

Random Forest–Based Integrated Crop Recommendation and NPK-Driven Fertilizer Decision Support System for Precision Agriculture

Sampiruthi R.¹, Sarumathi M.², Sneha C.³, Zeshan Ali.⁴, GnanaRubini R.⁵

^{1,2,3,4}Student, ⁵Assistant Professor, Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Anna University, India.

E-mail: ¹727722eucs145@skcet.ac.in, ²727722eucs156@skcet.ac.in, ³727722eucs170@skcet.ac.in, ⁴727722eucs209@skcet.ac.in,

⁵gnanrajrad@gmail.com

Abstract

Precision agriculture calls for data-driven methods to determine the optimal choice of crops and fertilizers depending on soil conditions and climate. In this paper, an integrated decision support system is presented, which is based on Random Forest crop classification and NPK-based fertilizer recommendation system. The proposed system works with seven agronomic features, including nitrogen, phosphorus, potassium, temperature, humidity, soil pH, and rainfall, from crop recommendation dataset with 2,200 entries and 22 classes of crops. After processing data and separating them into training set (80%) and testing set (20%), a Random Forest classifier with 100 decision trees was built to classify the crops. The classified crop is matched with the nutrient knowledge base, and the current soil NPK content is compared to the nutrient needs of the crop. Finally, the rules-based decision engine finds out nutrient deficiencies and makes fertilizer recommendations. The performance of the classifier on the test dataset was 97.6% accuracy, 97.1% precision, 96.8% recall, and 96.9% F1-score.

Keywords: Precision Agriculture, Multiclass Crop Classification, Random Forest Ensemble, Soil Nutrient Profiling, Agroclimatic Feature Integration, NPK-Based Fertilizer Recommendation, Agricultural Decision Support.

1. Introduction

The productivity of agriculture is highly dependent on the proper interaction among the selection of crop, soil, nutrients, and environment. Farmers have to find what type of crop is best suited for that particular farm while analyzing the nutrient level before farming. Variables such as Nitrogen (N), Phosphorous (P), Potassium (K), soil pH, temperature, humidity, and rain affect the selection of crops and the environment required for growth. Traditional decision making based on experience and physical estimation becomes tough when several interrelated variables have to be taken into consideration. This results in reduced productivity and increased cost of cultivation.

The emergence of agricultural datasets and improvements in machine learning technology, automation of such decision-making process has become possible. Machine learning algorithms can discover complicated associations between soil and environmental factors and then utilize these associations to find out which crops are more appropriate to particular conditions of planting. The most effective among ensemble learning methods in agricultural classification problem is the Random Forest algorithm, which is based on the combination of several decision trees and allows working efficiently with non-linear associations between heterogeneous input data.

However, selecting the right crop will not be enough in the context of efficient farming. Once the crop is selected, the nutrient needs of that particular crop have to be considered to see whether the soil condition is sufficient or whether any additional fertilizers need to be added. Crop prediction and fertilizer recommendation are thus both related tasks in the field of agriculture which have to be considered together. An efficient process should start by assessing the crop suitability in the given soil environment and then determine the nutrient needs.

To address this requirement, a Random Forest based Decision Support System has been proposed to predict crop as well as suggest the appropriate fertilizer. This system analyses various parameters including N, P, K, pH, temperature, humidity, and rainfall and suggests the most suitable crop by employing Random Forest classification model. After predicting the crop, another component for nutrient based recommendation analyses the soil condition with respect to nutrient requirement of the predicted crop and provides fertilizer suggestions.

2. Related Works

The application of Machine Learning (ML) has been seen in many aspects of agricultural decision making, such as crop selection, yield prediction, soil analysis, and precision farming. Previous research works on predicting crop yield has shown that computational models can be used to decide appropriate cultivation options using agricultural parameters [1]. Machine learning has been studied in conjunction with wireless sensor networks for assisting precision agriculture [2].

ML methods have been recently used for soil analysis and classification, proving their efficiency in pattern finding in data associated with soils [3]. A larger study on the application of machine learning in precision agriculture has proven the efficacy of predicting algorithms in helping manage crops [4]. The intelligent approaches to agriculture which use IoT and data mining have highlighted the importance of using data for resource management and sustainable agriculture [5].

