A Soil-Aware Hybrid AI Model for Precision Crop Recommendation and Yield Forecasting

Shilpa Mangesh Pande
Department of Computer Science and Engineering (VTU-RC), CMR Institute of Technology, Bengaluru, affiliated to Visvesvaraya Technological University, Belagavi, India
shilpa.v.deshpande@gmail.com (corresponding author)

Prem Kumar Ramesh
Department of Computer Science and Engineering (VTU-RC), CMR Institute of Technology, Bengaluru, affiliated to Visvesvaraya Technological University, Belagavi, India
premkumarramesh@gmail.com

Received: 31 January 2026 | Revised: 17 March 2026, 31 March 2026, and 12 April 2026 | Accepted: 17 April 2026

 DOI: https://doi.org/10.48084/etasr.17876

ABSTRACT

Matching crops to their specific soil type remains a major challenge for Indian farmers. This study proposes a model that uses crop yield, soil, and climatic data to guide optimal crop selection, improving productivity and lowering errors. Relevant environmental factors are identified through feature engineering to enhance yield prediction accuracy using the Sequential Forward Feature Selection Algorithm (SFFSA), Random Forest Variable Importance Algorithm (RFVIA), and Sequential Back Elimination Feature Selection Algorithm (SBEFSA). This study predicts the best crop to grow given a specific set of environmental conditions by employing a Machine Learning (ML) approach. A unified AI-enabled framework is created to help Indian farmers select the optimal crop to cultivate and project the Crop Yield (CY) using a hybrid system that combines Artificial Neural Network (ANN) and Multiple Linear Regression (MLR). This proposed approach employs a residual learning framework that separates linear agronomic effects from nonlinear interactions. ANN performs hyperparameter sensitivity analysis to assess the impact of learning rate and hidden neuron count to validate the performance. The results indicate that the Hybrid algorithm approach outperforms standard ML algorithms, with improvements up to 2.2-fold in RMSE and around 1.5-fold in MAE. The proposed work even predicts the best-suited fertilizer to grow the crops based on soil contents.

Keywords-hybrid algorithm; crop yield; crop recommendation; Multiple Linear Regression (MLR); Artificial Neural Network (ANN); feature selection; Machine Learning (ML)

I. INTRODUCTION

In Indian agriculture, traditionally, crop selection has often relied on inherited farming practices rather than a data-driven assessment of critical environmental and soil variables. This mismatch between traditional practices and current environmental conditions leads to sub-optimal crop choices, directly impacting a farmer's financial viability. To mitigate this risk, there is a need for modern decision support systems to provide an estimated Crop Yield (CY) that allows farmers to project expected average income and ensure that their output meets market requirements. Crop productivity is also governed by numerous factors, including vital nutrients, such as Phosphorus (P), Potassium (K), and Nitrogen (N), playing an indispensable role in plant growth and yield maximization. Currently, crop recommendation and yield prediction are two problems that are quite different in terms of problem formulation, modeling, and solution; they also have significant overlap in terms of data. Existing research aims to address one of these two in depth. However, there are no solutions that take advantage of the commonalities to address them both under one umbrella.

Addressing the limitations of traditional crop selection, initial research utilized Internet of Things (IoT) frameworks combined with Machine Learning (ML) to provide appropriate crop recommendations based on soil types and parameters. For example, an enhanced distribution-oriented Chicken Swarm Optimization (CSO) algorithm coupled with a weighted LSTM to forecast crop suggestions is proposed in [1].

However, the rapid influx of environmental data presents challenges regarding model complexity and computational efficiency. Including non-contributory features in a prediction model increases its time and space complexity without a proportional gain in accuracy. Multiple investigations have aimed to optimize the feature set. Authors in [2] highlighted the necessity of ML algorithms for choosing optimal features that directly influence crop yield, contrasting it with the pitfalls of including all raw data features. Authors in [3, 4] emphasized

the implementation of feature engineering algorithms, such as the hybrid attribute selection approach with FMIG-RFE and correlation-based filtering, to optimize attribute selection for yield prediction. These approaches focus on improving predictive accuracy without giving much importance to model stability and generalization.

Beyond basic classification, research has evolved toward robust and high-accuracy predictive frameworks. Authors in [5] utilized an ensemble model to boost both crop productivity and model accuracy, aiding farmers in choosing appropriate seeds and crop cultivation based on soil needs and projected profit. Similarly, authors in [6] developed a classifier considering soil fertility factors such as pH, K, P, N, along with rainfall, humidity, and temperature using Multilayer Perceptron, Decision Trees (DT), and JRip. The JRip model attained an accuracy of 98.22% and a Weighted Average Receiver Operating Characteristic of 1. The higher accuracy reported here stems from simplified validation settings, while the proposed approach is prone to overfitting and sensitive to data variance.

Recent hybrid ensemble efforts include the Deep Learning (DL) architecture. Authors in [7] introduced a Concurrent Excited Gated Recurrent Unit (CEGRU) architecture, fine-tuned with Hunter-Prey-Optimization (HPO), to process complex non-linear relationships in temporal and environmental data. Authors in [8] proposed an RFO-driven ensemble Recurrent Neural Network (RNN) model that synthesizes LSTM, BiLSTM, and GRU to address the dual goals of crop recommendation and yield forecasting by capturing both forward and backward temporal dependencies. Authors in [9] used a three-tiered stacking ensemble technique with various base learners such as DT, Random Forest (RF), and Support Vector Machine (SVM) to refine predictions through a meta-layer. Although CEGRU and RNN-based ensemble models achieved high accuracy, this raises the risk of overfitting when used with agricultural data sets of small size.

The efficacy of prediction models is heavily influenced by contextual factors and the explainability of the model's output. Studies emphasize the importance of specific agricultural practices. Authors in [10] constructed the Vegsyst model v3 to propose specific nutrient additions (Ca, P, Mg, K) for crops cultivated in Spain, while authors in [11] demonstrated that optimizing crop rotation planning in organic farming can enhance soil quality and increase yield by 20%. Authors in [12] analyzed the effect of rainfall during El Niño years, demonstrating an improved CY trend and providing a framework for policymakers to develop contingency plans. Given the opaqueness of many ML models, research has focused on transparency. Authors in [13] leveraged Explainable AI (XAI)-CROP with LIME to provide comprehensible DT-based recommendations, and authors in [14] further stressed user-centric interpretability through a Streamlit interface integrating LIME, SHAP, and DiceML visualizations.

