

ARTICLE

Solar-Powered IoT–ML Framework for Energy-Aware Crop Suitability Prediction

Yusra Mansoor¹, Huma Jamshed^{1,*}, Mohammed Khouj², Muhammad I. Masud^{2,*}, Urooj Waheed¹, Abdul Wahid Memon³, Najeeb Ur Rehman Malik^{4,*} and Touqeer Ahmed Jumani⁵

¹Department of Computer Science, DHA Suffa University, Karachi, Pakistan

²Department of Electrical Engineering, College of Engineering, University of Business and Technology, Jeddah, Saudi Arabia

³Department of Computer Systems Engineering, QUEST, Nawabshah, Pakistan

⁴Department of Information Technology, DHA Suffa University, Karachi, Pakistan

⁵Department of Electrical Engineering and Computer Science, College of Engineering, A'Sharqiyah University, Ibra, Oman

*Corresponding Authors: Huma Jamshed. Email: huma.jamshed@dsu.edu.pk; Muhammad I. Masud. Email: m.masud@ubt.edu.sa; Najeeb Ur Rehman Malik. Email: najeeb.malik@dsu.edu.pk

Received: 21 May 2026; Accepted: 30 July 2026; Published: 15 September 2026

ABSTRACT: The increasing demand for sustainable and energy-aware agricultural practices due to climate change, limited natural resources, and increasing food requirements has accelerated the adoption of Internet of Things (IoT) and machine learning (ML) technologies in smart farming. This study proposes a solar-powered IoT and ML-based framework for energy-aware crop suitability prediction in precision agriculture. The system employs low-power IoT sensors to continuously monitor important environmental and soil parameters, including temperature, humidity, soil moisture, soil temperature, heat index, nitrogen (N), phosphorus (P), and potassium (K) levels. To reduce dependence on conventional energy sources, the framework is powered by solar energy, making it suitable for deployment in remote and resource-constrained agricultural environments. The collected sensor data are transmitted to a cloud-based platform for storage and real-time monitoring. Different ML models, including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), and Neural Network (NN), are implemented for crop suitability classification while evaluating their predictive performance and computational overhead. A real-world dataset containing 11,644 initial sensor records was collected from agricultural regions of Sindh, Pakistan, covering four major crops: cotton, rice, sugarcane, and wheat. After preprocessing, 11,643 valid records were retained for model training and evaluation. Since the original dataset was imbalanced, particularly for the Sugarcane class, tabular augmentation was applied exclusively to the training set, while the test set remained unchanged to ensure an unbiased evaluation. Experimental results demonstrate that the proposed framework provides reliable crop suitability prediction and supports data-driven agricultural decision-making under resource-constrained field conditions. Among the evaluated models, RF achieved the highest held-out test performance, with an accuracy of 94.3331%, precision of 94.5985%, recall of 94.3331%, and F1-score of 94.3727%. By integrating solar-powered IoT infrastructure with ML-based analytics, class-imbalance handling, and computational-efficiency evaluation, the proposed framework offers a practical and scalable solution for energy-aware suitability assessment in resource-constrained farming environments.

KEYWORDS: Solar energy; Internet of Things (IoT); machine learning; precision agriculture; energy-aware systems; smart agriculture; crop suitability prediction; class imbalance; sensor data analytics

1 Introduction

In many developing countries, agriculture remains an major economic sector, but unsustainable practices such as excessive fertilizer and pesticide use degrade soil quality and harm the environment [1]. This creates a need for energy-efficient farming practices that improve resource use and decision-making [2]. Internet of Things (IoT) devices enable real-time monitoring of soil and environmental conditions, while machine learning (ML) algorithms support agricultural decision-making [3,4]. Solar energy can further reduce dependence on conventional electricity sources in agricultural systems [5].

Cloud Computing (CC), Artificial Intelligence (AI), Data Analytics (DA), and IoT collectively support cloud-based storage, processing, and analytics for smart agriculture, enabling crop suitability patterns to be identified from soil and environmental data [6–9].

Although numerous studies have investigated IoT and ML-based smart agriculture systems, most focus primarily on monitoring or accuracy-focused crop prediction. Limited research has integrated solar-powered IoT sensing, real-field sensor data preprocessing, class-imbalance handling, and computational-efficiency evaluation in a single crop suitability prediction framework. In this study, crop suitability prediction refers to supervised classification of the suitable crop class from historical field-sensor observations. Since the dataset does not include explicit yield, profitability, or economic-return labels, the proposed model does not claim to optimize crop selection based on these objectives.

This study proposes a solar-powered IoT–ML framework for crop suitability prediction using humidity, temperature, heat index, soil moisture, soil temperature, nitrogen, phosphorus, and potassium data. The framework applies median imputation, training-only class-imbalance handling, and evaluates Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Neural Network (NN) using prediction and execution metrics such as training time, prediction time, latency, and throughput.

The major contributions of this study are: (1) a solar-powered IoT framework for crop suitability prediction using real-field sensor data; (2) invalid sensor data preprocessing and training-only class-imbalance handling; and (3) system-level evaluation of ML models in terms of prediction performance, latency, throughput, and energy-aware deployment considerations. Accordingly, the contribution is positioned as an engineering integration and evaluation framework rather than a fundamentally new predictive algorithm.

The existing studies related to IoT and ML in the agriculture sector are reviewed in [Section 2](#). The proposed methodology is presented in [Section 3](#). The experimental setup, description of the dataset and model configuration are discussed in [Section 4](#). The results and discussion are presented in [Section 5](#) and the conclusion and future work are presented in [Section 6](#).

2 Related Work

Crop productivity is strongly influenced by soil composition, temperature, humidity, and related environmental conditions [10–12]. In many farming regions, crop selection is still based on farmer experience or trial-and-error practices, which may reduce decision precision and increase production risk. Therefore, data-driven agricultural systems are increasingly needed to support crop suitability assessment using real-time soil and environmental information [13,14].

Considerable research has been conducted on energy-efficient farming technologies, IoT-based monitoring, and cloud-supported analytics to enhance agricultural productivity and resource management [15–17]. Current IoT, and WSN-based systems are primarily used for environmental monitoring, irrigation automation, farm management, and communication efficiency [18–20]. Likewise, ML-based studies have been conducted for crop prediction, yield estimation and crop suitability analysis with soil, weather and environmental attributes [21–23]. However, most of the previous works focus on monitoring

infrastructure or predictive accuracy, with limited consideration of real-field sensor preprocessing, class imbalance, computational efficiency and solar-supported deployment [24]. The key differences between previous studies and the proposed framework are summarized in Table 1.