DL techniques have also been investigated for estimating crop yields through spatio-temporal observations of agriculture, which shows the capability of more sophisticated learning models to deal with complex agricultural data [6]. Agricultural recommendation systems have shown how computational decision support systems can help the farmers choose the right agricultural practices [7]. Smart farming techniques that use data mining have also shed light on the significance of automated analysis in agriculture [8].

However, sustainable agriculture has also been influenced by the use of data visualization and spatial technologies, which could aid in interpreting agricultural data and making decisions [9]. The latest research into crop recommendations has also shown the suitability of machine learning in choosing appropriate crops based on environmental and soil factors [10].

Despite these developments, existing research has concentrated on specific agricultural operations like crop prediction, soil testing, yield calculation, or general recommendation. There is hence a need for a holistic decision support system that combines machine learning-based crop prediction with fertilizer recommendation based on nutrients. The current work fulfils this need by integrating Random Forest-based crop prediction with a fertilizer recommendation system. .

3. Proposed Methodology

The proposed model works with an integrated processing approach from collecting agricultural inputs to generating crop and fertilizer decisions, which is shown in Figure 1 below. The proposed model takes N, P, K temperature, humidity, soil pH, and rainfall as the input parameters and then carries out data cleansing, feature preparation, and feature scaling. Further, the features set is fed to the Random Forest classifier consisting of 100 decision trees for multi-class crop classification in order to generate the predicted crop based on the learning of the agricultural features and crop classes. Further, the generated crop is given as an input parameter to the fertilizer recommendation module, where the soil NPK values are compared to those of the predicted crop NPK requirements, obtained from the nutrient knowledge base. The deficiencies are further processed using the rule-based decision engine to generate the fertilizer recommendation

