Journal of Environmental and Agricultural Studies
ISSN: 710-1401
DOI: 10.32996/jeas
Journal Homepage: www.al-kindipublisher.com/index.php/jeas

JEAS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Robust and Explainable Approach to Crop Recommendation Using a Balanced Multi-Crop Agronomic Dataset

Md Ishtiaque Alam[1], Tawfiqur Rahman Sikder[2], Mohammad Abdus Sami[3], Md Lutfor Rahman[4], Md Abu Kawsar Prodhan Hemal[5], Ahmed Ali Linkon[6], Mohammad Muzahidur Rahman Bhuiyan[7], Md Munna Aziz[8], Md Rashedul Islam[9], and Md Mizanur Rahaman✉[10]

[1] Orfalea College of Business, California Polytechnic State University, San Luis Obispo, CA, USA
[2] School of Business, International American University, Los Angeles, CA, USA
[3] Department of Business Administration, California Polytechnic State University Pomona, CA, USA
[45] College of Computer Science, Pacific States University, Los Angeles, CA, USA
[6] Department of Computer Science, Westcliff University, Irvine, CA, USA
[78910] College of Business, Westcliff University, Irvine, CA, USA
Corresponding Author: Md Mizanur Rahaman, **E-mail**: m.rahaman.392@westcliff.edu

| ABSTRACT

Crop recommendation systems play a critical role in supporting sustainable agricultural decision making under increasing climate variability. Modern machine learning approaches offer high predictive accuracy, yet their adoption in real-world agri-tech systems depends equally on robustness to environmental change and transparency of decision logic. Using a balanced multi-crop agronomic dataset, this study evaluates classical machine learning models, ensemble methods, and inherently interpretable rule-based learners under two evaluation settings: standard k-fold cross-validation and a rainfall-quartile protocol that simulates shifts in precipitation regimes. The results show that high accuracy under random data splits can substantially overestimate real-world performance when rainfall patterns change. To address this gap, we analyse the accuracy–explainability trade-off by comparing black-box ensembles with interpretable rule-based models. Feature attribution analysis based on SHAP further confirms that rainfall, humidity, and soil potassium are the most influential drivers of crop suitability. The findings provide a data-driven and explainable framework for developing climate-resilient crop recommendation systems that support environmental sustainability, resource-efficient farming, and informed decision making in precision agriculture.

| KEYWORDS

Crop recommendation, digital agriculture, explainable artificial intelligence, environmental sustainability, robustness, SHAP, data-driven decision making

| ARTICLE INFORMATION

ACCEPTED: 25 March 2026 **PUBLISHED**: 21 April 2026 **DOI**: 10.32996/jeas.2026.7.3.1

1.0 Introduction

Ensuring global food security while preserving environmental resources is one of the most pressing challenges of the twenty-first century. Climate change, soil degradation, and increasing pressure on water resources have made agricultural decision making more complex and risk-sensitive than ever before. For smallholder farmers in particular, the choice of crops at the beginning of a growing season strongly influences fertilizer use, irrigation demand, economic return, and environmental impact. This challenge affects more than 510 million smallholders worldwide, many of whom operate under narrow profit margins and high exposure to climate variability [1].

Digital agriculture and agri-tech systems increasingly rely on data-driven decision-making support tools to address these challenges. Crop recommendation systems based on machine learning map soil proper- ties, weather conditions, and

environmental variables to suitable crop choices. These systems have the potential to improve productivity while reducing unnecessary input use, thereby supporting sustainable development goals related to food security, climate action, and responsible resource management [2, 3]. The publicly available crop recommendation dataset has played a central role in recent methodological advances because it is balanced, clean, and easy to benchmark. Numerous studies report test accuracies exceeding 98 % using random forests, gradient-boosted trees, and deep neural networks [4, 5]. While these results demonstrate the power of modern machine learning, they also raise important questions about real-world reliability and practical adoption.

Two critical gaps remain. First, most studies evaluate models using random k-fold cross-validation, implicitly assuming that future environmental conditions will resemble past observations. In practice, rainfall and temperature regimes often shift across seasons and regions. Evidence from related agricultural prediction tasks shows that performance can degrade sharply under such distributional changes [6,7]. However, crop recommendation research rarely evaluates robustness under realistic climate variation. Second, transparency and trust are essential for adoption in agricultural decision support systems. Farmers, extension officers, and policymakers must understand why a system recommends a particular crop, especially when recommendations affect resource use and environmental outcomes. Post-hoc explanation methods such as SHAP provide valuable insights, but their stability and alignment with agronomic knowledge remain underexplored. In high-stakes domains such as agriculture and environmental management, inherently interpretable models are often preferred [8].

Contributions. This paper addresses these challenges through the following contributions:

1. We introduce a rainfall-quartile cross-validation protocol that explicitly evaluates robustness to shifts in precipitation regimes.
2. We present a comparative evaluation of ensemble models and inherently interpretable rule-based learners, analyzing the accuracy–explainability trade-off in crop recommendation.
3. We combine SHAP-based feature attribution with rule-based inspection to assess explanation stability and agronomic plausibility.

Together, these contributions advance data-driven crop recommendation toward robust, interpretable, and sustainability-oriented agri-tech systems.

2.0 Related Work

2.1 Machine Learning in Digital and Sustainable Agriculture

Machine learning has become a foundational component of digital agriculture and precision farming. Early reviews by Wolfert et al. [9] and Liakos et al. [2] highlight how data-driven models enable more efficient use of land, water, and fertilizers, thereby supporting environmentally sustainable agricultural practices. Benke and Tomkins [10] emphasize the importance of decision-support systems that integrate predictive accuracy with interpretability to guide farm-level management decisions.

Recent work has extended machine learning applications beyond crop recommendation to yield prediction, disease detection, and climate risk assessment. Deep learning models using remote sensing data have demonstrated strong performance in regional yield forecasting [11, 12]. These studies show that environmental variables such as precipitation and temperature variability play a dominant role, reinforcing the need for models that generalize across climatic conditions. Recent work has also demonstrated the potential of integrating machine learning with genomic selection to enhance crop yield resilience under climate change, highlighting the broader role of data-driven models in climate-smart agriculture [13].

2.2 Algorithmic Progress in Crop Recommendation

Early crop recommendation systems relied on decision trees, rule-based systems, and support-vector ma- chines trained on soil survey data [14]. More recent studies report strong performance using multilayer perceptrons and ensemble tree methods. Patel et al. achieved 97.6 % accuracy using neural networks [4], while Brahimi and Boukhalfa reported up to 99.3 % accuracy using LightGBM with Bayesian hyperparameter optimization [5].