Table 1: Comparison of existing studies with the proposed framework.

Study	Focus	IoT/WSN	Solar/ Energy-Aware	ML	Key Limitation
[19,20]	Smart agriculture monitoring	Yes	Limited	Limited	Mainly monitoring-focused
[15–17]	Energy-efficient/cybersecure sensing	Yes	Yes/Limited	Limited	Limited crop recommendation
[12,21,23]	Crop/yield prediction	Limited	No	Yes	Limited IoT-energy integration
[24]	Crop suitability clustering	Limited	No	Yes	No solar IoT deployment
Proposed	Crop suitability prediction	Yes	Yes	Yes	Integrates solar-supported IoT sensing, preprocessing, training-set augmentation, and efficiency evaluation

From the literature reviewed, three gaps are identified. First, there are only a few studies that integrate solar-powered IoT sensing with ML-based crop suitability prediction with real-field agricultural sensor data. Second, the accuracy of crop prediction studies is the primary focus but the studies are not thorough enough when it comes to invalid sensor readings, class imbalance, latency, and throughput. Third, limited work assesses ML models in terms of both prediction and computation in cloud-supported IoT agriculture. The proposed framework addresses these gaps at the system-integration level by consolidating sensing, real sensor data preprocessing, training-only imbalance handling, and computational-efficiency analysis into a single solar-supported crop suitability prediction workflow.

3 Proposed Methodology

The proposed solar-powered IoT agricultural system enables real-time monitoring of environmental and soil conditions to support data-driven agricultural decision-making. The system consists of two components, the first being a hardware platform using IoT to collect sensor data and transmit it into the cloud, and the second being the software platform to which the IoT data is sent, where the ML models are applied to predict crop suitability in real time. The recommended workflow for the experiments is depicted in Fig. 1.

The proposed approach starts with data acquisition from IoT devices, as depicted in Fig. 1. The raw dataset, collected for wheat, sugarcane, cotton, and rice crops, is stored in a database and preprocessed to remove missing values, noise and outliers. It then undergoes transformation and feature selection. Categorical features are transformed into numerical labels and selected features are chosen to build a refined

dataset. The dataset is transformed into a feature vector and split into training and testing sets using a stratified train-test split. Since the dataset is imbalanced, tabular augmentation is applied only to the training set, while the test set remains unchanged for unbiased evaluation. For training, different ML models are applied to the training set and evaluated using multiclass classification metrics. The complete crop suitability prediction workflow is presented in Algorithm 1.

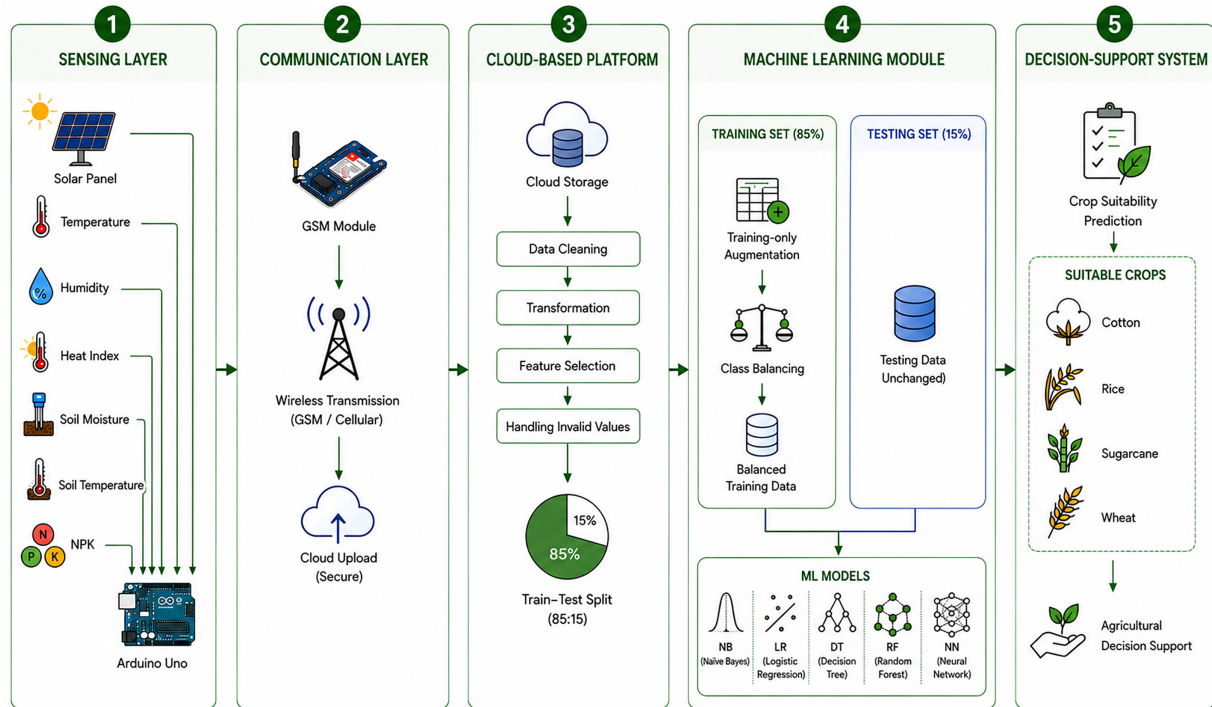


Figure 1: Proposed system architecture.

The hardware portion consists of a low power WSN system that measures environmental parameters by using low power communication. The network consists of sensors for measuring humidity, temperature, soil moisture, soil temperature, and NPK levels for continuous agricultural monitoring. This system employs the use of an Arduino Uno microcontroller as hardware platform to obtain data from these sensors in real time. After receiving data from the sensors, the microcontroller processes the information and transmits it to the cloud-based system via a GSM module for real-time monitoring and analysis. Energy-aware operation is supported by fixed-interval data transmission and microcontroller sleep mode during inactive periods. Solar panels are used to support field-node power supply and reduce dependence on conventional electricity sources. The cloud-based platform stores and processes the collected data, while the machine learning models are executed on the software platform for crop suitability prediction.

