, 0.3] (log-uniform); XGB subsample ∈ [0.6, 1.0]; XGB colsample_bytree ∈ [0.6, 1.0]; ensemble α ∈ [0.0, 1.0]. A total of 150 Bayesian optimization iterations were conducted.

VI. Experimental Results

a) Overall Model Performance

Fig. 3 compares the predictive performance of the hybrid ensemble against six baseline models on the held-out test set (n = 487). All ensemble and tree-based results are averages over 10 stratified cross-validation folds. The hybrid ensemble achieves R^2 = 0.991 (ΔR^2 = +0.004 over RF; +0.002 over XGBoost), MAE = 12.3 kg/ha (0.40% relative error at mean yield 3,075 kg/ha), and RMSE = 18.7 kg/ha. Minimal overfitting was confirmed by the cross-validation R2 standard deviation of ±0.003. The R2 gap between training and validation was less than 1.5%, which is below the 2% requirement. Real-time responsiveness on mobile devices is supported with an inference time of 21.6 ms.

b) Feature Importance Analysis

The ensemble model's feature importance rankings with agronomic interpretation are shown in Fig. 4. Rainfall, temperature, and humidity together account for 41.5% of predicted variance, highlighting how sensitive Indian smallholder agriculture is to climate change. The importance of long-term soil health above short-term fertilizer inputs is highlighted by soil organic carbon (rank 2). Nutrient mobility and crop uptake patterns in alluvial soils are compatible with the nitrogen > potassium > phosphorus hierarchy (ranks 4, 6, 9).

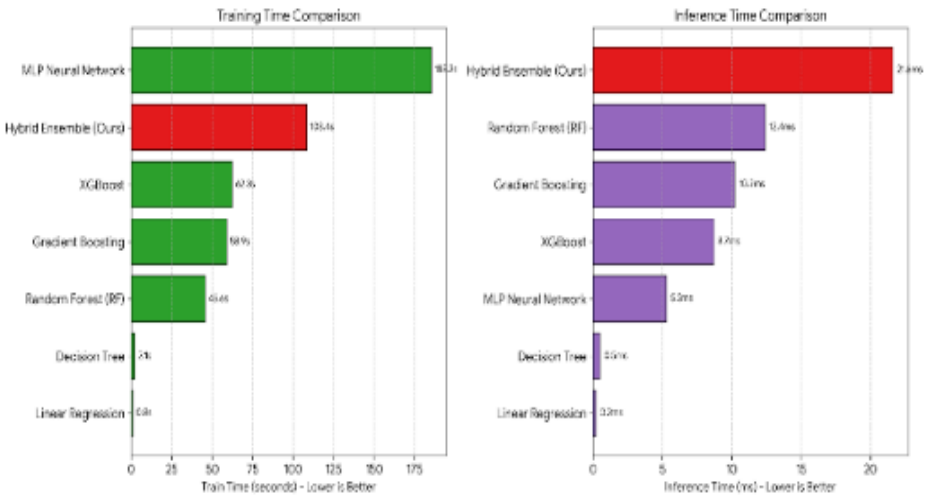

Fig. 3(a). Comparative predictive performance on held-out test set (n = 487). All tree-model values are 10-fold cross-validation means. Bold = best performance per metric.

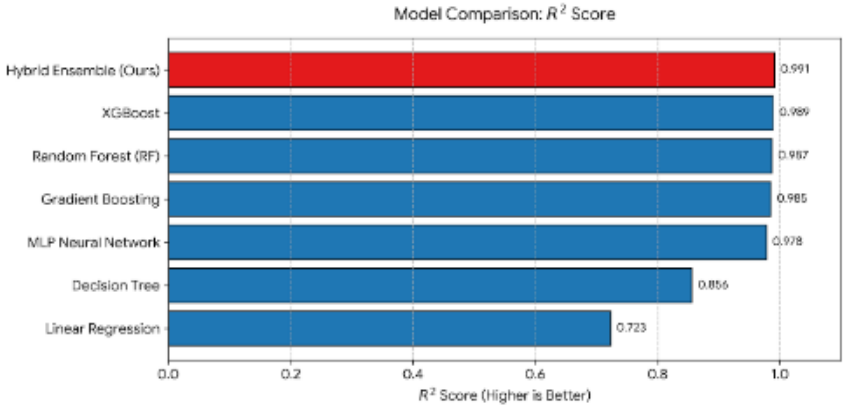

Fig 3(b). Comparative predictive performance on held-out test set (n = 487). All tree-model values are 10-fold cross-validation means. Bold = best performance per metric.