While contemporary XAI tools enhance model transparency through subsequent explanations, they do not fundamentally incorporate domain knowledge into model architecture. Consequently, the subsequent explanations may get detached from underlying agronomic principles, reducing the model's effectiveness.

Further extending the scope, research has explored related predictive domains. Hybrid models combining CNN and BiLSTM have been developed to predict crop prices [15], while hybrid models address the demand-supply chain for multiple crop commodities [16]. Authors in [17], focused on site-specific nitrogen recommendations for maize based on soil contents. Models combining YOLO and CNN-LSTM have been proposed to suggest proper harvesting timing [18] and CY [19, 20].

From the literature, it is evident that existing approaches mainly focus on either high-accuracy crop recommendation using ensemble and DL models or interpretable statistical models with predictive capacity. Existing hybrid ML approaches for agricultural prediction combine models employing ensemble voting, stacking, or sequential frameworks. However, these methods often do not explicitly model the residual errors produced by simpler models. The present study proposes a hybrid residual-learning framework combining Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) for CY prediction.

The MLR model detects the primary linear relationships between environmental variables and crop yield, while the ANN learns the residual errors of the MLR model to identify nonlinear patterns. This residual-learning strategy improves forecasting accuracy while maintaining interpretability, distinguishing the proposed approach from conventional hybrid architectures. However, a unified framework that simultaneously (i) performs agronomically grounded feature selection, (ii) provides interpretable recommendations for crops and fertilizers, and (iii) delivers robust yield prediction with controlled model complexity remains underexplored. In particular, the integration of linear agronomic relationships with nonlinear residual learning has not been sufficiently investigated in the context of precision agriculture.

The novelty and primary contributions of the study are:

- Identifying the optimal feature set for forecasting the most suitable crop to cultivate.
- Recommending the optimal crop based on localized soil properties, temperature, and rainfall data.
- Proposing the most appropriate fertilizer by analyzing the current N, P, and K soil component content.
- Employing both stand-alone ML and an AI-based hybrid modeling method to accurately forecast the final CY based on the recommendation.

II. METHODOLOGY

The proposed system is an integrated decision support framework for agriculture, designed to optimize crop selection, fertilizer recommendation, and yield forecasting. The core objective is to improve agricultural output by offering farmers information-driven advice based on soil and climatic inputs. The methodology involves data acquisition, rigorous reprocessing, optimal feature subset selection, as well as the

implementation of ML and the integrated hybrid framework for prediction. The novelty of this work lies not in proposing a new learning algorithm, but in the structured integration of domain-driven feature selection, interpretable linear modeling, and non-linear refinement through ANN within a single decision-support pipeline.

A. Datasets and Preprocessing

The data employed in this study are based on records acquired from data.gov.in [21-23]. The dataset utilized includes 2500 instances gathered between 2019 and 2023. The data included district-level agricultural records related to the state of Karnataka, India. The information focuses on important crops grown in the region, such as paddy, ragi, jowar, bajra, and maize. The dataset instances are balanced across all crop classes, ensuring adequate representation of each crop class for analysis and model development.

The additional processed dataset includes rainfall and climate statistics, soil health card records, and crop production reports, as listed in Table I. To avoid any overfitting that may arise, K-fold cross-validation was used. As no single dataset contained all the required attributes, extensive preprocessing was performed involving data cleaning, removal of irrelevant attributes, normalization, and integration of datasets to construct a unified dataset for analysis. Data preprocessing is essential to ensure high data quality and enhance model performance. Missing values were dealt with using statistically accepted imputation techniques. Continuous soil nutrient attributes (N, P, K) and rainfall were assigned with median values to reduce sensitivity to skewed distributions, while temperature attributes were computed using mean values. Categorical variables, such as soil type, were transformed into a numerical format using One-Hot Encoding and were imputed using the mode. Given the varying scales and units of measurement (e.g., rainfall vs. pH), the data were transformed using normalization methods to guarantee that all features contribute equally during model training, preventing features with large values from dominating the learning process.

B. System Architecture

The system architecture, illustrated in Figure 1, details the end-to-end flow of the unified framework. It begins with the collection of environmental and soil factors, followed by data preprocessing, training of ML algorithms, model validation, and deployment through a Flask connector to a user interface. This structure facilitates iterative model improvement via potential feedback loops from user recommendations. The model has an AI-enabled decision support system for agricultural planning that incorporates optimal crop suggestions, fertilizer guidance, and precise yield forecasts based on soil and environmental characteristics.

TABLE I. DATASET ATTRIBUTES AND CHARACTERISTICS

Attribute name	Unit	Description / type
Rainfall	mm	Total precipitation during the period
N	kg/hectare	Soil content of Nitrogen
P	kg/hectare	Soil content of Phosphorus
K	kg/hectare	Soil content of Potassium
Soil pH level (pH)	-	Soil acidity or alkalinity
Seed quantity	kg/hectare	Amount of seed used
Soil type	Categorical	Classification of soil (e.g., Black, Alluvial)
Minimum temperature (T_{min})	°C	Lowest temperature observed
Maximum temperature (T_{max})	°C	Highest temperature observed
Average temperature (T_{avg})	°C	Average temperature observed
Final yield	ton/hectare	Dependent Variable (target output)

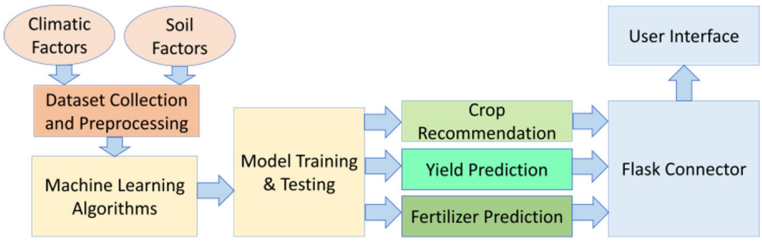

Fig. 1. System architecture.

C. Feature Engineering and Selection

To improve model accuracy, reduce computational overhead, and enhance interpretability, optimal features were selected from the raw dataset. Three established feature engineering algorithms were employed to identify the most significant subset of attributes influencing crop yield.

1) Sequential Forward Feature Selection Algorithm (SFFSA)

This iterative algorithm starts with an empty set and incrementally adds the feature that yields the greatest improvement in model performance, using the Akaike Information Criterion (AIC) as the primary stopping metric. SFFSA identified Rainfall, N, P, K, and pH as the optimal subset.

