Pertanika J. Sci. & Technol. 34 (4): 2127 - 2147 (2026)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Explainable Crop and Soil-Based Recommendation System for Next-Generation Smart Agriculture Using Model Classification

Himani Trivedi[1]*, Hetal Chauhan[2], Suresh Patel[2], Mahendra N Patel[3], Pradip Patel[4], Mahi Parmar[1], and Chrisha Dabhi[1]

[1]*Department of Computer Engineering, LDRP Institute of Technology and Research, Kadi Sarva Vishwavidyalaya, 382015 Gandhinagar, Gujarat, India*
[2]*Department of Information Technology, Government Engineering College, 382028 Gandhinagar, Gujarat, India*
[3]*Department of Information Technology, Government Engineering College, 383315 Modasa, District Aravalli, Gujarat, India*
[4]*Department of Information Technology, L. D. College of Engineering, 380015 Ahmedabad, Gujarat, India*

ABSTRACT

In India, selecting the right crop is essential due to diverse Agro-climatic conditions, such as soil types and socio-economic factors. Traditional crop recommendation methods have evolved and advanced by utilising the recent technologies in Artificial Intelligence (AI), Machine Learning (ML), the Internet of Things (IoT), and computer vision, which have enabled smart crop recommendation systems based on soil and climatic conditions. This review compares the methodologies, input features, performance evaluation metrics, limitations, and practical applicability by examining the publicly available agricultural datasets, including soil-characteristics, environmental factors, and previously harvested crops available in the open agricultural datasets from sources like Open Government Data Platform India (n.d.) and Kaggle. This study identifies key research limitations, including the drawbacks of model generalisation or integrating varied data, explainability, scalability, and real-world applications in the agricultural field. By this structured literature, best performing model was identified on classification of the standard dataset into classes as rice, maize, chickpea, kidney beans, pigeon peas, moth beans, mung bean, black gram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute and coffee, by comparing the state of art models viz. SVM, K-Nearest Neighbours, Random Forest, Gaussian Naive Bayes, XGBoost, Gradient Boosting, AdaBoost, LightGBM, Bagging, and CatBoost.

ARTICLE INFO

Article history:
Received: 22 May 2026
Accepted: 16 July 2026
Published: 13 August 2026

DOI: https://doi.org/10.47836/pjst.34.4.08

E-mail addresses:
himani_ce@ldrp.ac.in (Himani Trivedi)
hetalchauhan@gecg28.ac.in (Hetal Chauhan)
sbpatelit@gecg28.ac.in (Suresh Patel)
mahendra.patel@gecmodasa.ac.in (Mahendra N Patel)
pradippatel@ldce.ac.in (Pradip Patel)
mahi_24548@ldrp.ac.in (Mahi Parmar)
chrisha_24489@ldrp.ac.in (Chrisha Dabhi)
* Corresponding author

e-ISSN: 2231-8526

These models were evaluated on the following performance metrics as precision, recall, F1-score, accuracy and model inference time amongst which Gaussian Naive Bayes results to be accurately precise and fastest, making it effective for real-time applications.

Keywords: Boosting, climate, crop prediction, deep learning, machine learning

INTRODUCTION

Agriculture contributes significantly on nations' economy and sustenance, making the need for precision and efficiency in crop recommendation more pronounced. The parameters on which the best crop cultivations depend are humidity, P-N-K nutrient levels, rainfall, temperature and pH values of soil. Farmers face problems in selecting the most suitable crop for optimal yield and profitability because of India's vast diversity in the types of soil, agroclimatic conditions and socio-economic factors (Afzal et al., 2025). This necessity has become increasingly evident with continuous variation in climate patterns, fluctuating market demands, and the imperative to ensure food security for a burgeoning population. The traditional soil prediction techniques used by farmers are based on their intuitions and experiences, which sometimes fail to manage the soil properly, due to which the crop production being less than expected (Boudouda & Boukallel, 2025). However, such methods are subjective, limited in scope, and may not always leverage the full potential of available data and technology. Additionally, the conventional approach does not adequately account for the impacts of climate change, market trends and advances in agronomic research. By using Machine learning for the integration of microbiome sequencing, a tool is used to increase the prediction of soil by 18% as compared to the traditional soil prediction techniques (Gambhir et al., 2025). Dahiphale et al. (2025) provide an overview of the highly accurate Neural Network Models by performing on the Kaggle dataset (Ingle, 2020). However, the accuracy still depends on clean, well-labelled data, which is not feasible in real life. It was observed that the existing research has explored machine learning approaches, which include XGBoost and Neural Network for a crop recommendation system, which demonstrated promising predictive performance measures (Gawade et al., 2023). As machine learning uses advanced data processing and prediction techniques, the patterns to study large sets of information like soil quality, weather, past crop results, market trends and farmer conditions can be improved to suggest the best crop choices for production, which presents challenges that result in enhancing agricultural productivity, sustainability and resilience.

LITERATURE REVIEW

Agriculture plays a crucial role in global food security as crop management is highly essential for maximising crop yield and sustainability. With the advancement in technologies like Machine Learning, Deep Learning, and Internet of Things, intelligent systems are proficient in analysing complex agricultural data, including the variations in climatic conditions and crop characteristics. Suggestions for crop cultivation, yield forecasting, soil classification, and resource management are made using these intelligent systems. By integrating sensor-based data acquisition techniques, the evaluation of existing solutions featuring selection methods and predictive model techniques has obtained highly accurate and reliable outcomes. Table 1 is the literature review showcasing the assessment of the existing approaches, articulating their key algorithms, datasets, performance metrics and identifying the gaps for future research in the agricultural sector.

The study on crop prediction, recommendation, and soil analysis was done by demonstrating diverse applications of machine learning (ML) and deep learning (DL) techniques combined with IoT-enabled data collection. Models such as Random Forest, XGBoost, and LightGBM ensembles various research by achieving high predictions, often outperforming traditional models' decision trees, KNN, and nonlinear regression (Bouni et al., 2022; Cema & Kaliappan, 2022; Kiruthika & Karthika, 2023; Priyadharshini et al., 2021). Meta-heuristic optimisation and feature selection methods, including IDSCO, CJO, SMOTE, and Grey Wolf Optimisation, further enhance selection performance and reliability (Alshahrani et al., 2023; Gautam et al., 2022; Raja et al., 2022). Enabling models such as WLSTM, ANN, and SVM to adapt to regional conditions by IoT-based soil and environmental data was frequently leveraged for real-time or near-real-time analysis (Chauhan et al., 2023 ; Teja et al., 2022). While many approaches achieved remarkable accuracy, most were limited by dataset size, regional specificity, or the number of features considered, indicating potential challenges in generalisation (Jeyanathan et al., 2023; Ramya & Sankar, 2023; Swetha et al., 2023). These studies demonstrate the efficiency of hybrid ML-DL integrated with an optimised data preparation pipeline and sensor-based data collection for precise crop recommendation, soil classification, and yield prediction, offering insights for the development of scalable, intelligent agricultural decision support systems.