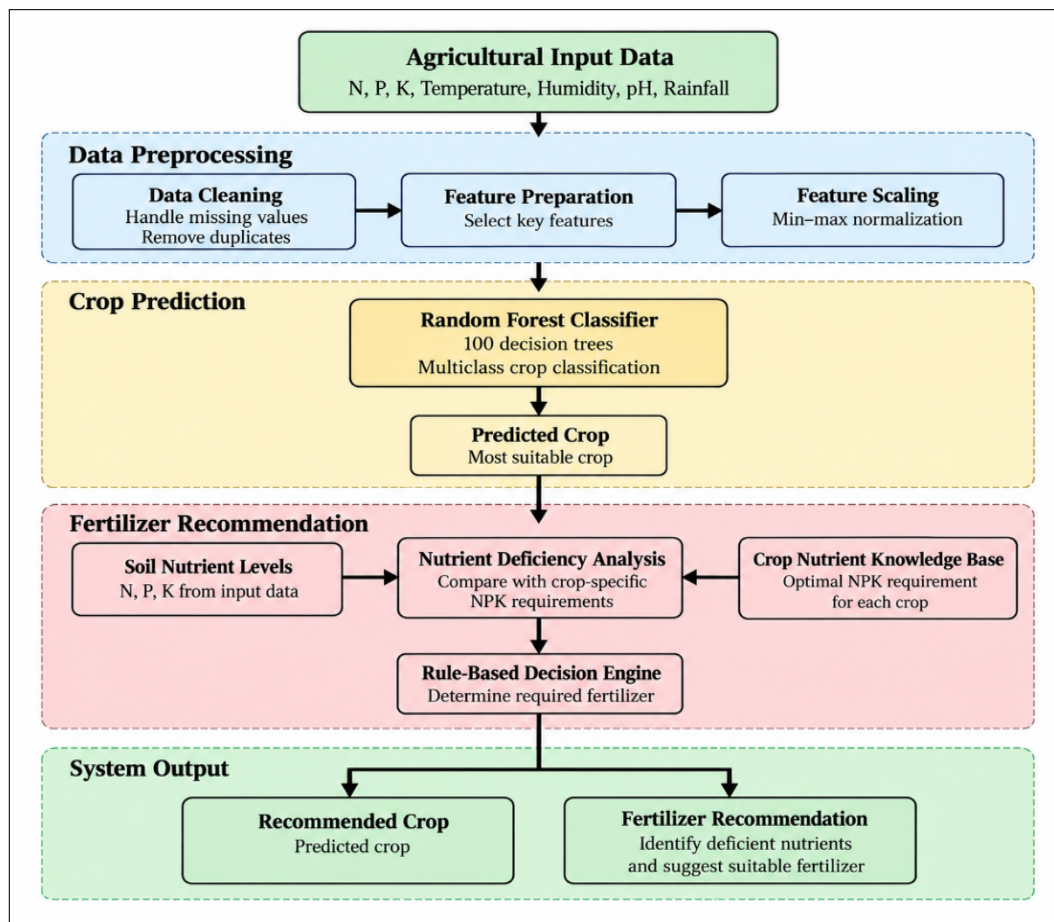


Figure 1. Workflow of the proposed crop recommendation and fertilizer decision support system

3.1 Data Collection and Dataset Description

The process of data collection supplies the information about agriculture for building and testing the crop prediction algorithm. The data set contains soil and environmental characteristics for different crop classes. Such characteristics are the primary factors affecting crop suitability, and the machine learning algorithm is used to learn the connection between such agriculture conditions and crop class.

Table 1. Dataset and experimental configuration

Parameter	Description
Dataset	Crop Recommendation Dataset [11]
Total observations	2,200
Crop classes	22
Input features	7
Features	N, P, K, Temperature, Humidity, pH, Rainfall
Target variable	Crop type
Learning task	Multiclass classification
Training data	1,760 observations (80%)
Testing data	440 observations (20%)
Classifier	Random Forest

The properties of the dataset, as shown in Table 1, provide the experimental foundation on which the processing and training phases are performed. The transformation of the collected data from the agriculture sector into the Random Forest model is shown in Figure 2.

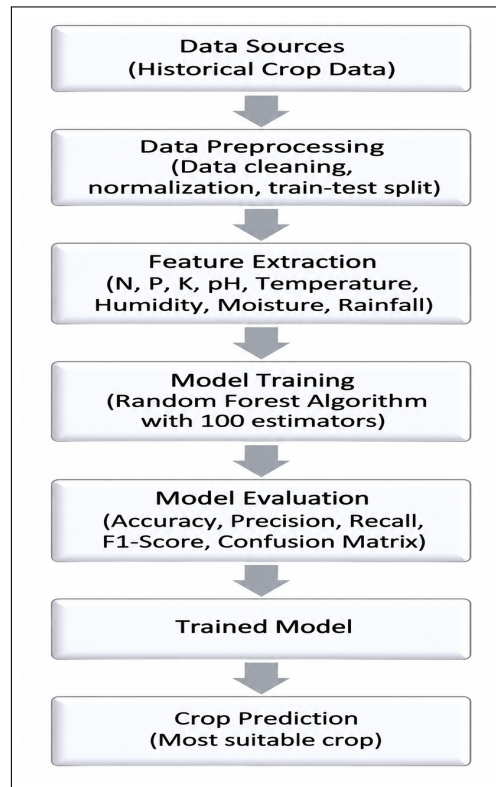


Figure 2. Random forest-based crop prediction architecture

The Crop Prediction dataset [11] will be used as the main dataset for the crop prediction experiment. This dataset has 2,200 samples which are categorized into 22 different types of crops. Each observation contains seven input attributes: (N), (P), (K), temperature, humidity, soil pH, and rainfall. All the above attributes provide a comprehensive combination of both soil fertility and climate conditions. The type of crop for each sample will be taken as the target variable. An input for the crop for a single observation can be presented as:

$$X = N, P, K, T, H, pH, R \quad (1)$$

where N,P, and K represent the soil macronutrient levels, T represents temperature, H represents humidity, pH represents soil acidity or alkalinity, and R represents rainfall. The corresponding crop label is represented by Y. This representation helps to the Random Forest algorithm understand the interaction between both the soil parameters and the environmental variables as a whole and not individually. The data is then split into two different sets – training set and test set. An 80:20 split implies that there are 1,760 samples in the training set and 440 samples in the testing set.

3.2 Data Pre-Processing

The pre-processed data from the agricultural records is fed into the system for feature selection and modeling. Pre-processing of the dataset involves three key processes in the order that is shown in Figure 2.

Initially, there is an inspection of the dataset for any missing values, duplicates, invalid numerical values, or any inconsistent classification labels. Entries that have missing or invalid values are cleaned up before developing a predictive model to ensure that the observation matrix includes valid values for all the attributes. Crop classification labels are kept as categorical target values.