2) Sequential Back Elimination Feature Selection Algorithm (SBEFSA)

This method begins with the full feature set and iteratively removes the least significant feature until the best subset is

obtained, again guided by the AIC. SBEFSA also converged on the set of Rainfall, N, P, K, and pH.

3) Random Forest Variable Importance Algorithm (RFVIA)

This ensemble learning approach assesses feature significance based on the reduction in node impurity across all trees. RFVIA selected a slightly larger subset: Rainfall, N, P, K, pH, SD, and T_{max} (MxT). Based on observations, the set of Rainfall, N, P, K, and pH are the top five predictors.

D. Forecasting Algorithms

The system employs a dual prediction strategy. First, classification using multiple standard ML techniques (SVM, XGBoost, Naïve Bayes, DT, RF, and LR) for crop recommendation is performed, and second, regression using a novel method combining MLR and ANN models for CY forecasting for the proposed crop (Paddy, in this study). DT employs a tree-based model that separates the total instances based on feature values, whereas RF results in multiple DTs from random sample selection to combine their forecasts. Naïve Bayes classifier uses probabilistic models that presume input attribute independence, whereas LR makes use of logistic regression. XGB utilizes a boosting approach to construct DTs to minimize forecasting errors. SVR translates information into a high-dimensional space with the help of a kernel function [24, 25]. Agronomic variables such as rainfall, soil pH, and macronutrient concentrations exhibit both linear and nonlinear relationships with crop yield. MLR effectively captures dominant linear trends and provides coefficient-level interpretability aligned with agronomic knowledge. However, residual nonlinear interactions, arising from complex soil-climate interactions, are inadequately modeled by linear techniques alone.

The proposed Hybrid MLR-ANN framework leverages MLR for interpretable baseline modeling, while the ANN learns nonlinear residual patterns, thereby improving prediction accuracy without significantly increasing model complexity. This design ensures a balance between interpretability, robustness, and predictive performance. Unlike purely ensemble-based or deep recurrent models, the proposed Hybrid MLR-ANN architecture emphasizes interpretability and robustness over architectural complexity, making it better suited for medium-sized agricultural datasets and decision-support applications.

1) Multiple Linear Regression (MLR)

MLR is used as a foundational model to capture the linear relationship between the forecasted variable (crop yield) and the predictors. The CY y_i is formulated as a linear function of the selected features:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \quad (1)$$

$$y_i = \beta_0 + \sum_{i=1}^{p} \beta_i \, x_i + \varepsilon \quad (2)$$

where y_i is the predicted crop yield, x_i are the regressor (Rainfall, N, P, K, pH), β_0 is the intercept (bias term), the slope is represented by β_i , and the error is indicated by ε. The computed coefficients and bias are listed in Table II. These coefficients offer crucial insights into the linear relationship between the explanatory variables and the target (crop yield):

- Dominant predictor: The Rainfall coefficient (0.8142) is the largest, confirming that precipitation is the most influential factor in determining CY within the scope of the model.
- Nutrient hierarchy: Among the macronutrients, K exhibits the highest positive impact, followed by N and P. This quantitative ranking is essential for optimizing fertilizer application strategies.
- Inverse relationship: The slight negative coefficient for pH indicates that as soil acidity decreases (pH increases), the model predicts a decrease in yield, although its magnitude is minimal compared to Rainfall or K.
- Model baseline: The Bias serves as the baseline yield predicted when all independent variables are hypothetically zero, establishing the minimum output of the linear model.

TABLE II. MLR ALGORITHM COEFFICIENTS AND VARIABLE DESCRIPTIONS

Parameters	Values	Description
Bias	0.0181	The intercept term (β_0) represents the predicted yield when all other factors are zero.
N	0.0312	Coefficient for the Nitrogen content (kg/hectare) in the soil.
K	0.1782	Coefficient for the Potassium content (kg/hectare) in the soil.
P	0.0051	Coefficient for the Phosphorus content (kg/hectare) in the soil.
Rainfall	0.8142	Coefficient for the Rainfall amount (mm). This shows the largest positive impact on yield.
pH	-0.007	Coefficient for the soil pH level (negative sign indicates a slight inverse relationship with yield).

2) Artificial Neural Network (ANN)

ANNs are employed for their capability to model intricate, non-linear relationships. Although neural networks typically benefit from large datasets, the proposed ANN employs a 5-1-1 architecture with limited trainable parameters, reducing data requirements and overfitting risks. Furthermore, extensive preprocessing, feature selection, and normalization were applied prior to model training.

The network topology consists of three layers: an input layer (5 neurons for the features), only one hidden layer, and an output layer (1 neuron for crop yield). The weighted sum input to the hidden layer neuron h is calculated as:

$$HiddenSum_h = \sum_{i=1}^{n} x_i^{in} \; w_i^{in} + b_i^{in} \quad (3)$$

where x_i^{in} are the input values, w_i^{in} are the corresponding weights, and b_i^{in} is the bias term. Furthermore, (4) gives the output of the hidden layer and is determined by applying a non-linear activation function to the weighted sum:

$$HiddenSum_{h_{out}} = function(HiddenSum_h) \quad (4)$$

In this model, yield is calculated using an activation function:

$$OSum_o = \sum_{i=1}^{n} x_i^h \, w_i^h + b_i^h \quad (5)$$

The structure of the designed ANN for yield prediction is illustrated in Figure 2.

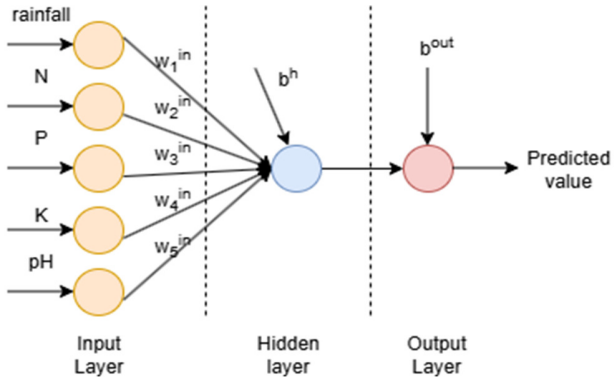

Fig. 2. Conceptual ANN architecture designed for yield prediction.

