scientific reports

OPEN

Incorporating soil information with machine learning for crop recommendation to improve agricultural output

Hadeeqa Afzal[1], Madiha Amjad[1]✉, Ali Raza[2], Kashif Munir[1], Santos Gracia Villar[3,4,5], Luis Alonso Dzul Lopez[3,4,6] & Imran Ashraf[7]✉

The agriculture field is the basis of a country's change and financial system. Crops are the main source of revenue for the people. One of the farmer's most challenging problems is choosing the right crops for their land. This critical decision has a direct impact on productivity and profit. Wrong crop selection not only reduces yields but also causes food shortages, creating more problems for farmers. The best crop depends on many parameters such as illustration humidity, N, K, P, pH, rainfall, and temperature of the soil. Getting advice from experts is not an easy task. This requires intelligent models in crop recommendations that use machine-learning models to suggest suitable crops for soil and other environmental conditions. Temperature, humidity, and pH are important data for growing crops in agriculture. In this study, we gather and preprocess relevant data. To recommend the most suitable crop, we propose a novel ensemble learning approach called RFXG based on random forest (RF) and extreme gradient boosting (XGB) to suggest the best crop out of the twenty-two major crops. To measure the capability of the proposed approach, various machine learning models are utilized including extra tree classifier, multilayer perceptron, RF, decision trees, logistic regression, and XGB classifiers. To get the best performance, optimization of hyperparameter, and K-fold cross-validation procedures are performed. Experimental outcomes show that the proposed RFXG technique achieves a recommendation accuracy is 98%. Specifically, the proposed solution provides immediate recommendations to help farmers make timely decisions.

Keywords Agricultural economics, Smart agriculture, Soil parameters, Crop recommendation, Smart farming, Machine learning, Ensemble model

Agriculture is an important part of the economic environment and sustains people's lives in many regions. The general prosperity of many regions depends heavily on the annual harvest, This highlights the important role agriculture plays in connecting food supply and distribution. Challenges faced by farmers due to unpredictable weather conditions require changes in agriculture[1]. Different regions have different climates, so it is important to consider the different environments in each region. This decision is important in determining the most suitable place for growing different types of crops. Precipitation patterns also vary greatly across the region and play an important role in agriculture. Too little or too much rain can damage crops while good rainfall helps produce the best crops. As soil properties and weather conditions affect agriculture significantly, crop selection based on accurate information, including soil composition and weather conditions, is crucial to increasing crop yields. However, access to this information is often hindered by the complexity and change of environmental factors. Improvements must be made to provide accurate data and better understand soil and climate dynamics and their effects on crops. The use of data has the potential to provide rich information, solving problems that will replace

[1]Institute of Information Technology, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan. [2]Department of Software Engineering, University of Lahore, Lahore 54000, Pakistan. [3]Universidad Europea del Atlántico, Isabel Torres 21, 39011 Santander, Spain. [4]Universidad Internacional Iberoamericana, Campeche 24560, México. [5]Universidade Internacional do Cuanza, Cuito, Bié, Angola. [6]Universidad de La Romana, La Romana, República Dominicana. [7]Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea. ✉email: madiha.amjad@kfueit.edu.pk; imranashraf92@gmail.com

modern technology and provide more accurate and important information. Although there are difficulties in processing complex data, technology can make good decisions in agriculture[2].

This research focuses on recommending optimal crops based on soil nutrients as illustration humidity, N, K, P, pH, rainfall, and temperature. Considering these parameters, the most suitable crop is recommended to get the optimal crop yield with respect to the soil nutrient conditions. It is expected to help farmers make better and more informed decisions for sustainable and smart farming thereby improving crop yield and farmers' financial gains. The following contributions are made in this study.

- A machine learning-based approach is presented in this study to recommend the most suitable crop with reference to particular soil type from 22 varieties of crops including chickpeas, maize, rice, kidney beans, broad beans, pigeonpeas, mungbeans, lentils, blackbeans, pomegranates, mangoes, bananas, grapes, watermelon, apples, melon, oranges, papaya, cotton, coconut, coffee, and jute.
- An ensemble approach is introduced for improved accuracy in suggesting the optimal crop which integrates random forest (RF) and extreme gradient boosting (XGB) models. In comparison, several machine learning models are implemented like extra trees classifier (ETC), multilayer perceptron (MLP), RF, decision tree (DT) classifier, logistic regression (LR), and XGB for comparative performance. Models are fine-tuned for optimal performance.
- Exploratory data analysis is carried out to analyze feature relationship analysis. The proposed model is evaluated in comparison to existing approaches and its performance is further validated using a k-fold cross-validation approach.

We organized this paper as follows. Section 2 offers insights into the existing work. Section 3 provides details of the approach, dataset, data preprocessing, and exploratory analysis. Section 4 provides discussion and results. Finally, the conclusion and the future work of the paper are presented in Section 5.

Literature review

In this section, we study the relevant literature concerning our proposed fact-finding study. We analyze previous studies that have been practiced for crop suggestion. This research entails evaluating the methodologies utilized, and the achieved accuracy/error rates, and analyzing the constraints inherent in current approaches.

Iniyan et al. introduced machine-learning techniques in foreseeing crop yields inside the[3]. This forecast includes an analysis of factors like soil quality, weather, and historical data. The accuracy of these predictions depends on the use of regression models. Farmers face the challenge of accurately predicting harvests in India. To solve this problem, the authors used satellite images and weather data to estimate crop yields for 18 years in Maharashtra. The first method involved using multiple linear regression (MLR), but it presented some limitations. To improve prediction accuracy, the authors tested various methods and evaluated their performance using metrics such as R-squared. The long short-term memory (LSTM), achieved an accuracy rate of 86.3%. The main aim was to provide farmers with useful tools to help them choose the most suitable crop to get the best harvest.

Garanayake et al.[4] proposed a system to suggest the best crops to grow based on N, K, P soil pH, temperature, precipitation, and rainfall. The study predicted the yield of Five crops in India, including ragi, gram, potato, rice, and onion. The system used various learning techniques including Naive Bayes (NB), RF, DT, LR, support vector machine (SVM), and XGBoost to recommend the most suitable crop for multiple regions of India. With this strategy, 94.78% accuracy in onion yield was achieved.

The authors examined the climate impact on wheat production in Ukraine in[5]. They used daily bulk of minimum, average temperature, maximum, and precipitation. Provided by the Global Historical Climatology Network for flake coverage data from 1985–2018 calm from the 190 meteorological stations across Ukraine. The authors used machine learning, specifically random forest models, which have proved to be more effective than traditional methods. Average weather conditions and extreme weather conditions were found to affect 53–62% and 36–40% of crop yield respectively. The analysis highlights the characteristics of the climate at different stages of wheat growth and emphasizes the importance of considering seasonal changes. This study analyzes how regional differences in critical temperature, weather conditions, and climate affect crop production and reveals the need for strategic change. These findings offer insights for farmers and crop breeders, highlighting the importance of understanding and addressing regional differences in the impact of climate on crops. This knowledge could help ameliorate climate change and improve world grain production.

A complete solution to predict crop yield is provided by Pande in[6]. The prompted system connects farmers through portable applications and determines the user's location with the help of GPS. The user presents the area and land variety as input. They proposed a model that predicted the crop yield by studying characteristics like rainfall, temperature, area, season, soil type, etc. The system also determines the best point in time to use fertilizers. Machine-learning models allow the selection of the most commercial crop names to predict fruitage yields for the user's chosen crop and guide them on using the right fertilizers. The researchers used past information from trusted sources like Kaggle and the Indianwaterportal, concentrating on the Karnataka and Maharashtra regions. They suggested 20 different crops like Arhar, Bajra, Black pepper, Castor, Cowpea, Dry chilies, Dry ginger, Gram, Groundnut, Horsegram, Jowar, Linseed, Maize, Moong, Nigerseed, Onion, Rabi pulses, Peas, Potato, and Ragi. They used machine learning models like SVM, RF, multivariate linear regression, artificial neural network (ANN), and k nearest neighbor (KNN) to predict the crop yield. The RF gave the most accurate result with 95% accuracy.