The proposed framework is based on an energy-aware sensing and decision-support workflow. The sensing layer collects field parameters with low-power sensors. The communication layer is responsible for sending the sensor data to the cloud at regular time intervals using the Arduino Uno and GSM module. The collected data is stored, cleaned, and transformed in the cloud layer for ML processing. The ML models are trained and evaluated in the intelligence layer for crop suitability prediction. The final decision-support layer outputs the predicted suitable crop class to assist farmers in field-level decision-making.

Overall, IoT data transmission, cloud storage, and ML-based analytics together support crop suitability prediction in the proposed system.

3.1 Data Preprocessing and Feature Preparation

3.1.1 Data Preprocessing

The analytical component of the proposed system utilizes data collected from IoT sensors. Let D represent the dataset containing n tuples, as expressed in Eq. (1)

$$D = \{d_1, d_2, d_3, \dots, d_n\} \quad (1)$$

The dataset consists of environmental and soil parameters collected from wheat, sugarcane, cotton, and rice crops for crop suitability prediction. The raw dataset is preprocessed to handle missing values, outliers, and inconsistencies before training the ML models. Inconsistent crop labels are standardized, while invalid negative sensor readings, such as -1 and -5 , are treated as missing values rather than valid agronomic measurements. These missing values are handled using median imputation to preserve crop-class representation and avoid removing complete crop categories. After preprocessing, the selected sensor attributes are combined into a feature vector, as shown in Eq. (2).

$$v_i = [f_1(x_i), f_2(x_i), \dots, f_m(x_i)] \quad (2)$$

The final feature vector includes humidity, temperature, heat index, soil moisture, soil temperature, nitrogen, phosphorus, and potassium values, while the target variable represents the crop type.

3.1.2 Data Split and Class-Imbalance Handling

The dataset was divided into training D_{train} and testing D_{test} sets using stratified sampling, as expressed in Eq. (3)

$$D_{\text{train}}, D_{\text{test}} = \text{StratifiedSplit}(D, [\text{training ratio}, \text{testing ratio}], \text{seed}) \quad (3)$$

Different train-test ratios were evaluated, and the best performance was achieved using 85% training data and 15% testing data. A fixed random seed was used to ensure reproducibility of the experimental results. The original dataset showed class imbalance, particularly for the Sugarcane class. Therefore, tabular data augmentation was applied only to the training set. Minority classes were oversampled using replacement, and small Gaussian noise was added to numerical sensor features to introduce limited sensor-level variation during training. This Gaussian training augmentation was retained in the main pipeline because it preserves the original training-set structure while adding only small sensor-level variation for minority classes. The testing set was not augmented and was kept unchanged to ensure unbiased evaluation. To avoid dependence on a single imbalance-handling strategy, SMOTE-like resampling was also included as a sensitivity comparison. The augmented training dataset is generated from the original training subset, as expressed in Eq. (4).

$$D_{\text{train}}^{\text{aug}} = \text{Augment}(D_{\text{train}}) \quad (4)$$

where $D_{\text{train}}^{\text{aug}}$ represents the balanced training dataset used for model learning.

3.1.3 Model Selection and Configuration

Five supervised ML algorithms, namely NB, LR, DT, RF, and NN, were used for crop suitability prediction. NB, LR, and DT were included as baseline classifiers, RF was added as an ensemble-based model

to evaluate the effect of tree aggregation, and NN was included to examine nonlinear learning behavior from sensor-based crop data. These algorithms were selected because they are commonly used for multiclass classification and can be evaluated in terms of both predictive performance and computational overhead in cloud-supported IoT environments.

NB is a probabilistic classifier based on Bayes theorem, represented in Eq. (5)

$$P(y|V) = \frac{P(V|y) \cdot P(y)}{P(V)} \quad (5)$$

LR is used for binary and multiclass classification problems. The sigmoid activation function used in LR is represented in Eq. (6)

$$f(x) = \frac{1}{1 + e^{-x}} \quad (6)$$

DT utilizes hierarchical decision structures for classification based on feature splitting and impurity reduction. Entropy for DT classification is represented in Eq. (7). RF extends DT by aggregating predictions from multiple decision trees. For multiclass classification, the RF prediction is obtained through majority voting, as expressed in Eq. (8).

$$S = \sum_{i=1}^Y -p_i \log_2 p_i \quad (7)$$

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_T(x)\} \quad (8)$$

where $h_T(x)$ represents the prediction of the T -th decision tree, and $\hat{y}$ is the final class predicted by majority voting.

NN consists of interconnected layers that process nonlinear relationships between input and output variables. The hidden layer computation is represented in Eq. (9).

$$h_j = f\left(\sum_{i=1}^n w_{ij}^{(1)} v_i + b_j^{(1)}\right) \quad (9)$$

Table 2 compares the selected ML models. The algorithms were evaluated using accuracy, precision, recall, and F1-score. In addition, execution-related metrics, including training time, prediction time, latency, and throughput, were recorded to assess computational suitability for cloud-supported IoT-based crop suitability prediction.

Table 2: Comparison of ML algorithms and final hyperparameter settings.

ML Algorithm	Type	Criteria	Final Hyperparameter Settings
LR	Linear classifier	Regularized multinomial log-loss optimization	family = multinomial; maxIter = 100; regParam = 0.01; elasticNetParam = 0.0
NB	Probabilistic classifier	Bayesian probability principles	modelType = multinomial; smoothing = 1.0

(Continued)

Table 2 (continued)

ML Algorithm	Type	Criteria	Final Hyperparameter Settings
DT	Tree-based classifier	Impurity measures	impurity = entropy; maxDepth = 10; seed = 42
RF	Ensemble tree-based classifier	Bootstrap aggregation and majority voting	numTrees = 100; maxDepth = 10; impurity = entropy; seed = 42
NN	Neural network classifier	Backpropagation optimization	layers = [8, 16, 8, 4]; maxIter = 200; blockSize = 128; seed = 42

Different configurations of NB, LR, DT, RF, and NN were explored during model optimization. Hyperparameters such as smoothing for NB, regularization for LR, tree depth for DT, number of trees and tree depth for RF, hidden-layer size, activation function, learning rate, and maximum iterations for NN were evaluated. Model comparison was based on both classification performance and computational overhead.

