International Journal of Intelligent Computing Systems
July-2026| IJICS | Volume-2 Issue-2 | ISSN: 3107-7218

Predictive Analytics For Optimal Crop Growth Using Machine Learning And Soil Data

D Siva Naga Jyothi[1],S Sarojini[2]
Student[1],Assistant Professor[2]
Priyadarshini Institute of Technology & Science for Women, Chintalapudi, Andhra Pradesh, India[1,2]

ABSTRACT-Nitrogen-rich soil is a necessity for plant growth and productivity, and knowing the fertility level of the soil according to crop type is important. This research seeks to develop a stable model for predicting soil fertility through Internet of Things data and federated learning-based feature selection techniques. The model was conditioned through the use of federated learning algorithms. The minimum square error (LSE) was calculated with a Gradient Descent Curve (GDC) for maximizing prediction precision. Decision trees combined with logistic regression were better than other models with 87% precision, 87% accuracy, 87% recall, and 86% F1-score. The LSE was utilized to find the optimal weight and bias after collecting two client system model outputs from a server model. Subsequent research will apply a recursive approach to improve the performance of the model by applying a feedback cycle.

KEYWORDS-Soil fertility, IoT sensors, Federated learning, machine learning

I.INTRODUCTION

Identifying soil fertility is important for soil health; it is essential for crop production and nutrient deficiencies can be overcome. Using methods like variance threshold, chi-square, PCA loadings, random forest importance, mutual information, lasso coefficients, and recursive feature elimination, a prediction model is created to ascertain the fertility state of the soil. A wise combination of federated learning techniques with IoT sensors increases the centralized server model's efficiency and guarantees data security for all clients. Centralized model: It collects weights and biases from training models on local clients, aggregates them and gives it back to enhance performance.

II. LITERATURE REVIEW

To identify the crops to be cultivated in the soil and to anticipate crop requirements, soil fertility study is crucial. Soil fertility has been studied using a variety of machine learning techniques, including support vector machines, random forests, weighted K-mean algorithms, and multiple linear regression (MLR). Deep learning and random forests turned out to be superior.

A framework using an artificial neural network (ANN) was proposed for soil nutrient quality analysis, with an average accuracy of 90%.

A variety of machine learning methods were used to determine the fertility of the soil and increase crop yields. Random Forest's forecast accuracy was 100%.

Decision tree and support vector machine techniques achieved 99% accuracy. The Gaussian radial basis function achieved 90% accuracy, optimizing the use of nutrients and reducing fertilizer purchasing costs.

To sum up, soil analysis has been carried out utilizing a variety of learning approaches and procedures in order to determine appropriate crops and monitor fertility levels. In the end, this improves crop yields by lowering fertilizer use and preserving soil health.

III. METHODS OF THE FEATURE SECTION

The client and server employed different strategies. The client's current dataset was consistently subjected to the model. We used feature selection techniques on a dataset for every client in order to improve model performance, reduce complexity, and improve data quality. The best model fits for an optimized study were then obtained by applying a model that integrated logistic regression, decision trees, naive Bayes, and their combinations. The server-end model received the encrypted results after that.

Feature Section Techniques

We used various features and evaluated how well they worked. Using the formula provided in Equation (1), the features whose variance surpassed the threshold value were chosen using the variance threshold:

$$\boxed{\text{var } X_j > \text{threshold}}$$

To determine the relationships between a feature and the target variables displayed in Equation (2), the chi-square was computed:

$$\boxed{X^2 = \sum (O_i - E_i)^2 / E_i}$$

where "Ei" represents the predicted frequency and "Oi" represents the observed frequency. Based on the model coefficient, the least significant features were eliminated via recursive feature elimination.

Furthermore, the PCA loadings were used to pinpoint the dataset's key characteristics. Equation (3) illustrates how random forest was used to determine the significance of particular traits based on entropy:

International Journal of Intelligent Computing Systems
July-2026| IJICS | Volume-2 Issue-2 | ISSN: 3107-7218

$$X_j = \sum_{t \in T} \Delta I_t X_j$$

where the decrease in impurity for feature X_j in tree t is denoted by $\Delta It\ X_j$. We used mutual information to gauge how much information our target variables held. Regularization was used to reduce the coefficients to zero in order to obtain the lasso coefficients. To determine the ideal amount of features, recursive feature elimination with cross-validation support was employed. To find the pertinent features, random forest and Boruta assistance were used.