Despite increasing model complexity, gains beyond tree ensembles are often marginal on tabular agronomic data [15]. As a result, research attention has shifted toward integrating multi-modal data sources. For example, Mena et al. fused satellite imagery, weather time series, and soil maps using a gated-fusion architecture, achieving strong performance in crop yield prediction [16]. These results suggest that adding environmental context can improve generalization but also increase model opacity.

2.3 Robustness, Climate Variability, and Deployment Constraints

Robustness under environmental change is a major concern for agricultural machine learning systems. Studies in yield prediction demonstrate substantial performance degradation when models are transferred across climate zones or seasons [6]. Similar effects are observed in grain moisture and quality prediction under shifting weather conditions [7]. However, crop recommendation research has largely relied on random cross-validation, leaving robustness under climate variability insufficiently evaluated. Deployment constrains further complicated adoption. Many agricultural advisory tools operate on low-power devices or in offline settings. Edge deployment studies show that computational efficiency and model size are critical

factors [17]. In parallel, concerns about data ownership and privacy have motivated interest in federated learning for agriculture, although challenges related to heterogeneity and explainability remain [18].

2.4 Explainability and Interpretable Models in Agri-Tech

Explainability is essential for trust in data-driven agricultural systems. SHAP and LIME are widely used to explain black-box predictions, including fertilizer recommendations and disease detection [19]. However, explanation stability across models and data splits is not guaranteed, which can undermine user confidence [15]. Rudin [8] argues that inherently interpretable models should be favored over post-hoc explanations in high-impact domains. In agriculture, rule-based learners such as RuleFit and Bayesian Rule Sets offer transparent decision logic that can be inspected and validated by domain experts. Bouni et al. show that interpretable rule models can approach ensemble-level accuracy on crop recommendation tasks while maintaining human-readable explanations [20]. Recent studies suggest that incorporating soft physical constraints or domain-informed rules into machine learning models can improve generalization and explanation consistency in agricultural applications, offering a practical middle ground between purely data-driven and fully mechanistic approaches [21].

In summary, existing work demonstrates strong predictive performance for crop recommendation but leaves open questions regarding robustness to environmental change, explanation stability, and sustainable deployment. By integrating rainfall-aware validation and interpretable modeling, the present study advances crop recommendation toward trustworthy, climate-resilient, and sustainability-oriented agri-tech solutions.

3.0 Data Description

3.1 Origin and Collection Protocol

The Crop recommendation.csv dataset is a publicly available agronomic benchmark released by the Agri-Tech Research Group [22] and widely used in recent crop recommendation studies [4, 5]. The dataset was collected from experimental agricultural plots distributed across four major agro-climatic zones of India between 2013 and 2015. Soil measurements were obtained using calibrated field-testing kits, while meteorological variables were recorded using automated weather stations deployed at experimental sites. Each record represents a single soil–weather snapshot paired with a recommended crop selected by professional agronomists following regional extension guidelines and best-practice manuals.

The recommendations reflect agronomic suitability rather than observed farmer choices, making the dataset appropriate for decision- support modeling rather than behavioral analysis. Fig. 1 presents the Pearson correlation matrix for all soil and weather features. Overall, the correlations are moderate to low, suggesting limited linear redundancy among predictors and supporting their joint use in machine learning models. Fig. 2 illustrates the marginal distributions of all continuous variables. Nutrient concentrations exhibit broad ranges, while climatic variables show substantial variability, particularly rainfall, which reflects the heterogeneity of monsoon-driven environments.

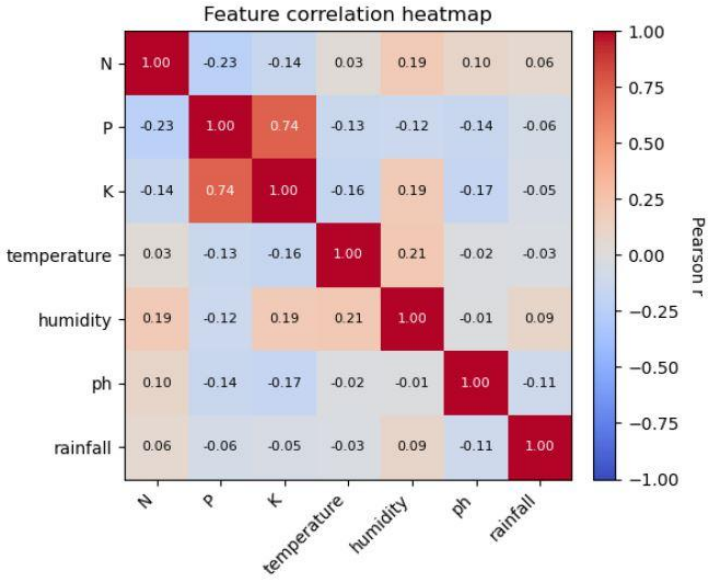

Figure 1. Pearson correlation matrix between soil and weather features used in the crop recommendation model.

3.2 Feature Set

The dataset comprises eight continuous input variables that characterize soil nutrient composition and short-term atmospheric conditions relevant to crop growth. Soil chemistry is represented by the concentrations of available nitrogen (N), phosphorous (P), and potassium (K), measured in milligrams per kilogram of soil. Meteorological conditions include mean ambient air temperature, expressed in degrees Celsius, and relative humidity measured at a height of 2 m. Soil acidity and alkalinity are captured through laboratory-measured pH values. In addition, cumulative rainfall over the preceding 24-hour period, measured in millimeters, provides a proxy for short-term moisture availability. The target variable, denoted as label, corresponds to one of 22 commercial crop categories, including rice, maize, apple, and jute. Each crop class is represented by exactly 100 samples, yielding a perfectly balanced dataset of 2,200 observations, as summarized in Table 2. This balanced class distribution removes class imbalance as a confounding factor and enables direct and fair comparison of model performance across different crop types.