Algorithm 1: Crop suitability prediction

Require: Sensor dataset D containing crop labels and soil–environmental features

Ensure: Predicted crop suitability class $\hat{y}$ for new sensor input

- 1: Standardize crop labels in D
 - 2: Select input features: humidity, temperature, heat index, soil moisture, soil temperature, N , P , and K
 - 3: Treat invalid negative sensor readings as missing values
 - 4: Apply median imputation to missing numerical values
 - 5: Convert selected features into vector form X and crop labels into target vector y
 - 6: Split (X, y) into D_{train} and D_{test} using stratified 85:15 sampling
 - 7: Balance only D_{train} by minority-class oversampling with small Gaussian noise
 - 8: Keep D_{test} unchanged for final held-out evaluation
 - 9: **for** each classifier $M \in NB, LR, DT, RF, NN$ **do**
 - 10: Train M using the augmented training set
 - 11: Predict crop labels on the unchanged test set
 - 12: Compute accuracy, precision, recall, F1-score, latency, and throughput
 - 13: **end for**
 - 14: Compare model performance using held-out testing and complementary cross-validation
 - 15: Select the final model according to the defined evaluation setting
 - 16: Predict crop suitability class $\hat{y}$ for new sensor-based input
-

4 Experiment Setup

4.1 Data Description

The dataset was collected from different agricultural regions of Sindh, Pakistan, including Shah Matiari, Shah Alam Shah, and Tando Allahyar, between 2023 and 2024. The selected crops include cotton, rice, sugarcane, and wheat. The dataset contains environmental and soil parameters such as nitrogen (N), phosphorus (P), potassium (K), temperature, heat index, soil moisture, soil temperature, and humidity. In total, the dataset comprises 11,644 initial sensor records. After removing one record with a missing crop label

and applying median imputation for invalid negative sensor readings, 11,643 valid records were retained for analysis. After preprocessing and feature analysis, 85% of the data was used for training, while the remaining 15% was reserved for testing. [Table 3](#) presents the summary of the crop suitability prediction dataset.

Table 3: Summary of the crop suitability prediction dataset.

Attribute	Description
Data source	Real-field IoT sensor data
Location	Shah Matiari, Shah Alam Shah, and Tando Allahyar, Sindh, Pakistan
Collection period	2023–2024
Initial records	11,644
Valid records after preprocessing	11,643
Crops	Cotton, Rice, Sugarcane, Wheat
Input features	Humidity, Temperature, Heat Index, Soil Moisture, Soil Temperature, N, P, K
Target variable	Crop type
Preprocessing method	Negative sensor readings treated as missing values and handled using median imputation
Train-test split	85:15

4.2 Class Distribution and Training-Set Augmentation

The original dataset was imbalanced, especially for the Sugarcane class. There were 5598 Cotton records, 4366 Wheat records, 1434 Rice records and 245 Sugarcane records. This class imbalance may cause the ML models to overfit the majority classes while underfitting the minority crop classes. Thus, tabular data augmentation was used only for the training set. Replacement was used to oversample minority crop classes and small Gaussian noise was added to numerical sensor features to introduce limited sensor-level variation. The test set was not expanded and remained unchanged to provide unbiased evaluation. The class distribution before and after training-set augmentation is shown in [Table 4](#).

Table 4: Class distribution before and after training-set augmentation.

Crop	Original Records	Training before Augmentation	Training after Augmentation	Testing Records
Cotton	5598	4758	4758	840
Wheat	4366	3711	4758	655
Rice	1434	1219	4758	215
Sugarcane	245	208	4758	37

After augmentation, the training dataset contained 19,032 records, with 4758 records for each crop class. The testing dataset contained 1747 original records and was used only for final model evaluation.

4.3 Experimental Environment and Model Configuration

The experiments were carried out using Apache Spark and PySpark MLlib to facilitate scalable data processing and the development of ML models. Spark was chosen due to its in-memory processing functionality, which supports efficient processing of sensor-based agricultural data. Five supervised ML models were developed and trained for crop suitability prediction: LR, DT, NB, RF, and NN. These models were selected for multiclass classification and allow comparison between probabilistic, linear, tree-based, ensemble-based, and neural learning approaches.

The models were trained on the augmented training set and tested on the original test set. Various model configurations were investigated to enhance the prediction performance, such as regularization parameters for LR, depth adjustment for DT, smoothing parameters for NB, number of trees and tree depth for RF, and learning rate and hidden layer parameters for NN. In the final configuration, RF was implemented using 100 trees with a maximum depth of 10, while the NN used two hidden layers with 16 and 8 neurons, respectively. To reduce split-dependent bias, 5-fold cross-validation was conducted, and accuracy and weighted F1-score were reported as mean $\pm$ standard deviation. For statistical comparison, paired t-tests were applied to fold-level weighted F1-scores obtained from the same five folds. Since only five folds were available and multiple pairwise comparisons were performed, the p -values were interpreted cautiously as supportive evidence rather than conclusive proof. In addition, the training-set augmentation strategy was compared with SMOTE-like resampling to examine the effect of imbalance handling. The accuracy, precision, recall, F1-score, training time, prediction time, latency and throughput were used to evaluate the final models.

In order to allow a fair computational comparison, prediction time was recorded after forcing Spark to execute using an action operation. Latency was computed as the mean prediction time for each test record and throughput was defined as the number of sensor records classified per second. The reported latency values therefore represent Spark-executed batch inference time normalized per test record, rather than end-to-end field-device communication latency.

5 Results and Discussion

5.1 Evaluation Metrics

The proposed crop suitability prediction task was evaluated as a multiclass classification problem with four crop classes: wheat, rice, cotton, and sugarcane. PySpark MLlib evaluators were used to compute classification and execution-related metrics. The selected input feature set included Humidity, Temperature, Heat Index, Soil Moisture, Soil Temperature, Nitrogen, Phosphorus, and Potassium.