- **Sensors that use an arduino board**

  We collected soil nutrient values using an Arduino board and Internet of Things sensors, which served as test samples for analysis. Sensors were connected

  to the Arduino board in order to get data on soilnutrients.

We collected potassium, phosphorus, and nitrogen using the NPK sensor. The soil's hardness or acidity was determined using the pH sensor, and its electrical conductivity was determined using the EC sensor. The quantity of organic carbon in the soil was determined using the quantity of sulphur measured by the NIR sensor, and the quantity of zinc, iron, copper, manganese and boron measured by the colorimetric RGB and X-ray fluorescence (XRF) sensors.

Questions For Research:

RQ1: What effects do various feature selection methods have on the machine learning model's performance?

RQ2: Can hybrid model using different classifiers be applied to enhance soil fertility prediction for the overall improvement in predictive accuracy and resilience?

RQ3: How can client model results be generated and sent to the central model in a private manner?

RQ4: What optimization techniques can enhance the ability of the primary model to make predictions?

IV. PROPOSED MODEL

The local model and training it with the data at each client system forms Phase I of the proposed model, as depicted in Figure 1. In Phase II, a number of techniques are used to enhance the accuracy of the server-end model. Moreover, Phase III focuses on collecting sample data with a sensory system to test the centralized model. The detailed explanation of how each local model works is provided in figure 1. A nutrient dataset is used to train a machine learning algorithm that determines whether or not the soil is fertile.

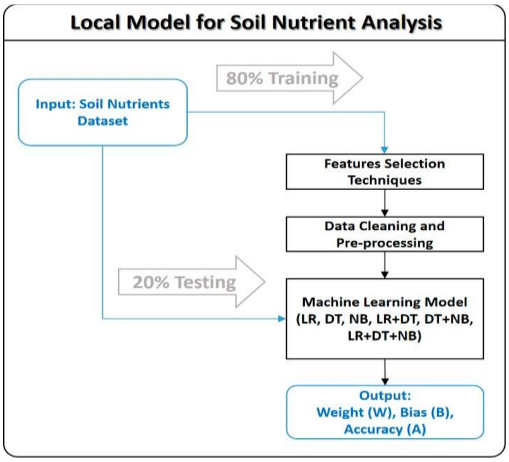


Figure 1: Proposed client model for analysing soil nutrients

- **PHASE I:**

  - The local client computer processed a conventional soil nutrient dataset with 12 unique features, including N, P, K, pH, EC, OC, S, Zn, Fe, Cu, Mn, B, and output.The dataset contains 11,440 items that show whether the binary classification is applicable.

  The feature output may be good or bad and is a measure of the fertility of the soil. The data has been cleaned by removing noise, outliers, duplicate and missing data and some features have been normalized for the binary classification. Following that, feature selection strategies were used for increasing the accuracy and to avoid overfitting.

  We computed: the RF significance, mutual information, lasso coefficients, RFECV support, Boruta support, chi square, RFE rank, PCA loadings, and variance threshold. Moreover, we utilized the naïve Bayes, logistic regression, decision trees and its stacking. Then, the ranking of the approaches based on their accuracy, precision, recall and F1 score. The best method to most accurately estimate the soil fertility level was then chosen based on the ranking, displayed in Figure 1.

  There are multiple local models of different clientsthat generate local results that do not give the server-end model the weight, bias and accuracy—the results. So, once they are achieved, they are put in each local computer

International Journal of Intelligent Computing Systems
July-2026 | IJICS | Volume-2 Issue-2 | ISSN: 3107-7218

without sharing with others.

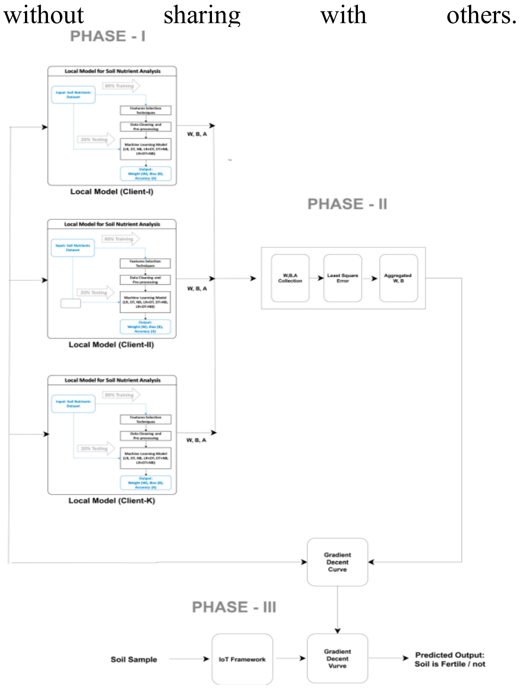


Figure 2: Proposed Centralized Federated Learning Framework for Combining, Analyzing, and Aggregating Client Results