Table 1
Literature review

Dataset	Performance Measures	Key Points	Limitations
Soil dataset public; crop dataset from Kaggle (Siddharth, 2021).	Accuracy (Crop Recommendation): RF 98.63%, Soil Fertility: RF 92.61%	Random Forest outperforms all other models.	Limited area-specific crop diversity.
Dataset built from augmented rainfall, climate, and fertiliser data from India (Siddharth, 2021).	Accuracy: ANN -87.71%, IDCSO-WLSTM-92.68% Precision: ANN-82.06%, IDCSO-WLSTM 90.88% Recall: ANN-84.78%, IDCSO-WLSTM-91.98% Execution time(sec): ANN-250.47, IDCSOWLSTM 241.0484	Improves prediction accuracy up to 95% through proper feature selection.	A dataset limited to one region lacks generalisation.
Public datasets for yield, cost, soil nutrients, and rainfall (Kaggle, n.d.).	Accuracy: DT-81%, KNN-85%, KNN-CV-88%, LR-88.26%, NB-82%, NN-89.88%, SVM-78%	Considering environmental and soil factors, training time is reduced.	Focuses on profit over sustainability.
Private dataset collected via sensors.	Accuracy: KNN 90%	Crop forecast improves KNN prediction accuracy (Sivanandam et al., 2023).	Prediction depends on data quality.
Dataset based on a specific region in Tamil Nadu (Cema & Kaliappan, 2022).	Random Forest - MAE:0.158, MSE:3.982, R2:0.99, Accuracy:99.98%	Random Forest achieves high accuracy and low error metrics.	Needs attention to soil health characteristics.
Soil characteristics obtained from Madurai district lab; multiple crops included.	Accuracy: DRL = RF > DT > NB > KNN; DRL/RF = 99.9%	DRL accuracy improves with larger datasets; RF also performs well.	Data reliability affects DRL performance.
Kaggle dataset (Ingle, 2020).	Accuracy = 99.5%, Recall/F1=99%, Validation Accuracy =97.73%	Highly accurate Neural Network Model (Meenachi et al., 2022).	Depends on clean, well-labelled data, which isn't available in the real world.
Private Dataset.	Accuracy = 95.7%, Recall = 96.3%, F1=95%, AUC=0.97	Accurate Prediction shows significant improvement.	Relies on region-specific datasets, and multiple ML models make it computationally heavy and complex for real life.

Table 1 (*continued*)

Dataset	Performance Measures	Key Points	Limitations
Dataset (Bhattacharya & Solomatine, 2006).	Accuracy-99.47%, Precision-94.60%, Recall 94.93%, F1-94.71%	Outperforms recent approaches with higher accuracy.	Not tested on large-scale real-time datasets.
Manually collected dataset, not publicly available.	Accuracy: RF 87.43%, Kappa 85.16%, Precision 90.34%, Recall 89.12%, F1 89.72%, AUC 92.39%	RF has the highest accuracy, precision, and specificity.	Forecasting may need a broader application.
Public dataset. (Ingle, 2020)	Accuracy: XGBoost 99.32%	XGBoost outperforms all models (Desai & Ansari, 2023).	Limited crop variety.
Open Government rainfall dataset (Open Government Data (OGD) Platform India, n.d.)	Accuracy: XGBoost 99.545%	High test-set accuracy (Sagar et al., 2023).	Depends on historical rainfall data.
Dataset with Rain, Temp, soil pH, moisture, crop type, crop yield (Kaggle, n.d.).	Random Forest: Accuracy 92%	RF gives the best results among three Tested algorithms (Kavitha et al., 2022).	Lack of comparison and quantisation.
Public generic crop dataset (Bhavana & Rao, 2025).	Accuracy: ANN 93%, LSTM 97%, Proposed SVM 99%	The proposed SVM model achieves the highest accuracy.	Models have different parameters; comparison is limited.
Dataset based on a specific region in Tamil Nadu	Random Forest - MAE: 0.158, MSE: 3.982, R2: 0.99, Accuracy: 99.98%	Random Forest achieves high accuracy and low error metrics.	Needs attention to soil health characteristics.
The dataset includes soil (N, P, K), rainfall, temperature, and pH.	Naive Bayes 89.3%, SVM 76.6%, RF 99%	RF highest accuracy (Ramachandra et al., 2023).	No real-time data is used.
Private dataset collected using sensors combined with public datasets.	Random Forest: Accuracy 99.39%	Provides detailed management guidance and forecasts.	Limited crop database.

Table 1 (*continued*)

Dataset	Performance Measures	Key Points	Limitations
Not disclosed	XGBoost: Crop 99%, RF Fertilizer 95.7%, MobileNet Disease 92%	Automates recommendations with high accuracy (Indira et al., 2022).	Needs additional sensors/ automation for full scaling.
Kaggle dataset (Ingle, 2020)	Naive Bayes / Random Forest: Accuracy, Precision, Recall, F1 = 99%	High-accuracy models (Meenachi et al., 2022).	Can be enhanced with IoT.
Public dataset with multiple crop types.	AUC=1, CA=99.6%, F1=99.6%, Precision=99.6%, Recall=99.6%	Extremely high accuracy.	Requires mobile deployment.
In laboratory-condition images with a simple background, the authors note that future work should add field-collected images across varying quality and growth stages. PlantVillage public apple subset, 3,171 images, 4 classes, 70-20-10 split, 2,228 train, 634 validation, 319 test	Accuracy 98%, Precision 0.97, Recall 0.97, F1 0.97, ROC-AUC 0.99. Per-class AUC Scab 0.993, Black Rot 0.998, Cedar Rust 0.998, Healthy 0.996.	Benchmarked against 9 pretrained models, VGG-19, ResNet-152, DenseNet-201, MobileNetV2, ResNet-50, VGG-16, InceptionV3, Xception, and MobileNet, needs only 121,964 parameters and 11 MB storage, enabling handheld and edge deployment (Vishnoi et al., 2023).	Trained only on laboratory-condition images with a simple background, the authors note that future work should add field-collected images across varying quality and growth stages.
12,000 IoT sensor records, 7,500 field-collected in India and 4,500 from public sources, 4-class crop stress labels, 80/20 split	Reported accuracy is inconsistent. Abstract and Conclusion state 98.54%, Results report 96.73% with Precision 95.80%, Recall 95.21%, F1 95.50%, significant versus SVM, Random Forest, LSTM-only, and CNN-only baselines, p less than 0.05, 94% for a separate 4-class breakdown. 96.73% is best supported since it is the only figure tied to a full metrics and baseline comparison. Sustainability gains, water use efficiency 1.8 to 2.6 kg per cubic meter, water usage reduced 29.05%, fertiliser usage reduced 21.28%, energy consumption reduced 22.12%, Sustainability Index improved 0.54 to 0.76, a 40.74% increase	Combines spatial and temporal IoT sensor fusion with Whale Optimisation-based adaptive threshold tuning, validated on real Indian farm deployments, ties accuracy to measurable sustainability outcomes (Ramu et al., 2026).	Depends on low-cost sensor accuracy limits, DHT11 plus or minus 2 degrees Celsius and or minus 5% relative humidity, FC-28 plus or minus 3%, the accuracy discrepancy between abstract and results tables is a genuine inconsistency in the source paper