3) Hybrid MLR-ANN Algorithm

Unlike purely ensemble-based or deep recurrent hybrid models, which often rely on multiple black-box learners, the proposed approach adopts a residual-learning paradigm that explicitly separates linear agronomic effects from nonlinear interactions. This structural design enhances interpretability, reduces overfitting risk, and advances existing hybrid modeling approaches by prioritizing robustness and domain transparency alongside predictive performance.

The hybrid algorithm initializes weights and bias of the ANN using the coefficients obtained from the MLR model (Table II). This integration addresses the common challenge of random initialization in ANNs, providing a better starting point that captures the existing linear trend in the data and potentially accelerating convergence. The overall function of an ANN neuron with p inputs and weights w_j is given by:

$$y = \text{function}\left(\sum_{j=1}^{p} w_j \, x_j + w_0\right) \quad (6)$$

A hyperparameter sensitivity analysis was performed to evaluate the effect of learning rate and hidden neuron count on validation performance. The complete process is detailed in Algorithm 1. Specifically, the MLR coefficients (β_1 to β_m) are utilized to set the input-to-hidden layer weights (w_1 to w_m), and the MLR intercept (β_0) is assigned as the initial bias (w_0) for the ANN's input layer. This initialization is applied to the crop dataset defined as:

$$\text{Crop} = \{(x^{(m)}, y^{(m)}) \mid m = 1,2,3,\dots,M\} \quad (7)$$

where $y^{(m)}$ represents the CY and $x^{(m)} = \{x_i^{(m)}, \; i = 0,1,\dots,P\}$ is the selected feature vector. This methodology provides the basis for evaluating the proposed hybrid model.

Algorithm 1: Hybrid Algorithm for Crop Yield Forecasting
Input: The crop data D_{crop}
Output: Predicted Crop yield Y_{pred}

1. Read the crop data.
2. Clean and normalize the data D_{norm}.
3. Partition D_{norm} into training (D_{train}) and testing (D_{test}) subsets.
4. Implement the MLR algorithm on D_{train} to calculate the coefficients (β_i) and intercept (β_0).
   $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon$
5. Create ANN model and configure hyperparameters:
   Parameter 1: Input layer neurons
   Parameter 2: Hidden layer neuron
   Parameter 3: Output layer neuron
   Parameter 4: Learning rate, η
   Parameter 5: Epochs
6. Configure the input layer bias (w_0) using the MLR intercept (β_0). Set up the input layer weights (w_i) with the MLR coefficients (β_i) from Step 4.
7. Perform Feed Forward to calculate the hidden layer weighted sum:
   $\text{HiddenSum}_h = \sum_{i=1}^{n} x_i^{in} \, w_i^{in} + b_i^{in}$
8. Initialize the hidden layer weight and bias with small random values.
9. Calculate the output layer weighted sum.
   $OSum_o = f(\sum_{i=1}^{n} x_i^h \, w_i^h + b_i^h)$
10. Calculate the prediction error E
11. Perform Back Propagation (Chain Rule) to update weights:
    $\frac{\partial E}{\partial w_i} = \frac{d\,E}{d\,HiddenSum_h} \cdot \frac{d\,HiddesnSum_h}{d\,OSum_o} \cdot \frac{d\,OSum_o}{dw_i}$
12. Adjust input layer weights (w_i) and bias (w_0) using the calculated gradient and learning rate. Repeat Feed Forward (Steps 8–10) and Back Propagation (Steps 11–12) for all the epochs or until the error is minimized
13. Calculate the final yield prediction Y_{pred} on D_{test}.

III. RESULTS AND DISCUSSION

The outcomes and performance metrics of the feature selection and forecasting algorithms demonstrated the efficacy of the proposed hybrid approach for CY forecasting.

A. Feature Selection Analysis

Out of the ten initial crop features, three feature selection methods, viz., SFFSA, SBEFSA, and RFVIA, were employed to determine the optimal subset of attributes for crop recommendation and yield prediction. The consensus set of Rainfall, N, K, P, and soil pH level (pH) was identified as highly significant. Table III summarizes the features selected by each algorithm.

1) SFFSA and SBEFSA Metrics

Both SFFSA and SBEFSA utilize the AIC, which balances model goodness-of-fit against complexity, serving as a robust measure against overfitting.

- SFFSA: Starting from an empty set ($\{\emptyset\}$), features were added sequentially. The process stopped at the set $\{RF, N, P, K, pH, MxT\}$ because the AIC value reached its minimum (-3618.27), as presented in Table IV. The subsequent addition of *ST* resulted in an increased AIC (-3617.11), signaling a decrease in model parsimony and effectiveness.
- SBEFSA: Starting with all features, the algorithm eliminated the least significant features. The process stopped at the set$\{RF, N, P, K, pH, ST\}$, where the AIC was minimized (-3619.27), as shown in Table V. The removal of the next feature, *ST*, led to an increased AIC (-3618.34), thus identifying the optimal subset.

TABLE III. FEATURES SELECTED BY VARIOUS ALGORITHMS

Feature methods	Rainfall	N	P	K	pH	SD	Soil type	Min. temp	Max. temp	Avg. temp
SFFS	√	√	√	√	√		√		√	
SBEFS	√	√	√	√	√		√			
RFVI	√	√	√	√	√	√			√	√

TABLE IV. AIC VALUE FOR THE SFFSA SELECTION METHOD

Feature set	AIC	Stopping criteria
{}	-2645.33	↓
{RF}	-3567.45	↓
{RF, N}	-3598.32	↓
{RF, N, P}	-3613.62	↓
{RF, N, P, K}	-3615.21	↓
{RF, N, P, K, pH}	-3617.36	↓
{RF, N, P, K, pH, MxT}	-3618.27	Stop
{RF, N, P, K, pH, MxT, ST}	-3617.11	↑

TABLE V. AIC VALUE FOR SBEFSA SELECTION METHOD

Feature set	AIC	Stopping criteria
{RF, N, P, K, pH, SD, ST, MT, MxT, AT}	-3615.23	↓
{RF, N, P, K, pH, SD, ST, MT, MxT}	-3615.23	↓
{RF, N, P, K, pH, SD, ST, MxT}	-3617.12	↓
{RF, N, P, K, pH, SD, ST}	-3618.47	↓
{RF, N, P, K, pH, ST}	-3619.27	Stop
{RF, N, P, K, pH}	-3618.34	↑

2) RFVIA Metrics

- RFVIA assesses feature importance based on the average reduction in *Node Purity* (or Gini impurity/variance) across the forest. Table VI shows that Rainfall, N, P, K, and pH exhibit the highest purity reduction values, indicating their dominant influence on the prediction outcome. Features such as ST and MnT contributed minimally.