- **PHASE II:**
  - In this stage, each client's weight, bias, and accuracy were sent to a centralized model on a server, which combined them and stored them. These details were gathered, and the error data for each sample was determined using the least squares error approach.
  - A gradient descent curve was used to confirm the bias values that were produced using an aggregation approach to the proper weighting and a minimum error. The dataset was then repeatedly tested by feeding it back to the client models.
  - The phase II model suggested here is based on a centralized server that combines the output of multiple client-side models in a federated setup. Each client contributes only the encrypted values of weight, bias, and accuracy—without revealing raw data—to maintain data privacy. The server accumulates such parameters and uses the Least Square Error (LSE) technique to analyse and reduce prediction errors. Theoptimized parameters are returned to the client models, facilitating a feedback loop that iteratively improves prediction performance without compromising confidentiality of data.
- **PHASE III:**
  - The main objective of this stage was to collect soil nutrients in the fields using sensing devices. The sensors were an NPK, pH, EC, NIR (near-infrared reflectance), colorimetric RGB, ion-selective electrode and sulphur sensor connected to an Arduino board to collect various nutrients, namely nitrogen, phosphorus, potassium, pH, electrical conductivity, NIR, ion-selective electrode, colourimetric RGB (red-green-blue), sulphur, and zinc, and iron, copper, manganese, and boron: We used the above mentioned sensors to carefully collect the soil nutrients required to analyse the soil sample with the power supply of 5 volt power to Arduino.
  - The core model has the least error and the optimal weight and bias. With more accuracy, the central model determined the soil's fruitful state after testing and analysing the nutrients gathered.

V.RESULTS AND DISCUSSION

In order to confirm their part in increasing the fertility analysis's performance, we first used feature selection techniques. While examining various attributes, A ranking table addresses the research questions and the response during implementation.

- **RESEARCH QUESTION #1**

Is it reasonable to analyse soil fertility using the feature selection technique? If yes, what effects do the various feature selection methods have on the machine learning algorithm's performance?

Solution: The different point selection methods can be used to discuss soil fertility. The RFECV (recursive point elimination with cross confirmation) was the best fashioned point as shown in our trial results because of its iterative property. This FS technique removed the least important features from our data set. We went through this process again to find the most suitable features in our data. These traits are robust as they are present in many contexts, such as traits N, P, pH, and Cu. It performs both removal and model verification, so it's the classy one. We may ascertain their relative rankings by looking at Table 1, which shows the outcomes of seven different methodologies applied to 12 criteria.

To address RQ1, the researchers examined the effectiveness of the machine learning models with varying feature selection strategies for predicting soil fertility. To boost the efficiency and exactitude of the model, feature selection is crucial for identifying which variables are the most informative, and discarding redundant or irrelevant information.

RFECV Support, and Boruta. These techniques helped in evaluating the relative importance of features like nitrogen (N), phosphorus (P), pH, and copper (Cu), which consistently appeared as significant contributors across multiple methods. Among all, RFECV emerged as the most effective method, offering both feature elimination and validation in a crossvalidated setup. The findings confirmed that the right choice of feature selection not only reduced overfitting and training time but also enhanced predictive accuracy, making the model more robust and generalizable.