Table 1 (*continued*)

Dataset	Performance Measures	Key Points	Limitations
Limited to tomato leaf diseases only, reliance on curated Kaggle imagery may limit generalisation to uncontrolled field conditions.	Accuracy 99.11%, Precision, Recall, and F1 approximately 99%, outperformed MobileNet 87.44% and VGG19 95.50%, 10-fold cross-validation training accuracy 99.41%, validation accuracy 98.88%	Integrates Grad-CAM, Grad-CAM++, LIME, and SHAP for explainability, deployed as a Flask web application (Assaduzzaman et al., 2025).	Limited to tomato leaf diseases only, reliance on curated Kaggle imagery may limit generalisation to uncontrolled field conditions.
Crop and Fertiliser Dataset, Western Maharashtra, India, IoT-enabled nitrogen, phosphorus, potassium, temperature, and humidity data	TabNet achieved 95.24% accuracy for fertiliser recommendation and 96.21% accuracy for crop recommendation, outperforming conventional classifiers.	Unifies fertiliser and crop recommendation in one interpretable system, attention mechanism removes the need for manual feature selection (Venkateswara & Padmanaban, 2025).	Single-region dataset limited to Western Maharashtra, with a higher computational cost than simpler classifiers used elsewhere in this table
Regional meteorological station data, including temperature, solar radiation, and other variables	Random Forest performed best, RMSE 0.52 mm per day, R-squared 0.96, temperature and solar radiation identified as key predictors	Systematically tests reduced-input scenarios for data-scarce regions, addressing the heavy data requirements of the FAO-56 Penman-Monteith method (Farag, 2025).	Static, non-sequential machine learning models may miss temporal weather dependencies captured by LSTM-based approaches, validated on a single region only
IoT-based soil salinity sensor data, including electrical conductivity, total dissolved solids, and pH from salt-affected fields.	Proposed leaching-based reclamation is shown to more effectively reduce soil salinity than the traditional reclamation method, but specific accuracy or error values are not detailed in the available abstract.	Addresses a soil management challenge largely absent from mainstream crop-recommendation literature, combining IoT salinity sensing with an ensembled LSTM for site-specific guidance (Khan et al., 2022).	Focused specifically on saline soils, not generalizable to standard agricultural contexts, requires specialised soil sensors beyond typical sensor suites.

DATASET

Exploratory Data Analysis was performed to understand the dataset before proceeding with model comparison, to identify the statistical data of the variables and to inspect the features and crop-specific environmental patterns. This provided a comprehensive overview of the dataset and helps in a reliable machine learning framework for crop recommendation.

Dataset Description

This review utilises the publicly available Crop Recommendation Dataset, which is a benchmark dataset for crop recommendation systems. The dataset was available on Kaggle, and it contains 2,200 rows of observations, which are equally distributed amongst 22 crop classes (Singh, 2023). The dataset is a balanced distribution of observations, having 100 samples for each crop category, which reduces the data preprocessing and provides an unbiased comparison among different classification algorithms.

This dataset contains seven numerical predictor variables having soil nutrient composition and climatic conditions, such as Nitrogen (N), Phosphorus (P), Potassium (K), temperature, humidity, soil pH, and rainfall, with the crop classes as the target variable. These predictor and target variables together showcase the soil fertility and environmental factors required for the recommendation of the most suitable crop.

The target variable comprises 22 crop classes, such as apple, banana, blackgram, chickpea, coconut, coffee, cotton, grapes, jute, kidneybeans, lentil, maize, mango, mothbeans, mungbeans, muskmelon, orange, papaya, pigeonpeas, pomegranate, rice and watermelon.

Before analysing the data, the dataset was examined to identify the data quality issues, including missing values, duplicate records, variation in data types, and invalid observations. No missing values or inconsistent data were found, resulting in the dataset being completely clean and consistent. Certain observations contain zero values for the nitrogen variable, but this leads to valid measurements. Hence, no observations were removed during the data preprocessing stage.

Figure 1 displays the distribution of the 22 crop classes. Each crop category contains exactly 100 observations, indicating a balanced multi-class dataset. This reduces the biases towards any crop class and contributes to fair comparative evaluation of all machine learning classifiers.

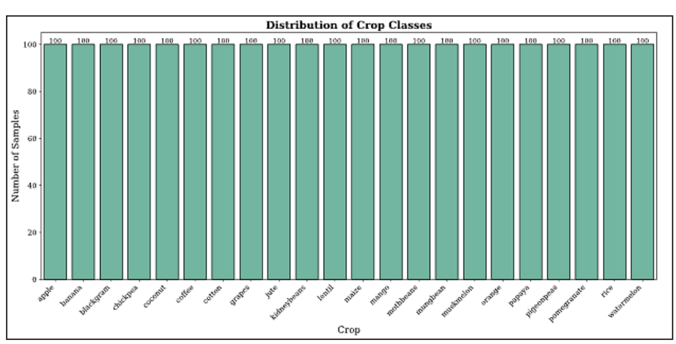

Figure 1. Crop class distribution

Statistical Characteristics of the Dataset

The corresponding feature distribution is mentioned in Figure 2.

