FRONTIERS IN
ENGINEERING | SCIENCE | MANAGEMENT
MNS-UET, MULTAN

FRONTIERS IN ENGINEERING, SCIENCE & MANAGEMENT

https://journal.mnsuet.edu.pk/

Research Article

A Comparative Analysis of Machine Learning Models for Crop Classification Using Soil Nutrients and Environmental Data

Muhammad Aamir [a, b, *], Muhammad Waseem Iqbal [c], Muhammad Kashif [d], Muhammad Yasir Bilal [e]

[a] Department of Computer Science, COMSATS University Islamabad, Sahiwal, Punjab, Pakistan
[b] Department of Computer Science, Superior University Lahore, 54000, Pakistan
[c] Department of Software Engineering, Superior University, Lahore, 54000, Pakistan
[d] Department of Computer Science & Information Technology (FCSIT) University of Southern Punjab, Multan
[e] Department of Computer Science MNS UET Multan

ARTICLE INFORMATION

Received	17 July 2026
Received in revised form	29 July 2026
Accepted	29 July 2026
Available Online	15 Sep 2026

*CORRESPONDING AUTHOR

Muhammad Aamir
Department of Computer Science, COMSATS University Islamabad, Sahiwal, Punjab, Pakistan.
Department of Computer Science, Superior University Lahore, 54000, Pakistan.

Email muhammadaamir@cuisahiwal.edu.pk

ABSTRACT

Purpose: Agriculture is essential to the growth of the world economy. Food availability today largely depends on whether crops have been harvested properly. The machine learning methods described in this research use soil nutrient and environmental data to classify crops.
Methodology: This study focuses on using only current soil and weather data to classify suitable crops. Several machine learning algorithms, such as light gradient boosting, decision trees, random forests, and logistic regression, are applied to evaluate crop classification models on soil and environmental data.
Findings: To calculate accuracy, precision, recall, and f1-score, this research used the available data from Kaggle. The decision tree and random forest algorithms achieve an approximately 99% accuracy by using the cross-validation method under the applied evaluation setting.
Originality: A comparison with other existing models reveals that the decision tree and random forest algorithms get higher results among the applied models. These models achieved the highest accuracy because they effectively capture complex, nonlinear relationships between soil nutrients and environmental factors, and reduce overfitting.
Practical implications: The findings of this work may support future research and decision-support tools for crop classification models.

Keywords: light gradient boosting, decision tree, random forest, logistic regression, soil nutrients, agricultural product

1. Introduction

The global economy relies on agriculture to handle food security and limit the impact of climate change as the world population grows. Accurate crop classification is currently the most important challenge in agriculture(Agrawal et al., 2025). The kind of weather, as well as things like rainfall and temperature, and the amount of nutrients used, strongly determine how agriculture performs (Eddamiri et al., 2024; Pande et al., 2021). The changes in agriculture system are most important to enhance crop quality and quantity. The requirement for the food is going day by day with the increase in the population (Dai et al., 2025; Kavita and Mathur, 2020; Li et al., 2024). Automatic crop classification systems are the need of the hour, using soil nutrients and environmental data. It looks like a big problem in crop classification while still regulating the destructive environmental effects of micro-nutrients (Chen et al., 2024; Debicka et al., 2023).Many methods are used to classify crops, respectively designed for the unique problems in different agricultural areas (Elbasi et al., 2023; Gawade et al., 2026).

Since agriculture is key to many economies, it is now more important than ever to get better results from crops while using pesticides responsibly (Raju et al., 2024). The machine learning models help farmers and experts make better decisions through useful tools for crop classification. One can find complex patterns and rules within different data types and use those findings in many sectors (Jorvekar et al., 2024). The various industries use ML techniques, such as credit analysis, image processing, meteorology, medical research, fraud detection, customer relationship management, bioinformatics, and agriculture (Pallathadka et al., 2023).

Machine learning within Artificial Intelligence helps improve crop classification by using different data types. Machine learning deals chiefly with finding patterns and making correlations in data sets. These models to work well, they are first trained on datasets that show past learning results. Models are built, parameters are chosen carefully, and the models' accuracy in predictions is improved using past data (Tripathi and Biswas, 2026). Although many studies have used machine learning for crop recommendation, there is still a need to compare common algorithms using soil nutrient and environmental features with consistent evaluation metrics (Pooja et al., 2025). It becomes possible to classify crop categories, helping select and choose the best farming strategies during growth (Bali and Singla, 2022).

As vital nutrients, potassium (Azadnia et al., 2023), phosphorus (Wang et al., 2023), and nitrogen (Iatrou et al., 2021) have different functions in boosting crop development and yields than conventional fertilizers. Increased yields result from potassium's beneficial effects on plant health, disease resistance, and fruit and vegetable qualiThe nitrogen is essential for crops like rice and maize for healthy plant development, foliage growth stimulation, and increased yields.

This research uses four machine learning (ML) algorithms and a publicly available dataset containing 2,200 entries of agrarian data, with dimensions for eight key variables: nitrogen (N), phosphorus (P), and potassium (K); temperature (°C); humidity (%); rainfall (mm) for weather conditions; soil pH (acidity); and crop type labels (22 categories). The important contributions are:

- This study compares four machine-learning algorithms for crop classification using soil nutrients and environmental features

- To calculate accuracy, precision, recall, and f1-score, this research used the available data from Kaggle.
- The decision tree and random forest algorithms achieve approximately 99% accuracy

The body of the remaining paper is as follows: the first one is the literature review section that enlightens the related work to the topic, the second is the methodology part that describes the overall methodology used, the third is the results part that clarifies the overall results, and the fourth one is the comparison and discussion segment. The last one is the conclusion, which defines the paper conclusion.

2. Literature Review

Previous research has discovered numerous characteristics of crop classification systems. The use of machine learning methods in agriculture, exactly in crop classification, has garnered significant attention in recent years (Yao et al., 2022) . Many farmers are now using cutting-edge machine learning methods like regression, decision trees, and neural networks to improve agricultural performance successfully (Yadav and Swarnkar, 2024). Many scholars have established the effectiveness of random forests in forecasting crop categories, whereas some other works focus on the impact of specific soil nutrients on crop health (Vanarase et al., 2021).

A novel methodology to classify rice using machine learning methods, such as decision trees, neural networks, and linear regression is presented in previous research (Hasan et al., 2023). This work has a precision above 85%. The models achieve an average accuracy of over 90%, demonstrating the remarkable strength of data-driven approaches in predicting crop classes. On the other hand, recall rates of an astounding 80% indicate that the models successfully replicate real-world yield outcomes.

The work in (Hageltoum et al., 2024; Mella and Pentakoti, 2022) grants an innovative method for crop classification, and the outcomes are inspiring. The random forest model provides enormously accurate global and regional crop classification forecasts. On the other hand, another research (Aliwi and Al-Tuwaijari, 2025) shows how to classify sugarcane yields. Specialists can classify sugarcane crops with over 90% accuracy using the random forest algorithm. Moreover, the model achieves an accuracy exceeding 80% when predicting actual yields. These discoveries show how these models could revolutionize sugarcane farming. A comparison of the existing literature is also given in Table 1.