The numerical properties are normalized to bring the input variables to a single numerical scale. The min-max normalization is carried out in case it is necessary with the help of:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (2)$$

where x is the original feature value, $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the respective feature, and x' is the transformed value. These transformation parameters are the same for the training and testing data.

The preprocessed dataset is split into 80% training dataset and 20% testing dataset. 80% is the training dataset which consists of 1,760 entries, whereas 20% testing dataset consists of 440 entries. The training set is applied for Random Forest model training, whereas testing set is used later during the evaluation stage of the model.

3.3 Feature Extraction

The feature extraction process involves selecting and arranging the agricultural features that the Random Forest classification algorithm needs. The pre-processed data set comprises seven independent variables: N, P, K, Temperature, Humidity, Soil pH, and Rainfall. The seven variables remain as the input variables for the model since they are the agricultural features in the chosen crop recommendation data set.

The feature values extracted from each record are stored in an ordered manner:

$$X = N, P, K, T, H, pH, R \quad (3)$$

where N,P, and K represent the soil macronutrient measurements, T represents temperature, H represents humidity, pH represents soil acidity or alkalinity, and R represents rainfall. The corresponding crop class is stored independently to be used as the class label.

3.4 Random Forest Model Training

The feature vector matrix obtained and the labels for crops are fed into the Random Forest Classifier. The algorithm uses 100 estimators (decision trees), which means that each decision tree is built using a bootstrap sample of the observations. Also, a subset of the input features is selected to split nodes.

For an input feature vector X , each tree h_m produces a crop-class prediction. The Random Forest algorithm's final output is generated using majority voting:

$$\hat{Y} = \text{mode} \{ h_1(X), h_2(X), \dots, h_{100}(X) \} \quad (4)$$

where $h_m X$ is the prediction of the m^{th} tree, and $\hat{Y}$ is the final predicted crop class.

The trained ensemble thus classifies the agricultural features into one of the crop classes in the training set. The trained model is thus kept aside for the prediction phase of crops, whereas the test data set is separately used for the validation purpose.

3.5 Crop Prediction

A trained Random Forest classifier is applied to predict the crop type in a new agriculture data set. The input is taken in the same order of features as while training the model: nitrogen, phosphorus, potassium, temperature, humidity, pH of the soil, and rainfall. The trained classifier applies the input to each of the 100 decision trees to obtain predictions.

The output is given by the aggregation of individual trees using the majority vote method. This class indicates the crop that is predicted using the available soil and environmental parameters.

The process of prediction can be expressed as follows:

$$C_p = f_{\text{RF}}(X) \quad (5)$$

where X is the input feature vector and C_p is the predicted crop class generated by the trained Random Forest model.

The projected crop is moved to the fertilizer recommendation phase next. The fertilizer model uses this crop classification to determine the nutrients required and then proceeds to carry out nutrient comparisons and deficiencies.

3.6 Fertilizer Recommendation

The recommendation phase gets the crop type provided by the Random Forest classifier and uses it to compute the nutrient needs of the chosen crop. The recommendation process involves requirement acquisition, nutrient comparison, nutrient deficiency check, and fertilizer recommendation using rules, shown in Figure 3.

The required amounts of nitrogen, phosphorus, and potassium for the predicted crop c_p are obtained from the fertilizer knowledge base and then compared with the amount present in the soil. Let the measured nutrient vector be $S = N_s, P_s, K_s$ and the crop requirement vector be $R = N_r, P_r, K_r$. The nutrient deficit is calculated as:

$$D_j = \max(0, R_j - S_j) \quad (6)$$

where D_j denotes the deficiency of nutrient j , R_j is the required level, and S_j is the measured soil level.

The deficiency values are calculated and then evaluated separately for nitrogen, phosphorus, and potassium. Positive values indicate that the nutrients available are lower than the required nutrient requirement for the crop, while zero values indicate that there is no deficiency of the particular nutrient.

output obtained from the deficiency analysis is then used by the decision engine based on rules to recommend the corresponding fertilizer type. As such, the fertilizer recommendation depends on the crop predicted as well as the nutrient requirement

3.7 Nutrient Deficiency Analysis and Rule-Based Decision

The nutritional deficiency values derived from the comparison stage are analyzed by the rule-based decision-making engine. The decision-making process identifies if any deficiency in nitrogen, phosphorous, or potassium needs to be addressed for the particular crop. Each nutrient is analyzed individually with respect to their requirements.