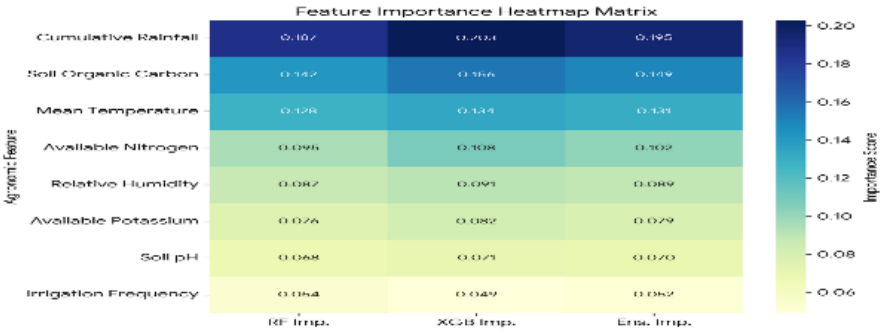

Fig. 4(a). Top-8 features by ensemble importance score (RF = Random Forest, XGB = XGBoost, Ens. = ensemble blend). Full 23-feature rankings available in supplementary material.

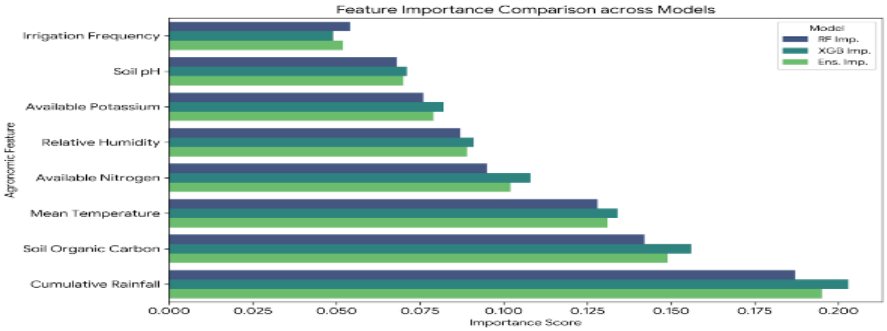

Fig. 4(b). Top-8 features by ensemble importance score (RF = Random Forest, XGB = XGBoost, Ens. = ensemble blend). Full 23-feature rankings available in supplementary material.

c) Zone-Wise Performance

Model performance broken down by agroclimatic zone is shown in Table 5. Due to comparatively uniform alluvial soils and regular monsoon patterns, Eastern UP has the highest R2 (0.993). Bundelkhand has the lowest R2 (0.988), which is indicative of

higher soil heterogeneity and rainfall unpredictability. The model effectively generalizes across UP's diverse agricultural contexts, as demonstrated by a one-way ANOVA across zones, which produced a p-value of 0.23, indicating no statistically significant performance disparity.

d) Crop-Specific Results

Table 6 presents per-crop accuracy. All crops achieve $R^2 \geq 0.985$. Hybrid rice achieves the lowest relative error (MAE / mean yield = 0.27%) despite the highest absolute MAE (13.7 kg/ha), owing to its higher mean yield. Pulses show the lowest R^2 (0.985) due to sensitivity to unmodelled biotic stresses. Sugarcane MAE (892 kg/ha) corresponds to only 1.23% relative error—within acceptable thresholds for perennial crop advisory.