Table 1 - Existing Literature Comparison

Reference	Methodology	Dataset	Results	Limitations
(Ahmad et al., 2018)	RF, DT, SVM	Private	Accuracy: 97%	Limited dataset, reliance on environmental conditions.
(Filippi et al., 2019)	RF	Western Australia case study dataset	Accuracy: 92%	Need for integration of more data sources.
(PS, 2019)	ANN, SVR, KNN, RF	Private	Accuracy: 94%	Lack of diverse datasets with environmental features.
(Khosla et al., 2020)	ANN, SVR	Private	Accuracy: 96%	Limited feature selection in this research.
(Leo et al., 2021)	RF. GBM	A public dataset of cotton	Accuracy: 96%	Potential for improvement with larger feature-rich datasets.
(Pant et al., 2021)	GBT, RFR, SVR, DT	A public dataset of maize, potatoes, rice, and wheat	Accuracy: 96%	Future enhancements with more relevant features.
(Paudel et al., 2021)	KNN, GBT, SVR	Private	Accuracy: 93%	More features can be added to train the model for better accuracy, which results in improved accuracy by adding more features.
(Lischeid et al., 2022)	RF, SVR	A public dataset of maize	Accuracy: 97%	Need for inclusion of environmental features.
(Paudel et al., 2022)	KNN, GBT, SVR	Private	Accuracy: 93%	Potential enhancement with spatial and temporal features.

A detailed comparative assessment of crop classification studies shows that many methods, datasets, and results contribute to field growth. In study (Ahmad et al., 2018) employs random forest, decision trees, and support vector machines on private datasets; they vary significantly in results, with (Ahmad et al., 2018) attaining a high accuracy of 97% and. The limitation of (Ahmad et al., 2018) contains need for limited eco-friendly conditions data. In work (Filippi et al., 2019) used RF on a Western Australia dataset with 92% accuracy but called for integrating more databases (PS, 2019) combined many techniques on a private dataset, achieving 94% accuracy but recommending further datasets featuring environmental factors. The studies in (Khosla et al., 2020) and (Leo et al., 2021) display similar accuracy of 96%, yet (Khosla et al., 2020) is limited by a distinctive feature set, and (Leo et al., 2021) inspires the use of larger, feature-rich datasets for development. In, (Pant et al., 2021) attains 96% accuracy using gradient boosting trees, random forest regression, support vector regression, and decision trees on public datasets of multiple crop types and looks to augment with related features, whereas (Paudel et al., 2021) employs KNN, GBT, and SVR on a private dataset, attaining 93% accuracy, and proposes further feature integration. In, (Lischeid et al., 2022) and (Paudel et al., 2022), accuracy reached 97% and 93% with RF, SVR, KNN, and GBT on maize datasets. These studies collectively underscore the wide variety of methods to predict crop classes, each present valuable insights and avenues for enhancement in accuracy and feature integration. Although many studies have used machine learning for crop classification, there is still a need for a clear comparison of common machine learning models using soil nutrient and environmental features on a public crop recommendation dataset.

3. Methodology

This study uses the public Kaggle crop classification dataset, which contains 2,200 records, seven input features, and one crop label. The data were pre-processed, divided into training and testing sets, and four machine learning models were compared for crop classification. The details of the applied algorithms are given below.

3.1. Dataset Details

This work uses a publicly available Kaggle dataset (kaggle, 2023) holding 2,200 entries of agrarian data, with dimensions for seven input features and one target label: nitrogen (N), phosphorus (P), and potassium (K), temperature (°C), humidity (%), and rainfall (mm) for weather situations, soil pH (acidity), and crop type labels (22 categories). The dataset contains 137 unique nitrogen values, 117 phosphorus values, and 73 potassium values, while temperature, humidity, pH, and rainfall have complete records for all entries. The illustration of data from the start and end of the dataset is given in Tables 2 and 3, which present a snapshot of the inconsistency in soil and eco-friendly factors analysed for this work.

Table 2 - Head samples of the dataset

Sr.	Soil Nutrients			temperature	Humidity	ph	rainfall	Label
	N	P	K					
0	90	42	43	20.87	82.00	6.50	202.93	Rice
1	85	58	41	21.77	80.31	7.03	226.65	Rice
2	60	55	44	23.00	82.32	7.84	263.96	rice
3	74	35	40	26.49	80.15	6.98	242.86	rice
4	78	42	42	20.13	81.60	7.62	262.71	rice

Table 3 - Tail samples of the dataset

Sr.	Soil Nutrients			tempera-ture	Humidity	ph	rain-fall	label
	N	P	K					
2195	107	34	32	26.77	66.41	6.780	177.77	coffee
2196	99	15	27	27.41	56.63	6.08	127.92	coffee
2197	118	33	30	24.13	67.22	6.36	173.32	coffee
2198	117	32	34	26.27	52.12	6.75	127.17	coffee
2199	104	18	30	23.60	60.39	6.77	140.93	coffee

3.2. Dataset pre-processing

In the data pre-processing stage, the data frames are fed into the model, containing N, P, K, temperature, humidity, pH, rainfall, and label. Finally, the data was split into training (70%) and testing (30%) sets via Scikit-Learn train_test_split. The cross validation as k-fold (k=5) validation, and general hyperparameters such as learning rate 0.001, batch size (4,8,16), number of epochs (40), optimizer type (Adam) were configured to run the models.

3.3. Models Explanation

In this work, four machine learning algorithms are used for classification of crop categories and evaluating the soil nutrients. At the first step, the dataset is divided into train and test sets. The data preprocessing for each model is done by preparing data on weather and soil nutrients (Addula and Sajja, 2024). The applied models, like decision trees and logistic regression present more straightforward, understandable predictions. Moreover, the LGBM and random forests construct ensembles of decision trees to increase accuracy. The models are trained, tuned, and assessed using common statistical metrics, and examined to determine the most important variables.

3.4. Experimental Setup and Hyperparameters

Experiments were carried out in the Python 3.11 environment within Anaconda Navigator running on Windows 11 (64 bit) platform in the Jupyter Notebook environment. The Decision Tree, Random Forest, Logistic Regression models and model evaluation were implemented using Scikit-learn 1.5.1, while the gradient boosting classifier was implemented using LightGBM 4.5.0. Furthermore, the data was numerically processed and preprocessed using NumPy 1.26.4 and Pandas 2.2.2, respectively, and graphically represented and analyzed using Matplotlib 3.9.0 and Seaborn 0.13.2.