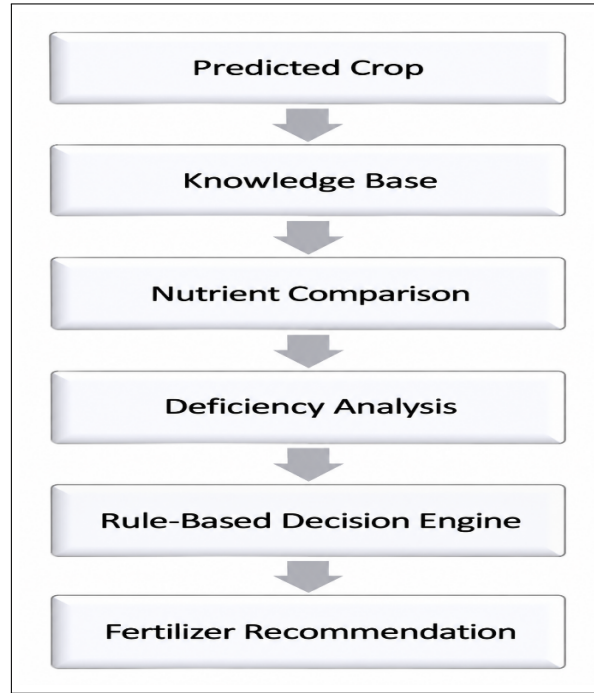


Figure 3. Rule-based fertilizer recommendation architecture

The decision rules are stated as follows:

$$\begin{aligned}
 D_N > 0 &\rightarrow \text{Nitrogen fertilizer} \\
 D_P > 0 &\rightarrow \text{Phosphorus fertilizer} \\
 D_K > 0 &\rightarrow \text{Potassium fertilizer}
 \end{aligned}
 \tag{7}$$

where D_N , D_P , and D_K denote the calculated nitrogen, phosphorus, and potassium deficiencies, respectively. If more than one deficiency is found, the decision engine applies the rules and generates a suitable fertilizer recommendation accordingly. In case the nutrient level detected meets the desired threshold level, no fertilizer is recommended for that particular nutrient.

3.8 Model Evaluation

The trained random forest model is tested with the help of the testing subset of 440 observations which was set aside during the pre-processing stage. The model's capacity for classification of the agricultural observations into different crops is measured during the testing phase.

Model performance will be measured based on accuracy, precision, recall, and F1 score. Accuracy is the total percentage of accurately classified data. Precision refers to the percentage

of accurate prediction for those observations predicted as belonging to a certain crop type, while recall refers to the percentage of accurately identified observations that belong to the certain class.

The proposed task belongs to the domain of multi-class classification problem, involving 22 classes. The evaluation measures will be calculated via the use of an adequate averaging strategy for multiclass case. Besides the numerical performance measures, a confusion matrix is constructed to analyze the class-specific classification performance. Correct classifications are highlighted by the matrix diagonal while misclassifications - out-of-diagonal.

4. Results and Discussion

Testing of the proposed system was carried out in order to ascertain the performance of the crop prediction module and the fertilizer recommendation module that followed it. This was done using the test data set obtained during the pre-processing stage. The performance of crop classification was measured using accuracy, precision, recall, F1 score, and confusion matrix analysis while the fertilizer recommendation module was tested using nutrient deficiency decisions

4.1 Crop Prediction Performance

The proposed Random Forest classifier achieved 97.6% accuracy, 97.1% precision, 96.8% recall, and 96.9% F1-score on the multi-class crop classification problem. Accuracy indicates the overall percentage of correct classifications, while precision and recall indicate correctness of classification and coverage of class identification, respectively. The F1 score is indicative of the trade-off between precision and recall, and its value is 96.9%.