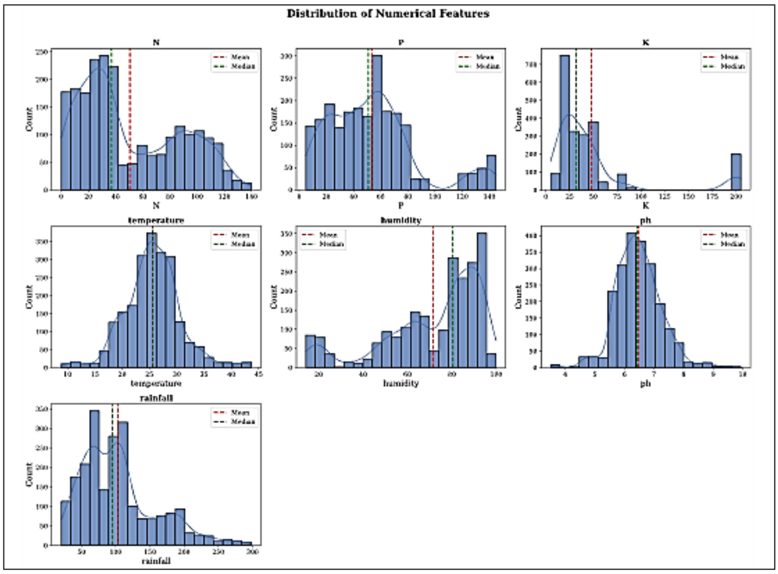

Figure 2. Distributive representation of statistical features

Figure 2 contains numerical predictor variables that show diverse statistical characteristics exhibiting variations in environmental factors. Potassium and rainfall show relatively higher positive skewness, whereas temperature and soil pH contain comparatively symmetric distributions. The Shapiro-Wilk normality test was performed to show the deviation of these numerical features from a normal distribution ($p < 0.05$).

Correlation Analysis

Pearson Correlation Analysis was performed to inspect the inter-feature relations of variables. The correlation matrix is presented in Figure 3.

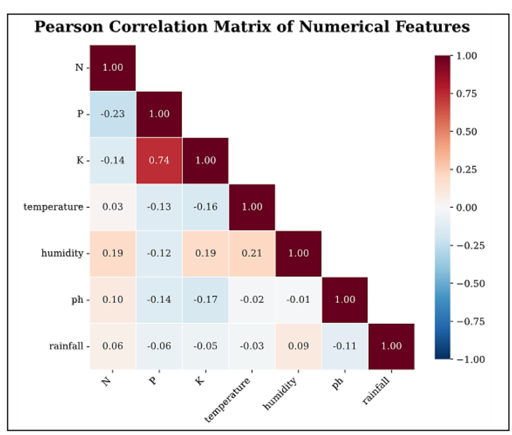

Figure 3. Pearson correlation matrix for inter-feature relationships

Weak linear correlation is exhibited by most of the feature pairs, which indicates limited multicollinearity within the dataset. It is observed that the highest positive correlation is between Phosphorus and Potassium ($r = 0.74$), while the other remaining feature pairs show relatively weak positive or negative correlations. All predictor variables were retrained for subsequent model development, as no correlation coefficient exceeded the commonly accepted multicollinearity threshold (|r|>= 0.80)

Identification of Outliers and Analysis

Figure 4, illustrating the Box Plots, and Table 3 displaying the Interquartile Range (IQR) were used to investigate the variability and presence of potential outliers within the numerical features.

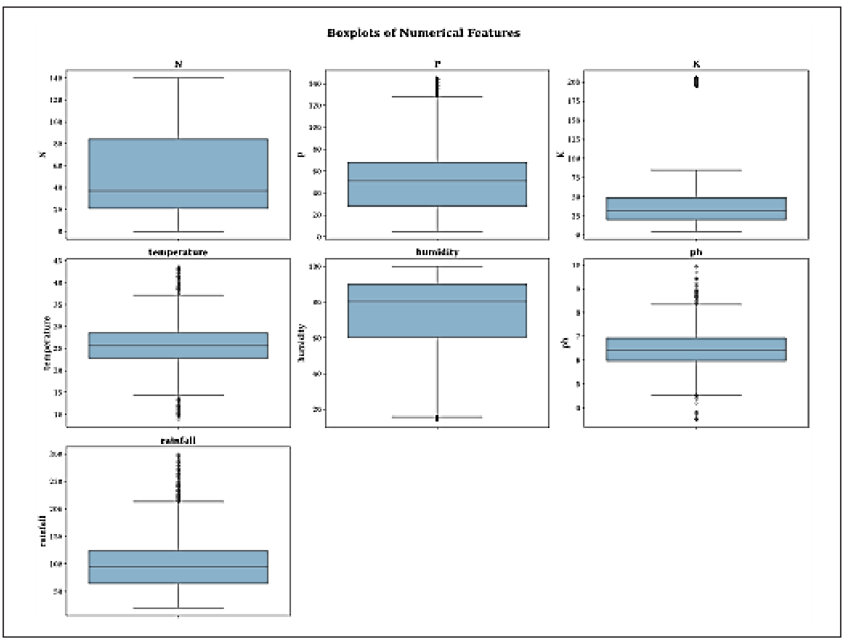

Figure 4. Box Plot showing median, quartiles and possible outliers for numerical features

Table 2
Overview of the interquartile range outlier for the dissemination of median features

Feature	Q1	Q3	IQR	Lower Bound	Upper Bound	No. of Outliers	Percentage (%)
N	21	84.25	63.25	-73.875	179.125	0	0
P	28	68	40	-32	128	138	6.27
K	20	49	29	-23.5	92.5	200	9.09
temperature	22.769	28.562	5.792	14.081	37.25	86	3.91
humidity	60.262	89.949	29.687	15.732	134.479	30	1.36
pH	5.972	6.924	0.952	4.544	8.352	57	2.59
rainfall	64.552	124.268	59.716	-25.022	213.841	100	4.55

Table 2 contains the outliers primarily observed in potassium, phosphorus and rainfall, where it was observed that these refer to the naturally occurring environmental variability rather than the false measurements and were therefore retrained. Dataset values were preserved adequately to improve the practical applicability of the developed crop recommendation models for diverse agricultural conditions.

Feature Relevance Analysis

Mutual Information (MI) analysis was performed to estimate the contribution of each predictor variable toward crop classification, whose result feature ranking can be seen in Figure 5.

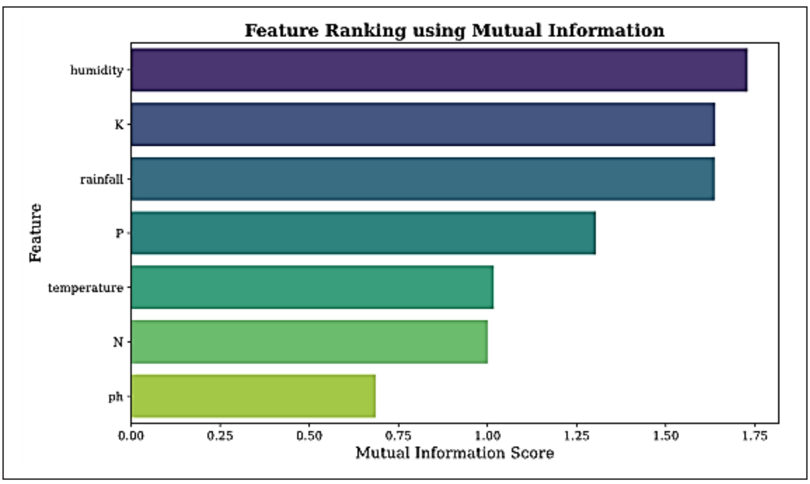

Figure 5. Ranking of mutual information of features showcases a stronger relationship or dependency between features and target variables

The most informative predictor identified was humidity, followed by potassium, rainfall, phosphorus, temperature, nitrogen and pH of soil. It was observed from the rankings that both climate and soil nutrient variables contribute substantially to crop, which supports the inclusion of all 7 predictor variables in the comparative machine learning study.