TABLE VI. NODE PURITY BASED ON RFVIA SELECTION METHOD

Feature	Node purity
Rainfall	0.068
N	0.066
P	0.067
K	0.065
pH	0.053
SD	0.008
ST	0.003
MnT	0.0052
MxT	0.0054
AvT	0.0053

B. Feature Selection Impact on Model Performance

Table VII compares the performance of the selected feature subsets across MLR (Algorithm 1) and ANN (Algorithm 2) models using MAE, RMSE, and Adjusted R^2. The RFVIA subset yielded the lowest RMSE for the MLR model (0.012), suggesting that its slightly larger set of features resulted in the most accurate linear fit. The Adjusted R^2 values, which represent the fraction of variance explained by the MLR model, were consistently high (≥0.85). Given that SFFSA and SBEFSA yielded minimal error rates and high R^2 values, and to maintain a parsimonious model, the set of five core features (Rainfall, N, P, K, pH) was finalized for subsequent yield prediction.

TABLE VII. MEASURES FOR FEATURE SELECTION

Measures	SFFS		SBEFS		RFVI	
	Algo. 1	Algo. 2	Algo. 1	Algo. 2	Algo. 1	Algo. 2
RMSE	0.014	0.087	0.014	0.087	0.012	0.93
MAE	0.008	0.071	0.008	0.071	0.009	0.062
Adj. R^2	0.86	NA	0.85	NA	0.85	NA

C. Crop Recommendation Performance

The system's first stage involves crop recommendation using various classification algorithms. Figure 3 illustrates the comparative performance of MLR vs ANN across the feature selection algorithms. It is observed that MLR consistently outperforms ANN in all scenarios. This inference helps in ruling out ANN-based crop recommendations.

Crop recommendation using standard classifiers (DT, RF, Naïve Bayes, XGBoost, SVR, and LR) is illustrated in Figure 4, which shows that the RF algorithm achieves the highest accuracy, 93.5%. This superior performance validates the use of ensemble tree-based methods for multi-class classification in agricultural contexts.

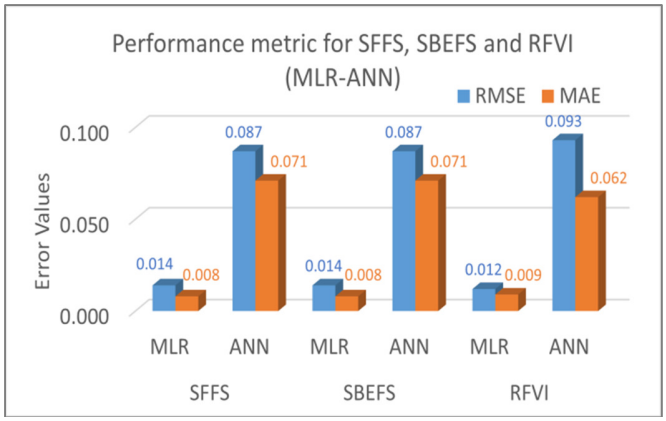

Fig. 3. Performance metrics for feature selection.

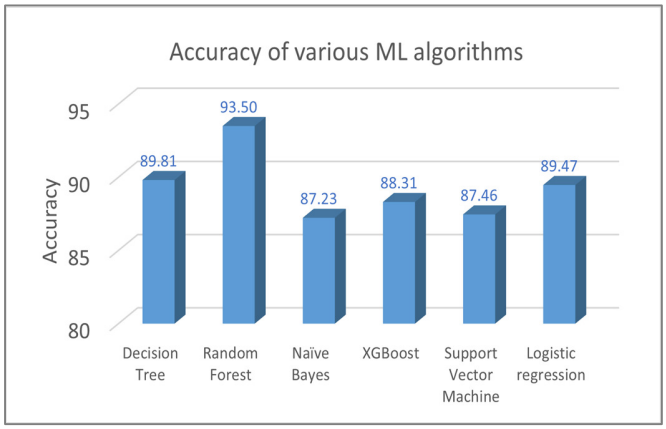

Fig. 4. Performance metrics for crop recommendation system.

D. Crop Yield Forecasting

Yield forecasting for the proposed crop (paddy) was conducted using KNN, RF, MLR, SVR, ANN, and the combined MLR-ANN model. The data were divided into 75% for training and 25% for testing. To ensure robustness and mitigate potential overfitting, yield prediction models were evaluated using 5-fold cross-validation. In each fold, the model was trained on 80% of the data and tested on the remaining 20%. Performance metrics were averaged across folds. This strategy reduces dependency on a single train-test split and provides a more reliable estimate of generalization performance.

1) MLR Model Interpretation

The MLR model provides an interpretable equation for CY based on the calculated coefficients, as presented in Table II:

$$CY = 0.018 + 0.03N + 0.1782K + 0.0051P + 0.8142R - 0.007pH \quad (8)$$

The coefficients reveal that rainfall *(R)* has the largest positive impact (0.8142) on crop yield, followed by Potassium *(K)* (0.1782). Conversely, soil *pH* exhibits a small negative correlation (-0.007). This quantification is crucial as it aligns with agricultural science, confirming the importance of adequate water and essential macronutrients (K, N, P) for crop production.

2) Hybrid MLR-ANN Performance

The Hybrid MLR-ANN model was developed by using the MLR coefficients (β_i) as the initial weights (w_i) and the MLR bias (β_0) as the starting bias (w_0) for the ANN's input layer. The optimal training configuration was determined as a 0.1 learning rate over 500 epochs. A hyperparameter sensitivity analysis was conducted to investigate the influence of the learning rate and hidden neuron count on validation performance. As illustrated in Table VIII, increasing network complexity beyond a single hidden layer with five neurons did not yield consistent RMSE improvements and led to higher variance, motivating the selection of a compact 5-1-1 architecture. To address concerns regarding model generalizability, overfitting, and reliance on a single train-test split, additional validation analyses were conducted.