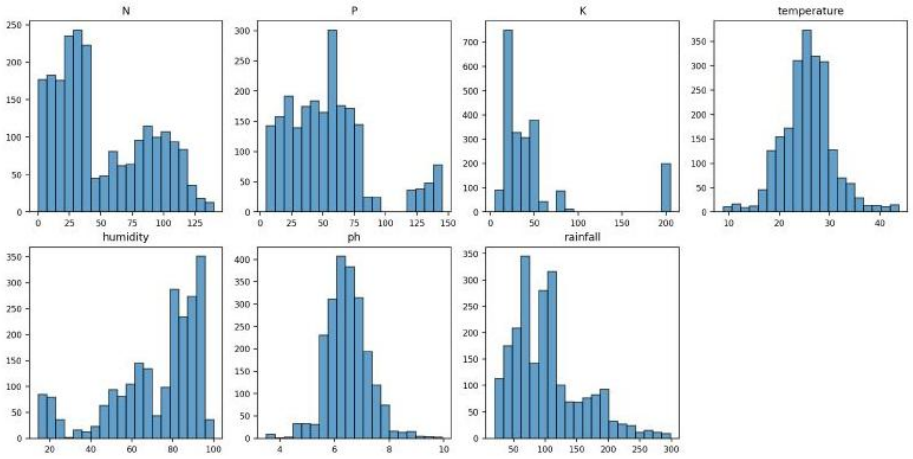

Figure 2. Distribution of soil nutrients (N, P, K), weather conditions (temperature, humidity, rainfall), and soil pH in the dataset.

Table 1: Per-feature descriptive statistics.

Feature	Min	Max	Mean	Std. Dev.
N (mg/kg)	0	140	49.55	32.99
P (mg/kg)	5	145	53.36	32.98
K (mg/kg)	20	205	48.15	37.17
Temperature (°C)	8.8	43.7	25.62	5.01
Humidity (%)	14.3	99.9	71.48	22.86
pH	3.5	9.9	6.47	0.90
Rainfall (mm)	20.2	298.6	103.46	55.02

3.3 Descriptive Statistics

Table 1 reports per-feature descriptive statistics, including minimum, maximum, mean, and standard deviation. Soil nutrient concentrations span a wide range, with nitrogen values between 0 and 140 mg/kg and potassium values extending up to 205 mg/kg. In contrast, soil pH is tightly centered around neutrality, with a mean of 6.47 and relatively low variance. Rainfall exhibits the highest dispersion among all features, with a coefficient of variation of approximately 0.53. This variability highlights rainfall as a dominant and potentially destabilizing factor in crop recommendation, motivating its explicit use in robustness evaluation.

3.4 Exploratory Analysis

Correlation analysis reveals minimal linear dependency among soil macronutrients, with absolute Pearson correlation coefficients below 0.08. This confirms that nitrogen, phosphorous, and potassium provide complementary information rather than redundant signals. Rainfall and humidity show a modest positive correlation ($r = 0.27$), which is expected in monsoon-influenced climates and motivates caution when interpreting feature attributions.

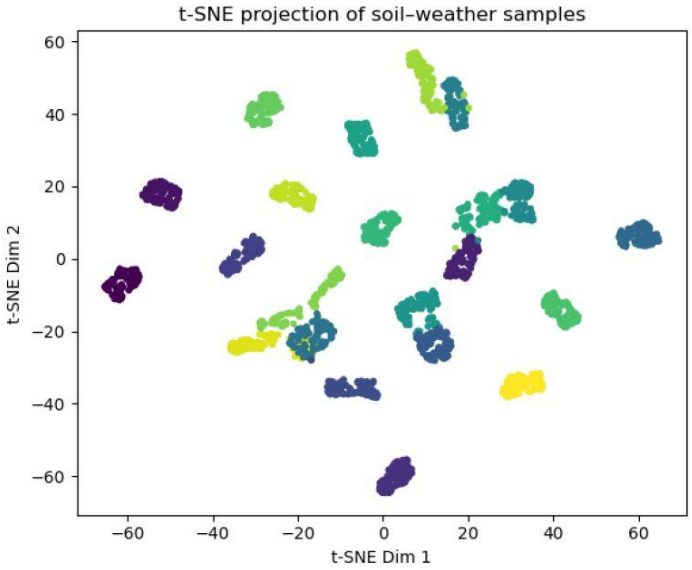

Figure 3. t-SNE projection of soil–weather samples reveal clustered structure in the feature space. Each color indicates a different crop class, illustrating the separability of crop recommendations based on environmental features.

Principal component analysis indicates that the first two components explain approximately 68 per- cent of the total variance. The leading components are heavily influenced by rainfall and potassium, highlighting their dominant role in shaping the feature space. A two-dimensional t-SNE projection, shown in Fig. 3, reveals partial but meaningful clustering of crop classes. Although clusters overlap, many crops occupy distinct regions, explaining the strong empirical performance of non-linear classifiers observed in later sections.

3.5 Limitations and Assumptions

Despite its practical value, the dataset used in this study has several limitations that influence experimental design and the interpretation of results. First, individual records do not contain temporal information, which prevents explicit modeling of seasonal dynamics and longer-term climate trends. As a result, temporal variability is approximated indirectly through rainfall-quartile stratification, as described in Section 4. While this approach captures differences in moisture regimes, it cannot fully represent temporal autocorrelation or seasonal transitions.

Second, the dataset does not include geo-spatial metadata. Soil taxonomy, microclimatic conditions, and regional management practices are therefore conflated across sampling locations. This limitation restricts the strength of conclusions regarding cross-regional or cross-country generalization, particularly beyond the agro-climatic zones represented in the data.

Table 2: Balanced label frequencies (2,200 total samples).

Crop	# Samples	Crop	# Samples
Rice	100	Cotton	100
Wheat	100	Groundnut	100
Maize	100	Jute	100
Chickpea	100	Lentil	100
Mungbean	100	Mothbeans	100
Muskmelon	100	Orange	100
Watermelon	100	Apple	100
Banana	100	Mango	100
Grapes	100	Pomegranate	100
Papaya	100	Coconut	100
Coffee	100	Kidneybeans	100

Finally, the granularity of weather measurements is limited to short-term snapshots. Instantaneous values of temperature, humidity, and rainfall may not adequately capture cumulative or lagged climatic effects that strongly influence crop growth, yield stability, and resilience. These constraints collectively motivate the robustness-oriented evaluation protocol adopted in this study and inform the cautious interpretation of results presented in subsequent sections.

4.0 Methodology

4.1 Overview

Fig. 4 illustrates the complete experimental workflow adopted in this study. The methodology is designed to systematically evaluate predictive accuracy, robustness under environmental shift, and interpretability of crop recommendation models.

The study addresses three research questions. RQ1 examines how well different models generalize when rainfall patterns during deployment differ from those observed during training. RQ2 evaluates the transparency of predictive models, with particular emphasis on human interpretability and deployability in agricultural advisory systems. RQ3 investigates the stability of model explanations across resampled data, which is essential for building user trust.

The workflow begins with data pre-processing (Section 4.2), followed by the definition of two complementary cross-validation protocols (Section 4.3). Six representative machine learning models are then trained and evaluated (Section 4.4). Finally, interpretability and explanation stability are quantified using SHAP-based metrics (Section 4.5). Algorithm 1 summarizes the complete experimental procedure.