Feature	Variance Threshold	Chi-Square	RFE Ranking	PCA Loadings	RF Importance	Mutual Information	Lasso Coefficients	RFECV Support	Boruta Support
N	5982.230	11,447.547	9	0.107	0.550	0.494	0.005	TRUE	TRUE
P	482.034	1347.885	7	0.108	0.127	0.144	0.006	TRUE	TRUE
K	15,413.778	101.166	11	0.107	0.029	0.012	0.000	TRUE	FALSE
pH	0.216	0.223	2	0.143	0.048	0.077	0.000	TRUE	TRUE
EC	0.020	0.025	3	0.004	0.026	0.026	0.000	TRUE	FALSE
OC	0.710	1.420	10	0.151	0.024	0.000	0.000	TRUE	FALSE
S	19.551	10.896	4	0.166	0.026	0.000	0.001	TRUE	FALSE
Zn	3.584	14.134	5	0.167	0.029	0.005	0.000	TRUE	FALSE
Fe	9.661	7.822	6	0.177	0.038	0.028	0.003	TRUE	TRUE
Cu	0.217	3.891	1	0.113	0.039	0.049	0.000	TRUE	TRUE
Mn	18.459	5.506	8	0.179	0.034	0.000	0.000	TRUE	TRUE
B	0.325	1.350	12	0.120	0.030	0.032	0.073	FALSE	FALSE

Table 1: Feature Ranking Analysis of the Dataset

Moreover, we attempted to depict the relevance scores in the figure 3 chart that lists the importance ratings of several methods. It displays each feature's ranking based on the various feature selection methods.

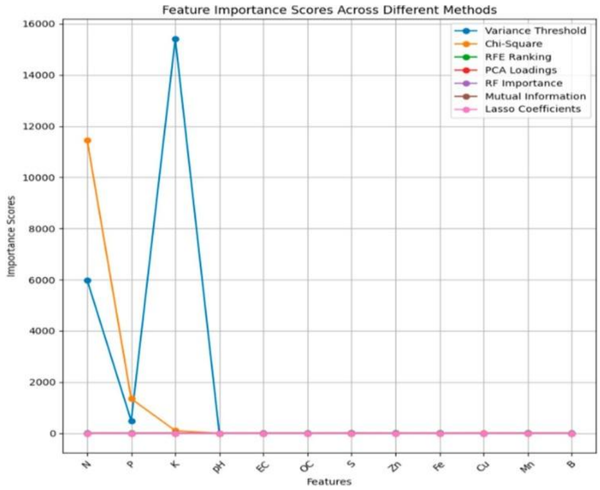


Figure 3: Feature Importance Comparison Across Different Techniques

A heat map (Figure 4) was created using a multiple feature selection method, which ranked the significance of features by comparing the outcomes of different approaches. It emphasized the importance of soil nutrients, with potassium and nitrogen standing out as crucial elements. By concentrating on these characteristics, the predictive model's accuracy can be improved.

Figure 4: Feature Ranking Using Different Techniques

- **RESEARCH QUESTION #2**

Can a hybridization model be constructed by considering the independent soil fertility models? Can the overall predictive strength and accuracy of a soil fertility prediction system be enhanced by a hybrid model that integrates all these classifiers?

Solution: Baseline classifiers were utilized in this study and hybrid models such as LR + DT, LR + DT + NB, and DT + NB were experimented with. Of these, the combination LR + DT proved to be the best with enhanced accuracy and precision. The ensemble model suggested that the best model for improving soil fertility prediction was LR + DT.

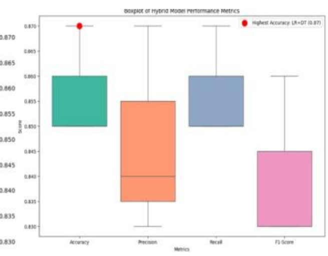
International Journal of Intelligent Computing Systems
July-2026 | IJICS | Volume-2 Issue-2 | ISSN: 3107-7218

Table 2 illustrates the methods of logistic regression, DT, and NB along with their ensemble methodologies used for data analysis.

Techniques	Logistic Regression	Decision Tree	Naive bayes	LR+DT	DT+NB	LR+DT+NB
Accuracy	0.83	0.82	0.49	0.87	0.85	0.85
Precision	0.78	0.81	0.59	0.87	0.84	0.83
Recall	0.83	0.82	0.49	0.87	0.85	0.85
F1-Score	0.81	0.81	0.41	0.86	0.83	0.83

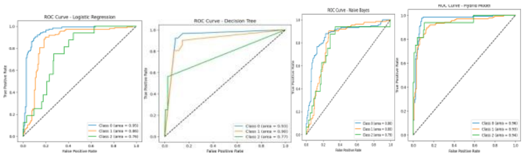
Table 2: Performance Analysis of Individual Algorithms and Their Ensemble Approaches

