Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

An Interpretable Machine Learning Approach for Soil Fertility Prediction from NPK and pH

Theepak T
Department of Computing Technologies
School of Computing, SRMIST
Kattankulathur, Chennai, India

Akshayakumar A. S.
Department of Computing Technologies
School of Computing, SRMIST
Kattankulathur, Chennai, India

Dr. S. Yazhinian
Assistant Professor, Dept. of Computing Technologies
School of Computing, SRMIST
Kattankulathur, Chennai, India

Abstract **- Soil fertility assessment is central to sustaining agricultural productivity, because the availability of nutrients such as Nitrogen (N), Phosphorus (P) and Potassium (K), together with soil pH, directly governs crop growth and yield. Conventional laboratory-based soil testing is accurate but slow, costly and dependent on specialised equipment and trained personnel, which limits how often farmers, particularly smallholders, can monitor their fields. Machine learning offers a faster, data-driven alternative for estimating soil fertility from routinely measured nutrient parameters. This paper presents a supervised learning approach that classifies soil samples into Low, Medium and High fertility classes using N, P, K and pH as input features. Two tree-based classifiers are implemented and compared: a Decision Tree, chosen for its transparent and easily interpretable decision rules, and a Random Forest, an ensemble of decision trees that reduces variance and the risk of overfitting. The workflow covers data collection, preprocessing, feature selection, model training and evaluation using accuracy, precision, recall and F1-score. Descriptive statistics, box plot analysis and correlation analysis are used to characterise the dataset before modelling, and feature importance scores are used to examine the contribution of each soil parameter. The experimental results show that both classifiers perform well and that the Random Forest achieves higher accuracy, precision and recall than the Decision Tree. The proposed approach provides a low-cost, interpretable decision-support tool for soil management and precision agriculture.**

Keywords - Decision Tree; Machine Learning; Random Forest; Soil Fertility Prediction; Soil Nutrient Analysis; Supervised Learning; Sustainable Agriculture

I. INTRODUCTION

Agriculture underpins food security and economic development, and soil fertility is one of its most basic determinants. The concentrations of Nitrogen (N), Phosphorus (P) and Potassium (K), together with soil pH, strongly influence plant growth and overall soil health, so accurate assessment of these parameters supports balanced fertilizer application and improved crop yield.

Conventional soil testing relies on laboratory chemical analysis. Although accurate, it is time-consuming, labour-intensive and often unaffordable for small-scale farmers, and periodic testing cannot provide the timely information that precision agriculture requires. These limitations have motivated data-driven approaches to soil fertility assessment.

Machine learning has shown considerable promise for agricultural prediction tasks. Early work tuned the J48 decision tree algorithm for soil fertility prediction [1], and later studies combined soil fertility analysis with crop prediction [2]. Random Forest classifiers have been applied directly to soil fertility prediction [3], machine-learning-based soil classification has been integrated with crop recommendation systems [4], and comparative studies have confirmed the suitability of supervised classifiers for categorising soil fertility [5].

Despite this progress, many existing systems couple fertility prediction with crop recommendation or IoT-based monitoring rather than focusing on fertility classification itself. There remains a need for focused models that predict soil fertility levels from essential nutrient parameters while remaining interpretable and computationally efficient.

In this work, Decision Tree and Random Forest classifiers are developed to classify soil samples into Low, Medium and High fertility classes using N, P, K and pH values. The main contributions are: (i) a simple, interpretable fertility classification workflow based on four routinely measured soil parameters; (ii) an exploratory analysis of the dataset using descriptive statistics, box plots and correlation analysis; and (iii) a comparative evaluation of a single tree and an ensemble model, supported by feature importance analysis. Section II reviews related work, Sections III and IV describe the algorithms and the proposed model, Sections V and VI present the dataset and workflow, Section VII discusses the results, and Sections VIII and IX conclude the paper and outline future work.

II. LITERATURE REVIEW AND AVAILABLE MODELS

Theoderaj [6] proposed a data-driven soil fertility classification framework that combines IoT sensing with machine learning models, while Ramesh and Venkatesan [7] applied machine learning algorithms to soil nutrient prediction. Latha and Kumaresan [8] used deep learning methods to evaluate soil fertility and recommend crops. Ensemble learning has received particular attention: Sunori et al. [9] assessed soil fertility using ensemble methods, and the authors of [10] combined soil fertility prediction with a fertilizer recommendation system. Prabavathi and Chelliah [11] proposed an optimized Gaussian extreme learning machine that predicts crop yield from soil factors.