The hyperparameters of each model were carefully set to standard values reported in the literature and preliminary experiments to facilitate a fair comparison and make the results reproducible. With 200 estimators, learning rate of 0.05, maximum tree depth of 10, 31 leaves and feature and bagging fractions of 0.8, it provides an effective balance between learning and overfitting control. The Decision Tree model had the following parameters: maximum depth of the tree at 15, minimum samples split at 5 and minimum samples per leaf at 2 to simplify the model and enhance the generalization. The Random Forest consisted of 200 decision trees with a maximum depth of 20, and bootstrap sampling, which helped to randomly sample the data and reduce variance in the final predictions. Implementation of the Logistic Regression model was done on the basis of the use of L2 regularization, the LBFGS optimizer, a maximum of 1000 iterations, with stable convergence and good classification performance. All models were run with a random state of 42 to ensure the results of the experiments were reproducible. The choice of these parameters was done to ensure that the model is not overfitting the data, and have a balance between accuracy, efficiency and generalisation.

4. Results

4.1. Light Gradient Boosting Model (LGBM)

The LGBM classification model confusion matrix is shown in Fig. 1, providing a clear picture of how well the model distinguishes between classes. The strong performance with few misclassifications is reflected in the overall score of 98%, which is determined by averaging the matrix factors. Overall, approximately 98.89% of positive and 99.42% of negative cases are accurately identified by the model. The equations 1-4 are used to calculate accuracy, precision, recall, and f1-score; the results are reported in Table 4. Though the variations in these assessments may point to areas for development, high precision, recall, and f1-scores show the model responsible classification for certain classes. Overall performance is given by the macro and weighted averages, particularly when dealing with unbalanced classes. The overall accuracy of 99% and continuously high evaluation ratings, the model in this instance performs remarkably well across almost all categories. Although having strong analytical capabilities, the LGBM model has some shortcomings. Its complicated structure can make interpretation challenging, and it runs the danger of overfitting if incorrectly adjusted, especially with limited or biased samples. Furthermore, the model could have trouble handling categorical variables with large cardinality.

$$\text{Precision}=\frac{tp}{(tp+fp)} \tag{1}$$

$$\text{Recall}=\frac{tp}{(tp+fn)} \tag{2}$$

$$\text{F1-score}=\frac{2(\text{recall} \times \text{precision})}{(\text{recall} + \text{precision})} \tag{3}$$

$$\text{Accuracy}=\frac{(tp+tn)}{(tn+tp+fp+fn)} \tag{4}$$

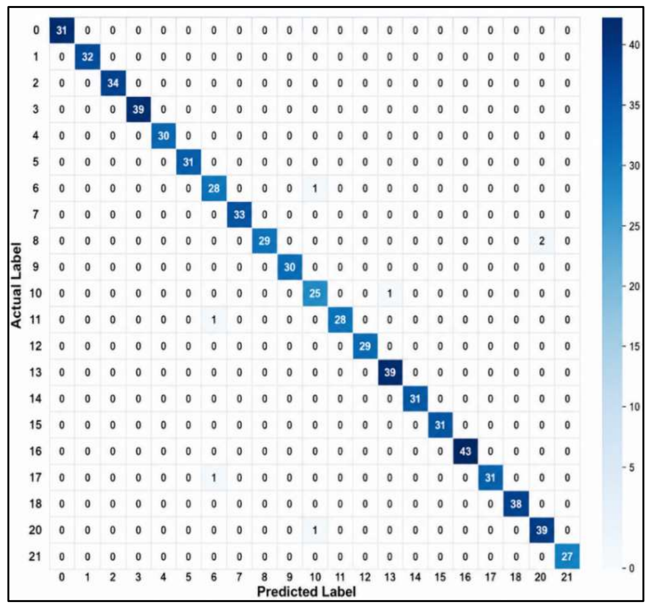

Fig. 1 - Confusion matrix for the LGBM model

Table 4 -Statistical values of the Light Gradient Boosting Model for crops

Label	Precision (%)	Recall (%)	F1-score (%)	Support
0	100	1.00	1.00	31
1	1.00	1.00	1.00	32
2	0.97	0.97	0.97	35
3	1.00	1.00	1.00	39
4	1.00	1.00	1.00	30
5	1.00	0.97	0.98	32
6	0.97	1.00	0.98	28
7	1.00	1.00	1.00	33
8	0.97	0.94	0.95	31
9	1.00	1.00	1.00	30
10	1.00	0.96	0.98	26
11	0.93	0.97	0.95	29

12	1.00	1.00	1.00	29
13	0.97	1.00	0.99	39
14	1.00	1.00	1.00	31
15	1.00	1.00	1.00	31
16	1.00	1.00	1.00	40
17	1.00	1.00	1.00	43
18	1.00	0.97	0.98	32
19	1.00	1.00	1.00	38
20	0.95	0.97	0.96	40
21	1.00	1.00	1.00	27
Accuracy	0.99			726
Macro Average	0.99	0.99	0.99	726
Weighted Average	0.99	0.99	0.99	726

4.2. Decision Tree (DT)

The confusion matrix used to evaluate the performance of the classification model is shown in Fig. 2. An overall score of 0.98 indicates the model correctly classifies almost all positive and negative situations, determining exceptional accuracy. With just a few miscalculations, it achieves 98.48% accuracy for the positive class and 99.18% for the negative class. The model strong and reliable performance across classes is added confirmed by the metrics of accuracy, precision, recall, and F1-score obtained from equations 1-4 and compiled in Table 5. The robust classification is supported by strong precision, recall, and F1-scores, while class imbalance is addressed by macro and weighted averages. However, overfitting, uncertainty, biased feature selection, and problems with missing data are some of the weaknesses of decision trees.

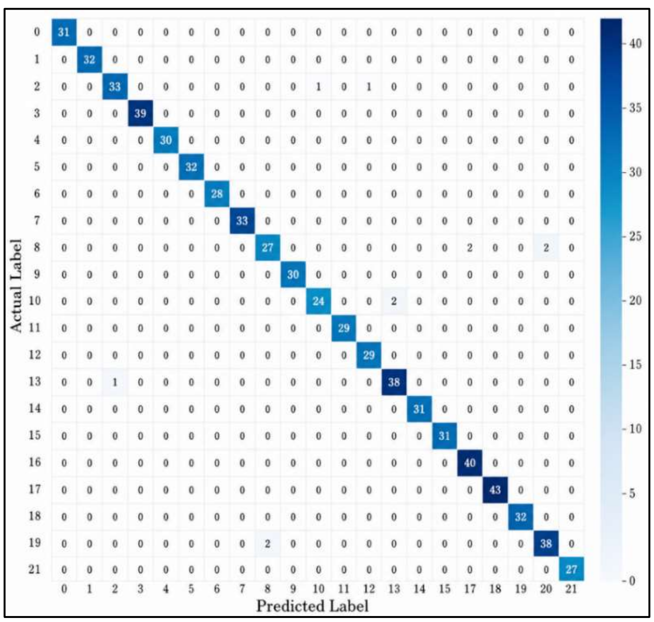

Fig.2- Confusion matrix for the DT model

Table 5- Statistical values of the decision tree for crops

Label	Precision (%)	Recall (%)	F1-score (%)	Support
0	1.00	1.00	1.00	31
1	1.00	1.00	1.00	32
2	0.97	0.94	0.96	35
3	1.00	1.00	1.00	39
4	1.00	1.00	1.00	30
5	1.00	1.00	1.00	32
6	1.00	1.00	1.00	28
7	1.00	1.00	1.00	33
8	0.93	0.87	0.90	31
9	1.00	1.00	1.00	30
10	1.00	0.92	0. 96	26
11	0.97	100.0	0.98	29
12	1.00	1.00	1.00	29
13	0.93	0.97	0.95	39
14	1.00	1.00	1.00	31
15	1.00	1.00	1.00	31
16	1.00	1.00	1.00	40
17	0.96	1.00	0.98	43
18	1.00	1.00	1.00	32
19	1.00	1.00	1.00	38
20	0.95	0.95	0.95	40
21	1.00	1.00	1.00	27
Accuracy	0.98			726
Macro Average	0.99	0.98	0.99	726
Weighted Average	0.98	0.98	0.98	726