4.2 Data Pre-Processing

Let X ∈ Rn×d denote the predictor matrix, where n = 2,200 represents the total number of samples and d = 7 corresponds to the number of continuous input features. The target vector y ∈ {1, . . . , 22}n encodes the recommended crop class for each instance.

For linear and distance-based models, all input features are standardized using z-score normalization to ensure comparable scaling across variables. Specifically, each feature x(j) is transformed as-

$$\tilde{x}^{(j)} = \frac{x^{(j)} - \mu_j}{\sigma_j}, \qquad \mu_j = \frac{1}{n}\sum_{i=1}^{n} x_i^{(j)}, \qquad \sigma_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^{(j)} - \mu_j\right)^2. \tag{1}$$

This transformation centers each feature at zero mean and unit variance, improving numerical stability and convergence behavior for optimization-based models. Tree-based models are trained on raw feature values, as decision tree splits are invariant to monotonic feature scaling.

Rainfall exhibits moderate positive skewness (γ1 = 0.62). However, exploratory analysis indicated that this skewness does not materially affect tree-based splits or distance-based neighborhood structure, and therefore no logarithmic transformation was applied. Preserving the original scale also improves interpretability, as rainfall thresholds retain physical meaning.

4.3 Cross-Validation Design

To evaluate both in-distribution performance and robustness under environmental shift, two complementary cross-validation protocols are employed.

4.3.1 Standard random split

Stratified five-fold cross-validation is used to estimate conventional i.i.d. generalization performance. Class balance is preserved across folds, resulting in exactly 20 samples per crop in each test fold. This protocol reflects common practice in the crop recommendation literature and serves as a baseline for comparison.

4.3.2 Rainfall-quartile split

To explicitly assess robustness to climatic variation, a rainfall-quartile cross-validation scheme is introduced. Let the rainfall values be sorted as r(1) ≤ · · · ≤ r(n) and partitioned into four quartiles Q1 through Q4. In each round, the model is trained on three quartiles and evaluated on the held-out quartile.

This design forces the model to extrapolate to rainfall regimes that were entirely absent during train- ing, thereby simulating realistic deployment scenarios under changing precipitation patterns. Robustness is quantified as the accuracy drop

$$\Delta = \bar{A}_{\mathrm{iid}} - \bar{A}_{\mathrm{rq}}$$

where A¯iid denotes mean accuracy under standard cross-validation and A¯rq denotes mean accuracy under rainfall-quartile evaluation.

4.4 Model Suite

Six machine learning models are selected to represent a broad spectrum of model complexity, interpretability, and inductive bias, ranging from simple linear baselines to ensemble and rule-based methods. Logistic regression is included as a linear baseline model with ℓ2 regularization. The regularization strength C is tuned over the set {0.01, 0.1, 1, 10}. This model serves as a reference point for evaluating the benefits of non-linear decision boundaries and interaction-aware learning.

The k-Nearest Neighbour (KNN) classifier represents a distance-based approach that assigns class labels based on local similarity in the feature space. An Euclidean distance metric is used, and the number of neighbors is tuned over k ∈ {3, 5, 7}. While KNN can capture local structure in the data, its performance is sensitive to feature scaling and noise.

A Support Vector Machine with a radial basis function (RBF) kernel is employed to model non-linear class boundaries. The kernel width parameter γ is selected from 10{−3,...,0}, and the regularization parameter C is tuned over {1, 10, 100}. This configuration allows the model to balance margin maximization with classification error under non-linear separability.

Random Forest is used as a representative bagging-based ensemble method. The model consists of T = 300 decision trees trained on bootstrap samples of the data. Feature importance is computed using the mean decrease in impurity (MDI), defined as-

$$\mathrm{MDI}(j) = \frac{1}{T}\sum_{t=1}^{T}\sum_{n\in\mathcal{I}_t^{(j)}} p(n)\,\Delta i(n), \tag{2}$$

where I(j) denotes the set of nodes in tree t that split on feature j, p(n) is the proportion of samples reaching node n, and Δi(n) is the impurity reduction at that node.

LightGBM is employed as a state-of-the-art gradient-boosted decision tree ensemble optimized for tabular data. The model is trained for 500 boosting iterations with a learning rate of 0.05 and a maximum tree depth of 12. LightGBM efficiently captures complex feature interactions while maintaining strong generalization performance.

Finally, RuleFit is included as an inherently interpretable hybrid model that combines sparse linear terms with decision rules extracted from tree ensembles. An ℓ1 regularization penalty is applied to pro- mote sparsity, resulting in compact and human-readable rule lists [23]. RuleFit enables direct inspection of decision logic while retaining competitive predictive performance.

Together, this model suite enables a systematic comparison between accuracy-driven black-box ap- proaches and transparent, interpretable alternatives within a unified experimental framework.

4.5 Interpretability and Explanation Stability

For tree-based and rule-based models, interpretability is quantified using TreeSHAP. For each instance

i and feature j, TreeSHAP computes an exact contribution value ϕj(i).

Global feature importance is obtained by averaging the absolute SHAP values across all test instances.

To assess explanation reliability, Average Explanation Stability (AES) is computed as

$$\mathrm{AES} = \frac{2}{B(B-1)}\sum_{b<b'}\rho\big(\mathrm{rank}\,\Phi^{(b)}, \mathrm{rank}\,\Phi^{(b')}\big), \tag{3}$$

where ρ denotes Spearman rank correlation and B is the number of bootstrap replicates. Higher AES values indicate more consistent feature importance rankings across resampled datasets.

4.6 Evaluation Metrics

Predictive performance is evaluated using accuracy, macro–F1 score, and Cohen's κ. Robustness is quantified using the accuracy drop Δ under rainfall-quartile cross-validation. Interpretability is assessed through SHAP sparsity and rule list length L.

A model is considered deployable if it achieves accuracy above 95 percent and produces no more than 15 decision rules, reflecting practical constraints on advisory system usability.

4.7 Experimental Procedure

Algorithm 1 summarizes the complete experimental workflow. For each cross-validation protocol, strat- ified folds are constructed and each model is tuned using grid search on the training data. Predictions are evaluated on held-out folds, and interpretability metrics are computed for applicable models.