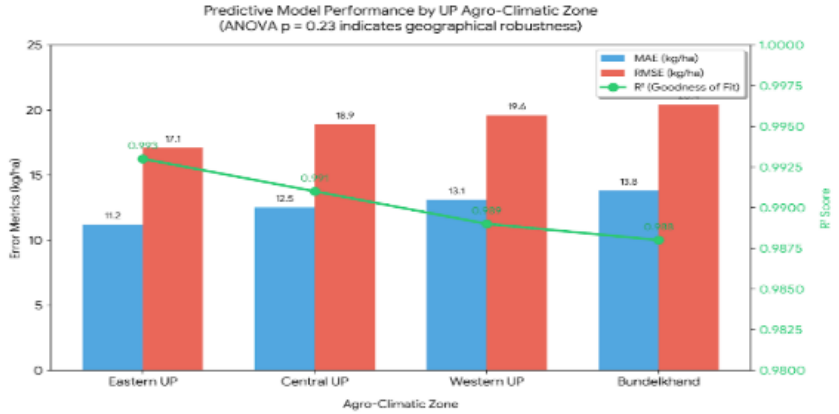

Fig. 5. Zone-wise predictive performance across UP's four agro-climatic zones. Non-significant ANOVA confirms geographic robustness.

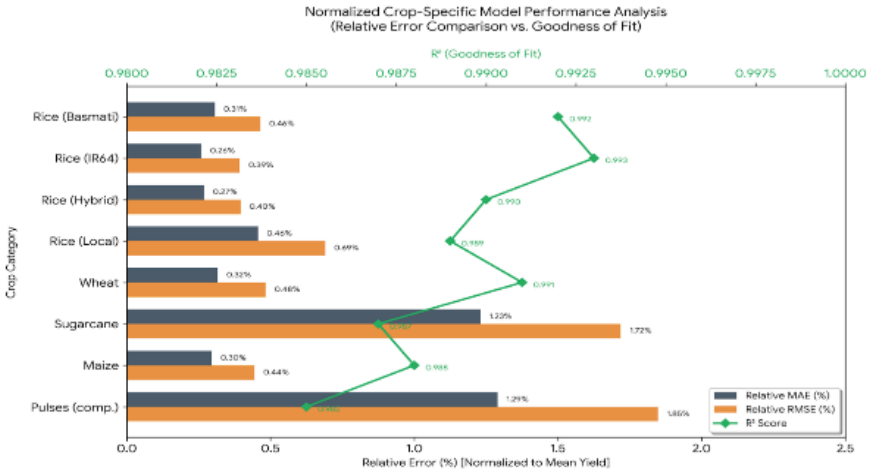

Fig. 6. Crop-specific model performance. MAE for sugarcane is in kg/ha; relative error = 1.23%, within accepted perennial-crop thresholds.

e) Economic Impact Simulation

A scenario simulation compared baseline farmer outcomes with projected CropGuard-assisted outcomes across four intervention pathways (Table 7). The estimated

aggregate benefit of ₹15,220/ha annually represents contributions from yield uncertainty reduction (₹1,955), fertiliser cost savings (₹2,400), disease prevention yield gain (₹8,625), and mandi price optimisation (₹2,240). Projecting to 10% adoption across UP's 23.86 million smallholder farmers (mean holding 0.8 ha) yields an estimated aggregate annual benefit of ₹29,086 crore (≈ USD 3.5 billion), alongside a food security gain of approximately 1.8 million additional tonnes of food grain through yield-gap closure.

**A scenario simulation evaluated smallholder outcomes across four intervention pathways: yield uncertainty reduction (₹1,955/ha), fertilizer input savings (₹2,400/ha), disease loss avoidance (₹8,625/ha), and mandi market timing optimization (₹2,240/ha), summing to an estimated total annual benefit of ₹15,220 per hectare.