Crop-wise Environmental Characteristics

The average feature values of each crop were standardised using Z-score normalisation and visualised as a heatmap to facilitate agronomic interpretation, which can be observed in Figure 6.

Crop classes vary as per the distinct environmental profiles. Highest standardised rainfall values are exhibited by rice, which confirms its dependence on high-water conditions, whereas characteristics of coconut are observed in elevated humidity and rainfall. Nutrient-rich soil requirements are observed in apples and grapes, as it requires comparatively higher phosphorus and potassium. Whereas lower humidity and temperature conditions are associated with chickpea, which requires relatively higher soil pH values. The dataset effectively captures meaningful variations in environmental conditions across different crop classes, which are consistent with the agroeconomic knowledge.

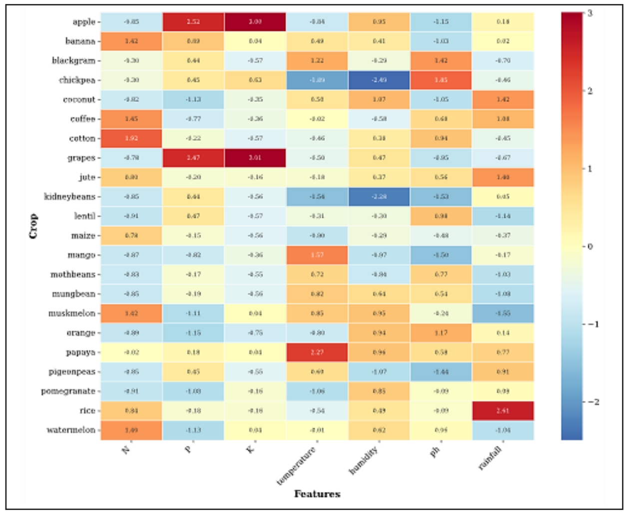

Figure 6. Heatmap illustrating standardised crop-wise soil and environmental characteristics

METHODOLOGY

The crop Recommendation dataset ensures balanced class distribution as the dataset comprises 2,200 instances equally distributed across 22 crop classes with 100 samples per class. The dataset was divided into 80:20 subsets of training and testing. To guarantee reproducibility of the experimental results, a fixed random seed (random_state = 42) was used during the dataset partitioning. Stratified 5-Fold Cross-Validation was performed to assess the robustness and stability of each classifier. The mean accuracy and standard deviation across the folds were recorded for comparative analysis.

10 Machine Learning models, viz. SVM, KNN, Random Forest, Gaussian Naive Bayes, XGBoost, Gradient Boosting, AdaBoost, LightGBM, Bagging and CatBoost were trained and implemented on the above discussed dataset. To avoid information leakage into the test set, hyperparameter tuning was carried out exclusively using GridSearchCV for models with relatively smaller search spaces and RandomisedSearchCV for ensemble boosting algorithms.

A non-parametric statistical hypothesis test was performed to determine whether the observed differences among the evaluated classifiers. To compare the overall performance of all classifiers across multiple evaluation folds, the Friedman test was performed. Pairwise Wilcoxon Signed-Rank tests were conducted as post-hoc analyses to identify significant differences between closely performing models when statistically significant differences were detected.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful classification algorithm that finds the optimal hyperplane to separate different classes in the feature space. SVM is particularly effective for crop recommendation systems when the decision boundaries between different crop types are complex. Its ability to work well in high-dimensional spaces ensures that even subtle differences in features such as soil nutrients and climatic conditions are considered. SVM makes the classification between the different crop types (classes) robust by maximising the margin, and tackling the complexity by running many small, accurate comparisons, which are highly reliable for suggesting the best crop according to the accurate efficiency showcased by this method.

K-Nearest Neighbours (KNN)

In the K-Nearest Neighbours (KNN) method, for accurate crop selection, a new data point based on the most similar existing farm data aligns with real-world conditions, that results into useful approach when diverse agricultural datasets are used instead of assuming the data. This adapts to the data distribution method naturally and is useful when specifically taking into consideration the local parameters like soil and climate. By cross-validation, the value of the nearest neighbour (k) for selection of the best crop fit for agriculture was determined. In this case, the highest accuracy was obtained at k equal to 4.

Random Forest

In Random Forest, multiple decision trees are combined to provide the best classification accuracy. It handles complex interactions among different variables, as crop recommendation depends on various factors which interact with each other in a complex manner, improving generalisation. The data is hidden, and the risk of overfitting is reduced by averaging multiple decision trees, resulting in a convoluted and non-linear relationship between the parameters. Large datasets with numerous features can be processed by Random Forest, which analyses them comprehensively. This potential identifies the most effective factors that affect better recommendations in crop selection.

Gaussian Naive Bayes (GNB)

Gaussian Naive Bayes is a programmable machine learning classifier that showcases output following Bayes' Theorem, which is a method used in probability. The classification can be done on smaller continuous datasets as well if the attributes follow a Gaussian distribution, requiring less computational resources, that results into fast predictions. It has a strong assumption of feature independence, which retains reasonably true value in practice, regardless of its simplicity. Studying the crops based on attributes' probability combinations, GNB provides fast and interpretable results by giving probabilistic outputs that help in understanding the certainty of each recommendation, sanctioning more informed decision-making.

XGBoost

In this study, several machine learning models such as decision trees, random forests, K-nearest neighbours, and neural networks are used to analyse soil and crop data, but XGBoost (Extreme Gradient Boosting) deliver the best results. XGBoost improves its predictions step-by-step by learning from previous errors, which allows it to accurately understand complex and nonlinear relationships between soil, climate, and crop parameters. XGBoost includes regularisation techniques that prevent overfitting, optimises performance of the model on new data, and prove to be highly effective for crop recommendation. Research on the usage of XGBoost along with other models, in comparison, helps to identify the most efficient approach for precise agriculture decision-making.