Gum et al. established a study on data spanning 51 years, in[2]. The study examined the use of machine learning to predict when rice will grow in Sanlıurfa city of Turkey. They analyzed daily weather data and predicted three planting dates (early, normal, late) using the nearest neighbor, support vector machine, and tree pruning. Climate variables included maximum daily temperature, average daily temperature, minimum daily temperature, daily humidity, and daily precipitation in the starting 300 days of each year. This study provides a genetic method for selecting important traits that leads to better results, In particular, when KNN is used for prediction, the correct classification reaches 37% and 92%, respectively. This approach provides modeling tools that can help farmers adapt to climate change and manage permaculture by accurately predicting planting dates.

Gulati et al. proposed a framework that explored crop production prediction in India using six different machine learning methods in[7]. The data used in the article includes states (West Bengal, Maharashtra, Uttar Pradesh, Gujarat, etc.), various crops (sugar, wheat, coconut, pulses, etc.), various seasons (rabbi, kharif, whole, summer, etc.), long-term crops over the years and other parameters like rainfall, temperature, pH, humidity. To predict crop yield, using gradient boosting regressors, ridge regression, linear regression, RF, DT, and polynomial regression. After the analysis, it was found that the accuracy of the supported gradient regressor in regression prediction was 87.90%. The random forest regressor performed better when predicting yield, with approximately 98.9% accuracy.

Proposed by Ali et al. a crop-recommended system using machine learning techniques is based on temperature for Pakistan's farmers in[8]. Analyzes data on rice, cotton, wheat, and sugar. They examined the yield of rice according to temperature and found that the maximum temperature of cotton declines within the range of 25°C to 35°C, while that of rice declines within the range of 12°C to 22°C, and the temperature range of sugar cane is from 20°C to 35°C and the maximum temperature of rice is in the range of 20°C to 35°C. They analyzed, measured, and predicted results for all crops (meat, cotton, sugar cane, and wheat) following almost the same pattern and basis. 90% accuracy is used as the identification algorithm.

The authors developed a method to predict crop yield based on historical data[9]. In this system, product selection depends on what is suitable for the selected area. Factors include location, soil, and food plants. Machine learning algorithms used to predict crop yield, recommend fertilization, and distribute soil are SVM and RF. The results show that random forest works best for land classification, achieving an accuracy of 86.35%, while SVM is good for the prediction of crops with an accuracy of 90.47%.

In[10], Dosi et al. proposed a strategy for recommending crops using machine learning algorithms. The system consists of two parts: i) a forecasting tool that indicates the suitable crop and ii) a rain gauge. It has considered five major crops i.e. bajra, maize, jowar, rice, and wheat, and fifteen minor crops; cotton, jute, barley, pulses, ground nut, other legumes, potatoes, rapeseeds and tur, mustard, ragi, sesames, soya beans, sunflowers. Characteristics like soil type, precipitation, aquifer thickness, soil pH, topsoil thickness, temperature, and location were taken into account. The system used different learning algorithms like DT, neural networks, KNN, and RF. It classified multiple tags and used a neural network algorithm to achieve a 71% accuracy for rain and over 91.00% accuracy for favorable crop prediction.

Another method of selecting the best crop according to climate and soil conditions is presented in[11]. The method is split into two parts: using a recurrent neural network (RNN) algorithm to predict the seasonal weather and using RF classification to come to a decision on which crop is the best. Five years of climate data from NRSA Hyderabad station, specifically in the agro-climatic zone of Telangana, were used in the study. The crop selection model looked at soil conditions, focusing on 10 different crops. The results showed that RNN works better than traditional AI, and random foresters recommend better harvests per acre. The method can also determine the best time to plant crops based on weather forecasts, creating a successful way to select crops based on the weather.

In another study, Manikrao et al. used the DT algorithm in their model and referred to its use in decision-making at each level of the binary tree in[12]. They trained samples of three major crops: soybeans, sugar, and tulpa. Their research has shown that approximately 50% of crop yield is unnatural by climate, such as humidity, temperature, precipitation, and solar intensity. They also include soil elements such as soil pH, soil organic matter, soil minerals, and soil moisture. Despite the existing approaches to crop recommendation for the agricultural sector, these studies lack the desired accuracy necessitating further research. A brief overview of the discussed literary works is presented in Table 1.

Ref.	Year	Parameters	Crop Studies	Data Source	Techniques Used	Shortcomings/limitations	Performance
8	2021	temperature-based	sugar, wheat, cotton, and rice	Government database	ML model Polynomial Regression	Limited number of crops (4) are considered and using old scheme with less accuracy	PR = 94%
3	2023	Soil, weather, past data	rice, wheat, maize, soybean, sugarcane	Satellite images, climate data	Multiple Linear Regression (MLR), LSTM with feature engineering.	5 crops are considered limited and using an old scheme with less accuracy	86.3%
?	2023	Weather data: rainfall, temperature, humidity, wind speed Soil data like pH and nutrient content	wheat, rice, maize, soybean, cotton	Andhra Pradesh Government	Random Forest, K-Nearest Neighbors, Support Vector Machine, and Weather forecasting using ARIMA model	5 crops are considered limited	RF=92%
12	2020	Soil conditions, weather conditions, crop suitability	10 crops	NRSA Hyderabad station	Recurrent Neural Network (RNN), Random Forest Classifier	10 crops are considered limited	93%
?	2022	Rainfall and temperature	Potatoes and Maize	NRSA Hyderabad station	Data mining, Random Forest, Polynomial Regression, and SVR	2 crops are considered limited	RMSE(potato=510.8 and maize= 129.9) R2(0.875 and 0.817
10	2018	temperature, rainfall, location, and soil condition	Twenty crops	No defined source	Multi-label classification, NN, KNN, DT, RF	20 crops are considered limited and using old schemes with less accuracy	NN=91%

Table 1. Comparison among the existing work on crop recommendation.

Research gap

- Previous studies on recommendation systems for crop suggestions have limited scope. These studies have often focused exclusively on either vegetables or fruits rather than considering a broader range of agricultural options.
- The existing scheme has worked on recommending a small number of crops. Furthermore, they have employed simpler models providing limited performance scores, which requires further exploration to design and test new hybrid models for improved accuracy.

The current study aims to provide a robust solution that covers a higher number of crops compared to existing studies. Moreover, the objective is to increase the accuracy and robustness of crop recommendations.

Proposed methodology

Despite recent attempts to find solutions, challenges still exist in providing effective crop recommendations. The proposed solution aims to address these challenges by developing machine learning models that consider vital parameters like N, K, P, rainfall, temperature, humidity, and pH, directly impacting farming. The objective is to suggest a broader range of suitable crops for the season, reducing farmers' difficulties in crop selection and ultimately increasing yield. The proposed model recommended the best crops for soil. The integration of agriculture and machine learning promises to advance the agriculture field by enhancing optimizing and yield resource utilization. The dataset undergoes comprehensive preprocessing, using a training split ratio is 80% and a testing ratio is 20%. The workflow of the proposed methodology is given in Figure 1. The methodology comprises several phases including collection of data, preprocessing, explorative data analytics, correlation, splitting of dataset, employing machine learning models, and crop recommendation.