TABLE VIII. SENSITIVITY ANALYSIS OF ANN HYPERPARAMETERS

Hidden neurons	Learning rate	Validation RMSE (mean ± std)
3	0.01	0.162 ± 0.018
3	0.05	0.121 ± 0.014
3	0.10	0.098 ± 0.011
5	0.01	0.158 ± 0.020
5	0.05	0.119 ± 0.016
5	0.10	0.095 ± 0.010
7	0.05	0.094 ± 0.026
7	0.10	0.093 ± 0.031

Figure 5 presents the fold-wise RMSE obtained using 5-fold cross-validation for paddy yield prediction. The consistency of RMSE values across folds indicates stable generalization performance and mitigates concerns associated with reliance on a single train-test split. The absence of large deviations across folds suggests that the proposed Hybrid MLR-ANN model does not suffer from overfitting. Figure 6 displays the relationship between actual and predicted yield values aggregated across all folds. The strong alignment of points along the diagonal reference line demonstrates high predictive accuracy and confirms that the hybrid model captures both linear and nonlinear relationships between agronomic features and yield.

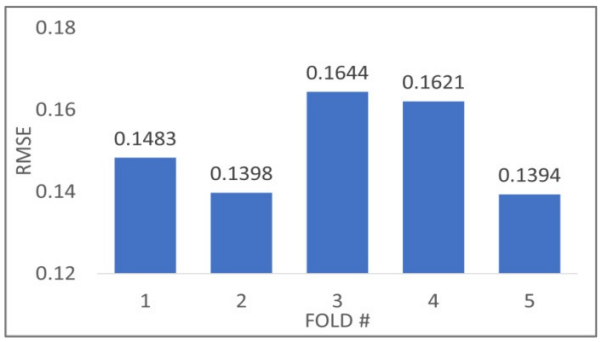

Fig. 5. Fold-wise RMSE distribution using 5-fold cross-validation.

Figure 7 shows the distribution of residual errors obtained from 5-fold cross-validation. The residuals are symmetrically distributed around zero with no visible skewness or heteroscedasticity, indicating that prediction errors are

randomly distributed and that no data leakage or systematic bias is present. Figure 8 presents the training and testing error curves of the Hybrid MLR-ANN model. The convergence of both curves with a small and stable generalization gap confirms that the model does not overfit and that the reported low RMSE values are not a consequence of data leakage. The performance metrics (RMSE, MAE, R-value) summarized in Table IX indicate that the Hybrid MLR-ANN model outperformed stand-alone MLR, KNN, SVR, RF, and ANN models. A 1.8 to 2.2-fold improvement was observed in RMSE, while around a 1.5-fold improvement was recorded in MAE.

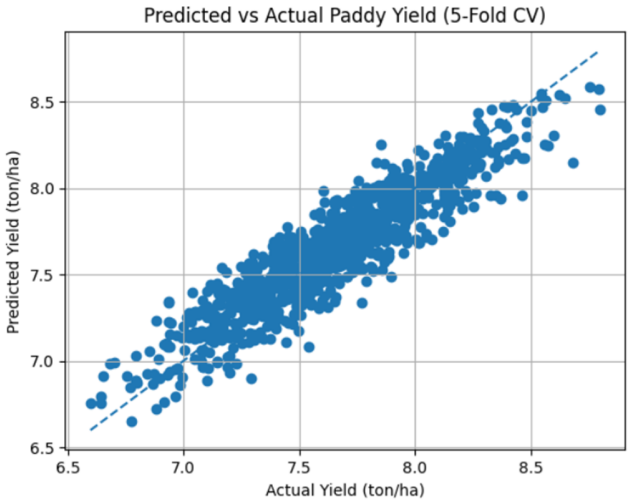

Fig. 6. Actual vs. predicted yield values across 5-fold cross-validation.

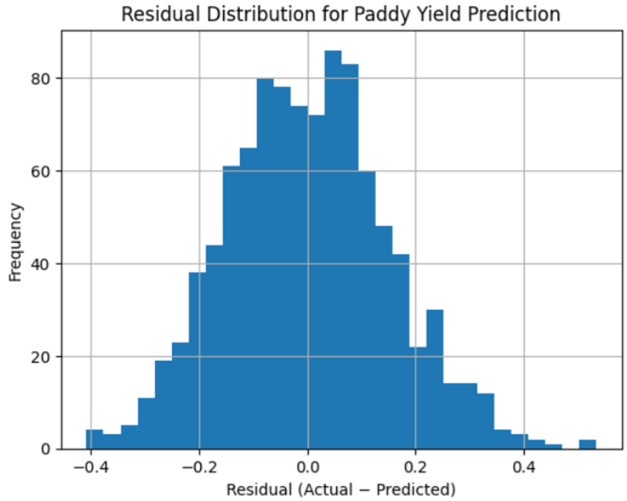

Fig. 7. Residual error distribution of yield prediction across cross-validation folds.

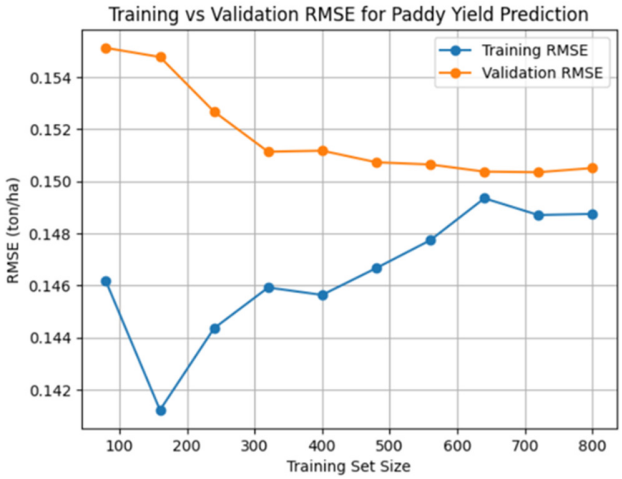

Fig. 8. Training vs testing error curve of Hybrid MLR-ANN.

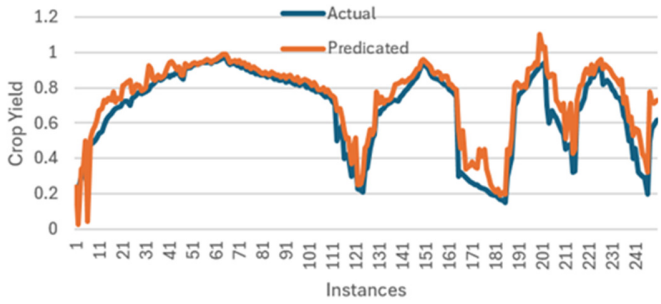

Fig. 9. Performance comparison of Hybrid MLR-ANN with other regression models.