4.3. Random Forest (RF)

The confusion matrix for a 22 classes classification model is shown in Fig. 3, determining exceptional performance with an overall score of 99%. With a false-positive rate of only 56%, the model accurately identifies 99% of positive cases. Furthermore, it accurately detects 99% of negative cases, with a false-negative rate of 16%. Table 6 summarizes the accuracy, precision, recall, and f1-score values obtained using equations 1–4. with outstanding precision, recall, f1-score values, and an astounding overall accuracy of 99%, these measures show the model regularly has good performance across most classes.

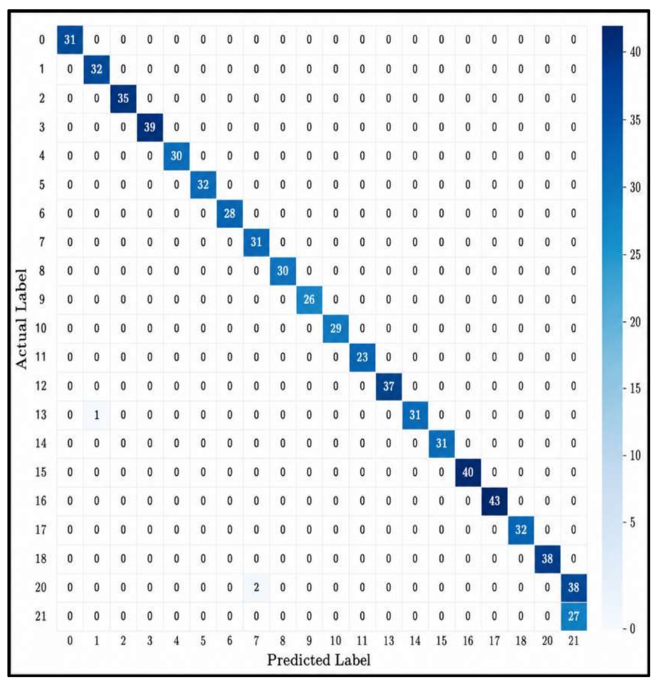

Fig. 3- Confusion matrix for RF model

Table 6- Statistical values of the Random Forest for crops

Label	Precision (%)	Recall (%)	F1-score (%)	Support
0	1.00	1.00	1.00	31
1	1.00	1.00	1.00	32
2	0.97	1.00	0.99	35
3	1.00	1.00	1.00	39
4	1.00	1.00	1.00	30
5	1.00	1.00	1.00	32
6	1.00	1.00	1.00	28
7	1.00	1.00	1.00	33
8	0.94	1.00	0.97	31
9	1.00	1.00	1.00	30
10	0.96	1.00	0.98	26
11	1.00	1.00	1.00	29
12	1.00	1.00	1.00	29
13	1.00	0.95	0.97	39
14	1.00	1.00	1.00	31
15	1.00	1.00	1.00	31
16	1.00	1.00	1.00	40
17	1.00	1.00	1.00	43
18	1.00	1.00	1.00	32
19	1.00	1.00	1.00	38
20	1.00	0.95	0.97	40
21	1.00	1.00	1.00	27
Accuracy	0.99	726		726
Macro Average	0.99	1.00	0.99	726
Weighted Average	0.99	0.99	0.99	726

4.4. Logistic Regression (LR)

Generally, the model performs well, accurately predicting outcomes 94.35% of the time, with error rates staying low at 4-6%, it detects 94% of positive instances and 96% of negative cases. The table 7 exhibits consistently strong recall and precision in most categories. However, the model works best with clean, balanced datasets and may struggle with more complex patterns. Its reliability in practical applications can be increased by improving data preparation and making certain modifications.

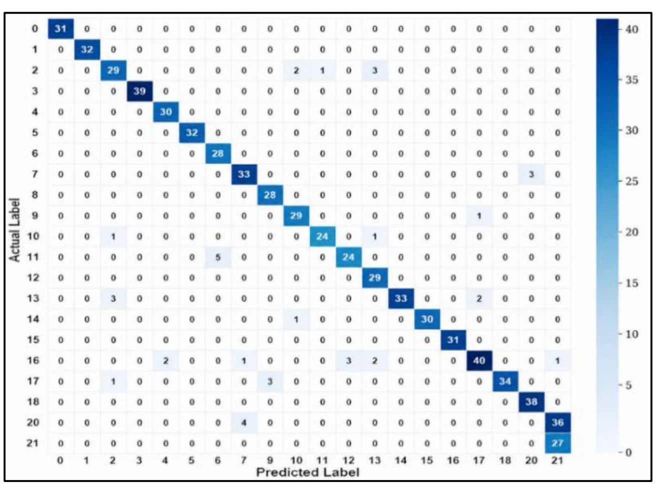

Fig. 4- Confusion matrix for the LR model

Table 7- Statistical values of Logistic Regression for crops

Label	Precision (%)	Recall (%)	F1-score (%)	Support
0	1.00	1.00	1.00	31
1	1.00	1.00	1.00	32
2	0.85	0.83	0.84	35
3	1.00	1.00	1.00	39
4	0.94	1.00	0.97	30
5	1.00	1.00	1.00	32
6	0.85	1.00	0.92	28
7	1.00	1.00	1.00	33
8	0.85	0.90	0.88	31
9	0.91	0.97	0.94	30
10	0.89	0.92	0.91	26
11	0.96	0.83	0.89	29
12	0.91	1.00	0.95	29
13	0.85	0.85	0.85	39
14	1.00	0.97	0.98	31
15	1.00	1.00	1.00	31
16	1.00	1.00	1.00	40
17	0.97	0.79	0.87	43
18	0.90	0.88	0.89	32
19	1.00	1.00	1.00	38
20	0.90	0.90	0.90	40
21	1.00	1.00	1.00	27
Accuracy	0.94			726
Macro Average	0.94	0.95	0.94	726
Weighted Average	0.95	0.94	0.94	726

4.5. Feature Analysis and Comparison

All 22 crops included in Table 8, such as apples, bananas, rice, and lentils, have 100 entries in our dataset, which is a balanced collection of examples. We recorded important growth factors for every crop, including soil pH levels, weather, and nutrients. The considerable trends are revealed by the data; for example, apples and grapes thrive in potassium-rich soils, while lentils and chickpeas do well even in lower nutrient concentrations. In addition, every crop has distinct requirements for humidity and temperature.