Vimaladevi et al. [12] identified suitable fertilizers from NPK levels using machine learning, and the authors of [13] developed a soil quality identification and monitoring approach for sugarcane cultivation. Interpretability has been addressed by Praveen et al. [14], who proposed an explainable AI (XAI)

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

assisted model for soil fertility prediction. Other studies have used IoT-based smart soil systems for crop prediction [15], sequential models to improve soil fertility prediction [16], Bayesian model averaging for crop prediction and soil fertility management [17], and machine-learning-based classifiers for soil fertility classification [18]. Most of these works link fertility prediction with crop or fertilizer recommendation; the present study focuses specifically on classifying soil fertility from essential nutrient parameters.

A. Traditional Soil Fertility Assessment

Conventional assessment determines nutrient concentrations through laboratory chemical analysis. These methods are accurate but require specialised equipment, skilled personnel and considerable time, which makes frequent monitoring impractical, especially for farmers in rural regions.

B. Machine Learning Methods Used in Previous Research

Decision Tree, Support Vector Machine, K-Nearest Neighbours and Random Forest algorithms have all been used to classify soil fertility from nutrient parameters [1], [3], [5]. These models improve prediction accuracy by capturing non-linear relationships between soil properties and fertility levels.

C. Limitations of Existing Models

Many existing studies predict suitable crops rather than soil fertility itself. Several models also offer limited interpretability or are trained on region-specific datasets, which reduces their wider applicability, and real-time, farmer-friendly systems remain relatively unexplored.

III. ALGORITHM SELECTION AND ANALYSIS

A. Selected Machine Learning Algorithms

The proposed system uses two supervised learning algorithms, Decision Tree and Random Forest. Both are well suited to structured tabular data and can model the relationships between basic soil nutrient parameters and fertility classes.

B. Decision Tree Model

A Decision Tree is a hierarchical classifier that recursively splits the data on feature thresholds chosen to maximise information gain. In soil fertility prediction, internal nodes represent decision rules on N, P, K and pH values, and leaf nodes represent fertility classes. Decision Trees are computationally efficient and produce rules that are easy to understand, but they may generalise poorly when trained on complex datasets.

Justification: The explicit decision rules show clearly how each soil parameter affects the fertility class, which is important in agricultural applications where transparent decisions are needed.

Limitation: A single tree can grow overly complex on varied data, which can lead to overfitting and reduced generalisation performance.

C. Random Forest Model

Random Forest is an ensemble algorithm that builds multiple Decision Trees on random subsets of the data and features, and determines the final prediction by majority voting across the trees.

Justification: Aggregating many trees reduces variance and the risk of overfitting, which improves the stability of predictions.

IV. PROPOSED MODEL

The proposed model predicts soil fertility levels using supervised machine learning. Instead of relying on complex black-box algorithms, it uses interpretable and robust tree-based methods that are practical for agricultural use. The model takes N, P, K and pH as input features. The Decision Tree provides transparent decision rules, while the Random Forest improves prediction accuracy and reduces overfitting, so the model balances interpretability with performance. The overall architecture is shown in Fig. 1.

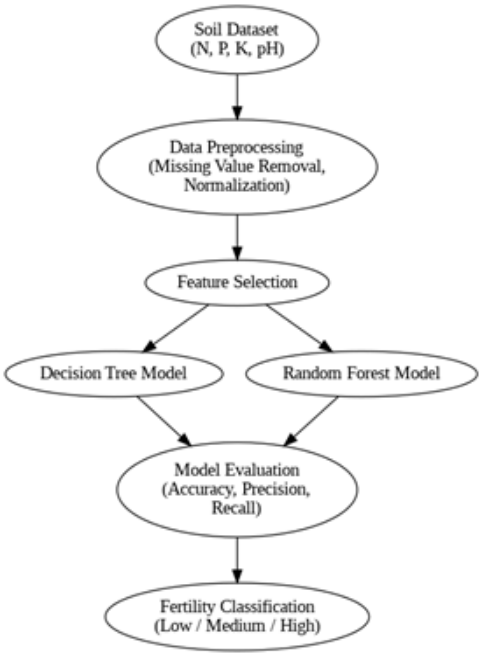


Fig. 1. Architecture of the proposed soil fertility prediction system

V. DATASET DESCRIPTION

The dataset consists of soil samples collected from different districts and obtained from public agricultural sources. Each sample contains four soil parameters: N, P, K and pH, which are essential indicators of soil nutrient composition and serve as the input features for classifying soil fertility into Low, Medium and High classes.