TABLE IX. COMPARISON OF ALGORITHMS BASED ON PERFORMANCE METRICS

Metric	SVR	KNN	RF	MLR	ANN	Hybrid MLR-ANN
RMSE	0.092	0.115	0.083	0.095	0.094	0.052
MAE	0.065	0.078	0.078	0.061	0.064	0.042
R^2	0.91	0.81	0.81	0.88	0.91	0.92

Figure 9 demonstrates the superior fitting capability of the Hybrid MLR-ANN model. The integration effectively uses the MLR's ability to model linear trends as an informed initialization for the ANN, allowing the network to quickly focus on complex, non-linear adjustments, leading to superior overall performance. Although advanced architectures such as CNN-LSTM, GRU, and attention-based models have shown promise, they typically require large-scale temporal datasets and higher computational resources. Given the tabular, non-sequential nature of the current dataset, the proposed hybrid approach offers a more interpretable and computationally efficient alternative.

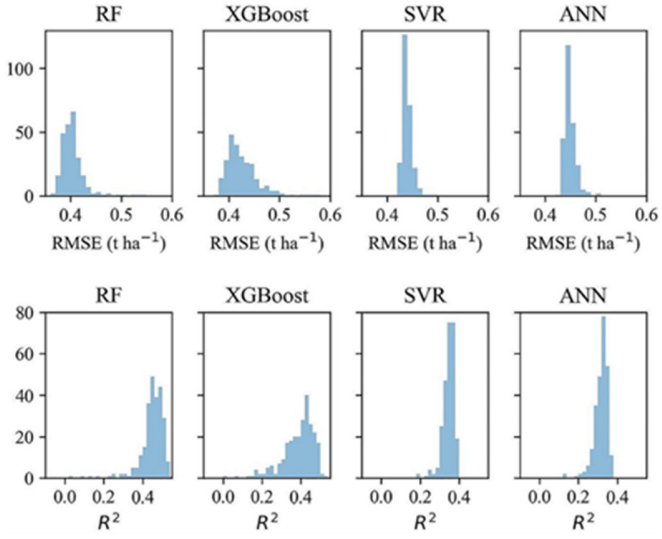

Fig. 10. Performance metrics for fertilizer recommendation system.

E. Fertilizer Recommendation System

The final component is a fertilizer recommendation system, which classifies the necessary fertilizer based on the current N, P, and K soil nutrient levels provided by the user. Figure 10 portrays the performance metrics for this classification task. Consistent with the crop recommendation results, the RF classifier also demonstrated the best performance for the fertilizer recommendation system, validating its robustness for classification tasks within this agricultural domain.

IV. CONCLUSION

This study presented an AI-enabled decision support system for agricultural planning, integrating optimal crop recommendation, fertilizer guidance, and accurate yield forecasting based on analyzed soil and environmental parameters. Through the application of the Sequential Forward Feature Selection Algorithm (SFFSA), the Sequential Back Elimination Feature Selection Algorithm (SBEFSA), and the Random Forest Variable Importance Algorithm (RFVIA), the minimal and most influential feature subset for prediction was identified as Rainfall, N, P, K, and soil pH.

The system provides farmers with two crucial recommendations:

- Crop Recommendation: Selected standard Machine Learning (ML) classifiers were applied to decide the best suitable crop based on soil characteristics, with the RF algorithm yielding the highest classification accuracy.
- Fertilizer Recommendation: A supplementary system was implemented to recommend suitable fertilizer types based on the current N, P, and K nutrient levels in the soil, also leveraging the superior performance of the RF classifier.

For forecasting paddy crop yield, a suite of standard regression models was evaluated against a hybrid Multiple Linear Regression (MLR)-Artificial Neural Network (ANN) approach. The hybrid model with optimized coefficients and the intercept values has a superior performance compared to standard ML models. The results indicate that the hybrid model's RMSE and MAE are improved by factors of 1.8–2.2-fold and 1.5-fold, respectively, while maintaining good R values.

This informed initialization, coupled with the iterative error minimization via the backpropagation algorithm, enabled the hybrid model to outperform all stand-alone algorithms. Future work will broaden the model's geographic scope using spatial and temporal data and incorporating explainable AI (XAI) to boost the transparency and trustworthiness of the crop recommendations for end-users.

DECLARATION OF COMPETING INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENT

Not applicable to this work.

DATA AVAILABILITY

The data used in this study were collected from [21-23]. Any additional data supporting the findings of this study can be made available upon reasonable request from the corresponding author.

REFERENCES

[1] S. Kiruthika and D. Karthika, "IoT-Based Professional Crop Recommendation System Using a Weight-Based Long-Term Memory Approach," *Measurement: Sensors*, vol. 27, Jun. 2023, Art. no. 100722, https://doi.org/10.1016/j.measen.2023.100722.

[2] S. P. Raja, B. Sawicka, Z. Stamenkovic, and G. Mariammal, "Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers," *IEEE Access*, vol. 10, pp. 23625–23641, 2022, https://doi.org/10.1109/ACCESS.2022.3154350.

[3] P. S. Maya Gopal and R. Bhargavi, "Selection of Important Features for Optimizing Crop Yield Prediction," *International Journal of Agricultural and Environmental Information Systems*, vol. 10, no. 3, pp. 54–71, Jul. 2019, https://doi.org/10.4018/IJAEIS.2019070104.

[4] M. Abdel-Salam, N. Kumar, and S. Mahajan, "A Proposed Framework for Crop Yield Prediction Using Hybrid Feature Selection Approach and Optimized Machine Learning," *Neural Computing and Applications*, vol. 36, no. 33, pp. 20723–20750, Nov. 2024, https://doi.org/10.1007/s00521-024-10226-x.

[5] R. Kavitha, M. Kavitha, and R. Srinivasan, "Crop Recommendation in Precision Agriculture Using Supervised Learning Algorithms," in *2022 3rd International Conference for Emerging Technology (INCET)*, Belgaum, India, May 2022, pp. 1–4, https://doi.org/10.1109/INCET54531.2022.9824155.

[6] K. Bakthavatchalam *et al.*, "IoT Framework for Measurement and Precision Agriculture: Predicting the Crop Using Machine Learning Algorithms," *Technologies*, vol. 10, no. 1, Jan. 2022, Art. no. 13, https://doi.org/10.3390/technologies10010013.