Table 8- Dataset labels with mean values for each environmental factor and soil nutrients

Label	Soil Nutrients			Humidity	ph	rainfall	temperature
	K	N	P				
0	199.89	20.80	134.22	92.33	5.92	112.65	22.63
1	50.05	100.23	82.01	80.35	5.98	104.62	27.37
2	19.24	40.02	67.47	65.11	7.13	67.88	29.97
3	79.92	40.09	67.79	16.86	7.33	80.05	18.87
4	30.59	21.98	16.93	94.84	5.97	175.68	27.40
5	29.94	101.20	28.74	58.86	6.79	158.06	25.54
6	19.56	117.77	46.24	79.84	6.91	80.39	23.98
7	200.11	23.18	132.53	81.87	6.02	69.61	23.84
8	39.99	78.40	46.86	79.63	6.73	174.79	24.95
9	20.05	20.75	67.54	21.60	5.74	105.91	20.11
10	19.41	18.77	68.36	64.80	6.92	45.68	24.50
11	19.79	77.76	48.44	65.09	6.24	84.76	22.38
12	29.92	20.07	27.18	50.15	5.76	94.70	31.20
13	20.23	21.44	48.01	53.16	6.83	51.19	28.19

14	19.87	20.99	47.28	85.49	6.72	48.40	28.52
15	50.08	100.32	17.72	92.34	6.35	24.68	28.66
16	10.01	19.58	16.55	92.17	7.01	110.47	22.76
17	50.04	49.88	59.05	92.40	6.74	142.62	33.72
18	20.29	20.73	67.73	48.06	5.79	149.45	27.74
19	40.21	18.87	18.75	90.12	6.42	107.52	21.83
20	39.87	79.89	47.58	82.27	6.42	236.18	23.68
21	50.22	99.42	17.00	85.16	6.49	50.78	25.591

4.6. Features of Soil Nutrients and Their Effect on Crops

Three types of soil nutrients, potassium (K), nitrogen (N), and phosphorus (P), are used in this research work. Fig. 5 shows the feature values of nitrogen (N) for different crops. In this case, cotton gains the highest nitrogen values, and apples and pomegranates get the lowest. On the other hand, Fig. 6 explains the feature values of potassium (K) for different crops. The apple and grapes required the highest, and the orange got the lowest value of K, while Fig. 7 describes the features of phosphorus (P) for different crops. The apple and grapes required the highest, and the coconut had the lowest P value.

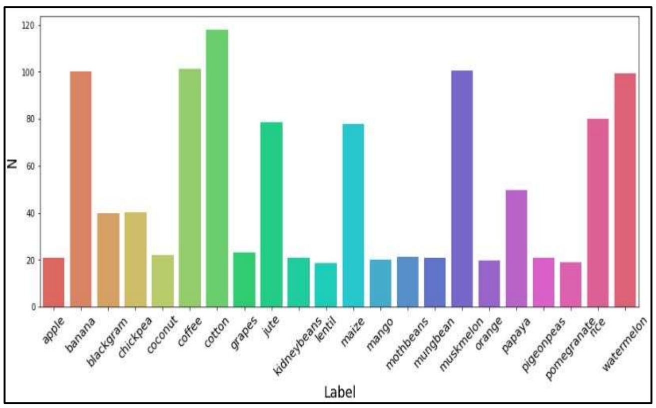

Fig. 5- Features of nitrogen (N) for different crops

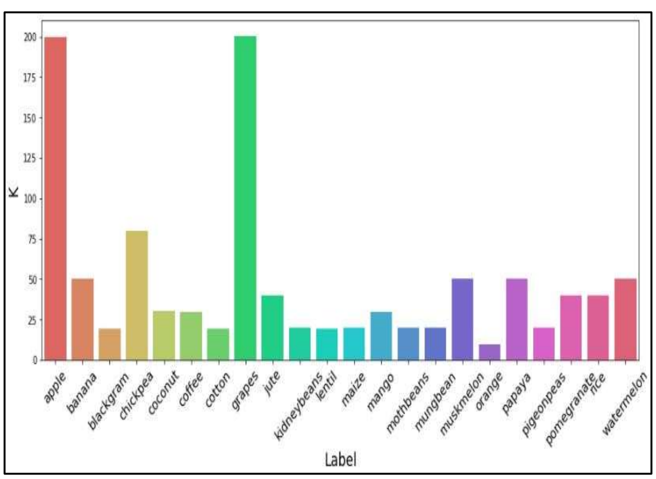

Fig. 6- Features of potassium (K) for different crops

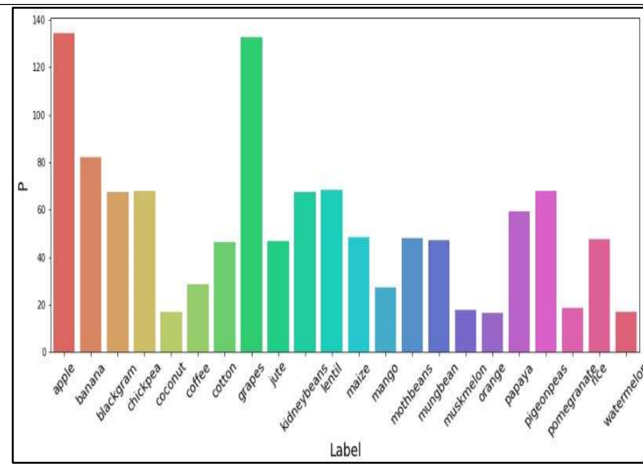

Fig. 7- Features of phosphorus (P) for different crops

4.6.1. Effects of Nitrogen (N)

Fig. 8 illustrates the nitrogen requirements of various crops measured in kilograms per hectare (kg/ha), listed in descending order of nitrogen requirement, from highest to lowest. Crops with the highest nitrogen requirements include Cotton (117.77 kg/ha), Coffee (101.2 kg/ha), Muskmelon (100.32 kg/ha), Orange (100.23 kg/ha), Banana (99.42 kg/ha), and Watermelon (79.00 kg/ha). Conversely, crops with the lowest nitrogen requirements comprise Pomegranate (19.58 kg/ha), Mango (20.07 kg/ha), Pigeon peas (78.4 kg/ha), Apple (77.76 kg/ha), Mungbean (77.76 kg/ha), Papaya (76.54 kg/ha), Mothbeans (75.89 kg/ha), Chickpea (75.34 kg/ha), and Coconut (74.79 kg/ha).