Descriptive statistics of the parameters are given in Table I. The average Nitrogen concentration is 24.77, with moderate variability (standard deviation 8.70). Phosphorus and Potassium have higher mean values (61.16 and 61.74), indicating relatively nutrient-rich samples. The pH ranges from 6.05 to 9.03 with a mean of 7.07, indicating slightly acidic to alkaline conditions across districts. This variability in nutrient concentration supports the use of machine learning for fertility classification.

TABLE I. DESCRIPTIVE STATISTICS OF SOIL PARAMETERS

Parameter	Mean	Std Dev	Minimum	Maximum
Nitrogen (N)	24.77	8.70	20.00	66.96
Phosphorus (P)	61.16	13.50	20.05	80.00
Potassium (K)	61.74	12.96	22.91	80.00
pH	7.07	0.28	6.05	9.03

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

A sample of the dataset records is shown in Table II.

TABLE II. SAMPLE SOIL DATASET RECORDS

District	Nitrogen (N)	Phosphorus (P)	Potassium (K)	pH
Yadgir	28.736	67.103	78.602	7.0954
Yadgir	22.229	62.822	75.221	7.3273
Yadgir	22.145	70.373	77.868	7.0311
Yadgir	26.321	72.488	76.697	7.0470
Yadgir	26.687	72.183	75.068	7.0078

Fig. 2 shows the distribution of each parameter. Nitrogen, Phosphorus and Potassium vary considerably in concentration, Nitrogen shows a number of high-value outliers, and pH lies within a much narrower range. Fig. 3 shows the correlation matrix of the four parameters. All pairwise correlations are weak (absolute values of 0.19 or less), which indicates that each parameter contributes largely independent information and that multicollinearity is not a concern.

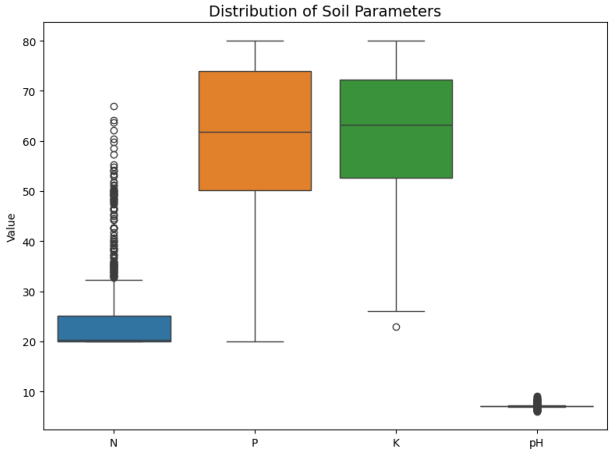


Fig. 2. Distribution of soil nutrient parameters (box plot)

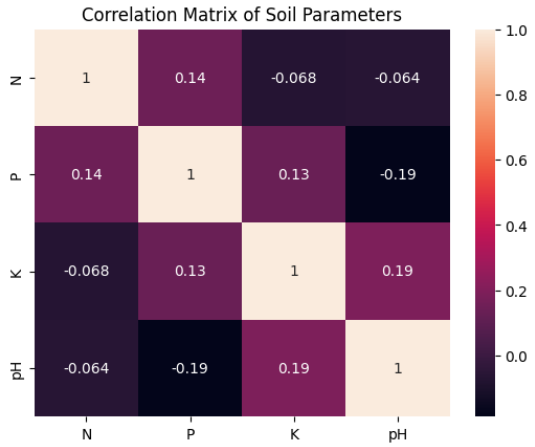


Fig. 3. Correlation matrix of soil parameters

VI. WORK FLOW OF THE PROPOSED SYSTEM

The workflow follows the architecture in Fig. 1. The soil dataset is acquired and preprocessed, relevant features are selected, and the processed data are used to train the Decision Tree and Random Forest classifiers.