Gradient Boosting

Gradient Boosting starts with the basic prediction and gradually learns from the mistakes. Based on factors like soil condition, weather, and past crop results, every new step solves the error that occurred previously, improving the recommendation over the passage of time, which goes well in the vibrant farming areas, resulting in more reliable suggestions to the farmers.

AdaBoost

Correcting prediction errors through iterative learning is focused by AdaBoost, though it requires further optimisation to fully leverage historical and environmental data for crop recommendations. Its iterative approach improves model performance over time, showing potential for enhancing accuracy with continued refinement and adjustment.

LightGBM

Large amounts of data are efficiently handled quickly by LightGBM because of its speed; it can provide crop recommendations in real time using the latest information. This makes it useful in farming situations where decisions need to be made fast to improve productivity and maintain sustainable practices.

Bagging

Bagging enhances the crop recommendation stability by averaging the predictions from multiple models trained on distinct data samples. It ensures the consistent performance across diverse agricultural practices, which improves the reliability in the decision-making process, resulting in a moderate impact of the outliers.

CatBoost

The management of the categorical variables commonly used in agricultural datasets is done by CatBoost. The effective incorporation of these features into the model optimises the prediction process of the suitable crop as per the real time scenarios. This ensures reliable recommendations, rendering it appropriate for precision agriculture applications where accurate integration of categorical data, essential for decision-making and resource optimisation, is obtained.

Table 3
Comparison of different models based on various METRICS

Model	Accuracy	Precision Macro (%)	Recall Macro (%)	F1 macro (%)	Model Inference Time
Bagging	99.55	99.57	99.55	99.55	0.012039958
Random Forest	99.32	99.35	99.32	99.32	0.042236809
CatBoost	99.32	99.37	99.32	99.32	0.002705756
Gaussian Naive Bayes	99.32	99.37	99.32	99.32	0.001809618
LightGBM	98.86	98.91	98.86	98.86	0.048731084
SVM	98.86	98.97	98.86	98.86	0.011766945
XGBoost	98.86	98.94	98.86	98.85	0.015371169
Gradient Boosting	98.64	98.75	98.64	98.64	0.032235853
KNN	98.18	98.23	98.18	98.17	0.008214612
AdaBoost	91.13	93.31	91.13	90.11	0.027248464

Gaussian Naive Bayes achieved one of the highest overall performances while maintaining the lowest inference time, demonstrating its suitability for real-time crop recommendation systems among the evaluated methods. Ensemble-based approaches such as Random Forest, XGBoost, and Bagging also produced competitive results, whereas AdaBoost exhibited comparatively lower predictive performance. Bagging achieved the highest predictive accuracy of 99.55% among the evaluated classifiers. However, Gaussian Naive Bayes achieved a comparable classification performance of 99.32% while exhibiting the lowest computational overhead in terms of both training and inference time. Even though Bagging provides the highest predictive performance, Gaussian Naive Bayes offers a more favourable trade-off between accuracy and computational efficiency, making it suitable for practical crop recommendation applications.

RESULTS AND DISCUSSIONS

As depicted in Figure 7, 10 various machine learning models, SVM, KNN, Random Forest, Gaussian Naive Bayes, XGBoost, Gradient Boosting, AdaBoost, LightGBM, Bagging, and CatBoost are trained on the same dataset with 7 features, with 22 crop types. As depicted in the results in Table 3, CatBoost, Gaussian Naive Bayes, Random Forest, and Bagging are the top performers with accuracy of 99.31%, 99.31%, 99.31%, and 99.54%, respectively. While LightGBM, SVM, XGBoost, Gradient Boosting, and KNN were the moderate performers, they reached the accuracy of 98.86%, 98.86%, 98.86%, 98.63% and 98.18%, respectively. AdaBoost achieved the least accuracy of 91.13%, respectively, because AdaBoost's reweighting points give priority to hard-to-classify samples, which are either overlapped or can disturb the generalisation. These models are also compared based on their inference time, which is shown in Figure 8. Hence, this shows that models like CatBoost or Naive Bayes are better comparatively. The Comparative analysis in Figure 8 shows that the inference time of all 10 models, where Gaussian Naive Bayes proves to have the least inference time of 0.00295s.

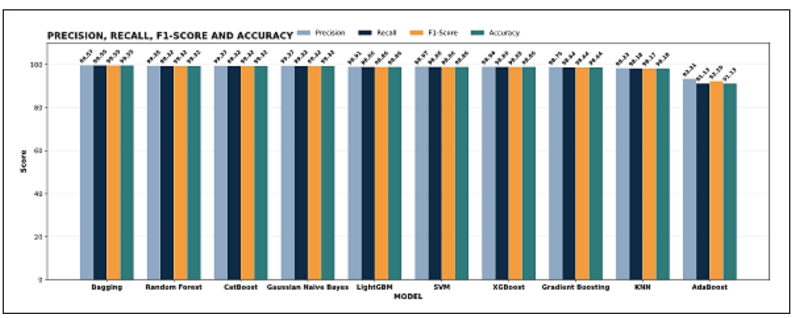

Figure 7. Visual representation of performance measures for various models

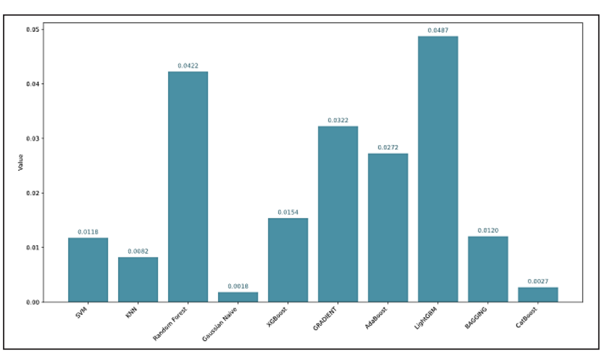

Figure 8. Comparison of inference time across ten classifiers

EXPLAINABLE AI ANALYSIS

To improve the interpretability of the comparative machine learning algorithms for crop recommendation systems, Explainable Artificial Intelligence (XAI) analysis was performed using Shapley Additive exPlanations (SHAP) on the Gaussian Naïve Bayes classifier. SHAP combines the contribution of each input feature utilised for model predictions and provides both global and local explanations. This provides a comprehensive reasoning for the quantitative performance metrics by showcasing how the agro-climatic factors affect the crop recommendation decisions.

Figure 9 presents the global Beeswarm Plot that provides a summary of the contribution of all input features across the dataset. From the plot, potassium (K) has the strongest overall influence on crop prediction, followed by rainfall and humidity. Temperature shows moderate importance, while nitrogen, phosphorus, and pH have comparatively less contribution. This plot indicates that the variations in both soil and climatic variables are the factors affecting model predictions for all crop classes.

