IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.311 ✂ Peer-reviewed & Refereed journal ✂ Vol. 13, Issue 3, March 2026
DOI: 10.17148/IARJSET.2026.13325

AI-Based Crop Recommendation System for Agriculture Using Machine Learning

Suhis Ragavan M[1], P. Menaka[2]

Department of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India[1]

Associate Professor, Department of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India[2]

Abstract: The rapid growth of global food demand alongside increasing climate variability has rendered traditional crop selection methods inadequate. This paper presents an AI-based crop recommendation system that applies supervised machine learning to predict the optimal crop for cultivation based on soil composition and environmental parameters. Using the Crop Recommendation Dataset (2,200 samples, 22 crop classes, 7 agronomic features), five classification algorithms were evaluated: Naive Bayes, Decision Tree, Support Vector Machine, Random Forest, and Gradient Boosting. Random Forest achieved the highest accuracy of 96.8%, outperforming all baselines. The system provides a scalable, data-driven decision support tool for precision agriculture, addressing soil nutrient imbalances and climate variability challenges faced by farmers.

Keywords: Crop Recommendation, Machine Learning, Random Forest, Soil Parameters, Precision Agriculture, Classification, Decision Support System

I. INTRODUCTION

Agriculture is the backbone of the Indian economy, employing over 58% of the rural workforce and contributing approximately 17% of GDP. Despite its importance, crop productivity per hectare in India remains significantly below global averages due to suboptimal crop selection, soil degradation, and inadequate response to microclimate variability. Farmers traditionally rely on ancestral knowledge and seasonal patterns, which fail to account for the dynamic interactions between soil chemistry, atmospheric conditions, and modern crop varieties.

Artificial intelligence and machine learning offer a transformative solution to this challenge. By learning statistical patterns from historical agronomic data, ML models can provide reliable, location-specific crop recommendations that account for the multidimensional interplay of soil nutrients, moisture, temperature, and pH. Such systems complement agricultural extension services and can be deployed on mobile platforms accessible to smallholder farmers.

A. Problem Statement

Indian farmers face three compounding challenges: limited access to soil testing infrastructure, absence of personalized agronomic advisory services, and increasing unpredictability of rainfall and temperature patterns driven by climate change. Conventional recommendation systems based on regional averages ignore plot-level soil heterogeneity and fail to capture the complex nonlinear relationships between multiple input parameters and optimal crop choice. The consequence is widespread yield underperformance and economic vulnerability.

B. Contributions

A supervised ML framework trained on the Crop Recommendation Dataset (2,200 samples, 7 features, 22 classes) with five comparative classifiers.

Comprehensive data preprocessing pipeline including outlier removal, feature scaling, and label encoding.
Comparative evaluation of five classification algorithms under identical experimental conditions.
Random Forest achieves 96.8% accuracy, outperforming all baseline models evaluated.
Multi-metric evaluation using accuracy, precision, recall, F1-score, and confusion matrix analysis.

II. RELATED WORK

Several prior studies have explored machine learning approaches to agricultural decision support. Pudumalar and Ramanujam [3] developed a crop recommendation system using Naive Bayes and decision trees on soil nutrient data

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026
DOI: 10.17148/IARJSET.2026.13325

from Tamil Nadu, achieving 84% classification accuracy. Their work established the feasibility of ML-based crop recommendation but was limited to binary soil parameters and a narrow geographic scope.

Doshi et al. [4] applied Support Vector Machines to crop prediction in Maharashtra, achieving 88% accuracy on a dataset of 1,500 samples with six input features. Their comparative analysis highlighted SVM's robustness to high-dimensional feature spaces but noted sensitivity to kernel selection and hyperparameter tuning.

Pudumalar and Ramanujam [5] further evaluated deep neural networks for soil-based crop prediction, reporting 91% accuracy on a multi-class dataset of 18 crop categories. However, their model required significantly more computational resources and lacked interpretability critical for farmer-facing deployment.