Step 1: Soil nutrient data (N, P, K and pH) are collected from agricultural datasets.

Step 2: The data are preprocessed to handle missing values and to scale the features to a common range.

Step 3: The relevant soil parameters are selected as input features.

Step 4: The Decision Tree and Random Forest models are trained on the processed data.

Step 5: Model performance is evaluated using standard classification metrics.

Step 6: Each soil sample is classified as Low, Medium or High fertility.

A. Data Collection

Soil records containing N, P, K and pH values are obtained from public agricultural sources.

B. Data Preprocessing

The dataset is cleaned to remove errors and missing values, and feature scaling brings all parameters to a comparable range so that the models train consistently.

C. Model Training

The dataset is divided into separate training and testing portions. Both models are trained on the training data and evaluated on the unseen test data so that performance is assessed fairly.

D. Fertility Classification

Using the patterns learned from the nutrient data, the trained models assign each sample to the Low, Medium or High fertility class.

Performance is measured using accuracy, precision, recall and F1-score. Accuracy is the proportion of correctly classified samples; precision is the proportion of predicted fertility labels that are correct; recall measures how many samples of each fertility class are correctly identified; and F1-score is the harmonic mean of precision and recall.

VII. EXPERIMENTAL RESULTS AND DISCUSSION

Fig. 4 compares the two classifiers. Both models score above 0.9 on accuracy, precision and recall, and the Random Forest achieves higher accuracy, precision and recall than the Decision Tree. Because it combines the predictions of many trees, it handles variation in the data more reliably than a single tree.

A. Performance Analysis of Decision Tree

The Decision Tree produces an easily interpretable set of rules based on the soil parameters and performs well overall. However, it is sensitive to small changes in the training data and is prone to overfitting as the data become more complex.

B. Performance Analysis of Random Forest

The Random Forest is more accurate and stable than the single Decision Tree. By aggregating many trees, it reduces variance and captures non-linear relationships among the soil parameters more effectively, which yields more reliable fertility predictions.

C. Comparative Discussion

The Random Forest outperforms the Decision Tree on the evaluated metrics, confirming that ensemble models are a

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

suitable choice for soil fertility prediction when accuracy and robustness are priorities. The Decision Tree remains useful where simple, transparent rules are preferred.

The contribution of each soil parameter was analysed using the feature importance scores of the Random Forest model, shown in Fig. 5. Nitrogen, Phosphorus and pH have the highest and nearly equal importance, while Potassium contributes the least. These results confirm the relevance of the selected input parameters for fertility classification.

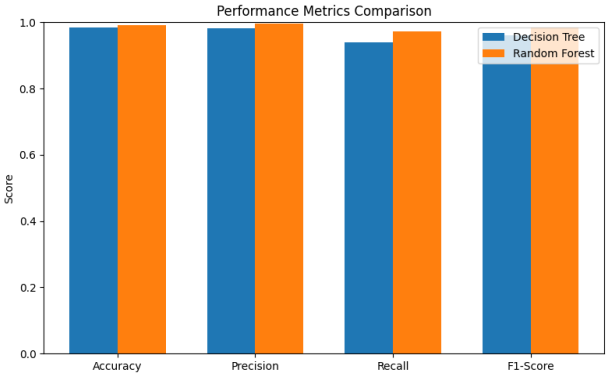


Fig. 4. Performance metrics of Decision Tree and Random Forest

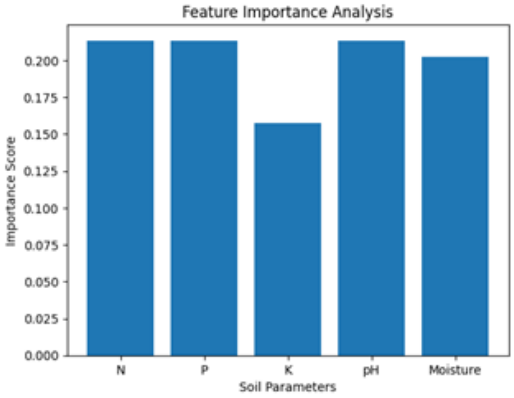


Fig. 5. Feature importance of soil parameters (Random Forest)

VIII. CONCLUSION

This paper presented a machine-learning-based approach for predicting soil fertility from four key soil parameters: Nitrogen, Phosphorus, Potassium and pH. The proposed system classifies soil fertility into Low, Medium and High categories, and the experimental results show that the Random Forest performs better than the Decision Tree. The approach offers a cost-effective complement to traditional soil testing and can support sustainable agricultural practices. The study is limited to four soil parameters and to data from the selected districts, so validation on larger and more diverse datasets is still required.