The input feature vector for the i -th sensor record is expressed in Eq. (10).

$$\mathbf{x}_i = [\text{Humidity}_i, \text{Temperature}_i, \text{HeatIndex}_i, \text{SoilMoisture}_i, \text{SoilTemperature}_i, N_i, P_i, K_i] \quad (10)$$

where $\mathbf{x}_i$ represents the selected feature vector for record i . The true output vector y and predicted output vector $\hat{y}$ are represented by Eqs. (11) and (12), respectively.

$$y = [y_1, y_2, \dots, y_n] \quad (11)$$

$$\hat{y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n] \quad (12)$$

Accuracy (ACC) measures the proportion of correctly predicted crop classes, as expressed in Eq. (13).

$$\text{ACC} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(\hat{y}_i = y_i) \quad (13)$$

Precision, recall, and F1-score were computed for each crop class and then summarized using weighted averages. For a class c , these metrics are expressed in Eqs. (14)–(16).

$$Precision(c) = \frac{TP_c}{TP_c + FP_c} \quad (14)$$

$$Recall(c) = \frac{TP_c}{TP_c + FN_c} \quad (15)$$

The F1-score is the harmonic mean of precision and recall and is expressed in Eq. (16).

$$F1(c) = \frac{2 \times Precision(c) \times Recall(c)}{Precision(c) + Recall(c)} \quad (16)$$

In addition to classification performance, computational efficiency was evaluated using training time, prediction time, latency, and throughput. In this study, latency represents Spark-executed batch prediction time normalized per test record, while throughput represents the number of sensor records classified per second. These metrics are expressed in Eqs. (17) and (18).

$$Latency = \frac{T_{pred}}{N_{test}} \quad (17)$$

$$Throughput = \frac{N_{test}}{T_{pred}} \quad (18)$$

5.2 Crop Suitability Prediction Performance

Table 5 presents the 5-fold cross-validation performance of the evaluated models using mean $\pm$ standard deviation. Neural Network achieved the highest mean cross-validation accuracy and F1-score, indicating stronger fold-level stability, followed by Random Forest. However, in the unchanged held-out test evaluation, Random Forest achieved the highest performance with 94.3331% accuracy and 94.3727% F1-score. Therefore, the cross-validation and held-out test results are interpreted as complementary rather than contradictory.

Table 5: Cross-validation performance of ML models for crop suitability prediction.

Model	Accuracy (%, Mean $\pm$ SD)	Precision (%, Mean $\pm$ SD)	Recall (%, Mean $\pm$ SD)	F1-Score (%, Mean $\pm$ SD)
Neural Network	95.88 $\pm$ 0.50	95.81 $\pm$ 0.49	95.88 $\pm$ 0.50	95.80 $\pm$ 0.49
Random Forest	94.55 $\pm$ 0.43	94.69 $\pm$ 0.36	94.55 $\pm$ 0.43	94.53 $\pm$ 0.44
Decision Tree	93.51 $\pm$ 0.33	93.55 $\pm$ 0.33	93.51 $\pm$ 0.33	93.48 $\pm$ 0.34
Logistic Regression	68.62 $\pm$ 0.54	68.79 $\pm$ 0.57	68.62 $\pm$ 0.54	65.85 $\pm$ 0.65
Naive Bayes	50.12 $\pm$ 0.54	53.08 $\pm$ 1.44	50.12 $\pm$ 0.54	42.75 $\pm$ 0.66

The comparative performance of the evaluated models and the confusion matrix of the Random Forest model are shown in Figs. 2 and 3, respectively. Random Forest achieved the highest accuracy, precision, recall, and F1-score on the held-out test set among the evaluated models. This superior performance suggests that ensemble-based aggregation can better capture nonlinear relationships among soil moisture, temperature, and NPK-related parameters compared with single-model baselines.

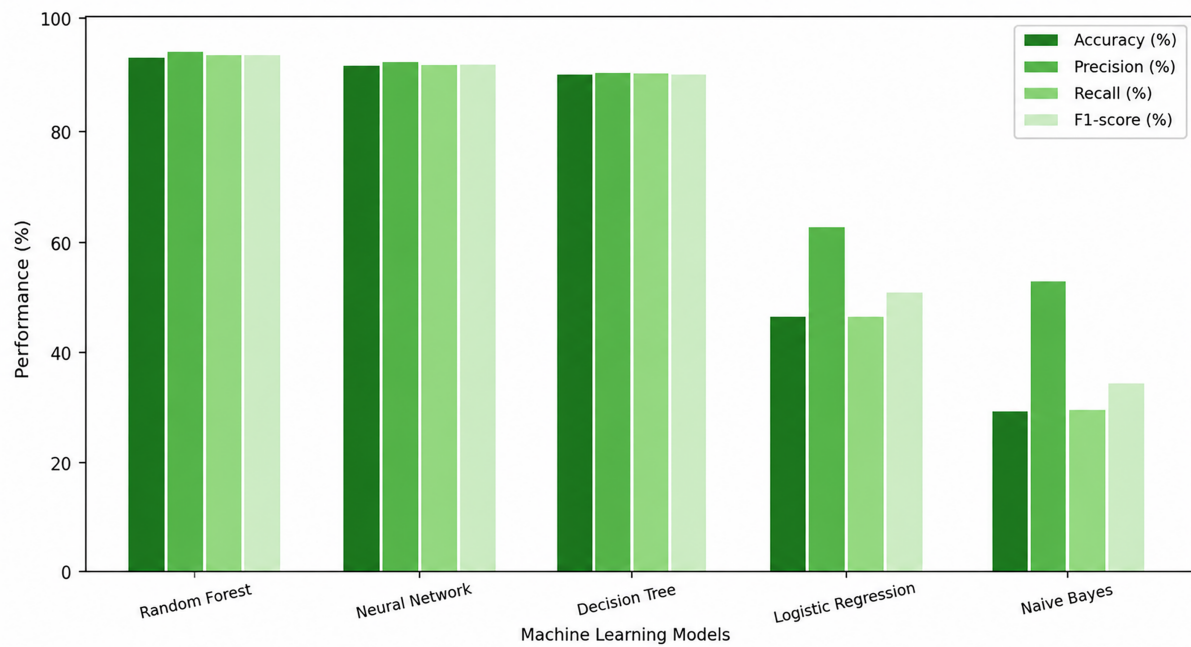


Figure 2: Performance comparison of ML models for crop suitability prediction.