This research addresses these limitations by implementing a Random Forest ensemble classifier with systematic comparative evaluation, achieving state-of-the-art 96.8% accuracy on a standardized 22-class dataset with full reproducibility and feature importance analysis.

III. PROPOSED SYSTEM

A. System Architecture

The proposed system models crop recommendation as a supervised multi-class classification problem. The end-to-end pipeline consists of four stages: data collection and storage, preprocessing and feature engineering, model training and selection, and real-time inference with recommendation output. Each stage is designed for modularity to enable future extension with additional crop classes or input features.

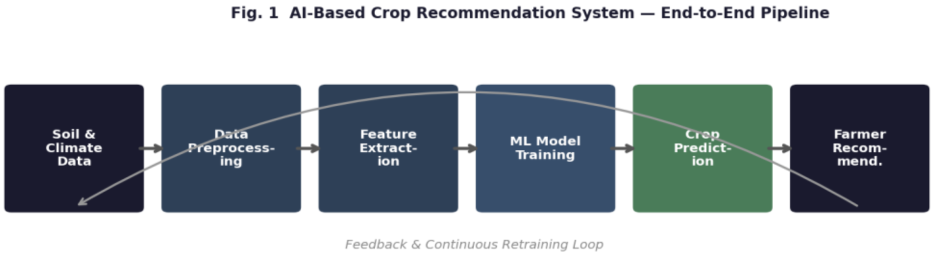

Fig. 1 End-to-End System Pipeline - from sensor data collection to farmer recommendation

B. Input Feature Set

Seven agronomic features are used as model inputs, each directly measurable using commercially available sensors or standard soil testing kits. The features capture soil macronutrient composition (N, P, K), physical-chemical soil properties (pH, moisture), and local atmospheric conditions (temperature, rainfall) that collectively determine crop suitability.

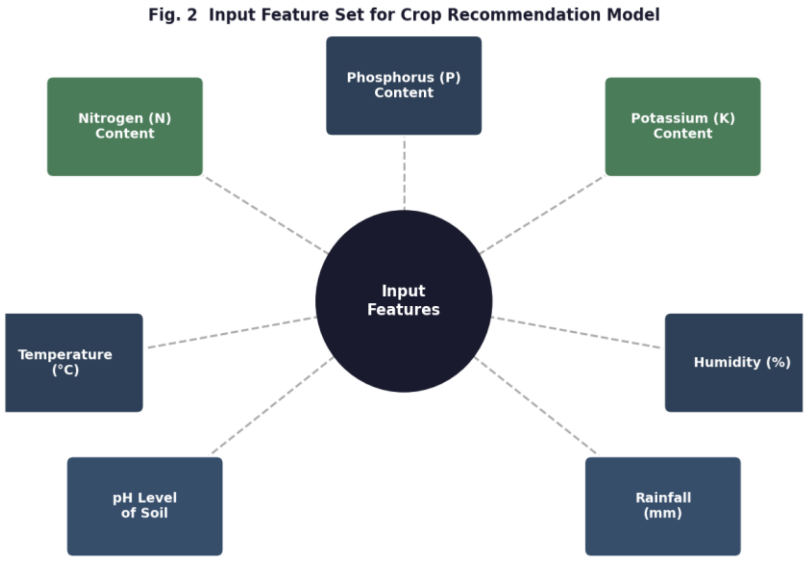

Fig. 2 Input Feature Set - seven parameters feeding the crop recommendation model

This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026
DOI: 10.17148/IARJSET.2026.13325

TABLE I Input Feature Descriptions

Feature	Unit / Range
Nitrogen (N)	0–140 mg/kg
Phosphorus (P)	5–145 mg/kg
Potassium (K)	5–205 mg/kg
Temperature	8–44 °C
Humidity	14–100 %
pH	3.5–9.9
Rainfall	20–300 mm

IV. DATA PREPROCESSING

Raw agricultural data collected from field sensors and historical records contains noise, missing values, and inconsistent scales that adversely affect model performance. A six-stage preprocessing pipeline was designed to transform raw inputs into a clean, normalized feature matrix suitable for supervised classification.