IX. FUTURE ENHANCEMENTS

Future work will integrate climatic parameters such as rainfall and temperature, develop a numerical Soil Fertility Index, extend the system towards crop recommendation and fertilizer advisory, and deploy the model as a web or mobile application for real-time use.

REFERENCES

[1] J. Gholap, "Performance Tuning of J48 Algorithm for Prediction of Soil Fertility," Proc. Int. Conf. on Advances in Communication and Computing, IEEE, pp. 1-5, 2015.

[2] M. S. Muneshwara, R. R. Raghavendra, K. R. Venugopal, "Soil Fertility Analysis and Crop Prediction using Machine Learning," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 9, no. 6, pp. 453-458, Apr. 2020.

[3] K. Kumar, R. Prasad and A. Sharma, "Soil Fertility Prediction using Random Forest Algorithm," Proc. IEEE Int. Conf. on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSM), pp. 98-103, 2018.

[4] N. H. Kulkarni and S. Srinivasan, "Machine Learning Based Soil Classification and Crop Recommendation System," Proc. IEEE Int. Conf. on Computing, Communication and Automation (ICCCA), pp. 1-6, 2019.

[5] M. Sujatha and C. D. Jaidhar, "Classification of Soil Fertility using Machine Learning Techniques," Proc. IEEE Int. Conf. on Secure Cyber Computing and Communications (ICSCCC), pp. 500-505, 2021.

[6] A. K. C. Theoderaj, "Data-Driven Soil Fertility Classification Using IoT Sensing and Machine Learning Models," in Proc. IEEE Int. Conf. Industrial Engineering and Engineering Management (IEEM), 2024, pp. 82-87.

[7] S. Ramesh, R. Venkatesan, "Application of Machine Learning Algorithms for Soil Nutrient Prediction," in the IEEE Int. Conf. on Artificial Intelligence and Smart Systems, ICAIS, pp. 210-215, 2022.

[8] P. Latha and P. Kumaresan, "Deep Learning Based Methods for Evaluating Soil Fertility and Recommending Crop for Production," in Proceedings of Access to the International Conference for Computers and Electronics, vol. 11, pp. 45678-45686, 2023.

[9] S. K. Sunori, P. B. Negi, N. C. Joshi, A. Mittal, and P. Juneja, "Soil Fertility Assessment Using Ensemble Methods in Machine Learning," in Proc. 2024

[10] M. R., D. R., E. E., G. R., and A. T., "Prediction of Soil Fertility Using ML Algorithms and Fertilizer Recommendation System," in Proc. 2023

[11] R. Prabavathi and B. J. Chelliah, "An Optimized Gaussian Extreme Learning Machine (GELM) for Predicting Crop Yield Using Soil Factors," in Proc. 2022

[12] M. Vimaladevi et al., "FertiForecast: Identification of Fertilizer Based on NPK Levels Using Machine Learning Algorithms," in Proc. 2024

[13] S. P. N. and H. P. M. Kumar, "Soil Quality Identifying and Monitoring Approach for Sugarcane Using Machine Learning Techniques," in Proc. 2022

[14] R. Praveen, V. Kale, and P. S. Srivastav, "Explainable AI (XAI) Assisted Prediction Model in Soil Fertility," in Proc. 2025

[15] G. S. Pravallika et al., "Proficient Smart Soil Based IoT System for Crop Prediction," in Proc. 2020

[16] K. P. and J. J., "Improving the Prediction of Soil Fertility with Sequential Models: A Pathway to Enhancing Crop Production," in Proc. 2024

[17] R. Dash, U. K. Kamila, and M. Sahu, "Improvement of Crop Prediction and Soil Fertility Management by Bayesian Model Averaging and Machine Learning," in Proc. 2025

[18] S. M. and J. C. D., "Classification of Soil Fertility Using Machine Learning Based Classifier," in Proc. 2021

(This work is licensed under a Creative Commons Attribution 4.0 International License.)