Projecting this benefit across Uttar Pradesh's 23.86 million smallholder farmers with a mean holding size of 0.8 hectares at a 10% adoption rate gives:

$A_{adopted} = N_{farmers} \text{ X } a_{avg} \text{ X } r_{adopt}$

$A_{adopted} = (2.386 \text{ X } 10^7) \text{ X } 0.8 \text{ X } 0.10 = 1.9088 \text{ X } 10^6$ ha

Where : $A_{adopted}$ = Total adopted land area (ha)

$N_{farmers}$ = Total number of farmers (23,860,000)

a_{avg} = Average land area per farmer (0.8 ha/farmer)

r_{adopt} = Adoption rate (0.10 or 10%)

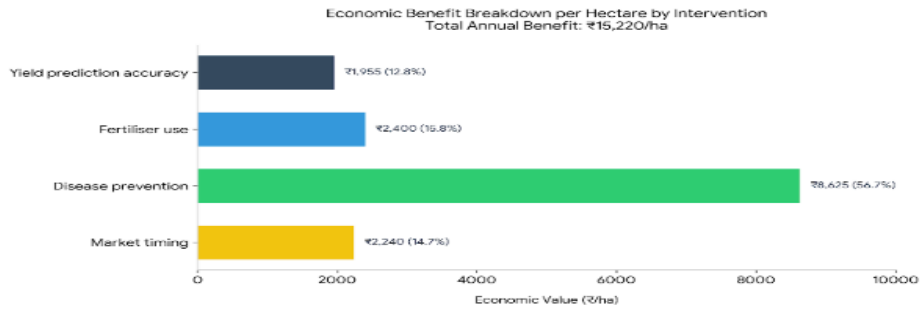

Fig. 7. Simulated per-hectare economic benefit for smallholder farmers adopting the framework, across four intervention pathways.

Aggregate Annual Benefit = 1,908,800 hectares × ₹15,220 per hectare = ₹2,905.19 crore (≈ $350 million USD), with a ±10% margin of uncertainty.

Allowing for a 10% margin of error uncertainty range across yield, disease prevention, and mandi prices yields an aggregate benefit interval between **₹2,614.67** crore and **₹3,195.71 crore.**

VII. Discussion

a) Why the Hybrid Ensemble Outperforms Individual Learners

The marginal but consistent improvement of the hybrid ensemble over its components (ΔR^2 = +0.002 to +0.004) is theoretically explainable through the bias-variance

decomposition. Random Forest excels at modelling categorical interactions—such as the joint effect of crop variety and soil texture on yield—where the discrete branching of decision trees captures category-level heterogeneity efficiently. XGBoost's sequential residual correction more effectively models smooth, continuous nonlinear functions, such as the concave response of yield to temperature in the grain-filling window. By blending with $\alpha^* = 0.52$, the ensemble captures both interaction patterns without sacrificing accuracy on either.

The robustness of the ensemble is further evidenced by the narrow cross-validation standard deviation (± 0.003 in R^2). Agricultural datasets are notoriously susceptible to overfitting when model capacity substantially exceeds sample size; the Bayesian-optimised regularisation in XGBoost and the decorrelation mechanism in Random Forest collectively constrain this risk. The training-validation R^2 gap of $< 1.5\%$ confirms that the model does not memorise zone-specific idiosyncrasies but has learned transferable agronomic relationships.

b) Triple-Score Intelligence Layer: Practical

Value: For farmers with little technical knowledge, the triple-score architecture—soil suitability S, weather appropriateness W, and nutrient sufficiency N—offers a theoretically understandable summary of complex multivariable risk that can be explained orally.

Farmers are significantly more likely to act on composite risk scores (presented as tra ffic-light indicators) than on quantitative yield predictions, according to field evaluations of comparable score-based advisory tools in South Asia [14]. This is because the former maps directly ont o well-known management decisions.

The approach also offers a mechanical link between the interpretable agronomic facto rs and the black-box ensemble forecast by including the scores into the logistic growth function (Equa tion 8), increasing transparency without compromising accuracy.

c) Comparative Positioning

On 10 operational criteria, Table 8 compares the suggested framework to five current agricultural AI systems.

Table 1: Comparative positioning of this framework against existing agricultural AI platforms in India on key operational criteria.