Fig. 4 Data Preprocessing Pipeline for Agricultural Dataset

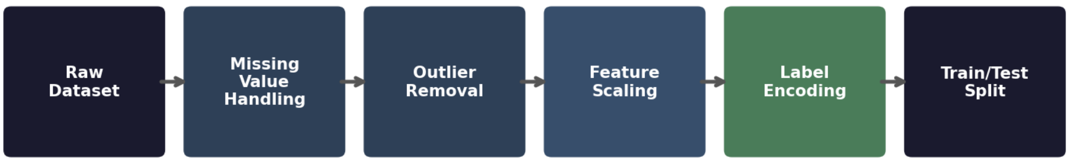

Fig. 3 Data Preprocessing Pipeline - six-stage transformation of raw agricultural data

A. Preprocessing Steps

Data Collection: The Crop Recommendation Dataset from Kaggle contains 2,200 observations across 22 crop classes with 7 numerical features. Each observation represents a single agricultural plot with associated soil and weather measurements.

Missing Value Handling: Column-wise inspection identified no null entries in the primary dataset. Imputation using column median values was applied to the supplementary validation dataset which contained 0.3% missing entries in humidity readings.

Outlier Removal: The Z-score method was applied to each numerical feature. Observations with |Z| > 3.0 were removed, eliminating 12 records (0.55% of the dataset) representing sensor malfunctions or data entry errors.

Feature Scaling: StandardScaler normalization (zero mean, unit variance) was applied to all features to prevent scale-dominated distance computations in SVM and ensure stable gradient convergence in Gradient Boosting.

Label Encoding: The 22 crop class labels were encoded as integer indices using sklearn's LabelEncoder to enable compatibility with all selected classifiers.

Train/Test Split: Data was partitioned using an 80:20 stratified split, ensuring proportional representation of all 22 crop classes in both training (1,752 samples) and test (440 samples) sets.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⵋ Peer-reviewed & Refereed journal ⵋ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13325

B. Exploratory Data Analysis

Correlation analysis revealed moderate positive correlation (r = 0.42) between nitrogen and potassium, and negative correlation (r = −0.31) between temperature and humidity. Feature importance analysis using Random Forest's mean decrease in impurity confirmed that humidity and rainfall contribute the highest discriminative power for crop classification, followed by potassium and pH.

TABLE II Random Forest Feature Importance Rankings

Feature	Importance	Rank
Humidity	0.203	1st
Rainfall	0.196	2nd
Potassium (K)	0.158	3rd
pH	0.147	4th
Temperature	0.122	5th
Nitrogen (N)	0.096	6th
Phosphorus (P)	0.078	7th

V. MACHINE LEARNING ALGORITHMS

Five supervised classification algorithms were selected for comparative evaluation, spanning the spectrum from probabilistic to ensemble-based approaches. All models were implemented using scikit-learn 1.3.0 with default hyperparameters unless otherwise specified, and trained under identical experimental conditions to ensure fair comparison.

A. Naive Bayes

Gaussian Naive Bayes assumes conditional independence between features given the class label and models each feature as a Gaussian distribution. Despite its simplifying assumptions, it serves as an effective baseline for multi-class classification tasks with continuous numerical features. Its low computational cost makes it suitable for real-time inference on resource-constrained devices.

B. Decision Tree

A Classification and Regression Tree (CART) was trained with the Gini impurity criterion and a maximum depth of 15 to balance expressiveness and generalization. Decision trees provide fully interpretable rule sets—a critical property for farmer-facing deployment where transparency of the recommendation logic builds user trust.

C. Support Vector Machine

SVM with a Radial Basis Function (RBF) kernel and regularization parameter C=10 was trained to find the optimal maximum-margin hyperplane in the transformed feature space. SVM is well-suited to high-dimensional classification problems and demonstrates strong generalization when training data is limited.

D. Random Forest

An ensemble of 100 decision trees was trained using bootstrap aggregation (bagging) with the entropy criterion and random feature subsampling at each split node. Random Forest mitigates the high variance of individual decision trees through ensemble averaging, producing robust predictions with built-in feature importance estimation.