Figure 10 demonstrates the ranking of the features based on their absolute SHAP values. This ranking correctly depicts that potassium is the most influential feature, followed by rainfall, humidity, and temperature. The less important features, nitrogen, phosphorus, and pH, show that these variables provide minimal contribution to the prediction. The similar result in the SHAP Beeswarm Plot and Bar Plots indicates the correct importance hierarchy of the features.

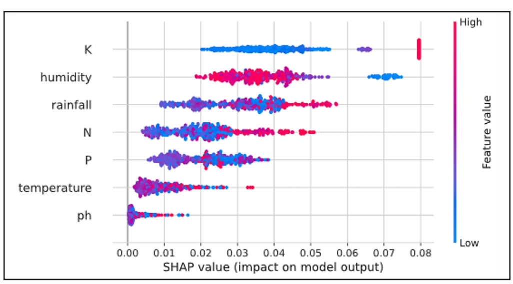

Figure 9. Dense summary of top features using SHAP Beeswarm Plot

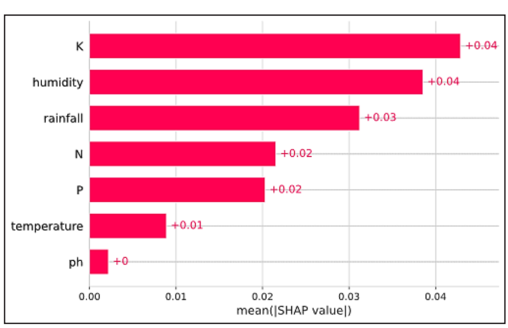

Figure 10. Global importance of each feature using the SHAP Bar Plot

Figure 11 illustrates how each feature contributes to the final decision of crop recommendation. It provides the overall contribution of soil nutrients and climatic variations that improves the overall interpretability of the evaluated models.

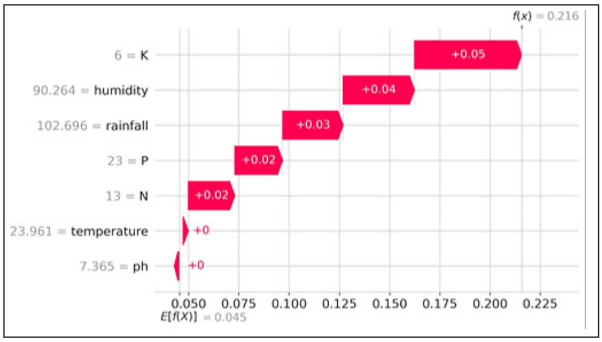

Figure 11. Proportional relationship between features and their contribution to the model

CONCLUSION

This research shows that machine learning assists farmers in determining which crops are to be grown, predicting accurate and tailored outcomes for each specified data characteristic. The diversity in datasets, like unpredictable weather and market fluctuation, is easily handled by this approach. Various models as SVM, K-Nearest Neighbours, Random Forest, Gaussian Naive Bayes, XGBoost, Gradient Boosting, AdaBoost, LightGBM, Bagging, and CatBoost unveils the significant enhancement in agricultural productivity, sustainability, and resilience. Bagging achieved the highest predictive accuracy (99.55%) among the evaluated classifiers. However, Gaussian Naive Bayes achieved comparable classification performance (99.32%) while exhibiting the lowest computational overhead in terms of both training and inference time. Even though Bagging provides the highest predictive performance, Gaussian Naive Bayes offers a more favourable trade-off between accuracy and computational efficiency, making it suitable for a practical crop recommendation application.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to their respective institutions for providing the necessary academic environment, infrastructure, and research facilities that made this work possible.

REFERENCES

Afzal, H., Amjad, M., Raza, A., Munir, K., Villar, S. G., Lopez, L. A. D., & Ashraf, I. (2025). Incorporating soil information with machine learning for crop recommendation to improve agricultural output. *Scientific Reports, 15*, Article 8560. https://doi.org/10.1038/s41598-025-88676-z

Alshahrani, H., Alkahtani, H. K., Mahmood, K., Alymani, M., Mohammed, G. P., Abdelmageed, A. A., Abdelbagi, S., & Drar, S. (2023). Chaotic Jaya optimisation algorithm with computer vision-based soil type classification for smart farming. *IEEE Access, 11*, 65849-65857. https://doi.org/10.1109/ACCESS.2023.3288814

Assaduzzaman, M., Bishshash, P., Nirob, M. A. S., Al Marouf, A., Rokne, J. G., & Alhajj, R. (2025). XSE-TomatoNet: An explainable AI-based tomato leaf disease classification method using EfficientNetB0 with squeeze-and-excitation blocks and multi-scale feature fusion. *MethodsX, 14*, Article 103159. https://doi.org/10.1016/j.mex.2025.103159

Bhattacharya, B., & Solomatine, D. P. (2006). Machine learning in soil classification. *Neural Networks, 19*(2), 186-195. https://doi.org/10.1016/j.neunet.2006.01.005

Bhavana, M., & Rao, K. S. (2025). Smart agriculture with Internet of Things for precise crop prediction using integrated machine learning and advanced stacking ensemble from soil parameters. *SSRG International Journal of Electronics and Communication Engineering, 12*(8), 74-90. https://doi.org/10.14445/23488549/IJECE-V12I8P107

Boudouda, S., & Boukallel, M. (2025). Intelligent soil classification for precision agriculture in Algeria: Integrating machine learning and image processing for optimised crop yield and recommendation. *Algerian Journal of Signals and Systems, 10*(2), 65-74. https://doi.org/10.51485/ajss.v10i2.264

Bouni, M., Hssina, B., Douzi, K., & Douzi, S. (2022). Towards an efficient recommender system in smart agriculture: A deep reinforcement learning approach. *Procedia Computer Science, 203*, 825-830. https://doi.org/10.1016/j.procs.2022.07.124

Cema, G., & Kaliappan, E. (2022). AI-based crop recommendations for intensive farming using WSN. In *2022 3rd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)* (pp. 1718-1723). IEEE. https://doi.org/10.1109/ICICICT54557.2022.9917962

Chauhan, A., Tsunduru, A., Parveen, K., Tokala, S., Hajarathaiah, K., & Enduri, M. K. (2023). A crop recommendation system based on nutrients and environmental factors using machine learning models and IoT. In *2023 International Conference on Information Technology (ICIT)* (pp. 453-458). IEEE. https://doi.org/10.1109/ICIT58056.2023.10226131

Dahiphale, D., Shinde, P., Patil, K., & Dahiphale, V. (2025). Smart farming: Crop recommendation using machine learning with challenges and future ideas. *Artificial Intelligence and Applications, 3*(4), 1-13. https://doi.org/10.47852/bonviewAIA52026214