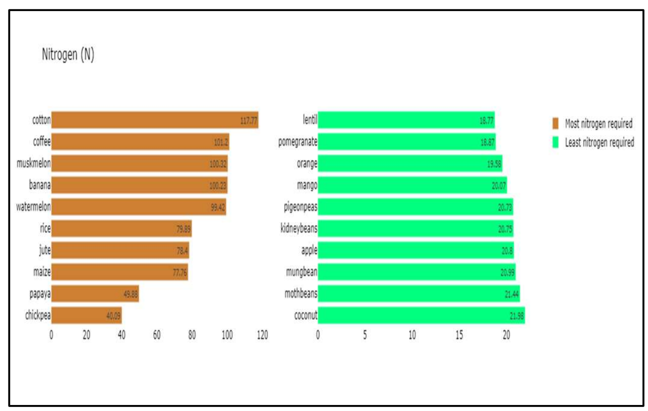

Fig. 8- Effects of Nitrogen on Different Crops

4.6.2. Effects of Phosphorus (P)

The phosphorus requirements of various crops, expressed in milligrams per kilogram (mg/kg) of dry matter and sorted from highest to lowest requirement, are shown in Fig. 9. The crops with the most significant phosphorus requirements include Apple (134.22 mg/kg), Apple (132.53 mg/kg), Chickpea (67.20 mg/kg), Pomegranate (67.73 mg/kg), Pigeon peas (67.73 mg/kg), and Mango (27.18 mg/kg).

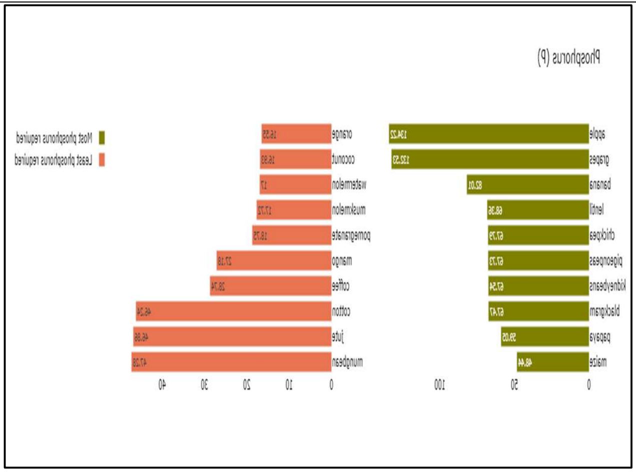

Fig. 9- Effects of Phosphorus on Different Crop

4.6.3. Effects of Potassium (K)

The potassium restrictions for various crops are shown in Fig. 10, placed from highest to lowest (kg/ha). The largest potassium supplies are for apples (199.89 kg/ha) and grapes (200.11 kg/ha), followed by watermelon (50.22 kg/ha) and papaya (50.04 kg/ha). On the other hand, considerably less potassium is needed for crops like orange (10.01 kg/ha), pigeon peas (20.29 kg/ha), and mango (29.92 kg/ha).

The NPK values of potassium, phosphorus, and nitrogen for numerous crops are denoted in kilograms per hectare in Fig. 11. These differences primarily highlight how crucial good nutrient control is for robust development and high output. The Fig. 12 analyses the impacts of temperature, humidity, and rainfall on crop development, with crops including rice, coconut, and jute being more exposed to these environmental elements.

Fig. 13 shows a correlation map that illustrates the relationship between temperature and relative humidity, with colour indicating an additional environmental condition like pH or rainfall. The plot shows an apparent inverse relationship: as temperature increases, relative humidity reduces. Overall, the figure shows how effectively temperature influences humidity in the examined environment.

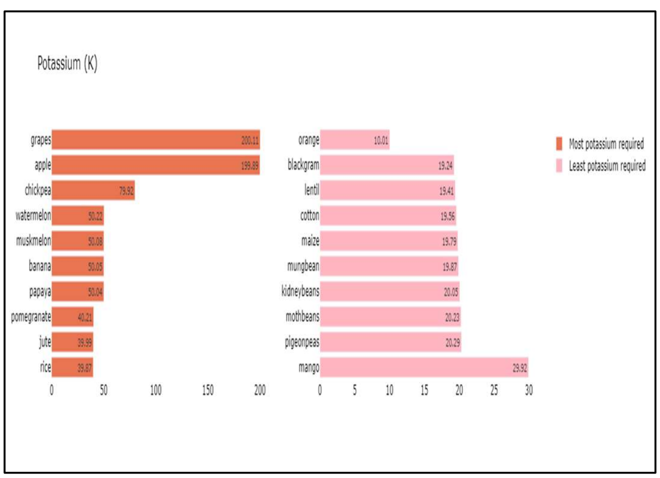

Fig. 10- Effects of Potassium on Different Crops

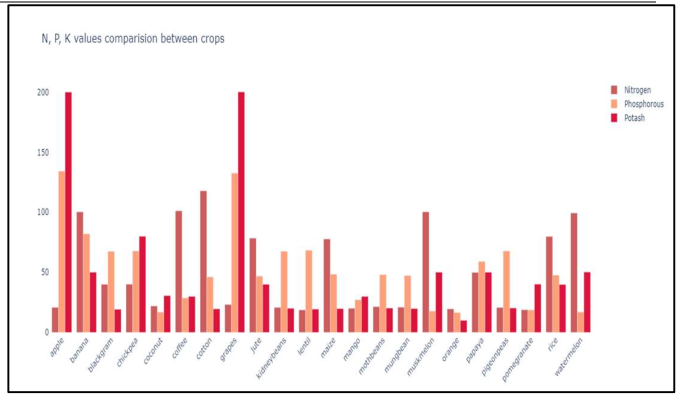

Fig. 11- NPK values comparison for crops

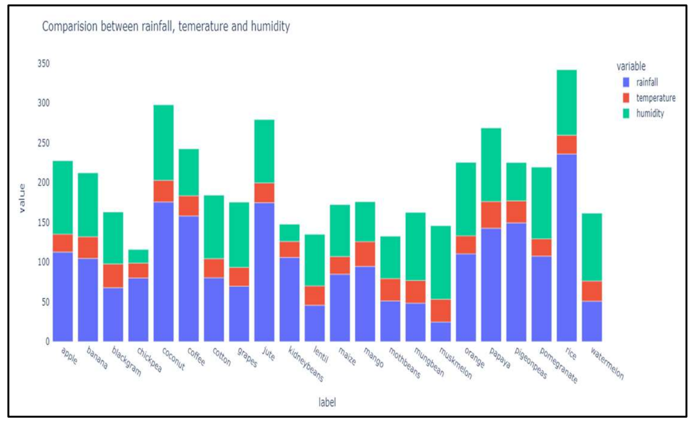

Fig. 12- Rainfall, temperature, and humidity values comparison for crops.