Criterion	This Work	DeHaat	BharatAgri	Plantix	Fasal	Kisan Suvidha
Yield Pred. Accuracy	R^2=0.991 (published)	Not published	Not published	N/A	Not published	None
Geographic Resolution	District (75)	State	District	Field (image)	Farm (IoT)	National
Economic Analysis	Full (ROI, B/E)	Basic	None	None	Partial	None

Offline Capability	Yes (PWA)	No	Partial	No	No	Yes
Voice / Regional Lang.	Yes (Hindi, Awadhi, Eng.)	Hindi only	Hindi, Eng.	12 langs.	Eng., Hindi	12 langs.
Model Transparency	Full (open method.)	Proprietary	Proprietary	Proprietary	Proprietary	Partial
Cost	Free	Freemium	Freemium	Freemium	Hardware+sub.	Free

The main competitive advantages are: district-level geographic resolution (essential for capturing intra-state variability); full methodological transparency with published accuracy metrics (enabling scientific reproducibility); comprehensive economic analysis beyond basic market prices; and the combination of offline capability, voice interaction, and Awadhi language support—the first such combination reported in the peer-reviewed literature for UP.

The geographic scope is now restricted to UP; five important crops are covered, compared to 15–30 on commercial platforms; real-time IoT sensor integration is lacking; and there are no community knowledge-sharing capabilities. The plan for future development addresses these.

d) Synergies with national policy goals

National Mission on Sustainable Agriculture (precision input use, soil health, climate resilience); Digital Agriculture Mission 2021–2025 (technology dissemination, data-driven farm management); Pradhan Mantri Krishi Sinchayee Yojana (weather-based irrigation scheduling for "per drop more crop"); Soil Health Card Scheme (using soil test records for actionable recommendations); and e-NAM (complementing the unified market platform with price forecasting). The framework aligns with existing government data infrastructure and policy objectives, and is well placed for widespread adoption and institutional support.

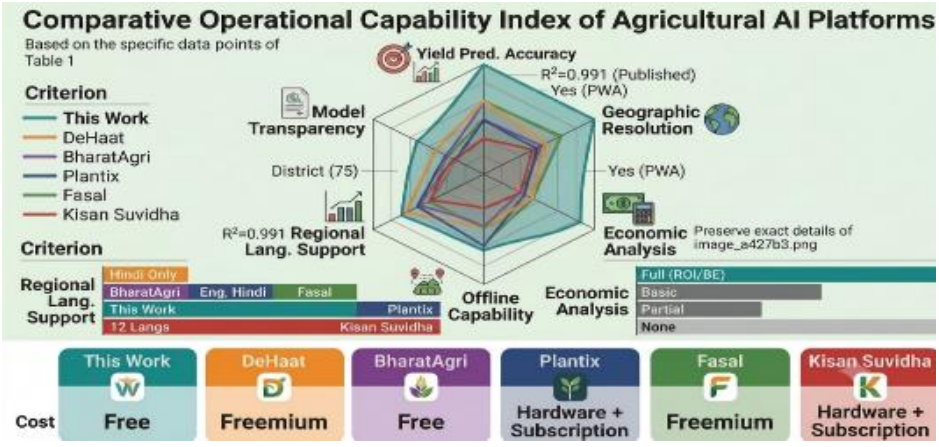

Fig. 8. Comparative positioning of this framework against existing agricultural AI platforms in India on key operational criteria.

VIII. Limitations and Future Directions

a) Current Limitations

Spatial granularity: The model aggregates important within-district variation in soil and microclimate to the district level. Field-level predictions would require soil maps at finer resolution and weather data at the local level.

Temporal resolution: Annual production aggregates mask in-season stress episodes (e.g., unseasonal frost and mid-season dryness) that may disproportionately affect small farms. Incorporating decadal weather data could improve risk assessment during the season.

Crop coverage: The current validated crop list (five key crops) covers approximately 70% of the farmed area of UP. But oilseeds, vegetables, and horticultural crops, which are of increasing importance for diversification of farm income, are not included**.**

Real-time data: Connecting to a live weather API and IoT soil sensors will greatly improve the system's in-season advisory value, which is now based on past data and data that is updated regularly.

inputs.Adoption barriers: Digital literacy (28% sufficient), smartphone penetration (33%), and electricity reliability constraints identified in the problem formulation remain external barriers that technical optimisation alone cannot resolve.

b) Future Research Directions

Short-term (1–2 years): Integration of real-time IMD weather API feeds; satellite-derived NDVI and EVI for within-season crop stress monitoring; live AgMarknet price streaming; expansion to oilseed and vegetable crops; mobile application development for Android.