Desai, M., & Ansari, N. (2023). An innovative method to increase agricultural productivity using machine learning-based crop recommendation systems. In *2023 2nd International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)* (pp. 645-651). IEEE. https://doi.org/10.1109/ICAISS58487.2023.10250531

Farag, A. A. (2025). Machine learning approaches for enhanced estimation of reference evapotranspiration (ETo): A comparative evaluation. *Scientific Reports, 15*, Article 38485. https://doi.org/10.1038/s41598-025-23166-w

Gambhir, S., Khan, M. A. S., Sharma, N., Nagar, M., Chawara, G., Shweta, & Prasad, A. (2026). Soil classification and crop recommendation with integrated microbiome and machine learning. In *Data*

Analytics and Management (Lecture Notes in Networks and Systems, Vol. 1603, pp. 412-426). Springer. https://doi.org/10.1007/978-3-032-03769-5_31

Gautam, V., Trivedi, N. K., Anand, A., & Kumar, A. (2022). Optimised a high-performance deep neural network for crop recommendation. In *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 838-843). IEEE. https://doi.org/10.23919/INDIACom54597.2022.9763156

Gawade, S., Rout, G., Kochar, P., Ahire, V., & Namboodiri, T. (2023). AGROFERDURE: Intelligent crop recommendation system for agricultural crop productivity using machine learning algorithms. In *2023 International Conference on Computer, Electronics & Electrical Engineering & Their Applications (IC2E3)* (pp. 1-9). IEEE. https://doi.org/10.1109/IC2E357697.2023.10262476

Indira, D., Sobhana, M., Swaroop, A. H. L., & Phani Kumar, V. (2022). KRISHI RAKSHAN: A machine learning-based new recommendation system for the farmer. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1798-1804). IEEE. https://doi.org/10.1109/ICICCS53718.2022.9788221

Ingle, A. (2020). *Crop recommendation dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

Jeyanathan, J. S., Medha, B., Sai, G. T. V., Kumar, R. B., & Sahu, V. (2023). Automated crop recommender system using pattern classifiers. In *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)* (pp. 567-572). IEEE. https://doi.org/10.1109/IDCIOT56793.2023.10053451

Kaggle. (n.d.). *Kaggle: Your machine learning and data science community*. https://www.kaggle.com/

Kavitha, R., Kavitha, M., & Srinivasan, R. (2022). Crop recommendation in precision agriculture using supervised learning algorithms. In *2022 3rd International Conference for Emerging Technology (INCET)* (pp. 1-4). IEEE. https://doi.org/10.1109/INCET54531.2022.9824155

Khan, A. A., Nauman, M. A., Bashir, R. N., Jahangir, R., Alroobaea, R., Binmahfoudh, A., Alsafyani, M., & Wechtaisong, C. (2022). Context-aware evapotranspiration (ETs) for saline soils reclamation. *IEEE Access, 10*, 110050-110063. https://doi.org/10.1109/ACCESS.2022.3206009

Kiruthika, S., & Karthika, D. (2023). IoT-based professional crop recommendation system using a weight-based long-term memory approach. *Measurement: Sensors, 27*, Article 100722. https://doi.org/10.1016/j.measen.2023.100722

Meenachi, L., Ramakrishnan, S., Sivaprakash, M., Thangaraj, C., & Sethupathy, S. (2022). Multi-class ensemble classification for crop recommendation. In *2022 International Conference on Inventive Computation Technologies (ICICT)* (pp. 1319-1324). IEEE. https://doi.org/10.1109/ICICT54344.2022.9850561

Open Government Data (OGD) Platform India. (n.d.). *Open Government Data (OGD) Platform India*. National Informatics Centre. https://data.gov.in/

Priyadharshini, A., Chakraborty, S., Kumar, A., & Pooniwala, O. R. (2021). Intelligent crop recommendation system using machine learning. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 843-848). IEEE. https://doi.org/10.1109/ICCMC51019.2021.9418375

Raja, S. P., Sawicka, B., Stamenkovic, Z., & Mariammal, G. (2022). Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers. *IEEE Access, 10*, 23625-23641. https://doi.org/10.1109/ACCESS.2022.3154350

Ramachandra, R. A., Ankitha, G. V., Divya, I., Vandana, P., & Jagadeesh, H. S. (2023). Crop recommendation using machine learning. In the *2023 International Conference on Data Science and Network Security (ICDSNS)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICDSNS58469.2023.10245154

Ramu, S. C., Raman, D., Ramana, K., Vijaya Krishna, A., Khan, A. A., Khan, S., & Abdu, S. M. (2026). A lightweight deep learning and whale optimisation framework for sustainable precision agriculture. *Discover Computing, 29*, Article 73. https://doi.org/10.1007/s10791-026-09952-8

Ramya, J., & Sankar, M. (2023). Crop recommendation using supervised learning techniques. In *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 1076-1084). IEEE. https://doi.org/10.1109/ICOSEC58147.2023.10276199

Sagar, B. M., Shashank, S. J., Shashank, V., Yashwanth Kumar, C., & Padmashree, T. (2023). Crop Connect: An approach to crop recommendation. In *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)* (pp. 1-6). IEEE. https://doi.org/10.1109/CSITSS60515.2023.10334150

Siddharth, S. (2021). *Crop recommendation dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/siddharthss/crop-recommendation-dataset

Singh, R. (2023). *Crop recommendation data* (Version 1) [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/4GBWFV

Sivanandam, K., Prasanth, M., Naveen, B., & Saran, S. (2023). An efficient machine learning approach for crop recommendation based on soil characteristics. In *2023 2nd International Conference on Electronics and Renewable Systems (ICEARS)* (pp. 71-76). IEEE. https://doi.org/10.1109/ICEARS56392.2023.10085361

Swetha, A. J., Kalyani, G., & Kirananjali, B. (2023). Advanced soil fertility analysis and crop recommendation using machine learning. In *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1035-1039). IEEE. https://doi.org/10.1109/ICOEI56765.2023.10125748

Teja, M. S., Preetham, T. S., Sujihelen, L., Jancy, S., & Selvan, M. P. (2022). Crop recommendation and yield production using the SVM algorithm. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1768-1771). IEEE. https://doi.org/10.1109/ICICCS53718.2022.9788274

Venkateswara, S. M., & Padmanaban, J. (2025). Interpretable deep learning models for independent fertiliser and crop recommendation. *Scientific Reports, 15*, Article 41721. https://doi.org/10.1038/s41598-025-26910-4

Vishnoi, V. K., Kumar, K., Kumar, B., Mohan, S., & Khan, A. A. (2023). Detection of apple plant diseases using leaf images through convolutional neural network. *IEEE Access, 11*, 6594-6609. https://doi.org/10.1109/ACCESS.2022.3232917