- In addition, a heat map and box plot (see Figure 5) were used to compare the hybridization of the following classifiers: naïve Bayes, decision trees and logistic regression.
- • The box plot was used to prove the accuracy of the best hybridization model. According to the accuracy of the performance measures, it selects the most appropriate model. In red the highest accuracy (LR + DT) is displayed.
- The precision, recall and F1-score comparison study is presented as a bar chart in Figure 5. It was observed that in all the measures the hybrid of logistic regression and decision tree performed better than naïve bayes.
- Furthermore, a heat map and box plot (Figure 6) were provided to compare and contrast the effectiveness of above discussed methods for the hybridization of naïve bayes, decision trees and logistic regression.
- The boxplot was used to show the accuracy of the best hybridization model. It picks the most suitable model based on the performance measures' correctness.
- The best accuracy (LR + DT) is indicated in red. Figure 6 shows the hybrid approach and the ROC curve and confusion matrices of the separate approaches.

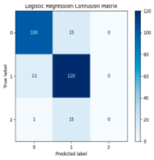


(a) Heatmap (b) Box plot

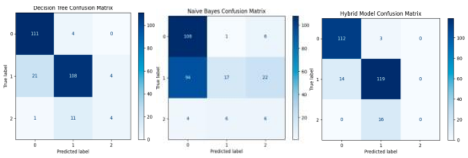
Figure 5: Heatmap and boxplot of hybrid models



(a) , (b) , (c) , (d)



(e)



(f) (g) (h)

Figure 6: ROC Curves and Confusion Matrices of Individual Models and the Proposed Hybrid Model: (a) Logistic Regression ROC Curve, (b) Decision Tree ROC Curve, (c) Naïve Bayes ROC Curve, (d) Proposed Hybrid Model ROC Curve, (e) Logistic Regression Confusion Matrix, (f) Decision Tree Confusion Matrix, (g) Naïve Bayes Confusion Matrix, and (h) Proposed Hybrid Model Confusion Matrix.

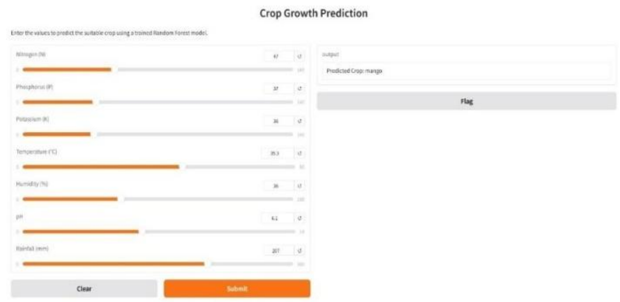


Figure 7: Prediction of Crop

VI. CONCLUSION AND FUTURE WORK

A series of techniques were employed to enhance model performance at each stage. At the client end, seven feature selection methods, including Boruta support, improved generalization, reduced overfitting, and boosted accuracy. Classifiers like logistic regression, decision tree, and their ensemble (LR + DT) were tested, with the ensemble achieving the highest accuracy of 87%. At the server end, least square error and a gradient descent curve were used to refine weights and biases via recursive feedback, which is under development. A sensory system was also integrated to test real-time soil nutrient samples.

VII. REFERENCES

1. Chandra, H. et al. (2023). Explainable AI for Soil Fertility Prediction. IEEE Access.
2. Kanade, P. (2023). Soil Analysis Using Machine Learning. Br. J. Multidiscip. Adv. Stud.
3. Madhumathi, R. et al. (2020). Soil Nutrient Analysis Using ML. CCCA-2020.
4. Pant, J. et al. (2021). Soil Quality Prediction in India Using ML. Int. J. Anal. Appl.
5. Rajamanickam, J. (2021). Soil Fertility Using Decision Tree. INFOCOMP J. Comput. Sci.
6. Yadav, J. et al. (2021). Crop Fertility Prediction Using ML. IRJCS.
7. BlesslinSheeba, T. et al. (2022). ML Classification in IoT Farms. J. Nanomater.
8. Suchithra, M.S. & Pai, M.L. (2020). Optimizing ELM Parameters. Inf. Process. Agric.
9. Chen, G. et al. (2014). Weighted K-Means for Soil Fertility. Springer.
10. Folorunso, O. et al. (2023). Review on Soil Nutrient ML Models. Big Data Cogn. Comput.
11. Kairouz, P. et al. (2021). Advances and Open Problems in Federated Learning. Foundations and Trends in ML.