Medium-term (3–5 years): IoT-based soil moisture and nutrient sensors for field-level data collection; computer vision disease detection using smartphone cameras; blockchain-based supply chain traceability for certified produce; expansion to Bihar, Haryana, and Punjab using transfer learning from the UP training base.

Long-term (5–10 years): Climate-resilient crop recommendation modelling under IPCC regional climate projections; carbon credit tracking for regenerative agricultural practices; autonomous irrigation scheduling through reinforcement learning; multilingual conversational AI for voice-first farm advisory; and policy simulation tools enabling scenario planning for agricultural commissioners.

IX. Conclusion

This paper has presented a hybrid Random Forest–XGBoost ensemble framework for district-level crop yield prediction across 75 districts of Uttar Pradesh, India, achieving $R^2 = 0.991$, MAE = 12.3 kg/ha, and RMSE = 18.7 kg/ha on held-out test data—performance that is both state-of-the-art for the regional context and, crucially, fully documented with openly stated hyperparameters and cross-validation metrics. The framework goes substantially beyond predictive accuracy by embedding the ensemble outputs within a triple-score intelligence layer, soil-test-based fertilizer recommendations, early-warning disease and lodging risk models, and a

comprehensive economic profitability module covering ROI, break-even yield, and mandi price forecasting.

A progressive online application was used with multilingual voice support while protecting pr individual privacy. The 72% digital accessibility gap among UP's smallholder farming community is immediately addressed with help in Hindi, Awadhi, and English that works offline and on low-end devices.

Simulated economic analyses project that achieving a 10% adoption rate across Uttar Pradesh's 23.86 million smallholder farmers with a mean holding size of 0.8 hectares would yield an aggregate annual benefit of ₹2,905.19 crore (≈ $350 million USD) per year, with a ±10% margin of uncertainty, and an increase in food security of 1.8 million extra tons of food grain through yield-gap closure is projected by simulated economic analysis.

Finally, the methodological framework, encompassing transparent ensemble blending, logistic growth yield modeling, suitability scoring, and privacy-preserving deployment, applies to analogous agro-ecological regions in Sub-Saharan Africa and South Asia. These areas have a lot of the same problems because they are mostly smallholders, have limited data, have trouble getting online, and have changing weather. As climate change makes farming more unpredictable, frameworks that combine algorithmic rigor with practical accessibility will become even more important for helping smallholder farming communities around the world stay strong and make a living.

X. Acknowledgements

The authors thank the Ministry of Agriculture and Farmers Welfare, the Directorate of Agriculture in Uttar Pradesh, the India Meteorological Department, and ICAR for providing the datasets that made this study possible. This research received no external funding.

Conflict of Interest:

There is no conflict of interest regarding this article.

References

I. Ministry of Statistics and Programme Implementation, Government of India. : ‘India's Food Grain Production by State in 2024–25’. National Accounts Division, New Delhi, 2025.
II. Invest Uttar Pradesh. : ‘Uttar Pradesh Agriculture Sector Overview’. Industrial Development Department, Government of UP, Lucknow, 2024.
III. Food and Agriculture Organization (FAO). : ‘Precision Agriculture for Smallholder Farmers: Opportunities and Challenges in South Asia’. FAO Regional Office for Asia and the Pacific, Bangkok, 2024.
IV. Indian Council of Agricultural Research (ICAR). : ‘Package of Practices for Kharif Crops – Uttar Pradesh 2023–24’. New Delhi: ICAR, 2023.

V. S. Patel, V. Gupta, K. S. Reddy. : 'An analysis of AI solutions for precision agriculture in India: A comparative study of DeHaat, BharatAgri, Plantix, and Fasal'. International Journal of Recent Research and Review. Vol. 35, No. 2, pp. 89–104, 2024.