Table 2. Performance of the proposed random forest crop prediction model

Metric	Value (%)
Accuracy	97.6
Precision	97.1
Recall	96.8
F1-score	96.9

As explained in Table 2, the difference between precision and recall equals 0.3, whereas the difference between the F1-score and recall equals only 0.1%. The precision equal to 97.1% is achieved at the expense of a small number of false positive detections. Similarly, the recall equal to 96.8% is achieved because of a small number of false negative crops. Overall, the F1-score equal to 96.9% keeps the classifier's performance between precision and recall, while the accuracy equals 97.6.

4.2 Confusion Matrix Analysis

The class-wise probability distribution of the proposed classifier Random Forest is given in Figure 4. This matrix consists of 440 test samples belonging to 22 crop classes, with the diagonal elements being the correct predictions while the off-diagonal elements are the misclassification cases.

It is evident that there is a clear concentration of the observations on the main diagonal, whereby most class-wise classifications get to the level of having 19-20 observations out of the 20 test observations. There are very few observations outside the main diagonal, meaning that there are very few classification mistakes. This distribution is consistent with the performance measures provided in Table 2, especially 97.6% and 96.9%.

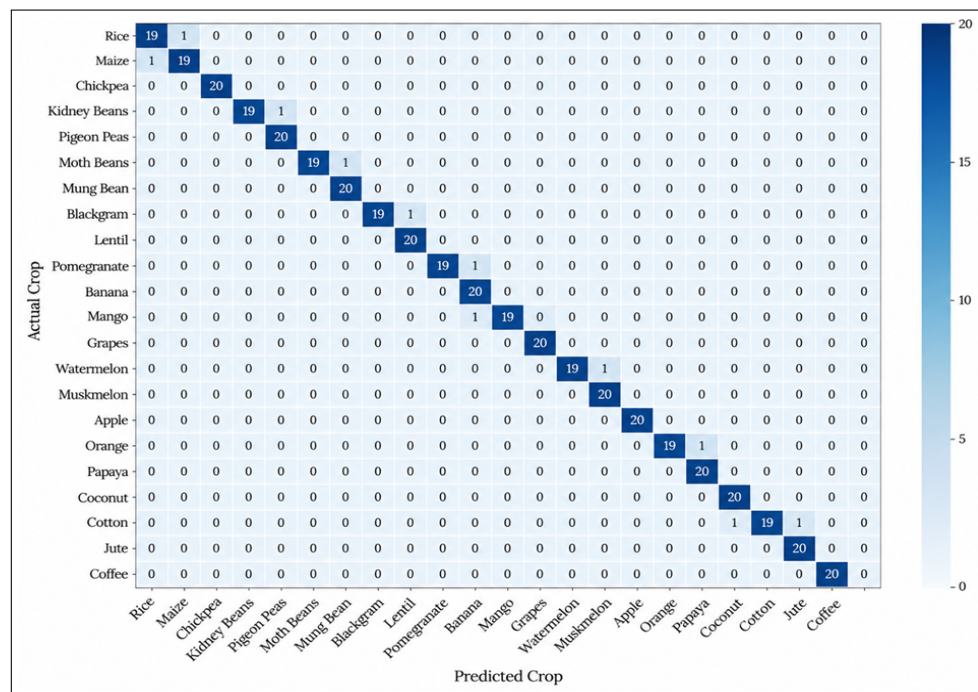


Figure 4. Confusion matrix of the proposed random forest crop classification model

4.3 Fertilizer Recommendation Results

The fertilizer recommendation sub-module generates a recommended fertilizer choice based on the estimated crop and the respective farming inputs. The crop prediction generated by the Random Forest classifier forms the basis for recommending fertilizer recommendations.

The obtained fertilizer production becomes the nutrient management suggestion for the crop that is predicted. Thus, there is an additional output that can be used for decision making, along with the crop prediction output.

4.4 Web Interface Implementation

This proposed web interface serves as an interactive interface for the suggested crop and fertilization recommendations system, as depicted in Figure 5. This interface consists of two sections: the field parameter entry panel and the prediction outcome panel. The field parameter entry panel takes as inputs the seven field parameters utilized by the machine learning algorithm, which includes nitrogen, phosphorous, potassium, temperature, humidity, soil pH value, and rainfall amount. Upon submission of the field values, the prediction will be obtained by passing the vector of the input field values through the Random Forest classification algorithm, where decision is made using majority voting.