[7] M. Venkatanaresh and I. Kullayamma, "Deep Learning Based Concurrent Excited Gated Recurrent Unit for Crop Recommendation Based on Soil and Climatic Conditions," *Multimedia Tools and Applications*, vol. 83, no. 24, pp. 64109–64138, Jan. 2024, https://doi.org/10.1007/s11042-023-18004-y.

[8] P. S. S. Gopi and M. Karthikeyan, "Red Fox Optimization with Ensemble Recurrent Neural Network for Crop Recommendation and Yield Prediction Model," *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 13159–13179, Jul. 2023, https://doi.org/10.1007/s11042-023-16113-2.

[9] C. Raju, D. V. Ashoka, and B.V. Ajay Prakash, "CropCast: Harvesting the Future with Interfused Machine Learning and Advanced Stacking

Ensemble for Precise Crop Prediction," *Kuwait Journal of Science*, vol. 51, no. 1, Jan. 2024, Art. no. 100160, https://doi.org/10.1016/j.kjs.2023.11.009.

[10] M. Gallardo, M. T. Peña-Fleitas, C. Giménez, F. M. Padilla, and R. B. Thompson, "Adaptation of VegSyst-DSS for Macronutrient Recommendations of Fertigated, Soil-Grown, Greenhouse Vegetable Crops," *Agricultural Water Management*, vol. 278, Mar. 2023, Art. no. 107973, https://doi.org/10.1016/j.agwat.2022.107973.

[11] S. Fenz, T. Neubauer, J. Heurix, J. K. Friedel, and M.-L. Wohlmuth, "AI- and Data-driven Pre-Crop Values and Crop Rotation Matrices," *European Journal of Agronomy*, vol. 150, Oct. 2023, Art. no. 126949, https://doi.org/10.1016/j.eja.2023.126949.

[12] N. Subash *et al.*, "Relevance of Climatological Information on Spatial and Temporal Variability of Indian Summer Monsoon Rainfall (ISMR) in Recent El Niño Years and Its Impact on Four Important Kharif Crops Over India," *Climate Services*, vol. 30, Apr. 2023, Art. no. 100370, https://doi.org/10.1016/j.cliser.2023.100370.

[13] M. Y. Shams, S. A. Gamel, and F. M. Talaat, "Enhancing Crop Recommendation Systems with Explainable Artificial Intelligence: A Study on Agricultural Decision-Making," *Neural Computing and Applications*, vol. 36, no. 11, pp. 5695–5714, Apr. 2024, https://doi.org/10.1007/s00521-023-09391-2.

[14] Y. Akkem, S. K. Biswas, and A. Varanasi, "Streamlit-Based Enhancing Crop Recommendation Systems with Advanced Explainable Artificial Intelligence for Smart Farming," *Neural Computing and Applications*, vol. 36, no. 32, pp. 20011–20025, Nov. 2024, https://doi.org/10.1007/s00521-024-10208-z.

[15] A. M. Joshi and S. Patel, "A CNN-Bidirectional LSTM Approach for Price Forecasting of Agriculture Commodities in Gujarat," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, Salem, India, May 2022, pp. 266–272, https://doi.org/10.1109/ICAAIC53929.2022.9793154.

[16] R. L. Manogna, V. Dharmaji, and S. Sarang, "A Novel Hybrid Neural Network-Based Volatility Forecasting of Agricultural Commodity Prices: Empirical Evidence from India," *Journal of Big Data*, vol. 12, no. 1, Apr. 2025, Art. no. 85, https://doi.org/10.1186/s40537-025-01131-8.

[17] V. Fassa, N. Pricca, G. Cabassi, L. Bechini, and M. Corti, "Site-Specific Nitrogen Recommendations' Empirical Algorithm for Maize Crop Based on the Fusion of Soil and Vegetation Maps," *Computers and Electronics in Agriculture*, vol. 203, Dec. 2022, Art. no. 107479, https://doi.org/10.1016/j.compag.2022.107479.

[18] P. E. Rubini and P. Kavitha, "Prediction of the Right Crop for the Right Soil and Recommendation of Fertiliser Usage by Machine Learning Algorithm," *International Journal of Computer Applications in Technology*, vol. 69, no. 2, 2022, Art. no. 163, https://doi.org/10.1504/IJCAT.2022.126885.

[19] P. Srinivas and A. Suresh, "CNN-LSTM Model for Cotton Yield Prediction Using Remote Sensing Data," in *2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, Tirunelveli, India, Jul. 2025, pp. 515–520, https://doi.org/10.1109/ICDICI66477.2025.11135132.

[20] N. M. Basavaraju, U. B. Mahadevaswamy, and M. Srikantaswamy, "Optimized Crop Yield Forecasting Using the Naive Bayes Regression Algorithm in Smart Agriculture," *Engineering, Technology & Applied Science Research*, vol. 15, no. 6, pp. 28995–29001, Dec. 2025, https://doi.org/10.48084/etasr.11856.

[21] "District-Wise Season-Wise Crop Production Statistics from 1997." Ministry of Agriculture and Farmers Welfare, Government of India, 2013, [Online]. Available: https://www.data.gov.in/resource/district-wise-season-wise-crop-production-statistics-1997.

[22] "Daily District-Wise Rainfall Data." India Meteorological Department, Ministry of Earth Sciences, Government of India, Sep. 2022, [Online]. Available: https://www.data.gov.in/resource/daily-district-wise-rainfall-data.

[23] "State/UT-wise Number of Soil Health Cards Issued to the Farmers in the Country under Soil Health Card (SHC) Scheme from 2019-20 to 2023-24." Ministry of Agriculture and Farmers Welfare, Government of India, Oct. 2024, [Online]. Available: https://www.data.gov.in/resource/stateut-wise-number-soil-health-cards-issued-farmers-country-under-soil-health-card-shc.

[24] S. P. Sudha and J. B. S. Loret, "A Review on Machine Learning-based Precision Agriculture Techniques for Crop Farming Monitoring with IoT," *Discover Environment*, vol. 4, no. 1, Jan. 2026, Art. no. 10, https://doi.org/10.1007/s44274-025-00305-8.

[25] M. Baishya and L. Dutta, "Tiny ML-Based Crop Recommendation System for Precision Agriculture 5.0," *Smart Agricultural Technology*, vol. 12, Dec. 2025, Art. no. 101247, https://doi.org/10.1016/j.atech.2025.101247.