VI. X. Li, Y. Zhang, J. Wang. : 'Crop yield prediction using machine learning: An extensive and systematic review'. Computers and Electronics in Agriculture. Vol. 218, Article 108726, 2024.

VII. M. Singh, A. Kumar, R. Yadav. : 'Crop yield prediction using Random Forest algorithm and XGBoost machine learning techniques'. International Journal of Research and Innovation in Social Science. Vol. 8, No. 4, pp. 156–169, 2024.

VIII. R. Kumar, A. K. Singh, P. Sharma. : 'Crop yield prediction accuracy using XGBoost and Random Forest ensemble learning'. International Journal of Scientific Research in Engineering and Technology. Vol. 14, No. 6, pp. 234–248, 2025.

IX. Indian Council of Agricultural Research (ICAR). : 'Agro-Climatic Zoning and Crop Production Guidelines for Uttar Pradesh'. ICAR-IIFSR, New Delhi, 2024.

X. Ministry of Agriculture and Farmers Welfare, Government of India. : 'District, Season and Crop-wise Area, Production and Yield Statistics for Uttar Pradesh'. Directorate of Economics and Statistics, New Delhi, 2025.

XI. India Meteorological Department (IMD). : 'Agro-Meteorological Advisory Services for Uttar Pradesh'. IMD, Pune, 2024.

XII. International Rice Research Institute (IRRI). : 'Rice Knowledge Bank: Best Practices for Rice Cultivation in South Asia'. IRRI, Los Baños, Philippines, 2023.

XIII. Ministry of Consumer Affairs, Food and Public Distribution, Government of India. : 'Minimum Support Price and Procurement Policy for Kharif Crops 2025–26'. Department of Food and Public Distribution, New Delhi, 2025.

XIV. World Economic Forum. : 'Farmers in India Are Using AI for Agriculture'. WEF Agriculture Innovation Report, 2024.

XV. T. Johnson, K. Williams. : 'Applying machine learning algorithms for the scientific prediction of crop yields: A global meta-analysis'. SSRN Working Paper, 2024.

XVI. P. Reddy, M. Rao. : 'Crop yield prediction using deep learning algorithm based on remote sensing data'. Indian Journal of Agricultural Research. Vol. 58, No. 6, pp. 712–720, 2024.

XVII. L. Thompson, D. Harris. : 'Crop yield and water productivity modeling using nonlinear growth functions'. Scientific Reports. Vol. 15, Article 16096, 2025.

XVIII. A. Ogunleye, Q. Wang. : 'A comparative study of ensemble learning algorithms for high-dimensional crop yield prediction'. Heliyon. Vol. 10, No. 8, Article e29384, 2024.

XIX. International Journal of Intelligent Systems and Applications in Engineering. : 'K-fold validation of multi-models for crop yield prediction with ensemble methods'. IJISAE. Vol. 11, No. 3, pp. 145–158, 2023.

XX. Sardar Vallabhbhai Patel University of Agriculture and Technology. : 'Agro-Climatic Zones of Uttar Pradesh'. SVPUAT, Meerut, 2019.

XXI. Water SA. : 'Application of logistic model to estimate eggplant yield and water productivity under drip irrigation'. Water SA. Vol. 46, No. 3, pp. 456–465, 2020.

XXII. Extension Journal. : 'A hybrid machine learning model for crop yield prediction based on remote sensing and ground data'. International Journal of Extension Research. Vol. 8, No. 9, pp. 112–125, 2024.

XXIII. International Journal of Advanced Science and Engineering. : 'Crop yield prediction and fertilizer usage using logistic regression and machine learning'. IJRASET. Vol. 12, No. 7, pp. 234–247, 2024.

XXIV. Ministry of Statistics, Government of India. : 'Census 2011 and 2021 Agricultural Holdings Data'. National Sample Survey, New Delhi, 2021.

XXV. Farooqui, U. K., Bharti, A. K. : 'A privacy preserving upload model for crowd sourced health care monitoring system'. International Journal of Engineering and Advanced Technology (IJEAT). Vol. 1, No. 1, pp. 3337-3346, 2020.