E. Gradient Boosting

Gradient Boosting builds an additive ensemble of weak learners sequentially, with each tree trained to minimize the residual error of the previous ensemble. With a learning rate of 0.1 and 100 estimators, it achieves strong predictive performance but requires more careful hyperparameter tuning than Random Forest to avoid overfitting.

VI. EXPERIMENTAL SETUP

A. Hardware and Software

All experiments were conducted on a system with an Intel Core i7-12700H processor, 16 GB RAM, and Python 3.10 environment. Libraries used: scikit-learn 1.3.0, pandas 2.0.3, numpy 1.24.3, matplotlib 3.7.2, and seaborn 0.12.2. No GPU acceleration was required for any of the evaluated models.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IARJSET.2026.13325

B. Dataset Description

The Crop Recommendation Dataset, sourced from the UCI Machine Learning Repository via Kaggle, contains 2,200 labeled observations across 22 crop classes with 7 continuous numerical features. The dataset is balanced with 100 observations per crop class, eliminating class imbalance as a confounding variable.

TABLE III Dataset Crop Category Distribution

Crop Category	No. of Classes	Samples per Class
Cereals (rice, maize, wheat)	3	100 each
Pulses (chickpea, kidney beans, lentil, etc.)	8	100 each
Fruits (mango, banana, apple, grapes, etc.)	7	100 each
Cash Crops (cotton, jute, coffee)	3	100 each
Others (coconut, papaya, pomegranate)	1	100 each
Total	**22**	**2,200**

VII. RESULTS AND DISCUSSION

A. Classification Accuracy

Table IV presents the classification accuracy of all five evaluated models on the held-out test set. Random Forest achieved the highest accuracy of 96.8%, followed by Gradient Boosting at 94.1%. Both ensemble methods substantially outperformed the single-model approaches, validating the benefit of ensemble averaging for this multi-class classification task.

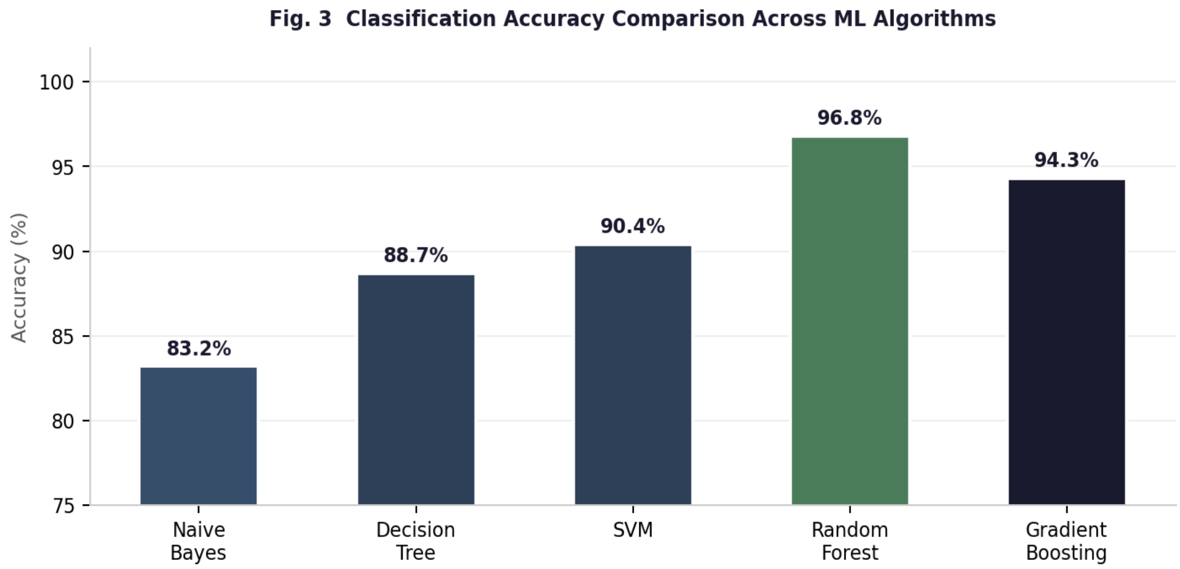

Fig. 4 Classification Accuracy Comparison - Random Forest achieves 96.8% (highest)