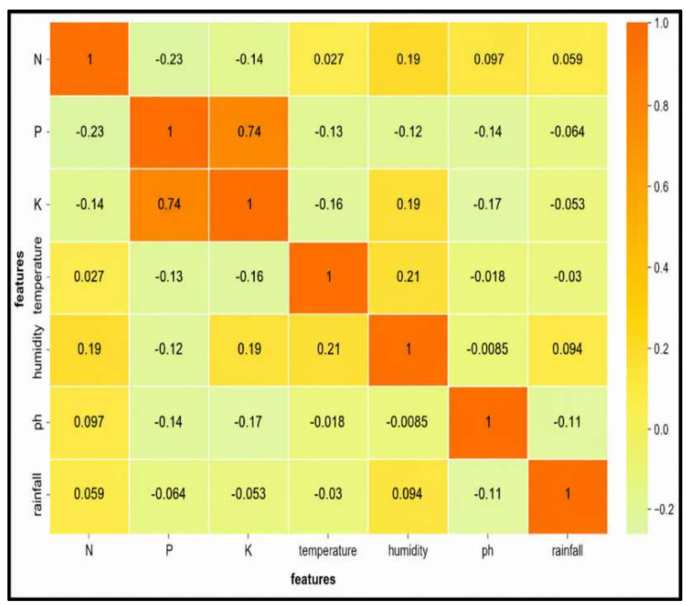

Fig. 13- Correlation between numerous features

Discussion

To mitigate potential overfitting and ensure model generalizability, k-fold cross-validation (k = 5) was performed for all models, including Light Gradient Boosting, Decision Tree, Random Forest, and Logistic Regression. As shown in Table 8, the models demonstrate stable performance across different folds with low standard deviation values, indicating that the high accuracy is consistent and not a result of overfitting.

Table 9 evaluates the performance of earlier available models with the model developed in this study, aiming at soil nutrients use classification across different methods and datasets. Previous studies from 2018 to 2022 used a variety of machine learning methods, such as RF, DT, SVM, ANN, SVR, KNN, and GBT on both public and mostly private datasets, describing strong accuracies up to 97%. In contrast, different studies evaluated their method solely on public data and still achieved impressive results, with accuracy ranging from 94% to 99% across various algorithms.

The main reason for the increased accuracy of the Decision Tree and Random Forest models is their capability of learning complex decision boundaries and non-linear relationships among input features. These models successfully recognize which features have the most informative content and which interactions can be considered in order to make more accurate predictions of the class. Especially, Random Forest is an ensemble learning method that integrates the prediction results of several decision trees, thereby decreasing the errors of prediction and enhancing generalisation by limiting the variance. The ensemble approach results in a less noisy and overfitting model compared with a single Decision Tree. Logistic Regression, on the other hand, is a linear decision boundary model and it may not be able to capture complex interactions between features, which can result in problems with classification accuracy. LightGBM, as a boosting algorithm, can be very powerful, but its performance relies on tuning the hyperparameters and the nature of the data to be optimized. The tree-based models were more appropriate for this data set and allowed for higher classification accuracy than the other algorithms evaluated, due to the nature of the data set itself.

Table 8- Five-Fold Cross-Validation Accuracy (%)

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Light Gradient Boosting	98.0	98.3	98.2	98.1	98.4
Decision Tree	99.2	99.1	99.7	99.5	99.6
Random Forest	99.0	99.3	99.1	98.9	99.2
Logistic Regression	95.1	95.5	94.2	95.4	94.3

Table 9- Comparison of all the model statistical values

Reference	Methodology	Dataset	Use of soil nutrients	Results
		Existing Model Results		
(Ahmad et al., 2018)	RF, DT, SVM	Private	Yes	Accuracy: 97%
(Filippi et al., 2019)	RF	Western Australia case study dataset	Yes	Accuracy: 92%
(PS, 2019)	ANN, SVR, KNN, RF	Private	Yes	Accuracy: 94%
(Khosla et al., 2020)	ANN, SVR	Private	Yes	Accuracy: 96%
(Leo et al., 2021)	RF. GBM	A public dataset of cotton	Yes	Accuracy: 96%
(Pant et al., 2021)	GBT, RFR, SVR, DT	A public dataset of maize, potatoes, rice, and wheat	Yes	Accuracy: 96%
(Paudel et al., 2021)	KNN, GBT, SVR	Private	Yes	Accuracy: 93%
(Lischeid et al., 2022)	RF, SVR	A public dataset of maize	Yes	Accuracy: 97%
(Paudel et al., 2022)	KNN, GBT, SVR	Private	Yes	Accuracy: 93%
		Our Model Results		
1	LGBM	Public dataset	Yes	Accuracy:98%
2	DT	Public dataset	Yes	Accuracy: 99%
3	RF	Public dataset	Yes	Accuracy: 99%
4	Lr	Public dataset	Yes	Accuracy: 94%

Conclusions

This study reveals that when machine learning models are applied to a public Kaggle dataset with 2,200 records, they accomplish exceptionally high accuracy. Even though all models performed strongly, decision trees and random forests were the top performers, achieving approximately 99% accuracy with the test crop dataset. These models help farmers to classify crop classes based on soil nutrient and environmental data. Future work could incorporate more external datasets, real field data, seasonal data, location data, soil type, irrigation, crop yield, and fertilizer application records to further strengthen model performance. As a limitation, it depends on one public dataset and needs external validation before real-world use.

Dataset Availability

The dataset used in this research is available at : https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset/data

Author Contributions

Author contributed as Conceptualization and methodology by Muhammad Aamir, Software,Validation,Formal Analysis by Muhammad Waseem Iqbal,Investigation,Resources by Muhammad Kashif,and Data Curation by Yasir bilal.

Acknowledgements

No acknowledgement

Conflicts of Interest Statement

No conflict of interest

Funding Information

No funding

AI Statement

NoAI tool used

References

Agrawal, N., Govil, H. and Kumar, T. (2025), "Agricultural land suitability classification and crop suggestion using machine learning and spatial multicriteria decision analysis in semi-arid ecosystem", Environment, Development and Sustainability, Vol. 27 No. 6, pp. 13689-13726.

Ahmad, I., Saeed, U., Fahad, M., Ullah, A., Habib ur Rahman, M., Ahmad, A. and Judge, J. (2018), "Yield forecasting of spring maize using remote sensing and crop modeling in Faisalabad-Punjab Pakistan", Journal of the Indian Society of Remote Sensing, Vol. 46, pp. 1701-1711.