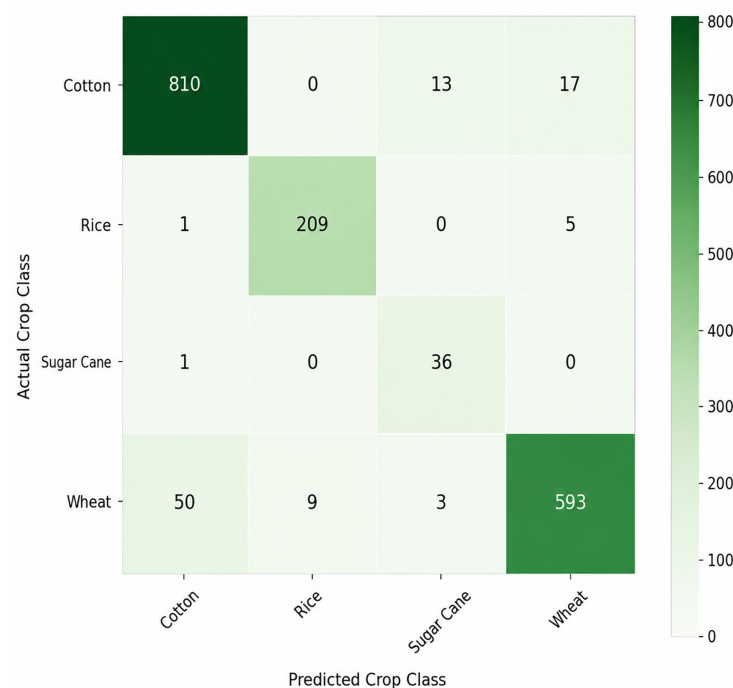


Figure 3: Confusion matrix of the best-performing Random Forest.

The per-class evaluation further shows that the Random Forest model achieved strong F1-scores for Cotton, Rice, and Wheat, with values of 0.9518, 0.9654, and 0.9339, respectively. For the minority Sugar

Cane class, the model achieved a recall of 0.9730 and an F1-score of 0.8090; however, its precision remained comparatively lower at 0.6923 due to the limited number of test samples for this class. This indicates that the model was able to detect most Sugar Cane instances, but some misclassification from other classes remained.

To further examine robustness rather than replace the final test-set ranking, 5-fold cross-validation and SMOTE-like resampling were also conducted. The unchanged held-out test set was treated as the primary final evaluation because it was not augmented and remained fixed across models. In cross-validation, NN achieved the highest mean F1-score of $95.80 \pm 0.49\%$, followed by RF with $94.53 \pm 0.44\%$, indicating stronger fold-level stability for NN, while RF remained the leading held-out test model. Paired t-tests based on fold-level weighted F1-scores suggested differences between NN and RF ($p = 0.0040$), DT ($p = 0.0004$), LR ($p < 0.001$), and NB ($p < 0.001$); however, these results are interpreted cautiously due to five folds and multiple pairwise comparisons. SMOTE-like resampling improved RF performance to 95.6497% accuracy and 95.6394% F1-score, indicating that alternative imbalance-handling methods may be preferable under some conditions, while the held-out test results are discussed separately to support final model interpretation.

5.3 Execution Time, Latency, and Throughput Analysis

In addition to prediction performance, execution-time metrics were evaluated to assess the computational suitability of the models for cloud-supported IoT-based crop suitability prediction. The augmented training dataset contained 19,032 records, while the original testing dataset contained 1747 records. The training data were balanced to 4758 records per crop, whereas the testing set remained unchanged to ensure unbiased evaluation. The execution time, latency, and throughput comparison of the evaluated ML models is presented in [Table 6](#).

Table 6: Execution time, latency, and throughput comparison of ML models.

Model	Training Time (s)	Prediction Time (s)	Total Time (s)	Latency (ms/record)	Throughput (records/s)
Random Forest	18.3124	1.2189	19.5313	0.6977	1433.2351
Neural Network	20.6272	0.2870	20.9142	0.1643	6086.9930
Decision Tree	3.2454	0.5627	3.8081	0.3221	3104.8671
Logistic Regression	11.0235	0.5177	11.5412	0.2964	3374.3628
Naive Bayes	3.8219	2.5452	6.3671	1.4569	686.3852

[Table 6](#) reports the computational performance of the evaluated models in terms of training time, prediction time, total execution time, latency, and throughput. Random Forest achieved the highest predictive performance, but required higher prediction time than Neural Network, Decision Tree, and Logistic Regression because it aggregates multiple decision trees during inference. Neural Network achieved the lowest latency of 0.1643 ms/record and the highest throughput of 6086.9930 records/s, while Naive Bayes showed the highest latency and lowest throughput. Findings highlight a trade-off between predictive accuracy and inference efficiency, where Random Forest is preferable for accuracy-focused deployment, whereas Neural Network provides faster batch inference under the evaluated Spark-based environment.

The combined latency and throughput comparison of the evaluated ML models is presented in [Fig. 4](#). Neural Network achieved the lowest latency and highest throughput, indicating faster batch inference under the Spark-based setting. Random Forest achieved the strongest predictive performance, but required higher inference latency because it aggregates multiple decision trees. Naive Bayes exhibited the highest latency and

the lowest throughput among the evaluated models under the same Spark-based experimental setting. These results indicate a trade-off between predictive performance and inference efficiency.

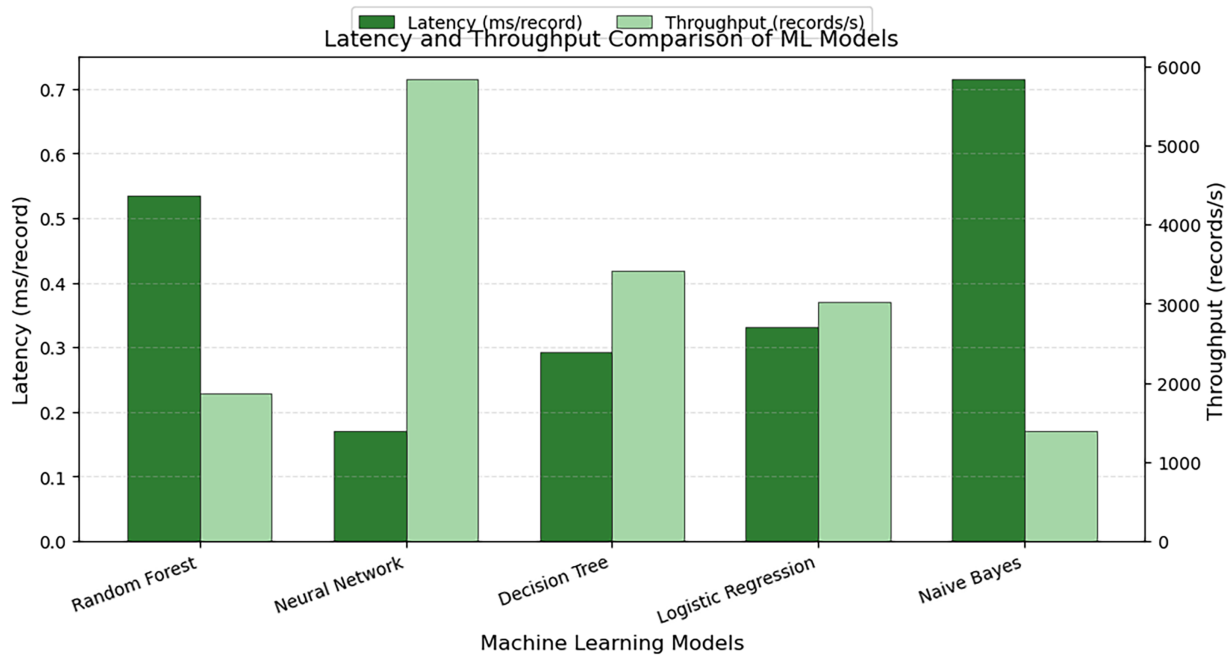


Figure 4: Combined latency and throughput comparison of ML models for crop suitability prediction.

5.4 Energy-Aware Design Analysis

The proposed framework supports energy-aware field monitoring through solar-powered sensing, low-power sensors, periodic data transmission, and cloud-supported ML processing. The use of solar panels reduces dependence on conventional electricity sources, while periodic transmission reduces unnecessary GSM communication overhead. In addition, shifting ML computation to the cloud reduces the computational burden on field-level IoT devices. The energy-aware design features of the proposed framework are summarized in [Table 7](#).

Table 7: Energy-aware design features of the proposed framework.

Component/Feature	Energy-Aware Role
Solar panel	Provides renewable power support for field deployment
Low-power sensors	Support lower sensing-level energy demand
Arduino Uno sleep mode	Supports reduced power use during inactive periods
Periodic transmission	Limits continuous GSM communication overhead
Cloud-based processing	Shifts ML computation away from resource-limited field nodes
ML models	Compares predictive performance and computational efficiency for cloud-supported crop suitability prediction

The energy-aware design makes the proposed framework suitable for resource-constrained agricultural environments where stable electricity supply may not always be available. However, the present study does

not report direct power consumption in watts or watt-hours. Therefore, the energy contribution is positioned as an energy-aware system design rather than a measured energy-saving evaluation.

5.5 Discussion

The results demonstrate that the proposed framework can support crop suitability prediction by integrating IoT-based sensing, data preprocessing, class-imbalance handling, ML-based classification, and computational-efficiency evaluation. Median imputation was used to preserve class distribution and prevent the significant loss of Sugarcane samples. Data augmentation was applied only to the training set to mitigate class imbalance and improve model learning. The test set remained unchanged to ensure an unbiased performance evaluation.

RF achieved the strongest performance on the unchanged held-out test set, which was used as the primary final evaluation setting. Its high accuracy and F1-score suggest that ensemble-based tree aggregation can capture nonlinear relationships among the selected sensor features. NN achieved the highest mean cross-validation F1-score, suggesting stronger fold-level stability; however, RF is reported as the held-out test leader rather than as a universally superior model across all validation settings. Logistic Regression and Naive Bayes performed comparatively lower, likely because they are less suitable for nonlinear and correlated agricultural sensor features such as temperature, soil moisture, and NPK values.

From the computational point of view, NN achieved the lowest latency and the highest throughput, whereas RF needed higher inference cost because of ensemble aggregation. Thus, the results should be interpreted as a trade-off among held-out test performance, cross-validation stability, and inference efficiency: RF is suitable when held-out test performance is prioritized, while NN is preferable when fold-level stability and Spark-based batch inference speed are emphasized.

The present study has some limitations as the number of real Sugarcane samples is small and the evaluation was conducted in selected areas of Sindh, Pakistan with four crop classes. While the training-set augmentation helps to balance the classes, further validation in other regions, seasons, crop types, and climatic conditions is needed before the results can be extended to other agricultural settings. The quantitative measurement of power consumption should also be added to future implementations to confirm the energy-saving advantages of the solar-powered IoT deployment.

6 Conclusion & Future Work

This study presented a solar-powered IoT and ML-based framework for energy-aware crop suitability prediction in precision agriculture. The proposed system monitors important environmental and soil parameters, including humidity, temperature, heat index, soil moisture, soil temperature, nitrogen, phosphorus, and potassium. A real-field dataset containing 11,644 initial sensor records was collected from agricultural regions of Sindh, Pakistan, covering four major crops: cotton, rice, sugarcane, and wheat. After preprocessing, 11,643 valid records were retained for model training and evaluation.

The proposed framework addresses practical challenges associated with real-world agricultural sensor data by handling invalid negative sensor readings, missing values, and class imbalance. Median imputation was used to handle invalid sensor readings, while training-set augmentation was applied to reduce minority-class imbalance without altering the original test set. The experimental results demonstrated that Random Forest achieved the highest held-out test performance among the evaluated models, with 94.3331% accuracy, 94.5985% precision, 94.3331% recall, and 94.3727% F1-score. Execution-time analysis further showed that the NN achieved the lowest latency and highest throughput, indicating a trade-off between predictive performance and inference efficiency in cloud-supported IoT-based crop suitability prediction.

Since the present evaluation is limited to selected regions of Sindh, Pakistan, and four crop classes, future work will extend the framework to additional crops, larger agricultural regions, and seasonal variations to improve model generalization. Future enhancements may include collecting more Sugarcane samples, integrating weather forecast data, deploying a farmer-facing mobile application, and conducting quantitative power-consumption analysis to further assess the energy-aware benefits of the solar-powered IoT deployment.

Acknowledgement: None.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Yusra Mansoor, Mohammed Khouj and Huma Jamshed; methodology, Yusra Mansoor, Huma Jamshed and Muhammad I. Masud; software, Mohammed Khouj, Touqeer Ahmed Jumani and Huma Jamshed; validation, Mohammed Khouj, Najeeb Ur Rehman Malik and Touqeer Ahmed Jumani; formal analysis, Urooj Waheed; investigation, Yusra Mansoor, Mohammed Khouj and Touqeer Ahmed Jumani; resources, Yusra Mansoor and Abdul Wahid Memon; data curation, Yusra Mansoor and Najeeb Ur Rehman Malik; writing—original draft preparation, Yusra Mansoor and Huma Jamshed; writing—review and editing, Huma Jamshed; visualization, Urooj Waheed; supervision, Abdul Wahid Memon and Najeeb Ur Rehman Malik; project administration, Abdul Wahid Memon and Muhammad I. Masud; funding acquisition, Muhammad I. Masud. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data available on request from the authors.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang X. Managing land carrying capacity: key to achieving sustainable production systems for food security. *Land*. 2022;11(4):484. doi:10.3390/land11040484.
2. Wakweya RB. Challenges and prospects of adopting climate-smart agricultural practices and technologies: implications for food security. *J Agric Food Res*. 2023;14(1):100698. doi:10.1016/j.jafr.2023.100698.
3. Syari MA, Rahardja U, Wellem T, Purnomo HD, Buaton R. IoT enabled smart farming system for optimizing crop management using sensors and machine learning. In: *Proceedings of the 2025 4th International Conference on Creative Communication and Innovative Technology (ICCIT)*; 2025 Aug 15–16; Kota Cirebon, Indonesia. p. 1–7. doi:10.1109/ICCIT65724.2025.11167551.
4. Bendahmane T, Rezeg K, Merizig A, Kazar O, Harous S. IoT and deep learning-based approach for an efficient land suitability prediction in smart farms. *Int J Comput Digit Syst*. 2025;17(1):1–14. doi:10.12785/ijcds/1571062172.
5. Dhanaraju M, Chenniappan P, Ramalingam K, Pazhanivelan S, Kaliaperumal R. Smart farming: Internet of Things (IoT)-based sustainable agriculture. *Agriculture*. 2022;12(10):1745.
6. Sawant NR, Kumar A, Pant S, Kotecha K. An IoT-driven machine learning system for real-time smart crop recommendation and optimization in precision agriculture. *Discover Artif Intell*. 2026;6(1):194. doi:10.1007/s44163-026-00896-y.
7. Dayioglu MA, Turker U. Digital transformation for sustainable future-agriculture 4.0: a review. *Tarim Bilimleri Dergisi*. 2021;27(4):373–99.
8. Shaikh TA, Rasool T, Lone FR. Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. *Comput Electron Agric*. 2022;198:107119. doi:10.1016/j.compag.2022.107119.

9. Ali A, Hussain T, Tantashutikun N, Hussain N, Cocetta G. Application of smart techniques, Internet of Things and data mining for resource use efficient and sustainable crop production. *Agriculture*. 2023;13(2):397. doi:10.3390/agriculture13020397.
10. Nalawade A, Kulkarni M, Kamble S. Smart agriculture: comparative analysis of ML models for crop yield prediction. In: *Proceedings of the 2025 6th International Conference on Electronics and Sustainable Communication Systems (ICESC)*; 2025 Sep 10–12; Coimbatore, India. p. 2114–9.
11. Roy P, Padhiary M, Hoque A, Sahu B, Roy D, Kumar K. Machine learning for precision agriculture and crop yield optimization: techniques and applications. Hershey, PA, USA: IGI Global; 2025. p. 189–234. doi:10.4018/979-8-3693-9208-9.ch007.
12. Screpnik CR, Zamudio E, Gimenez LI. Artificial intelligence in agriculture: a systematic review of crop yield prediction and optimization. *IEEE Access*. 2025;13:70691–7. doi:10.1109/access.2025.3560631.
13. Vemuri N, Thaneeru N, Tatikonda VM. Smart farming revolution: harnessing IoT for enhanced agricultural yield and sustainability. *J Knowl Learn Sci Technol*. 2023;2(2):143–8.
14. Condran S, Bewong M, Islam MZ, Maphosa L, Zheng L. Machine learning in precision agriculture: a survey on trends, applications and evaluations over two decades. *IEEE Access*. 2022;10():73786–803. doi:10.1109/access.2022.3188649.
15. Manono BO, Mwami B, Mutavi S, Nzilu F. Precision farming with smart sensors: current state, challenges and future outlook. *Sensors*. 2026;26(3):882. doi:10.3390/s26030882.
16. Liang C, Shah T. IoT in agriculture: the future of precision monitoring and data-driven farming. *Eig Rev Sci Technol*. 2023;7:85–104.
17. Križanović V, Grgić K, Spišić J, žagar D. An advanced energy-efficient environmental monitoring in precision agriculture using LoRa-based wireless sensor networks. *Sensors*. 2023;23(14):6332. doi:10.3390/s23146332.
18. Shafik W, Tufail A, Apong RA, De Silva LC. Internet of Things for smart agricultural practices. In: *Internet of Things applications and technology*. Boca Raton, FL, USA: Auerbach Publications; 2024. p. 190–217.
19. Hassan SI, Alam MM, Illahi U, Al Ghamdi MA, Almotiri SH, Su'ud MM. A systematic review on monitoring and advanced control strategies in smart agriculture. *IEEE Access*. 2021;9:32517–48.
20. Almalki FA, Soufiene BO, Alsamhi SH, Sakli H. A low-cost platform for environmental smart farming monitoring system based on IoT and UAVs. *Sustainability*. 2021;13(11):5908. doi:10.3390/su13115908.
21. Ahmed KR, Sarkar SC. Machine learning applications in modern agriculture: advancing precision farming through intelligent systems. Cham, Switzerland: Springer Nature; 2026. p. 233–55. doi:10.1007/978-3-032-12770-9_11.
22. Elbasi E, Mostafa N, AlArnaout Z, Zreikat AI, Cina E, Varghese G, et al. Artificial intelligence technology in the agricultural sector: a systematic literature review. *IEEE Access*. 2023;11():171–202. doi:10.1109/access.2022.3232485.
23. Yang X, Hua Z, Li L, Huo X, Zhao Z. Multi-source information fusion-driven corn yield prediction using the Random Forest from the perspective of agricultural and forestry economic management. *Sci Rep*. 2024;14(1):4052. doi:10.1038/s41598-024-54354-9.
24. Noor D, Jamshed H, Memon N, Iqbal S. Artificial intelligence-based approach in precision agriculture for crop yield prediction. *Int J Inf Syst Comput Technol*. 2025;5(1):65–79. doi:10.58325/ijisct.005.01.00144.