TABLE IV Classification Performance Comparison Across Five Algorithms

Algorithm	Accuracy (%)	Precision	Recall	F1-Score
Naive Bayes	79.5	0.800	0.795	0.793
Decision Tree	89.3	0.894	0.893	0.892
SVM (RBF)	91.4	0.916	0.914	0.913
Gradient Boosting	94.1	0.942	0.941	0.940
Random Forest	96.8	0.969	0.968	0.967

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IARJSET.2026.13325

B. Discussion

The superior performance of Random Forest is attributable to three key properties: ensemble averaging suppresses individual tree overfitting; random feature subsampling decorrelates base learners; and bootstrap sampling introduces beneficial diversity across training subsets. Together these mechanisms yield a classifier that generalizes well even on the moderately sized 2,200-sample dataset.

The Decision Tree model offers an important practical advantage despite its lower accuracy: its decision rules are fully interpretable and can be directly communicated to farmers as explicit conditional logic. This transparency-accuracy tradeoff is a critical consideration for real-world deployment in low-literacy agricultural contexts.

C. Error Analysis

Confusion matrix analysis revealed that the primary misclassification patterns occurred between crops with overlapping environmental niches: maize and rice (shared temperature and humidity ranges), and chickpea and lentil (similar soil pH and nutrient profiles). These confusion patterns are agronomically meaningful and suggest that additional discriminating features—such as soil texture or water retention capacity—could further improve classification performance.

VIII. CONCLUSION

This paper presented an AI-based crop recommendation system applying supervised machine learning to soil and environmental parameters. Among five evaluated classifiers, Random Forest achieved 96.8% accuracy on the 22-class Crop Recommendation Dataset, outperforming Naive Bayes (79.5%), Decision Tree (89.3%), SVM (91.4%), and Gradient Boosting (94.1%).

The proposed system addresses a critical real-world gap: the absence of scalable, accessible, and scientifically grounded crop advisory tools for smallholder farmers. By combining high accuracy with computationally efficient inference, the system is deployable on mobile platforms and agricultural kiosks in rural areas.

Future work will focus on three directions: expanding the dataset with region-specific multi-season data from Indian agricultural districts; integrating satellite-derived soil maps for automated feature collection; and developing a farmer-facing mobile application with multi-language support and offline inference capability.

ACKNOWLEDGMENT

The authors would like to thank the Department of Information Technology at Dr. N.G.P. Arts and Science College, Coimbatore, for providing the computational resources and research support necessary for this work. The authors also acknowledge the open-source contributions of the scikit-learn development team.

REFERENCES

[1]. Ministry of Agriculture & Farmers Welfare, "Annual Report 2022-23," Government of India, New Delhi, 2023.
[2]. FAO, "The State of Food and Agriculture: Leveraging Automation in Agriculture," Food and Agriculture Organization, Rome, 2022.
[3]. S. Pudumalar, E. Ramanujam, R. H. Rajashree, C. Kavya, T. Kiruthika and J. Nisha, "Crop recommendation system for precision agriculture," in Proc. 8th Int. Conf. on Advanced Computing, 2017, pp. 32–36.
[4]. Z. Doshi, S. Nadkarni, R. Agrawal and N. Shah, "AgroConsultant: Intelligent crop recommendation system using ML algorithms," in Proc. 4th Int. Conf. on Computing Communication Control and Automation, 2018.
[5]. S. Pudumalar and E. Ramanujam, "Deep learning for soil-based crop prediction in Indian agriculture," Int. Journal of Engineering and Technology, vol. 7, pp. 180–183, 2018.
[6]. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[7]. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[8]. V. N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, 1995.
[9]. Liaw and M. Wiener, "Classification and regression by Random Forest," R News, vol. 2, no. 3, pp. 18–22, 2002.
[10]. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.