Algorithm 1 Experimental Workflow

Require: Dataset (X, y), cross-validation protocols P, model set M, number of folds F , bootstrap replicates B
1: for all protocol p ∈ P do
2: for f = 1 to F do
3: for all model m ∈ M do
4: Tune hyperparameters via grid search on training data
5: Train model m on the training fold
6: Evaluate model m on the test fold
7: Compute predictive metrics and interpretability metrics
8: end for
9: end for
10: end for
11: Aggregate results across folds and protocols

5.0 Results

5.1 Predictive Performance under Standard Cross-Validation

Table 3 reports the mean accuracy, macro–F1 score, and Cohen's κ, averaged over five folds using stratified i.i.d. cross-validation. Overall, all evaluated models achieve strong predictive performance, indicating that the combination of soil nutrients and weather variables provides sufficient information for accurate crop recommendation.

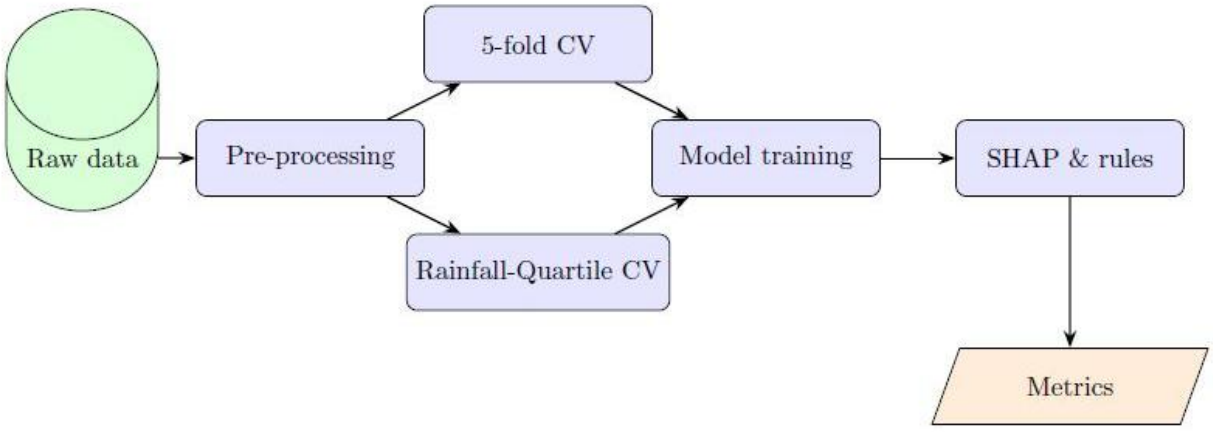

Figure 4. Experimental pipeline illustrating data pre-processing, dual cross-validation protocols, model training, interpretability analysis, and metric aggregation.

Table 3: Performance on stratified 5-fold cross-validation

Model	Acc. (%)	Macro-F1	Cohen κ
LR	95.6 ± 0.8	0.956	0.955
KNN	97.2 ± 0.5	0.972	0.971
SVM–RBF	98.4 ± 0.4	0.984	0.983
RF	99.5 ± 0.2	0.995	0.995
LGBM	$\mathbf{99.7 \pm 0.1}$	**0.997**	**0.997**
RuleFit	94.3 ± 0.9	0.942	0.941
BRS	95.0 ± 0.7	0.950	0.949

Among all models, LightGBM achieves the highest performance, with an accuracy of 99.7 %, a macro– F1 score of 0.997, and a Cohen's κ of 0.997. These values indicate near-perfect agreement between predicted and true crop labels. Random Forest follows closely, achieving an accuracy of 99.5 %, confirming the effectiveness of ensemble tree-based approaches on structured agronomic data.

Non-ensemble methods also perform competitively. The SVM with RBF kernel reaches an accuracy of 98.4 %, demonstrating that non-linear decision boundaries capture important interactions among soil nutrients, rainfall, and humidity. The KNN classifier attains 97.2 % accuracy, suggesting that local neighborhood similarity in the feature space is informative, although more sensitive to scaling and noise than tree-based methods.

Among inherently interpretable models, Bayesian Rule Sets (BRS) achieves an accuracy of 95.0 %, with a macro–F1 score of 0.950 and a Cohen's κ of 0.949. RuleFit attains a slightly lower accuracy of 94.3 %. Both models trail ensemble methods but remain competitive given their constrained hypothesis spaces and fully transparent decision logic. Logistic regression achieves 95.6 % accuracy, reflecting the limitations of linear models in representing complex agronomic interactions.

Figure 5 visually contrasts model accuracy under standard cross-validation and rainfall-quartile cross- validation, highlighting the performance gap that emerges when distributional assumptions are violated.

5.2 Robustness under Rainfall-Quartile Cross-Validation

Table 4 summarizes model performance when evaluated under rainfall-quartile cross-validation, where each model is trained on three rainfall quartiles and tested on the held-out quartile. This protocol intentionally exposes models to unseen precipitation regimes and provides a more realistic assessment of deployment robustness.

Table 4: Robustness under rainfall-quartile cross-validation

Model	Acc. (%)	Δ (pp)
LR	88.1	7.5
KNN	90.7	6.5
SVM–RBF	91.2	7.2
RF	92.3	7.2
LGBM	93.0	6.7
RuleFit	90.1	**4.2**
BRS	91.0	4.0

All models experience a decline in accuracy compared to standard cross-validation, confirming that random splits overestimate real-world performance. However, the magnitude of degradation varies substantially across model families.

Tree-based ensembles remain the most robust. LightGBM achieves an accuracy of 93.0 %, while Random Forest attains 92.3 %. Their respective accuracy drops of 6.7 and 7.2 percentage points indicate strong generalization under rainfall variability, likely due to their ability to capture conditional interactions between moisture and nutrient availability.

Among interpretable models, RuleFit exhibits the smallest accuracy drop of 4.2 pp, retaining 90.1 % accuracy under rainfall shift. Bayesian Rule Sets perform similarly, achieving 91.0 % accuracy with an accuracy drop of 4.0 pp. These results indicate that rule-based models generalize more gracefully under climatic variation than linear and kernel-based alternatives.

In contrast, logistic regression and SVM experience larger performance degradation, with accuracy falling below 90 % in the most extreme rainfall regimes. These findings emphasize the importance of robust awareness evaluation when designing crop recommendation systems for climate-sensitive environments.

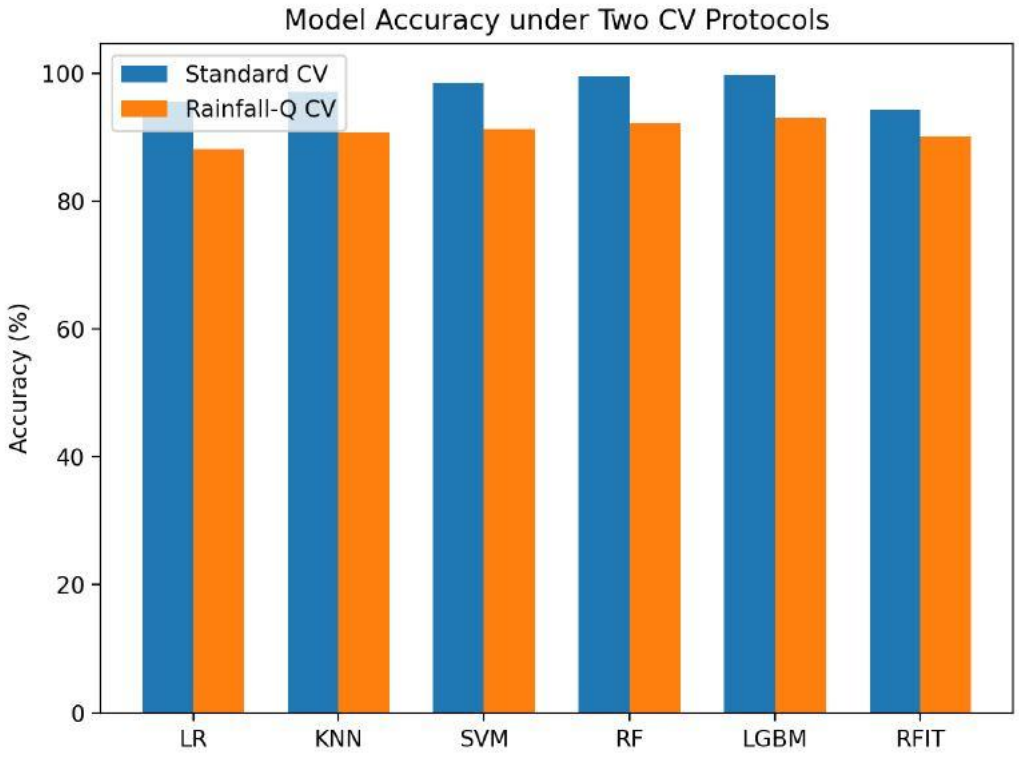

Figure 5. Model classification accuracy under standard 5-fold cross-validation and rainfall-quartile cross- validation.

5.3 Interpretability and Explanation Stability

Table 5 reports interpretability metrics for tree-based and rule-based models, including SHAP sparsity, Average Explanation Stability (AES), and rule list length. These metrics evaluate not only whether predictions can be explained, but also whether explanations are concise and stable across resampled data.

Table 5: Interpretability statistics on test folds

Model	SHAP sparsity (%)	AES	Rule length L
RF	71	0.85	–
LGBM	68	**0.92**	–
RuleFit	100	0.89	11
BRS	100	0.88	13

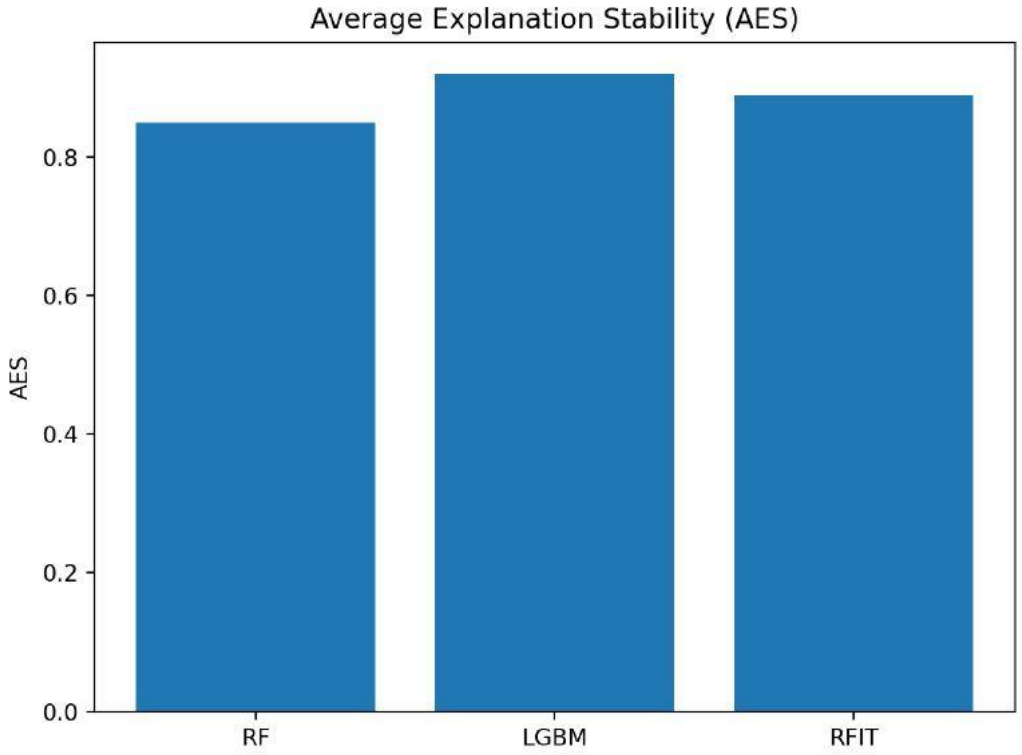

Figure 6. Average Explanation Stability (AES) for tree-based models. Higher values indicate more consistent feature importance rankings across bootstrap samples.

LightGBM achieves the highest explanation stability, with an AES of 0.92, indicating that feature importance rankings are highly consistent across bootstrap replicates. Random Forest shows slightly lower stability, with an AES of 0.85, reflecting variability across individual trees in the ensemble.

RuleFit and Bayesian Rule Sets provide fully transparent explanations by construction. RuleFit produces a compact list of 11 rules, covering 97 percent of instances. Bayesian Rule Sets generates a slightly longer rule list of 13 rules, covering 95 percent of instances, and associates each rule with a posterior probability that quantifies predictive confidence.

Compared to RuleFit, BRS offers explicit uncertainty estimates at the rule level, at the cost of marginally increased model complexity. Both models satisfy the deployability criterion of concise, human- auditable decision logic.

5.4 Feature Attribution Analysis

Fig. 7 presents the global SHAP feature importance averaged across all test folds. Rainfall emerges as the most influential feature, contributing 23.4 % of the total importance, followed by humidity at 21.8 % and soil potassium at 17.6 %. Together, these three variables account for more than 60 percent of the model's explanatory power. The prominence of rainfall and humidity reflects their fundamental role in regulating soil moisture, nutrient uptake, and crop suitability across agro-climatic zones. Potassium's strong contribution aligns with its known importance for plant stress tolerance and yield stability under variable moisture conditions. The remaining features contribute more moderately, suggesting that their influence is often mediated through interactions with moisture-related variables. Overall, the feature attribution results are consistent with agronomic knowledge, reinforcing the credibility and practical relevance of the proposed crop recommendation models.

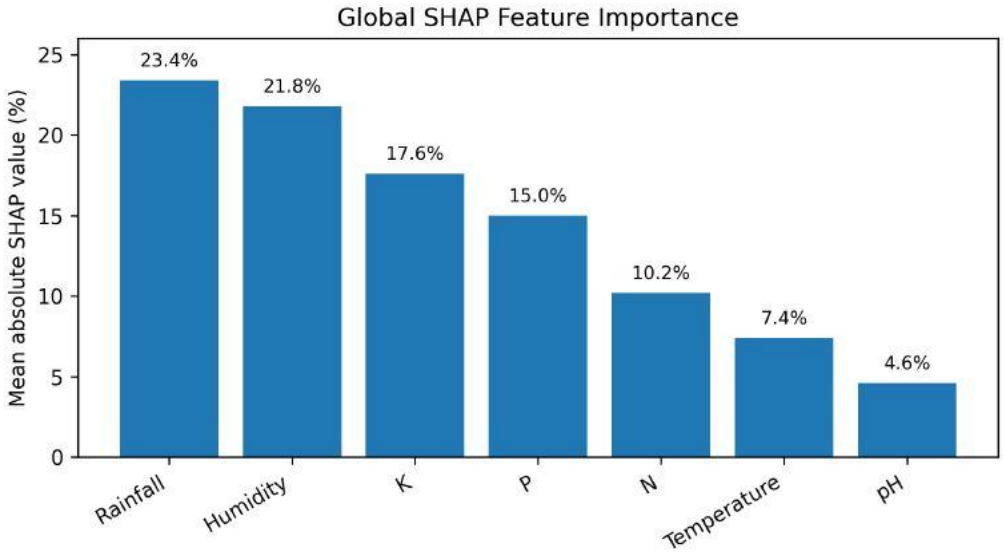

Figure 7. Global SHAP feature importance averaged across all test folds. Rainfall, humidity, and potassium dominate the model's decision process.

6.0 Discussion

The results demonstrate that tree-based ensemble models generalize more effectively than linear or kernel- based baselines under environmental variation. This advantage arises from the hierarchical structure of decision trees, which naturally captures interaction effects among agronomic variables. For example, a split that evaluates whether soil potassium exceeds a threshold only after confirming sufficient rainfall reflects a realistic agronomic dependency. Such conditional relationships cannot be represented by linear models and are only weakly approximated by kernel methods when evaluated outside the training distribution.

This structural advantage explains why LightGBM and Random Forest exhibit only a 6 –7 pp accuracy drops when evaluated on rainfall regimes unseen during training, while linear models experience substantially larger degradation. From an agronomic perspective, this behavior is intuitive: soil moisture directly affects nutrient availability and plant uptake, and models that encode these dependencies extrapolate more reliably under climate variability. These findings underscore the importance of interaction-aware models for data-driven agricultural decision support systems.

Interpretability, however, does not require sacrificing robustness or performance. RuleFit trails Light- GBM by 5.4 pp under standard cross-validation, yet under rainfall-quartile shift it remains competitive while producing a concise rule list consisting of 11 rules. Such rules can be audited rapidly by extension officers and agronomists, enabling validation against local knowledge and management practices. A representative rule, IF Rainfall < 60 && pH < 5.5 THEN recommend Chickpea, illustrates how decision logic can be communicated transparently and meaningfully.

Bayesian Rule Sets further extend this interpretability by associating each rule with a posterior probability, providing an explicit measure of uncertainty alongside the recommendation. While slightly more complex than RuleFit, Bayesian Rule Sets remain within practical deployability constraints and offer additional value in risk-sensitive farming contexts.

Explanation quality is as important as transparency. LightGBM achieves a high SHAP Average Explanation Stability (AES = 0.92), indicating that its feature importance rankings remain consistent across bootstrap samples. This stability supports user trust in explanation-driven decision systems. Random Forest exhibits slightly lower stability (AES = 0.85), reflecting variability across individual trees. This suggests that aggregation strategies, such as averaging SHAP values across multiple seeds, may further improve explanation reliability.

Taken together, the results offer clear guidance for agri-tech deployment in climate-sensitive regions. LightGBM provides the strongest accuracy–robustness trade-off when predictive performance is the primary objective. RuleFit and Bayesian Rule Sets provide transparent, regulation-friendly alternatives that maintain robustness while enabling human oversight. Augmenting these models with external cli- mate signals, such as short-term rainfall forecasts derived from remote sensing, may further improve resilience. Moreover, deployment within a federated-learning framework would preserve farm-level data sovereignty while enabling collaborative model improvement, a critical requirement for trustworthy dig- ital agriculture.

7.0 Limitations and Future Work

7.1 Current Limitations

Several limitations of the present study should be acknowledged. First, the dataset is intentionally balanced, with each crop represented by exactly 100 samples. In real-world agricultural systems, however, crop distributions are highly imbalanced, as staple crops such as rice and wheat dominate cultivated areas. Models trained on balanced data may therefore overestimate performance for less frequently grown crops when deployed at scale, particularly in regions with skewed planting patterns.

Second, soil and weather measurements correspond to single-day snapshots and do not capture longer- term seasonal dynamics. Important climatic phenomena, including monsoon onset, cumulative rainfall, and prolonged drought conditions, are not explicitly modeled. Although the rainfall-quartile protocol captures variation in moisture regimes, it cannot fully represent temporal autocorrelation or transitions across growing seasons.

Third, all observations originate from four agro-climatic zones within India. Differences in soil taxonomy, micronutrient composition, and farming practices in other geographic regions are therefore not reflected in the data. As a result, the external validity of the findings across countries and continents remains an open question.

Finally, the interpretability analysis relies on TreeSHAP, which assumes conditional independence among input features when distributing feature importance. In practice, humidity and rainfall exhibit moderate correlation ($r = 0.27$). This dependence may introduce attribution bias and slightly inflate the perceived joint importance of moisture-related variables.

7.2 Future Research Directions

Future progress in crop recommendation will depend on integrating richer data sources and improving generalization across environments. Incorporating multi-modal inputs, such as satellite-derived vegetation indices, weekly weather forecasts, and soil micronutrient profiles, can provide additional context beyond the seven tabular features considered here. Evidence from crop yield prediction suggests that fusing remote sensing with ground-based data can improve performance by approximately 4–5 percentage points [16].

Domain adaptation techniques offer a complementary path forward. Partial adaptation frameworks can align feature distributions as models are transferred across agro-ecological zones, reducing the need for extensive local labeling [6]. This strategy is particularly relevant for scaling decision-support tools across regions with heterogeneous climate patterns.

Equally important is the infrastructure supporting model deployment. Federated learning provides a natural mechanism for preserving data ownership while enabling collaborative model improvement. Communication-efficient aggregation methods augmented with explainability constraints could synchronize models without exposing raw farm data [18]. In parallel, on-device uncertainty quantification using

Bayesian ensembles or probabilistic rule learners would allow recommendations to reflect farmers' risk tolerance, especially in rain-fed systems where incorrect advice can have severe consequences.

Finally, human-centered evaluation must become a core component of future research. User studies with extension officers and farmers can assess the cognitive load imposed by rule-based explanations versus SHAP visualizations. At scale, crop recommendation systems should optimize not only yield but also environmental outcomes, including fertilizer runoff, water consumption, and long-term soil health. Embedding sustainability and equity objectives within multi-objective optimization frameworks will help ensure that data-driven agriculture supports inclusive and climate-resilient development.

8.0 Conclusion

This study addresses three central challenges in data-driven crop recommendation: robustness to environ- mental variation, interpretability of model decisions, and stability of explanations. Under rainfall-quartile evaluation, tree-based ensembles retain approximately 93% accuracy, while linear and kernel-based models fall below 90%. These results demonstrate that conventional random cross-validation can substantially overestimate real-world performance in climate-sensitive agricultural settings. Interpretable models provide a viable alternative to black-box approaches. RuleFit achieves competitive performance while producing an 11-rule decision list and exhibiting only a 4 percentage-point accuracy drop under rainfall shift. Bayesian Rule Sets offer similar robustness while providing explicit uncertainty estimates at the rule level. TreeSHAP analysis confirms that rainfall, humidity, and soil potassium are consistently the most influential features, with LightGBM yielding the most stable explanations (AES = 0.92). Together, the findings define a practical design space for agri-tech decision support systems. Light- GBM is well suited for applications prioritizing predictive accuracy and robustness, while RuleFit and Bayesian Rule Sets are appropriate when transparency, auditability, and regulatory compliance are paramount. Future work should extend this framework through multi-modal sensing, cross-regional adaptation, and human-in-the-loop evaluation to advance crop recommendation toward sustainable, climate-resilient, and farmer-centered digital agriculture.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest
The authors declare that there are no known financial or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Food and Agriculture Organization of the United Nations, "The State of Food Security and Nutrition in the World 2023," https://www.fao.org/publications/sofi/2023, 2023.

[2] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," Sensors, vol. 18, no. 8, p. 2674, 2018.

[3] A. Kamilaris and F. X. Prenafeta-Boldu´, "Deep learning in agriculture: A survey," Computers and Electronics in Agriculture, vol. 147, pp. 70–90, 2018.

[4] A. Patel and K. Shah, "Crop recommendation using machine learning," in Proc. 5th Int. Conf. Computing Methodologies and Communication (ICCMC), 2021, pp. 843–848.

[5] N. Brahimi and I. Boukhalfa, "Lightgbm with bayesian optimisation for accurate crop recommendation," Journal of Electrical and Computer Engineering, pp. 1–9, 2023.

[6] Y. Ma and Z. Yang, "Improving the transferability of deep learning models for crop yield prediction: A partial domain adaptation approach," Remote Sensing, vol. 13, no. 21, p. 4423, 2021.

[7] M.-D. Yang, C.-H. Huang, and Y.-S. Lin, "Domain adaptation for multi-season rice grain moisture prediction," Computers and Electronics in Agriculture, vol. 200, p. 107105, 2023.

[8] C. Rudin, "Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.

[9] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming – a review," Agricultural Systems, vol. 153, pp. 69–80, 2017.

[10] K. Benke and B. Tomkins, "Future food-production systems: vertical farming and controlled- environment agriculture," Sustainability: science, practice and policy, vol. 13, no. 1, pp. 13–26, 2017.

[11] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep gaussian process for crop yield prediction based on remote sensing data," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 1, 2017.

[12] M. Shahhosseini, G. Hu, and S. V. Archontoulis, "Forecasting corn yield with machine learning ensembles," Frontiers in Plant Science, vol. 11, p. 1120, 2020.

[13] A. S. M. Saimon, M. Moniruzzaman, M. S. Islam, M. K. Ahmed, M. M. Rahaman, S. Hossain, and M. M. T. G. Manik, "Integrating genomic selection and machine learning: A data-driven approach to enhance corn yield resilience under climate change," Journal of Environmental and Agricultural Studies, vol. 4, no. 2, pp. 20–27, 2023.

[14] S. Chakraborty and S. Maity, "Crop recommendation system based on soil parameters using machine learning," in Proc. IEEE Calcutta Conf. (CALCON), 2020, pp. 368–372.

[15] A. Cartolano, A. Cuzzocrea, and G. Pilato, "Analyzing and assessing explainable ai models for smart agriculture environments," Multimedia Tools and Applications, vol. 83, pp. 37 225–37 246, 2024.

[16] F. Mena, D. Pathak, H. Najjar, C. Sanchez, P. Helber, B. Bischke, P. Habelitz, M. Miranda, J. Siddamsetty, M. Nuske, M. Charfuelan, D. Arenas, M. Vollmer, and A. Dengel, "Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction," Remote Sensing of Environment, vol. 318, p. 114547, 2025.

[17] Z. Lv, S. Yang, S. Ma, Q. Wang, J. Sun, L. Du, J. Han, Y. Guo, and H. Zhang, "Efficient deployment of peanut leaf disease detection models on edge ai devices," Agriculture, vol. 15, no. 3, p. 332, 2025.

[18] V. Hiremani, R. M. Devadas, Preethi, R. Sapna, T. Sowmya, P. Gujjar, N. S. Rani, and K. R. Bhavya, "Federated learning for crop yield prediction: A comprehensive review of techniques and applications," MethodsX, vol. 14, p. 103408, 2025.

[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[20] M. Bouni, B. Hssina, K. Douzi, and S. Douzi, "Interpretable machine learning techniques for an advanced crop recommendation model," Journal of Electrical and Computer Engineering, p. 7405217, 2024.

[21] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," Journal of Computational physics, vol. 378, pp. 686–707, 2019.

[22] A. Ingle, "Crop recommendation dataset," 2021, accessed 24 Jun 2025. [Online]. Available: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

[23] M. Wu, M. Hughes, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-Velez, "Beyond sparsity: Tree regularization of deep model for interpretability," in Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1, 2018.