Phase 1: Dataset collection

Utilizing data from previous years plays a vital role in forecasting current performance. We collect historical data[13,14] from reliable sources like Kaggle and IEEE Dataport. The dataset was originally collected by the Agricultural research stations and weather stations from Islamabad Capital of Pakistan. The dataset incorporates information on N, K, and P levels in the soil, alongside temperature and rainfall measurements, elucidating their impact on crop growth. This dataset serves as a valuable resource for formulating data-driven recommendations to optimize nutrient and environmental conditions, ultimately enhancing crop yield. The dataset contains 2200 instances. It includes twenty-two different crops: maize, rice, chickpeas, pigeon peas, broad beans, kidney beans, mung beans, lentils, pomegranates, black beans, bananas, grapes, watermelon, mangoes, melon, oranges, cotton, papaya, apples, coffee coconut, and jute.

Phase 2: Data preprocessing

By using categorical values such as labels in the datasets are managed using the label encoding method. The analysis of descriptive features related to the dataset is presented in Table 2. It provides details on attributes, attribute types, and their corresponding descriptions[15,16].

The dataset has seven features while the 8th attribute is the label of the class which indicates the crop name. Each attribute shows different features of the dataset which helps to recommend a specific crop. The 'N' indicates Nitrogen content, 'P' shows the quantity of Phosphorus, and 'K' refers to the Potassium content. 'Temperature'

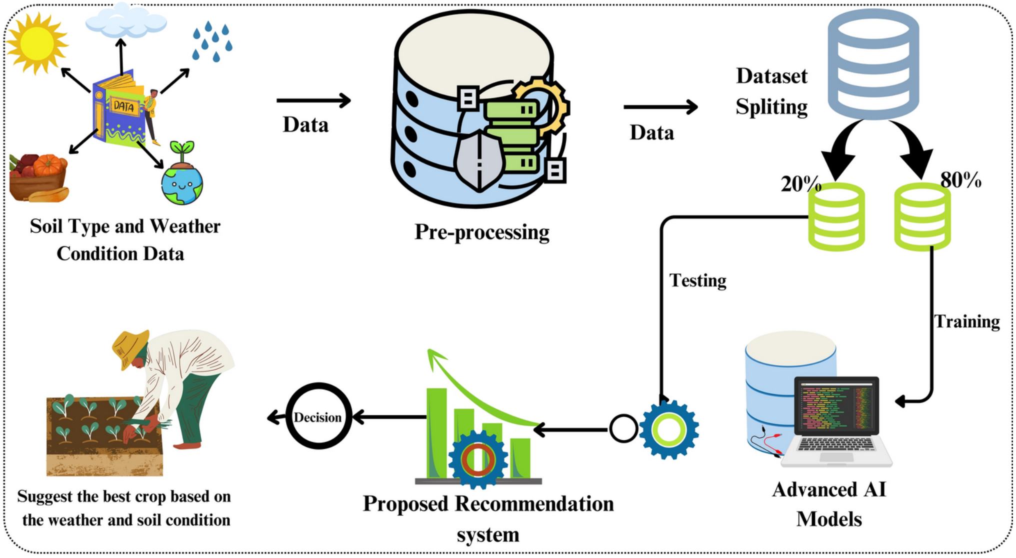
DATA
Soil Type and Weather Condition Data
Data
Pre-processing
Data
Dataset Spliting
20%
80%
Testing
Training
Advanced AI Models
Proposed Recommendation system
Decision
Suggest the best crop based on the weather and soil condition

Fig. 1. The methodological infrastructure of our proposed research suggests the best crop.

Feature	Non-Null Count	Data type	Description
N	2,200	int64	Contains the Nitrogen.
P	2,200	int64	Contains the Phosphorous
K	2,200	int64	Contains the Potassium
Temperature	2,200	float64	The average soil temperature for biological activity ranges from 50 to 75F.
Humidity	2,200	float64	the total amount of water, including the steam in an unsaturated soil.
pH	2,200	float64	The pH range 5.5 to 6.5 is best for plant growth as the availability of nutrients is ideal.
Rainfall	2,200	float64	A well-balanced rain and proper irrigation can lead to faster-growing plants.
Label	2,200	object	Contains different crop names

Table 2. Analysis of descriptive features in the soil type and weather dataset.

Features	Mean values/ per feature	Standard deviation	Min. values	25% values	50% values	75% values	Max. values
N	50.55	36.91	0.00	21.00	37.00	84.25	140.00
P	53.36	32.98	5.00	28.00	51.00	68.00	145.00
K	48.15	50.65	5.00	20.00	32.00	49.00	205.00
Temperature	25.62	5.06	8.83	22.77	25.60	28.56	43.68
Humidity	71.48	22.26	14.26	60.26	80.47	89.95	99.98
pH	6.47	0.77	3.50	5.97	6.43	6.92	9.94
Rainfall	103.46	54.96	20.21	64.55	94.87	124.27	298.56

Table 3. The soil type dataset features related information.

shows soil temperature in Fahrenheit, ’Humidity’ is the amount of water content, ’pH’ shows soil’s acidic or alkaline level and varies between 5.5 to 6.5 while the ’Rainfall’ feature shows the quantity of rain in mm.

Phase 3: Exploratory data analytics

Exploratory data analytics refers to the process of data analytics, using data sets, graphical representation, and content analytics. Feature relationship analytics and data visualization methods contribute to the suggestion process of the proposed models. In Table 3 are shown the results of the data analytics process. Special statistics calculated mean, standard deviation (std. varies), minimum, 25%, 50%, 75%, and maximum. Inspection shows that the file contains 2200 lines of data. This analysis provides insights related to the central tendency of data, its distribution, variability, and outliers.

Table 3 provides quantitative information about the dataset concerning standard deviation, minimum values, maximum values, etc. Often, such information is useful and can be used to improve the performance of models. The mean value indicates the average value of data for each feature while min., and max. values indicate the minimum and maximum values for the data given for each feature. Similarly, standard deviation is taken from all the data for each feature, and 25%, 50%, and 75% show the first, second, and third quartiles for each feature.

Figure 2 provides a visual comparison of different parameters (nitrogen, phosphorus, potassium). It helps understand the relationship between these parameters and the proposed crop, making it easier to make decisions such as crop selection and management in agriculture.

In Figure 3 the bar chart visually shows the relationship between three parameters (temperature, humidity, and pH) and their impact on crop recommendations. Each bar on the chart shows the average of one of these parameters.

In Figure 4the bar chart shows the relationship between one parameter rainfall and their impact on crop recommendations. Each bar on the chart shows the average of one rainfall. Exploratory analysis exhibits that soil pH, nitrogen, phosphorus, potassium, temperature, and rainfall contents are very important for farmers to determine, which kind of crops can be fully grown in soil type[17].

Phase 4: Correlation analysis

While designing machine learning models, it is important to select appropriate features for training models that are strongly related to the target variable. The chosen features correlation analysis is illustrated in Figure 5, indicating that all selected features exhibit a positive correlation with each other. Notably, the K and P features exhibit a high correlation of 0.74. This analysis signifies that the features within our research dataset exhibit favorable correlation values, making them well-suited for training machine learning models, particularly for crop recommendation.

The classification analytics of the bar plot is illustrated in Figure 5. The analytics of different features of the collected data. It shows that each feature has a different impact on the crop, demonstrating the importance of features in predicting each bar line.

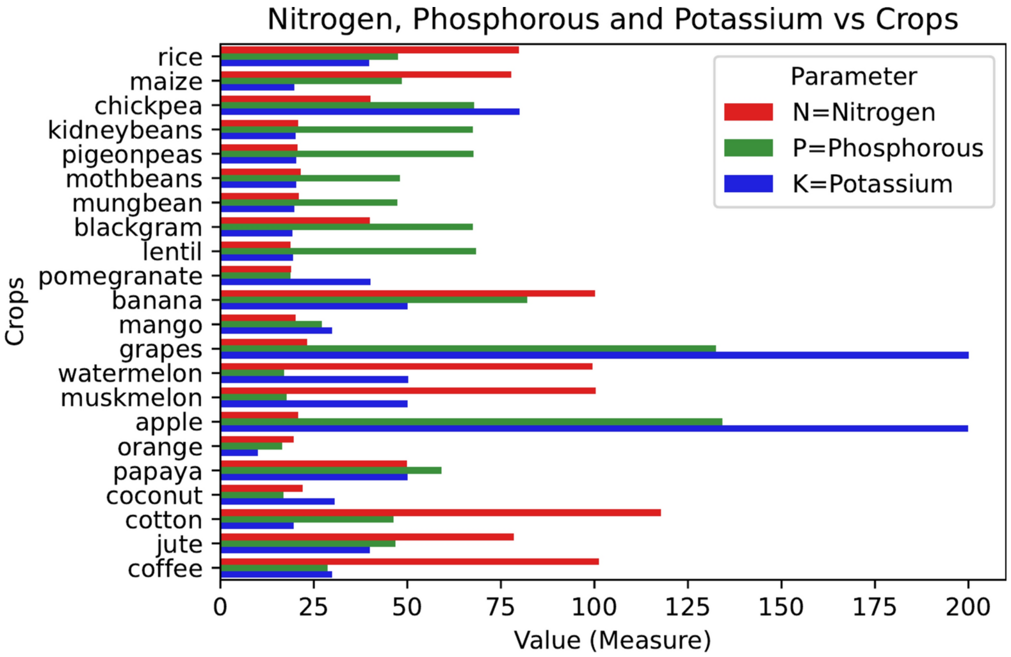

Fig. 2. Distribution of N, p, and K input analysis with respect to various crops.

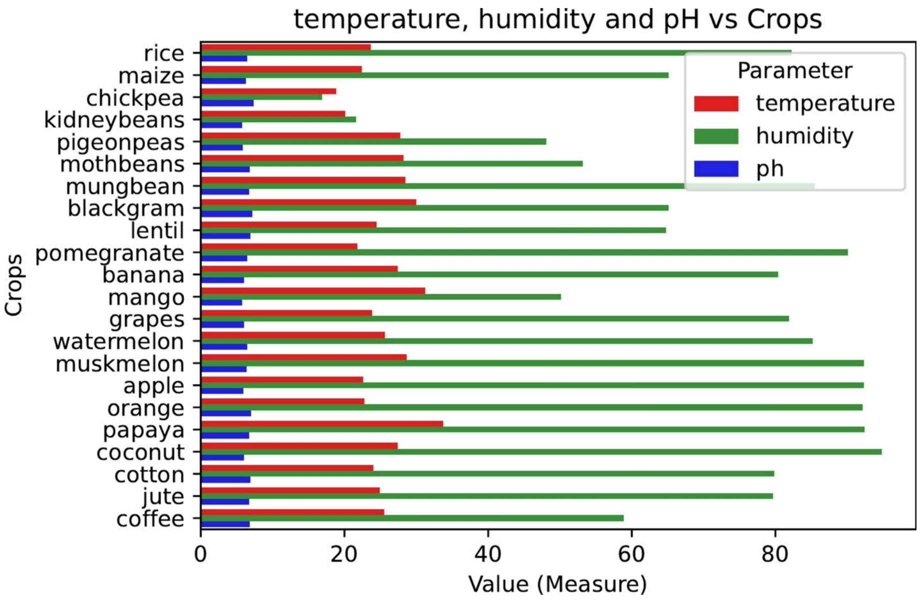

Fig. 3. Variation in temperature, humidity, and pH input analysis with respect to various crops.

Phase 5: Dataset splitting

To remove the overfitting of the model and assess the trained models on seen test data, data splitting is employed. The data is divided into 2 segments for testing and training using the train-test-split function of the sklearn library in Python, with an 80:20 ratio. The model is trained using the larger 80% portion, while a 20% portion of the dataset (2200 instances) is used for testing machine learning models' performance.

Phase 6: Utilized machine learning techniques

This section examines the machine-learning models[18–20] employed for best crop recommendations based on temperature and rainfall. The working mechanism of machine learning models is described, with a focus on six recommended machine learning models in our research study.

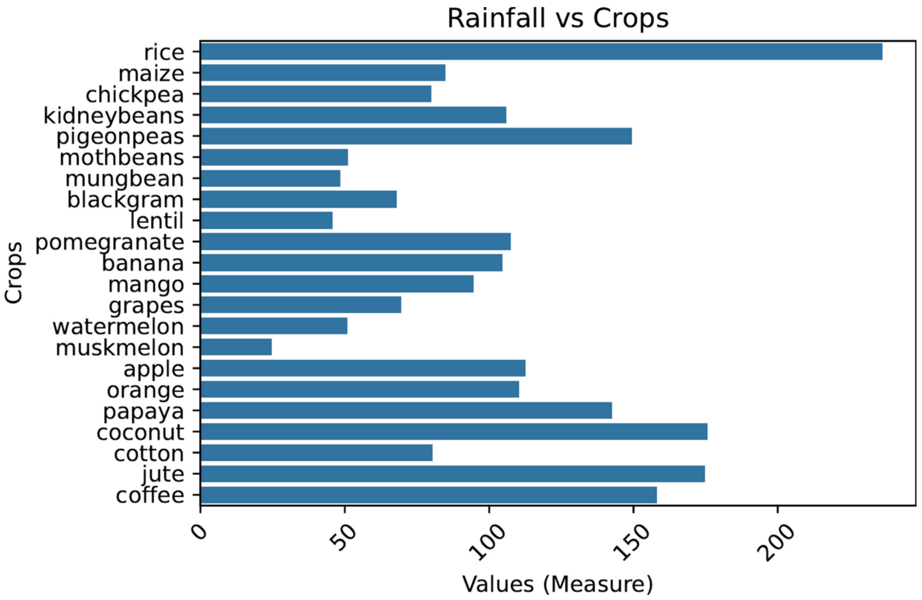

Fig. 4. The rainfall distribution and various crops.

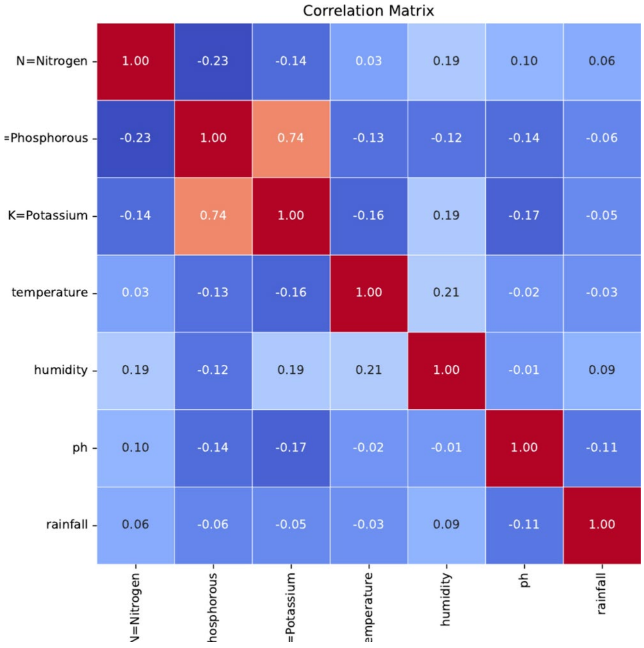

Fig. 5. Evaluation of the relationship between input features within a dataset bee-line correlation analysis.

Extra trees classifier

Similar to random forests, additive tree classifiers are a popular ensemble learning method used in agricultural systems to recommend crops based on soil type and precipitation[21]. This classifier builds multiple decision trees using a random subset of the training data and a random subset of the features. An additional tree classifier for crop recommendation can be expressed as:

$$Z(x) = \; frac1T \sum_{t=1}^{T} z_t(x) \tag{1}$$

where $Z(x)$ represents the expected cleanup recommendation for input rows x. Each $z_t(x)$ is a decision tree trained on a random subset of the training dataset and features. The final prediction is obtained by aggregating the predictions of all decision trees, often through majority voting.

Multilayer perceptron
The MLP is a type of artificial neural network commonly used in agricultural systems for crop recommendation based on features such as soil type and rainfall[22]. Here's the formulation of an MLP for crop recommendation.

Let X represent the input feature matrix, where every row corresponds to a plot of soil, and each column represents a specific feature such as soil type, rainfall, temperature, etc. Each row is denoted as $x_i \in \mathbb{R}^d$, where d is the number of features.

Let Y be the corresponding target class variable, indicating the recommended crop for each plot of land. Each target class variable is denoted as $y_i \in 0, 1, \ldots, C$, where C is the number of crop classes.

Decision tree
The DT model has emerged as a forceful tool in agricultural systems for crop recommendation based on soil type and rainfall. By leveraging a tree-like structure, the DT model effectively captures the relationships and patterns inherent in environmental features, facilitating accurate recommendations for crop selection. Through recursive partitioning, the model partitions the dataset into inferior subsets based on specific features, ultimately generating decision rules that enable the classification of plots of land as suitable for cultivating certain crops[23].

Consider a dataset $\mathscr{S}$ comprising M instances representing plots of soil, each characterized by a rest of features X_j, and labeled suitable $(Y_j = 0)$ or unsuitable $(Y_j = 1)$ for a particular crop. The objective is to build a DT model to classify landing instances of soil plots. At each decision tree node, the optimal splitting criterion needs to be determined. Gini impurity is one commonly used measure, denoted as

$$\mathrm{Gini}(\mathscr{G}) = 1 - \sum_{b=0}^{C-1} \left(\frac{|\mathscr{G}_b|}{|\mathscr{G}|} \right)^2 \tag{2}$$

where, C shows the number of classes, $\mathscr{G}_b$ denotes the subset of rows inclusion to class b, and $|\mathscr{G}|$ indicates the total number of rows in $\mathscr{G}$.

Logistic regression
LR[24,25] models are widely used in agricultural systems to recommend crops based on environmental factors like rainfall and soil type. LR models excel at binary classification tasks, accurately recommending crops that are suitable or unsuitable for cultivation based on input data.

In the context of crop recommendations, let us denote by X the characteristics of a particular piece of land, including soil type, precipitation, temperature, and other environmental variables. The binary variable Y indicates whether a particular crop is recommended to be grown in a particular region.

The LR model imagines an unbending relationship between traits and the likelihood of recommending a particular crop.

$$\log \left(\frac{L(y = 1 \mid x)}{1 - L(y = 1 \mid x)} \right) \tag{3}$$

Random forest
The RF algorithm has obtained significant interest and adulation in agricultural systems for crop recommendations based on soil type and precipitation[26,27]..The ensemble learning method has shown optimistic accuracy in efficiently processing complex agricultural data. Combining the DT model RF effectively captures the patterns, relationships, and correlations in environmental characteristics to accurately recommend crops suitable for cultivation. X shows the feature matrix as input. Here, every row consists of a piece of land and every column shows a specific characteristic such as soil type, precipitation, temperature, etc. Every instance is indicated by $X_j \in \mathbb{R}^f$. Where f is the number of objects. The corresponding target variable is shown with Y to the recommended crop for each land. Every target variable is indicated by $Y_j \in \{0; 1; \ ldots; B\}$. In which B is the number of crop information.

RF models comprise a decision tree T, denoted $r_t(x)$. Here $t = 1; 2; \ ldots; A$. Every decision tree is trained to use a randomly selected subset of the agricultural dataset. The RF model integrates the prediction from all using the decision trees to predict the recommended crop for a new piece of land, denoted x. A majority vote obtains the final recommendation, and the recommended crop in which the mentioned target class collective mostly votes beyond all decision trees. RF model for crop recommendation is indicated as follows.

$$K(x) = \ frac1R \sum_{j=1}^{R} h_j(x) \tag{4}$$

where $K(x)$ shows the expected cleanup recommendation for input instance x.

Extreme gradient boosting classifier
XGBoost classifier is an algorithm of machine learning widely used in agricultural systems for crop recommendation placed on soil type and weather. Renowned for its efficiency, speed, and accuracy, XGBoost belongs to the boosting algorithm family[28]. The XGBoost algorithm iteratively constructs decision trees, with each subsequent tree aiming to rectify the errors of its predecessors. It amalgamates the prediction of numerous weary learners (individual DT) to produce the final recommendation.

The formulation for the crop recommendation problem using XGBoost can be expressed as follows:

$$\text{Target} = \sum_{j=1}^{m} \text{Loss}(\hat{Y}_j, Y_j) + \lambda \sum_{p=1}^{P} \Omega(f_p) \quad (5)$$

where:

- $\text{Loss}(\hat{Y}_j, Y_j)$ signifies the inconsistency between the predicted values $(\hat{Y}_j)$ and the actual values (Y_j) and measures the prediction error.
- $\Omega(f_p)$ represents the complexity penalty term applied to each decision tree (f_p), penalizing the model's complexity.
- P indicated the total number of trees in the novel ensemble.
- λ serves as a regularization parameter, governing the trade-off between minimizing the prediction error and controlling the model's complexity.

Phase 7: Fine-Tuning hyperparameters for models

To promote the performance of machine learning models, a systematic hyperparameter tuning process is implemented. To ensure the best recommendations, we use the iterative method a k-fold cross-validation process, to determine the optimal hyperparameters which divides the data into training, validation, and testing. Table 4lists the hyperparameters selected for the proposed method. The results of the analysis show that the negative agreement we identified was successful in accepting crops and achieved good performance[29].

The primary reason for choosing systematic hyperparameter tuning for the proposed model was due to its complex architecture, which made it challenging to apply automated tuning methods effectively. We selected a manual, systematic approach to carefully explore and control the hyperparameters, such as learning rate, batch size, and number of hidden layers, before training the model.

The systematic tuning was guided by domain knowledge and iterative testing, allowing us to make informed decisions with relatively lower computational overhead. This approach helped us manage computational resources effectively by avoiding exhaustive searches or high-dimensional parameter spaces. Additionally, while there are many sophisticated hyperparameter tuning methods, such as grid search, random search, or Bayesian optimization, they can be computationally expensive. Since our primary goal was to establish a baseline performance, systematic tuning was a practical first step.

Phase 8: Proposed ensemble model

To improve decision-making by learning features, an ensemble model RFXG was implemented, combining more than one architecture, Like an RF and XGB model, and making an RFXG ensemble model. The experiments showed that RF and XGB performed well when we trained RF and XGB individually, so we combined RF and XGB to yield better results.

Combinations of RF and XGB were employed to model soil and weather parameters from the dataset using the hard-voting technique. We describe the operational rules of the implemented machine learning techniques ETC, MLP, DT, RF, XGB, and LR and introduce the ensemble model RFXG. Rather than encompassing all machine learning models, we strategically choose six benchmark models to represent various categories. These models encompass ETC for randomized decision trees, RF for ensemble learning, LR for binary classification, XGB for boosting ensemble, MLP for neural network-based learning, and DT for hierarchical classification. This selection aims to provide a comprehensive analysis, training each algorithm individually with the soil type data and evaluating their accuracy for comparison with the proposed ensemble learning model RFXG.

The RFXG model is proposed to improve prediction accuracy by accounting for data differences during model training. It combines RF and XGB algorithms using the advantages of simplicity, low data quality, and fast implementation. Unlike traditional methods, RFXG dynamically evaluates the importance of new data based on their similarity to the training context, without the need for an index. By integrating new machine learning technology in Figure 6. It combines the expression function of two algorithms to provide a unique hyperplane fusion. This method is capable of solving the overfitting problem and optimizing the dataset. RFXG's architecture facilitates direct application to new data based on machine learning training and testing procedures. Although the principle remains the same, tuning the hyperparameters that make up the algorithm can improve the performance of different data sets.

The ensemble model combines RF and XGB based on the following points:

Model	Hyperparameters
ETC	min_samples_split=2, max_depth=6, n_estimators=7, random_state=0
MLP	max_iter=100, hidden_layer_sizes=10, random_state=0
RF	max_depth=9, n_estimators=2, min_samples_split=9
DT	max_depth=4, min_samples_leaf=4, Criterion='entropy', min_samples_split=70, random_state=0
LR	solver='lbfgs', multi_class='auto', random_state=0, C=1.0, max_iter=10
XGB	random_state=0, n_estimators=6, max_depth=2, learning_rate=0.1

Table 4. Optimal hyperparameters for machine learning models for crop suggestion.

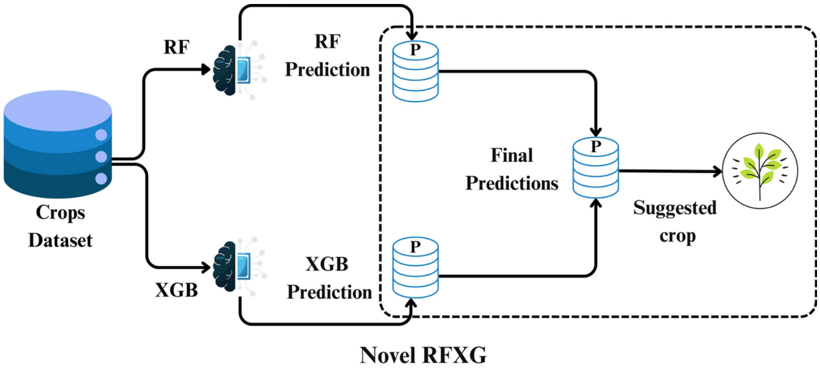

Fig. 6. Ensemble machine learning model RFXG.

- When analyzing the existing literature, tree-based models like RF and XGB are found to produce good performance for crop recommendations and similar tasks.
- Their ensemble makes sense because RF's ability to handle high-dimensional data and feature interactions when combined with XGB's efficiency in handling large datasets and its robustness to outliers can provide better results than using an ensemble of other models.
- Preliminary experiments showed that RF and XGB performed well as stand-alone models leading to joining them for better accuracy.
- When combined, RF and XGB have improved feature representation due to RF's feature importance and XGB's feature interaction.
- RF's ensemble approach and XGB's regularization can provide enhanced robustness.
- Better handling of non-linear relationships is possible using RF's decision trees and XGB's gradient-boosting approach.

The dataset consisted of twenty-two different crops, turning it into a multi-classification task. Crop suitability, in this context. These conditions encompass various factors such as soil properties, climate, temperature, precipitation, and sunlight[11]. Recommending the best suitable crop using soil factors and environmental conditions empowers farmers to select the most suitable increasing productivity and improving resource utilization.

Results and discussions

This section presents the analysis of experimental findings and evaluations. The comparative assessment of machine learning techniques is validated using performance metrics like Accuracy, recall, precision, and F1. These classification metrics are scrutinized through mathematical notations and computed scores for the employed learning techniques.

Experimental setup

This section outlines the experimental setup employed to develop machine learning techniques. The research utilizes the Python 3 programming tool for data analysis, model building, and evaluations. The Scikit-learn library module is employed for machine learning model construction and testing. TensorFlow and Keras library modules are utilized for building and testing machine learning models. The online Google Colab used for conducted research is in[30] frame in GPU settings. Using a backend equipped with 8GB of RAM to make a high-performance GPU. The experiments utilized a computational framework with a PC system an intel(R) Core(TM)i5 6300U, 2.40GHz, 2.50GHz CPU, and an HP processor.

Evaluation parameters

The machine learning algorithms for classification, including our innovative RFXG approach, are employed to suggest the best crop. The training process involves teaching these methods using specific data, allowing the model to grasp the data patterns and make predictions. The effectiveness of the machine-learning scheme is commonly determined using metrics metrics like accuracy, F1 score, precision, and recall. These metrics quantify different aspects of model performance, including the correctness of predictions, the strength to determine positive instances exactly, and the balance between recall and precision.

Accuracy represents the total integer of correct suggestions from the prediction model across all predictions. The formula is

$$Accuracy = \frac{TPR + TNR}{TPR + FPR + TNR + FNR} \tag{6}$$

where TPR is a true positive, indicating our predictive model is correct, FPR is a negative number the positive model is incorrect, and TNR is negative. The value indicating a positively identified Pattern predicts a negative sample and the FNR is the negative value representing an incorrectly identified negative sample[31].

Model	Precision	F1	Recall
ETC	0.88	0.87	0.88
MLP	0.41	0.43	0.52
RF	0.96	0.96	0.96
DT	0.59	0.63	0.72
LR	0.88	0.88	0.89
XGB	0.96	0.96	0.96
Proposed RFXG	0.98	0.98	0.98

Table 5. Accuracy for all machine learning models.

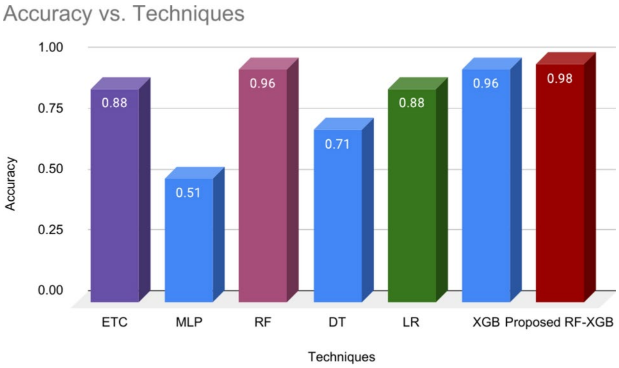

Fig. 7. Classifiers' performance of all machine learning models by employing various metrics and the accuracy score.

Precision is the ratio of the algorithm exactly predicted by the model to the overall required data quality. The formula is

$$Precision = \frac{TPR}{TPR + FPR} \tag{7}$$

Recall shows the preciseness of the model and using the formula is

$$Recall = \frac{TPR}{TPR + FNR} \tag{8}$$

F1 performance measures are based on the combination of recall and accuracy and are evaluated as the relationship between recall and accuracy. The F1 score equation is

$$F1 = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{9}$$

where *P* indicates precision, and *R* shows the recall.

Results of applied models

Table 5 shows the classification report of each machine learning model. ETC achieves 88% precision, 88% recall, and 87% F1 results. MLP has an accuracy of 41%, 52% is a recall rate, and an F1 score is 43%. RF achieves 96% precision, 96% recall, and 96% F1 accuracy. DT achieved 59% precision, 72% recall, and 63% F1 accuracy. LR achieves 88% precision, 89% recall, and 88% F1 accuracy. XGB has an accuracy is 96%, 96% is the recall rate, and 96% is the F1 score. RFXG concept achieves 98% accuracy, precision, recall, and F1 scoring.

In summary, RFXG consistently outperformed other models in predicting crop yield for all crop parameters analyzed based on N, K, P, precipitation, humidity, temperature, and pH notes. The RFXG consistently demonstrates good accuracy, precision, repeatability, and F1 scores compared to other models. This comparison shows that RFXG is better than existing ML models for existing datasets.

Figure 7 compares the accuracy scores of different machine learning techniques. The analysis reveals that our proposed approach attains a notable accuracy score of 98%, surpassing other techniques. Additionally, the RF and XGB techniques exhibit commendable accuracy scores. In contrast, the ET Classifier, DT Classifier, LR and MLP Classifier models achieve the lowest accuracy score of 51%. This comparison highlights the superior performance of our proposed approach.

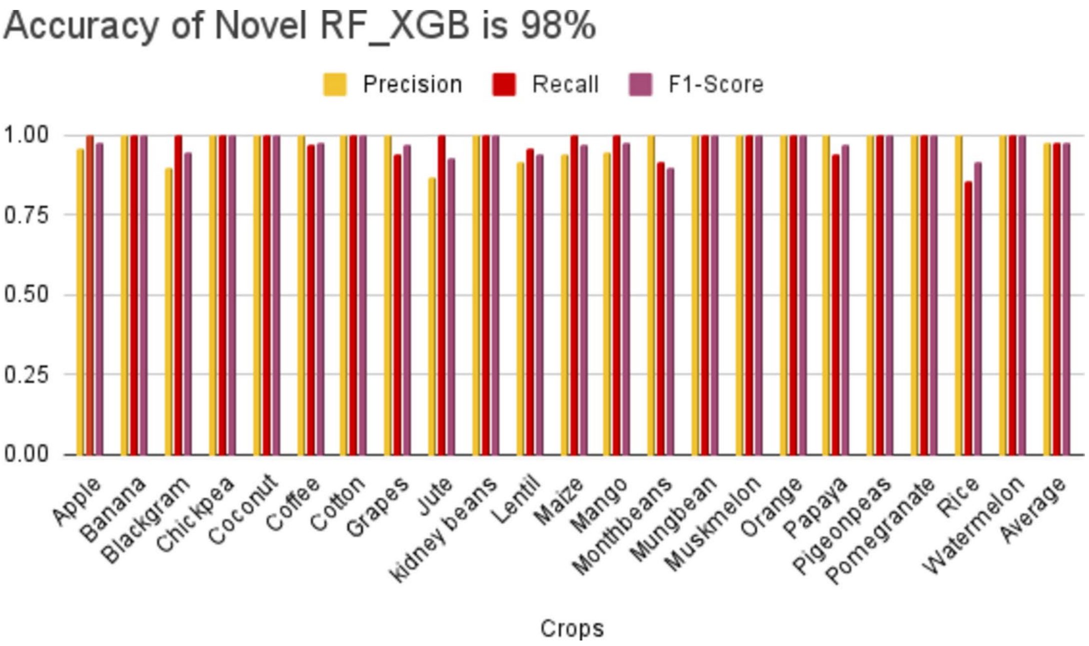

Fig. 8. Evaluation report on the classification performance of the proposed RFXG model for all classes.

Model	Accuracy score	Standard Deviation (±)
ETC	0.8827	0.0187
MLP	0.5350	0.0354
DT	0.6414	0.0253
RF	0.9595	0.0158
LR	0.8600	0.0250
XGB	0.9577	0.0169
RFXG	**0.9810**	**0.0119**

Table 6. Performance validations for machine learning models and proposed RFXG model.

Class-wise results of the proposed approach

In this portion, the accuracy of our newly proposed novel scheme RFXG is examined. Figure 8 shows the performance measurements. The results showed that the machine learning technique called RFXG achieved a good score of 0.98. Additionally, the performance of each class shows the best score. Papaya stood out with the lowest score of 0.90, and other crops highest score than papaya. In summary, the use of the new RFXG model has proven to be the best way to deliver the promise.

Cross-Validation analysis

In Table 6, the accuracy of 10 cross-validations K-fold for employed machine learning techniques is shown. The purpose of cross-validation is to analyze the performance of models for each where the models are trained on k-1 folds and tested on one fold. Cross-validation helps reduce the probability of model overfitting by training the model on multiple folds. Consequently, the model can generalized better and provide more robust results.

The cross-validation analysis classifies the accuracy of every applied algorithm. For this purpose, 10-fold cross-validation is applied for all the models. The outcome reveals that the LR model achieves an accuracy score of 0.86 with 10-fold, followed by DT, MPL, and ETC especially, the proposed RFXG technique attains a higher 0.98 accuracy. Overall, the study confirmed the good performance and generalization potential of each of the methods used for crop recommendation based on numerical soil data[32].

Crop recommendation with proposed approach

Using environmental and soil data to the initially trained for the proposed technique. Subsequently, based on the current year's weather and soil type data, the model suggests the best crop. The valued characteristics of sole crops are then compared with crop expectations for the particular area. If the soil type meets the criteria, the crop is recommended. The model, when comparing crops with the same harvesting period, recommends the most suitable crops for the specific location shown in Table 7.

N	P	K	Temperature (°C)	humidity (%rh)	pH	Rainfall (MM)	Crop Recommended
90	42	43	20.879	82.002	6.502	202.935	Rice
107	34	32	26.774	66.413	6.780	177.774	Coffee
73	58	21	19.972	57.682	6.596	60.651	Maize
14	67	22	23.825	24.754	5.624	84.641	Kidneybeans
26	11	11	13.703	90.955	7.609	106.294	Orange
78	58	44	26.800	80.886	5.108	284.436	Rice
31	29	35	27.187	92.199	6.137	141.322	Coconut

Table 7. Crop recommendation based on soil and weather parameters.

Reference	Techniques	Performance
[3]	Multiple Linear Regression (MLR), LSTM with feature engineering	86.3%
[33]	machine ensemble method (KRR)	MSE= 0.016, R^2=0.99
[34]	ARMA, SAR IMA, ARMA with exogenous variables compared, selecting the best for predicting rainfall and temperature	fuzzy logic model for crop yield prediction
[35]	MapReduce and K-means clustering	provides a mean result on the data.
Proposed	**Ensemble model RFXG**	**98%**

Table 8. Comparative analysis of the proposed research approach with other studies.

Comparison of proposed approach with state-of-the-art research

The results of the proposed approach are compared with the other leading studies, and the results are presented in Table 8. In this study, we introduced a novel ensemble model and implemented machine learning models including the ETC, MLP, RF, DT, LR, and XGB, which were applied to the obtained dataset. The assessment of comparative performance includes metrics like precision, accuracy, F1, and recall. The findings indicate that the novel approach outperforms the other state-of-the-art models, getting an impressive accuracy of 98%.

Discussion and limitations

This research proposes an ensemble approach for crop recommendation based on soil properties data. The proposed ensemble learning approach called RFXG, combines RF and XGB models to suggest the best crop out of twenty-two major crops with respect to various soil properties such as N, K, pH, humidity, etc. The model outperforms other stand-alone models concerning accuracy, robustness, and cross-validation results. Similarly, it shows better performance compared to models from existing studies.

Despite its excellent performance, the proposed approach has several limitations. One limitation is the complex architecture of the proposed ensemble model, which may increase computational cost. Additionally, the 98% accuracy of the proposed approach may further need to be increased. Transfer learning models may be integrated with the proposed approach to fill this gap. Current experiments were carried out on a balanced dataset and showed good performance. However, the model needs to be tested on the imbalanced dataset to analyze its behavior.

Conclusion and future work

In this study, we aim to design an ensemble model to recommend suitable crops given the characteristics of soil, expected rain, and temperature information. The initial analysis involved the application of six well-established classical machine learning algorithms and hyperparameter optimization to achieve the best results. Further, we developed the ensemble model RFXG, combining two models RF and XGB. A comprehensive analysis was performed to estimate the performance of our novel proposed model in comparison to the six classical models. The result exhibited that RFXG achieved the best values for precision, F1-score, and recall of up to 98%. These results indicate the potential of the proposed models to identify the best crop that can significantly improve crop management practices. Additionally, our research study reaffirmed the importance of rainfall, K, P, N, pH, precipitation, and temperature as crucial factors influencing crop growth, yield, and productivity, aligning with findings from existing literature. In the future, we aim to design deep learning and machine learning models to predict crop yield to better estimate the benefits of selecting a particular crop.

Data availability

The datasets can be requested from Hadeeqa Afzal.

Received: 11 July 2024; Accepted: 29 January 2025
Published online: 12 March 2025

References

1. Vempati, A. V. *et al.* Agricultural problems and technology-based sustainable solutions for an impoverished village of bihar, india. In *2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC)*, 1–5 (IEEE, 2020).
2. Gümüşçü, A., Tenekeci, M. E. & Bilgili, A. V. Estimation of wheat planting date using machine learning algorithms based on available climate data. *Sustainable Computing: Informatics and Systems* **28**, 100308 (2020).
3. Iniyan, S., Varma, V. A. & Naidu, C. T. Crop yield prediction using machine learning techniques. *Advances in Engineering Software* **175**, 103326 (2023).
4. Garanayak, M., Sahu, G., Mohanty, S. N. & Jagadev, A. K. Agricultural recommendation system for crops using different machine learning regression methods. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)* **12**, 1–20 (2021).
5. Schierhorn, F., Hofmann, M., Gagalyuk, T., Ostapchuk, I. & Müller, D. Machine learning reveals complex effects of climatic means and weather extremes on wheat yields during different plant developmental stages. *Climatic Change* **169**, 39 (2021).
6. Pande, S. M. *et al.* Crop recommender system using machine learning approach. In *2021 5th international conference on computing methodologies and communication (ICCMC)*, 1066–1071 (IEEE, 2021).
7. Gulati, P. & Jha, S. K. Efficient crop yield prediction in india using machine learning techniques. *International Journal of Engineering Research & Technology (IJERT)* **8** (2020).
8. Ali, S. M., Das, B. & Kumar, D. Machine learning based crop recommendation system for local farmers of pakistan. *REVISTA GEINTEC-GESTAO INOVACAO E TECNOLOGIAS* **11**, 5735–5746 (2021).
9. Bondre, D. A. & Mahagaonkar, S. Prediction of crop yield and fertilizer recommendation using machine learning algorithms. *International Journal of Engineering Applied Sciences and Technology* **4**, 371–376 (2019).
10. Doshi, Z., Nadkarni, S., Agrawal, R. & Shah, N. Agroconsultant: intelligent crop recommendation system using machine learning algorithms. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1–6 (IEEE, 2018).
11. Pereira, L. et al. Prediction of crop coefficients from fraction of ground cover and height. background and validation using ground and remote sensing data. *Agricultural Water Management* **241**, 106197 (2020).
12. Jain, S. & Ramesh, D. Machine learning convergence for weather based crop selection. In *2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 1–6 (IEEE, 2020).
13. Gholap, S., Singh, H., Rajesh, R. & Barshe, S. Crop recommendation, https://doi.org/10.21227/3ays-5h79 (2024).
14. KHAN, M. A. Real_time_crop_recommendation system. https://www.kaggle.com/datasets/altafk/real-time-crop-recommendation-system. (Accessed on 04/12/2024).
15. Rajeena PP, F., SU, A., Moustafa, M. A. & Ali, M. A. Detecting plant disease in corn leaf using efficientnet architecture-an analytical approach. *Electronics* **12**, 1938 (2023).
16. Wu, J., Olesnikova, A., Song, C.-H. & Lee, W. D. The development and application of decision tree for agriculture data. In *2009 Second International Symposium on Intelligent Information Technology and Security Informatics*, 16–20 (IEEE, 2009).
17. Deshmukh, S. *et al.* Application of data analytics in agriculture sector for soil health analysis: Literature review. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, 1–4 (IEEE, 2019).
18. Raza, A. *et al.* An improved deep convolutional neural network-based youtube video classification using textual features. *Heliyon* **10** (2024).
19. Akhtar, I. et al. Novel glassbox based explainable boosting machine for fault detection in electrical power transmission system. *Plos one* **19**, e0309459 (2024).
20. Raza, A. et al. Optimized virtual reality design through user immersion level detection with novel feature fusion and explainable artificial intelligence. *PeerJ Computer Science* **10**, e2150 (2024).
21. Tripathi, D. *et al.* An expert system using ensemble learning for crop yield prediction: Eescyp-i. In *2022 International Conference on Advances in Computing, Communication and Materials (ICACCM)*, 1–5 (IEEE, 2022).
22. Janrao, S. & Shah, D. Performance evaluation for identifying best feature selector and machine learning algorithm for optimizing crop recommendation system. In *2020 IEEE Congreso Bienal de Argentina (ARGENCON)*, 1–8 (IEEE, 2020).
23. Shams, M. Y., Gamel, S. A. & Talaat, F. M. Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making. *Neural Computing and Applications* 1–20 (2024).
24. Naseer, A. *et al.* A novel transfer learning approach for detection of pomegranates growth stages. *IEEE Access* (2024).
25. Raza, A., Qadri, A. M., Akhtar, I., Samee, N. A. & Alabdulhafith, M. Logrf: An approach to human pose estimation using skeleton landmarks for physiotherapy fitness exercise correction. *IEEE Access* (2023).
26. Shakoor, M. T., Rahman, K., Rayta, S. N. & Chakrabarty, A. Agricultural production output prediction using supervised machine learning techniques. In *2017 1st international conference on next generation computing applications (NextComp)*, 182–187 (IEEE, 2017).
27. Kulkarni, N. H., Srinivasan, G., Sagar, B. & Cauvery, N. Improving crop productivity through a crop recommendation system using ensembling technique. In *2018 3rd international conference on computational systems and information technology for sustainable solutions (CSITSS)*, 114–119 (IEEE, 2018).
28. Dey, B., Ferdous, J. & Ahmed, R. Machine learning based recommendation of agricultural and horticultural crop farming in india under the regime of npk, soil ph and three climatic variables. *Heliyon* **10** (2024).
29. Nti, I. K., Zaman, A., Nyarko-Boateng, O., Adekoya, A. F. & Keyeremeh, F. A predictive analytics model for crop suitability and productivity with tree-based ensemble learning. *Decision Analytics Journal* **8**, 100311 (2023).
30. Raza, A., Munir, K., Almutairi, M. S. & Sehar, R. Novel transfer learning based deep features for diagnosis of down syndrome in children using facial images. *IEEE Access* (2024).
31. Haq, S. I. U., Raza, A., Lan, Y. & Wang, S. Identifying of pest attack on corn crop using machine learning techniques. (2023).
32. Shehzad, F. *et al.* Optimizations of modified machine learning algorithms using k-fold cross validations for wheat productivity: a hyper parametric approach. (2022).
33. Hasan, M., Marjan, M. A., Uddin, M. P., Kardy, S. & Nam, Y. Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation. *Frontiers in Plant Science* **14**, 1234555 (2023).
34. Bang, S., Bishnoi, R., Chauhan, A. S., Dixit, A. K. & Chawla, I. Fuzzy logic based crop yield prediction using temperature and rainfall parameters predicted through arma, sarima, and armax models. In *2019 Twelfth international conference on contemporary computing (IC3)*, 1–6 (IEEE, 2019).
35. El Hachimi, C., Belaqziz, S., Khabba, S. & Chehbouni, A. Towards precision agriculture in morocco: A machine learning approach for recommending crops and forecasting weather. In *2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA)*, 88–95 (IEEE, 2021).

Author contributions

HA conceptualization, formal analysis, and writing - original draft. MA data curation, conceptualization, and writing - original draft. AR software, formal analysis and data curation. KM visualization, methodology, and project administration. SGV funding acquisition, methodology, visualization. LADL investigation, project administration and validation. IA supervision, validation and writing - review and editing. All authors reviewed the manuscript.

Funding

This research is funded by the European University of Atlantic.

Additional information

Correspondence and requests for materials should be addressed to M.A. or I.A.