Aliwi, H. R. and Al-Tuwaijari, J. M. (2025), "Crops Recommendation System Based on Rain Data Using Machine Learning Techniques", مجلة جامعة الكوت, Vol. 10 No. 2, pp. 243-256.
Azadnia, R., Rajabipour, A., Jamshidi, B. and Omid, M. (2023), "New approach for rapid estimation of leaf nitrogen, phosphorus, and potassium contents in apple-trees using Vis/NIR spectroscopy based on wavelength selection coupled with machine learning", Computers and Electronics in Agriculture, Vol. 207, pp. 107746.
Bali, N. and Singla, A. (2022), "Emerging trends in machine learning to predict crop yield and study its influential factors: A survey", Archives of computational methods in engineering, pp. 1-18.
Chen, J., Zhao, L., Wang, B., He, X., Duan, L. and Yu, G. (2024), "Uncovering global risk to human and ecosystem health from pesticides in agricultural surface water using a machine learning approach", Environment International, Vol. 194, pp. 109154.
Dai, J., Chen, X., Zhang, Y., Zhang, M., Dong, Y., Zheng, Q., Liao, J. and Zhao, Y. (2025), "Machine learning-enhanced color recognition of test strips for rapid pesticide residue detection in fruits and vegetables", Food Control, Vol. 174, pp. 111256.
Debicka, M., Jamroz, E., Bekier, J., Ćwieląg-Piasecka, I. and Kocowicz, A. (2023), "The influence of municipal solid waste compost on the tranformations of phosphorus forms in soil", Agronomy, Vol. 13 No. 5, pp. 1234.
Eddamiri, S., Bassine, F. Z., Ongoma, V., Epule Epule, T. and Chehbouni, A. (2024), "An automatic ensemble machine learning for wheat yield prediction in Africa", Multimedia Tools and Applications, pp. 1-27.
Elbasi, E., Chamseddine, Z., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Cina, E., Shdefat, A. Y. and Saker, L. (2023), "Crop prediction model using machine learning algorithms".
Filippi, P., Jones, E. J., Wimalathunge, N. S., Somarathna, P. D., Pozza, L. E., Ugbaje, S. U., Jephcott, T. G., Paterson, S. E., Whelan, B. M. and Bishop, T. F. (2019), "An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning", Precision Agriculture, Vol. 20, pp. 1015-1029.
Gawade, S. D., Bhansali, A., Chopade, S. and Kulkarni, U. (2026), "Optimizing crop yield prediction with R2U-Net-AgriFocus: A deep learning architecture with leveraging satellite imagery and agro-environmental data", Expert Systems with Applications, Vol. 296, pp. 128942.
Hageltoum, I. A., Abdallah, E. K., Alfeel, M. I. and Abbas, S. A. (2024), "Predictive modelling of crop rotation using data mining approaches", Journal of Artificial Intelligence and Computational Technology, Vol. 1 No. 1.
Hasan, M., Marjan, M. A., Uddin, M. P., Afjal, M. I., Kardy, S., Ma, S. and Nam, Y. (2023), "Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation", Frontiers in Plant Science, Vol. 14, pp. 1234555.
Iatrou, M., Karydas, C., Iatrou, G., Pitsiorlas, I., Aschonitis, V., Raptis, I., Mpetas, S., Kravvas, K. and Mourelatos, S. (2021), "Topdressing nitrogen demand prediction in rice crop using machine learning systems", Agriculture, Vol. 11 No. 4, pp. 312.
Jorvekar, P. P., Wagh, S. K. and Prasad, J. R. (2024), "Crop yield predictive modeling using optimized deep convolutional neural network: An automated crop management system", Multimedia Tools and Applications, Vol. 83 No. 14, pp. 40295-40322.
kaggle. (2023), "Crop Recommendation Dataset", available at: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset/data (accessed 30-10-2023).
Kavita, M. and Mathur, P. (2020), "Crop yield estimation in India using machine learning", in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*. IEEE, pp. 220-224.
Khosla, E., Dharavath, R. and Priya, R. (2020), "Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression", Environment, Development and Sustainability, Vol. 22, pp. 5687-5708.
Leo, S., De Antoni Migliorati, M. and Grace, P. R. (2021), "Predicting within-field cotton yields using publicly available datasets and machine learning", Agronomy Journal, Vol. 113 No. 2, pp. 1150-1163.
Li, Z., Nie, Z. and Li, G. (2024), "Integrating Crop Modeling and Machine Learning for the Improved Prediction of Dryland Wheat Yield", Agronomy, Vol. 14 No. 4, pp. 777.
Lischeid, G., Webber, H., Sommer, M., Nendel, C. and Ewert, F. (2022), "Machine learning in crop yield modelling: A powerful tool, but no surrogate for science", Agricultural and Forest Meteorology, Vol. 312, pp. 108698.
Mella, N. and Pentakoti, V. M. (2022), "Crop yield prediction and fertilizer recommendation using voting based ensemble classifier", J. Eng. Sci, Vol. 13 No. 8, pp. 262-270.
Pallathadka, H., Mustafa, M., Sanchez, D. T., Sajja, G. S., Gour, S. and Naved, M. (2023), "Impact of machine learning on management, healthcare and agriculture", Materials Today: Proceedings, Vol. 80, pp. 2803-2806.
Pande, S. M., Ramesh, P. K., ANMOL, A., Aishwarya, B., ROHILLA, K. and SHAURYA, K. (2021), "Crop recommender system using machine learning approach", in *2021 5th international conference on computing methodologies and communication (ICCMC)*. IEEE, pp. 1066-1071.
Pant, J., Pant, R., Singh, M. K., Singh, D. P. and Pant, H. (2021), "Analysis of agricultural crop yield prediction using statistical techniques of machine learning", Materials Today: Proceedings, Vol. 46, pp. 10922-10926.
Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C. and Athanasiadis, I. N. (2021), "Machine learning for large-scale crop yield forecasting", Agricultural Systems, Vol. 187, pp. 103016.
Paudel, D., Boogaard, H., de Wit, A., van der Velde, M., Claverie, M., Nisini, L., Janssen, S., Osinga, S. and Athanasiadis, I. N. (2022), "Machine learning for regional crop yield forecasting in Europe", Field Crops Research, Vol. 276, pp. 108377.
Pooja, M., Attavar, A., Manohar, A. L. and Himoshra, D. (2025), "Pattern Prediction System for Smart Farming in the Agriculture Sector Using Machine Learning: Smart Agriculture", in *Hybrid Soft Computing Techniques for Machine Learning and Optimization*. IGI Global Scientific Publishing, pp. 159-174.
PS, M. G. (2019), "Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms", Applied Artificial Intelligence, Vol. 33 No. 7, pp. 621-642.
Raju, C., Ashoka, D. and BV, A. P. (2024), "CropCast: Harvesting the future with interfused machine learning and advanced stacking ensemble for precise crop prediction", Kuwait Journal of Science, Vol. 51 No. 1, pp. 100160.
Tripathi, D. and Biswas, S. K. (2026), "A precise artificial intelligence system for multi-class crop yield prediction with ensemble learning and swarm optimization", Environment, Development and Sustainability, pp. 1-28.
Vanarase, V., Mane, V., Bhute, H., Tate, A. and Dhar, S. (2021), "Crop Prediction Using Data Mining and Machine Learning Techniques", in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, pp. 1764-1771.
Wang, Y., Shi, W. and Wen, T. (2023), "Prediction of winter wheat yield and dry matter in North China Plain using machine learning algorithms for optimal water and nitrogen application", Agricultural Water Management, Vol. 277, pp. 108140.
Yadav, A. K. and Swarnkar, S. K. (2024), "Exploring the effectiveness of decision trees for comprehensive detection of crop diseases in agricultural environments", in *Smart Agriculture*. CRC Press, pp. 111-127.
Yao, J., Wu, J., Xiao, C., Zhang, Z. and Li, J. (2022), "The classification method study of crops remote sensing with deep learning, machine learning, and Google Earth engine", Remote Sensing, Vol. 14 No. 12, pp. 2758.