Figure 5. Web-based crop and fertilizer recommendation interface

The interface also brings together the nutrient deficiency analysis and the crop prediction results. In the given test input, Jute has been determined to be the predicted crop, with 36 out of 100 trees taking part in the vote of the said class. The nutrient analysis is done by comparing the given NPK amounts with the requirements for that particular crop. It is observed from the given results that nitrogen is found to be deficient, while phosphorus and potassium meet their respective requirements.

5. Conclusion and Future Work

This study has developed an integrated decision support system based on Random Forests classification of crops and NPK-based fertilizer recommendation for precision agriculture. This suggested framework uses seven soil and environmental factors, namely, nitrogen, phosphorus, potassium, temperature, humidity, soil pH, and rainfall, to determine the suitable crops and enable nutrient management. The proposed Random Forest classifier, which achieves 97.6% accuracy on the test set, proves the effectiveness of the classifier in the classification of 22 crops in the selected dataset. The fertilizer recommendation module considers the nutrient requirement of the crops and soil NPK measurements to find out the nutrient deficiency and provide rule-based fertilizer recommendations. Despite all the promising outcomes, this system needs to be validated with large datasets from various locations to expand its utility across varied agricultural environments. Further development of this system must include other parameters, such as soil moisture, organic carbon, electrical conductivity, irrigation, and climatic factors. Combining the IoT-based soil monitoring, other advanced ensemble methods, and deep learning algorithms can add further flexibility and help in the fertilizer dose estimation.

References

- [1] Bhanose, Shreya S., Kalyani A. Bogawar, Aarti G. Dhotre, and Bhagyashree R. Gaidhani. "Crop and Yield Prediction Model." *International Journal of Advance Scientific Research and Engineering Trends*, 2016, vol. 1, no. 1: 23-28.
- [2] Mekonnen, Yemeserach, Srikanth Namuduri, Lamar Burton, Arif Sarwat, and Shekhar Bhansali. "Machine Learning Techniques in Wireless Sensor Network-Based Precision Agriculture." *Journal of the Electrochemical Society*, 2020, vol. 167, no. 3: 037522.

- [3] Rajeswari, V., and K. Arunesh. "Analysing Soil Data using Data Mining Classification Techniques." *Indian Journal of Science and Technology*, 2016, vol. 9, no. 19: 1-4.
- [4] Sharma, Abhinav, Arpit Jain, Prateek Gupta, and Vinay Chowdary. "Machine Learning Applications for Precision Agriculture: A Comprehensive Review." *IEEE Access*, 2020, vol. 9: 4843-4873.
- [5] Ali, Awais, Tajamul Hussain, Noramon Tantashutikun, Nurda Hussain, and Giacomo Cocetta. "Application of Smart Techniques, Internet Of Things and Data Mining for Resource use Efficient and Sustainable Crop Production." *Agriculture*, 2023, vol. 13, no. 2: 397.
- [6] Bose, Pritam, Nikola K. Kasabov, Lorenzo Bruzzone, and Reggio N. Hartono. "Spiking Neural Networks for Crop Yield Estimation Based on Spatiotemporal Analysis of Image Time Series." *IEEE Transactions on Geoscience and Remote Sensing*, 2016, vol. 54, no. 11: 6563-6573.
- [7] Kumar, Vikas, Vishal Dave, Rahul Bhadauriya, and Sanjay Chaudhary. "Krishimantra: Agricultural Recommendation System." In *Proceedings of the 3rd ACM Symposium on Computing for Development*, 2013, vol. 45: 1-2.
- [8] Chandak, Priyanka P., and A. J. Agrawal. "Smart Farming System using Data Mining." *International Journal of Applied Engineering Research*, 2017, vol. 12, no. 11: 2788-2791.
- [9] Latu, Savae. "Sustainable Development: The Role of GIS and Visualisation." *The Electronic Journal of Information Systems in Developing Countries*, 2009, vol. 38, no. 1: 1-17.
- [10] Priyadharshini, A., Swapneel Chakraborty, Aayush Kumar, and Omen Rajendra Pooniwala. "Intelligent Crop Recommendation System using Machine Learning." In *5th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, 2021, 843-848.
- [11] A. Ingle, "Crop Recommendation Dataset," Kaggle, 2021. Available: <https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset>