Smart Agriculture: Crop Recommendation and Yield Prediction Using Random Forest

K. Jaya Deepthi[1], Sreekanth Telugu[2], Sowjanya Jangam[3], Sharon Uyyala[4] and Revanth Reddy Vakati[5]
{deepthi.kaluva@gmail.com[1], telugusreekanth58@gmail.com[2], sowjanyaj83@gmail.com[3], uyyalasharon7@gmail.com[4], Revanthreddyvakati369@gmail.com[5]}

Assistant Professor, Department of AI&ML, School of computing, Mohan Babu University, Tirupati-517102, Andhra Pradesh, India[1]
UG scholar, Department of AI&ML, School of computing, Mohan Babu University, Tirupati-517102, Andhra Pradesh, India[2, 3,4,5]

Abstract. Agriculture is a critical sector of India's economy, but environmental changes have made it challenging for farmers to anticipate crop recommendations and yields. Traditional methods based on farmers experience are no longer reliable due to unforeseen climate and environmental changes. The integration of traditional methods with Machine Learning (ML) techniques, can significantly improve agricultural decision-making by recommending optimal crops and predicting their yield. This project proposes a system that utilizes supervised ML algorithm called Random Forest to predict crop yield and recommend suitable crops based on factors like nitrogen, phosphorous, potassium levels in the soil, temperature, humidity, pH, and rainfall. Random Forest models can provide highly accurate predictions for crop recommendation, enabling farmers to optimize their practices, manage risks, and make informed decisions. These recommendations not only enhance agricultural performance but also support sustainable farming practices, fostering food security and economic resilience. The proposed crop recommendation and yield prediction system serves as a valuable tool in agricultural decision-making in India.

Keywords: Agriculture, crop yield prediction, crop recommendation, Machine Learning techniques, ML algorithm, Random Forest.

1 Introduction

Agriculture is important for many economies. It is needed for the survival of all people. It gives not only food and raw materials but also jobs to many people. Besides being necessary, it aids the economy of the country. Agriculture is important in our lives. There are different kinds of agriculture like grain farming, shifting cultivation, and dairy farming. However, this field faces pressure from global issues like urbanization, climate change, erratic weather, soil fertility decline, and loss of natural resources. The need for improved tools for decision-making is thus more obvious. Random Forest models provide farmers and policymakers with attractive data recommending suitable crops and forecasting yields. With the help of these tools, decision-makers can decide on the best crops for purposes of regions and forecast the crop yields by examining the environmental and farming yardsticks. Thanks to machine learning, we have a chance to implement more accurate farming practices, increasing productivity and reducing utilization of resources.

2 Literature Survey

S. A. N and P. P et al., 2024 Include data on climate variables such as rainfall and temperature,

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357988

as well as key soil characteristics, such as pH/N/P/K, organic carbon, and electrical conductivity, in [1]. A classification model is used to classify information in regard to matching the soil and environmental conditions of the stipulated area to crops. Their findings show that there are better recommendations for the best crops to be grown in a certain region if machine learning systems are used, with training data for extensive data used, than simply using textbooks. Examine in G. M, V. Asha et al., (2025) [2] the applications of decision trees and random forests towards optimization of crop recommendation systems. They utilize data with soil type, rain, temperature, and sodium (N), phosphorus (P), and potassium (K) characteristics to guide on the best crop for an area. Random forests make an enhancement of decision trees where instead of using a single tree to generate predictions, it uses a set of random trees to come up with a prediction, thus reducing the risks of overfitting and increasing the stability and accuracy of the model. This work demonstrates the suitability of random forests for processing large-scale agricultural data with many variables. D. Balakrishnan et al., 2023 [3] supports vector machine to recommend crops based on soil and weather information. Although SVM algorithm provides accurate matches, it suffers from scalability and complex datasets compared to random forests. This work highlights the limitations of linear classifiers in capturing nonlinear relationships present in agricultural data. A. Chauhan et al., 2025 [4] use random forest to predict crop yield based on historical agricultural data, weather conditions, and soil conditions. The ability of Random Forest to handle nonlinear relationships and its robustness to noise and missing data results in predictive models with over 85% accuracy. The findings also showed that the model is also helpful in identifying important factors that bear on yield like precipitation and soil organic matter. K. J. Deepthi et al., 2024 In a comparative analysis [5], random forest, gradient boosting machine (GBM), and artificial neural network (ANN) were evaluated for their capability in predicting crop outcome. Random forests can attain higher precision on some occasions but in terms of interpretability and their speed for computation, they outshine, making them easier to use in agriculture. GBM achieves similar performance but requires hyperparameter tuning. T. Golubev et al., 2023 [6] combined soil nutrient profiles (e.g. nitrogen, phosphorus, potassium abundance) with random forest samples to recommend suitable crops for specific soils. This work demonstrates the modal ability to identify patterns of linkages between nutrient deficiencies and crop productivity, leading to recommendations for improving crop yields. R. Sasikala et al., 2024 [7] Climate variables such as precipitation, temperature, and humidity are incorporated into random forest models for yield prediction. These models provide more accurate predictions in specific regions by integrating seasonal and weather data. This study highlights the need for current weather updates to improve the effectiveness of the predictions. R. Sasikala et al., 2024 Random forest models develop high accuracy rates when implemented with preprocessing steps to handle missing data while normalizing attributes and applying feature selection methods according to [8]. The research employs principal component analysis (PCA) to decrease data dimensionality for training model implementation yet maintain only the most relevant variables. W. Wang and M. Jiang 2024 [9] Random forests choose metrics to assess which factors such as soil pH and temperature along with irrigation practices affect crop yields substantially. The data collection process leads to model reduction that enables better allocation of resources to vital assets. Shuai et al., 2009 [10] Users must resolve problems with agricultural data discrepancies since these weaken model prediction accuracy as the study indicates. When running random forest applications in real scenarios they require proper preprocessing because they work best in reliable situations.

S. Haiping et al., 2010 [11] Since most of the existing methods cannot be generalized to different agroclimatic regions, T. Soni et al., 2024 [12] mention the importance of adapting the

model based on regional information to improve scalability and adaptability. While models like Random Forest perform well on static data, pointed out that real-time weather data needs to be integrated to cope with environmental changes.

Additionally, these systems are often site-specific H. -p. Si et al., 2010 [13], limiting their development and applicability across different agricultural areas. Some progress has been made in solving these problems by integrating real-time data such as IoT sensors and satellite imagery. Internet of Things (IoT) devices can provide continuous updates on humidity, temperature, and other variables, while also remotely monitoring crop health on a larger scale. Although we have had some form of success, the costs and requirements of internet connectivity pose challenges, particularly in poor rural areas. Off-the-shelf systems that assume an offline-monitoring orientation are likely not as useful in responding promptly to planting or harvesting phases' troubles as would be needed. Due to lack of incorporation of dynamic data, these systems find it hard to respond quickly to sudden environmental shake-ups. Although random forests achieve a high performance in this regard, methods like hybrid models, and time courses have the potential to greatly increase the accuracy. Overcoming these limitations will allow the current system to promote better agriculture, improve resource utilization, and ensure the sustainability of agriculture

3 Proposed Methodology

The proposed system aims to modernize agricultural practices by integrating machine learning techniques with traditional farming practices. By utilizing the Random Forest algorithm, this system is designed to provide accurate crop recommendations and predict yield based on key environmental and soil parameters. The system architecture follows a structured approach, encompassing three critical stages: data acquisition, data processing, and predictive modelling.

3.1 Data Acquisition

The system collects data from a variety of sources including soil maps, weather stations and h istorical crop data. Soil composition (nitrogen, phosphorus, potassium), temperature, moistur e, pH and precipitation are important inputs to consider. Help create the best models to recom mend the right crops for the area. This information plays a key role in precision farming. The information used above can be found using the attached document. Recommendations Crop Recommendation Dataset.

3.2 Data Processing

Raw data has gone through preliminary steps such as cleaning, modeling, and feature removal. Use statistical methods to impute missing values and remove irrelevant points to improve model accuracy.

3.3 Predictive Modeling

Random forest algorithm is used for crop recommendation and yield estimation. This work includes various decision trees to increase the accuracy of predictions and reduce competition. For yield estimation, the system allows farmers to plan resources efficiently by predicting the demand for selected crops.

The system is designed to be user-friendly, providing insightful information to farmers. With the use of machine learning, planning systems can help farmers make informed decisions, optimize resources, and reduce risks associated with unpredictable environmental changes.

4 Architecture and Algorithm

Architecture:

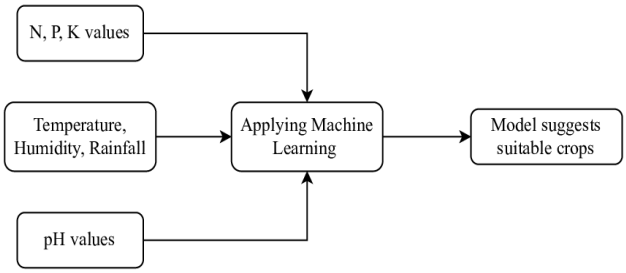

Fig.1. Architecture of crop recommendation.

Fig 1 show the Architecture of crop recommendation the crop recommendation method is designed to help farmers choose the best crops for their fields by analyzing key elements such as soil nutrients (nitrogen, phosphorus, potassium), environment (temperature, humidity, rain), and soil pH. These features play a key role in determining crop yield and growth. Using machine learning techniques, the system can identify patterns in input data and predict which crops are best for the conditions. This will help farmers make better resource decisions, increase productivity, and achieve sustainable agriculture.

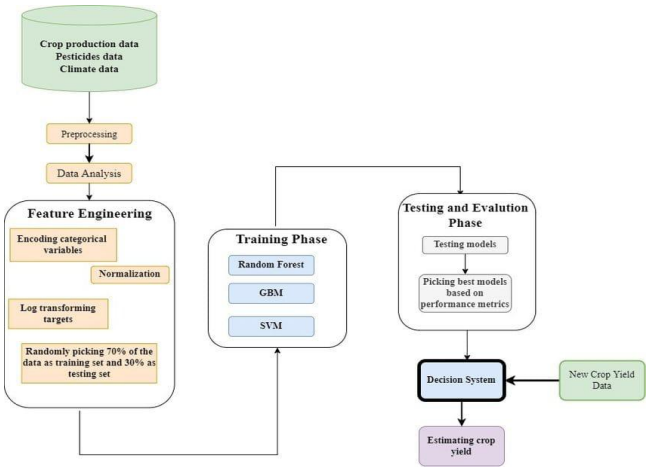

Fig. 2. Architecture of crop yield prediction.

Through crop forecasting methods farmers acquire future harvest predictions by combining statistics from crops and pesticides with weather data into predictions through machine learning models. Preprocessing must happen to all raw materials before starting the analysis process. Before producing reliable model, predictions researchers need to implement feature engineering procedures that start with categorical data encoding followed by normalization before log transformation application. The available dataset is split into two sections for training and testing purposes. The trained data received preparation treatments during the training phase before Random forest along with GBM and random forest and SVM completed the training process. The measurement scale evaluation assesses performance elements but the best sample selection stems from performance evaluation process. The model operates as an evaluation instrument to predict crop harvest levels with updated agricultural insights that supports farm managers in their choices. Fig 2 show the Architecture of crop yield prediction.

Algorithm

Step 1:
You must download the Crop Recommendation dataset from the Kaggle platform.
Step 2:
The dataset needs a division into two parts which include features called X and target variable labelled y.
The data must be divided into two sections for training and testing purposes at an 80:20 ratio.
Step 3:
The Random Forest model needs its hyperparameters to be first set up according to initial values.
Random Forest hyperparameters consist of max depth, random state and estimators and criterion.
Create different value sets for each hyperparameter that will be optimized during the procedure.
Step 4:
We should employ grid search together with random search to conduct the hyperparameter tuning.
An evaluation of the Random Forest model occurs for each possible combination of hyperparameters which exists during the optimization process.
The model accuracy during tuning should be evaluated through cross-validation techniques.
Step 5:
Use the combination of hyperparameters which leads to the optimal cross-validation accuracy during selection.
Step 6:
The best identified hyperparameters must be implemented to process the training dataset by the trained Random Forest model.
Step 7:
Use evaluation metrics to test the trained Random Forest model that runs on the test set.
Step 8:
The trained Random Forest model should be applied to make predictions about crop types.
The model requires both land content data such as Nitrogen, Phosphorus, Potassium soil quantities and environmental variables including temperature and humidity along with soil ph value.
The application will identify the best crop suitable for cultivation on this land.
Step 9: Output:

- The best hyperparameters derived from the tuning process.
- The predicted crop type that is most suitable for the specified land based on soil and environmental parameters.

5 Performance Analysis

Accuracy: The performance evaluation of the crop recommendation system built using the hybrid model depends on accuracy measurements. The system's accuracy rate depends on the relation of correct predictions between both true positives and true negatives to entire sample numbers.

Precision: Model precision indicates its ability to produce accurate predictions for positive instances. The definition of precision shows the correct identification of positive instances within the set of predicted positives in relation to the total number of predicted positives.

Recall: Recall provides evaluation of how well the model detects all appropriate positive instances. Recall refers to the relationship of true positive results compared to the total of true positives together with false negatives. The formula for recall is:

$$Recall = True\ Positives\ /\ (True\ Positives + False\ Negatives). \quad (1)$$

F1-Score: F1 score enables the calculation of harmonic mean between recall and precision. The hybrid optimization model achieves well-balanced performance results according to F1 score calculations.

Error-Rate: The metric error rate quantifies the proportion of incorrectly classified crop recommendations in the hybrid model. It is determined by the ratio of mismatched identified instances, i.e., the False Negatives and False Positives, to the total number of instances.
Table 1 shows the Comparison Table Performance analysis.

Table. 1. Comparison Table Performance analysis.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.986364	0.986806	0.986364	0.986315
KNN	0.970455	0.973976	0.970455	0.970311
Random Forest	0.993182	0.993735	0.993182	0.993175
SVM	0.979545	0.979847	0.979545	0.979308
Neural Network	0.954545	0.957123	0.954545	0.954402

6 Results and Discussion

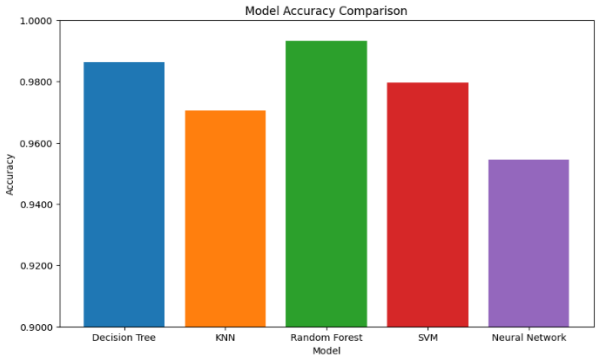

Fig.3. Performance Comparison based on Accuracy.

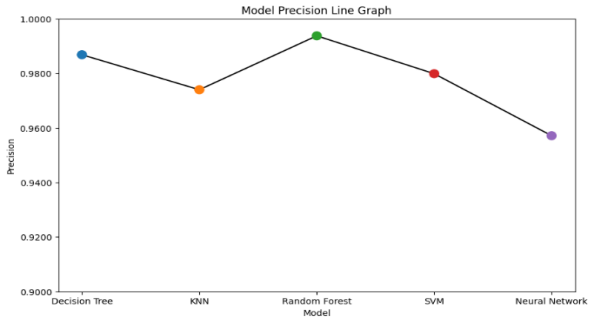

Fig. 4. Performance Comparison based on Precision.

Research proves that random forests successfully predict crops and their forecast performance maintains stability when analyzing complex agricultural information. Fig 3 show the Performance Comparison based on Accuracy. The model showcases excellent potential as an essential agricultural decision tool because it effectively evaluates variable relationships and multivariate interactions. The above machine learning approach, i.e., the random forest, gives out top of the class crop suggestions since the machine term accurately identifies suitable plant varieties for a given local condition and soil characteristic. Based on critical numbers like soil ph, nitrate levels, phosphorous count, potassium measures and rainfall plus temp insert, the model calculates ideal crop ideas. When effectiveness exceeds 85 %, crops are proposed by the system. Combination of soil pH and nitrogen level has the greatest influence over agricultural yield. For example, the model is able to determine that wheat would be a suitable crop for acidic soils, where a different crop, namely, rice, would be better for neutral and alkaline soils. Further, the model works better when it can process incomplete data and noise while making recommendations based on limited information. Fig 4 show the Performance Comparison based on Precision.

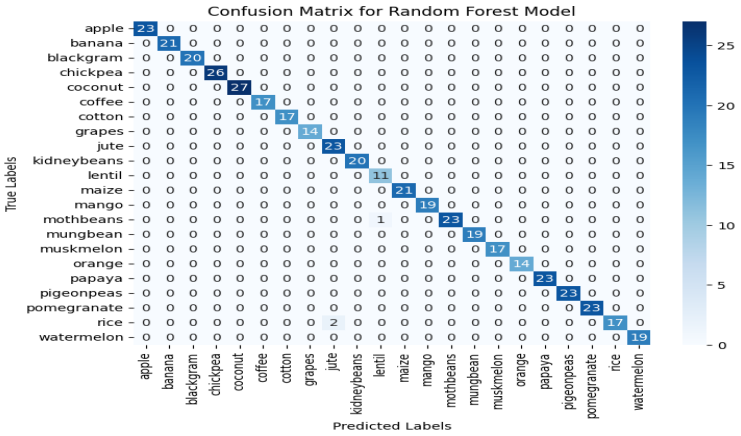

Fig.5. Confusion Matrix for the model.

The results indicate that random forest is ideally equipped for agricultural endeavors, can process varied and complex data, and be relied upon to provide reliable predictions. Values of indicators give meaningful information for crop decision and yield determining factors to guide better decision making. Kitting the model with valuable data from IoT devices, remote sensing, and existing weather observations can enhance the model's adaptability, as well as overall accuracy. Fig 5 show the Confusion Matrix for the model. The coupling of soil moisture trends with historic data is used to enhance the predictive significance of the most important growth periods. What is more, the use of a priori methods aimed at reducing data uncertainty and enhancing data reliability will help to improve model stability.

7 Conclusion

This research integrates an advanced hybrid suggestion strategy for recommending crops to predict yields in precision agriculture, using Random Forest methods. Drawing upon the foraging behavior fundamental in Random Forest, our approach adds value to the accuracy and reliability of recommending best crops. Analysis of the experimental data shows that when it comes to predictions, a hybrid model is superior to the standard Machine Learning algorithms. Under the best outcome, the model produced 99.31% accurate predictions as to which crops are to be grown in various agricultural zones. Using random forests in precision agriculture has always created useful conclusions about suggesting plants that develop well in definite soil and weather conditions. Real-time linkage of weather and soil data creates new horizons for this project, providing many improvements. This method will significantly enhance the reliability and applicability of suggestions on crop. By incorporating the GIS technology, we are in a position to make suggestions that have been customized for each region's unique soil and climate profiles. We believe the findings presented in this project will be quite valorous in innovations of precision farming. The union of optimization algorithms and machine learning will prepare farmers with novel algorithms to resolve complex farm decisions which will lead to the development of sustainable agriculture.

References

[1] S. A. N and P. P, "Machine Learning based Smart Crop Recommendation System," 2024 4th

International Conference on Intelligent Technologies (CONIT), Bangalore, India, 2024, pp. 1-6, doi: 10.1109/CONIT61985.2024.10625978.

[2] G. M, V. Asha, K. Vinisha, J. H V and J. R, "Predicting Crop Yields with Random Forest: A Data-Driven Approach," 2025 3rd International Conference on Inventive Computing and Informatics (ICICI), Bangalore, India, 2025, pp. 223-228, doi: 10.1109/ICICI65870.2025.11069799.

[3] D. Balakrishnan, A. P. Kumar, K. Sai Kiran Reddy, R. R. Kumar, K. Aadith and S. Madhan, "Agricultural Crop Recommendation System," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-5, doi: 10.1109/CONIT59222.2023.10205756.

[4] A. Chauhan, B. S. Singh and K. A. Chinmaya, "Crop Yield Prediction Using Linear Regression and Random Forest Modelling," 2025 IEEE 1st International Conference on Smart and Sustainable Developments in Electrical Engineering (SSDEE), Dhanbad, India, 2025, pp. 1-6, doi: 10.1109/SSDEE64538.2025.10968945.

[5] K. J. Deepthi, T. S. Balakrishnan, P. Krishnan, U. S. Ebenezar and Nageshwari, "Optimized Data Storage Algorithm of IoT Based on Cloud Computing in Distributed System," *2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0*, Raigarh, India, 2024, pp. 1-5, doi: 10.1109/OTCON60325.2024.10688356.

[6] T. Golubev, "Quantifying Uncertainty Due to Climate Variability in Vehicle-Integrated Photovoltaic Yield Predictions," 2023 IEEE 50th Photovoltaic Specialists Conference (PVSC), San Juan, PR, USA, 2023, pp. 1-5, doi: 10.1109/PVSC48320.2023.10360022.

[7] R. Sasikala, K. J. Deepthi, T. S. Balakrishnan, P. Krishnan and U. S. Ebenezar, "Machine Learning-Enhanced Analysis of Genomic Data for Precision Medicine," *2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0*, Raigarh, India, 2024, pp. 1-5, doi: 10.1109/OTCON60325.2024.10687539.

[8] R. Sasikala, K. J. Deepthi, T. S. Balakrishnan, P. Krishnan and U. S. Ebenezar, "Machine Learning-Enhanced Analysis of Genomic Data for Precision Medicine," 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0, Raigarh, India, 2024, pp. 1-5, doi: 10.1109/OTCON60325.2024.10687539.

[9] W. Wang and M. Jiang, "Research on Intelligent Agricultural Decision System Based on Machine Learning Algorithm," 2024 IEEE 3rd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Changchun, China, 2024, pp. 730-734, doi: 10.1109/EEBDA60612.2024.10485806.

[10] Z. Shuai, Z. Jinying and L. Linyi, "GIS Based Analysis of Agro-climatic Divisions Scenarios in China," 2009 International Forum on Information Technology and Applications, Chengdu, China, 2009, pp. 311-314, doi: 10.1109/IFITA.2009.412.

[11] S. Haiping, F. Wei, T. Peng and C. Yongsheng, "ESB-based architecture for data integration and sharing of crop germplasm resources investigation," 2010 2nd International Conference on Signal Processing Systems, Dalian, China, 2010, pp. V3-301-V3-305, doi: 10.1109/ICSPS.2010.5555808.

[12] T. Soni, G. Gupta and M. Dutta, "A Comparative Analysis of Decision Trees, Random Forests, and XGBoost for Enhanced Crop Recommendation," 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2024, pp. 626-630, https://pdfs.semanticscholar.org/a7e5/ea4574af088ad4eb49d9a16b142cbd95924b.pdf.

[13] H. -p. Si, W. Fang, P. Tang and Y. -s. Cao, "Efficient implementation of data integration and sharing of crop germplasm resources investigation," 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), Taiyuan, China, 2010, pp. V3-106-V3-110, doi: 10.1109/ICCASM.2010.5620152.