International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 8, Issue 3
DOI: https://doi.org/10.15157/IJITIS.2025.8.3.513-536
Received: 01.08.2025; Revised: 23.08.2025; Accepted: 28.08.2025

Research Article

Developing a Conceptual Framework for Soil Property Analysis and Crop Yield Prediction Using Machine Learning Techniques

Vimla Dangi*, Chandrashekhar Goswami, Prasun Chakrabarti

Faculty of Computing & Informatics, Sir Padampat Singhania University, India
***vimla.phd2023@spsu.ac.in**

Abstract

The most important single factor is soil fertility which influence crop sustainability and agricultural productivity. The necessity to use data-driven approaches to assess the health of the soil and propose the crops that should be grown in it has become a crucial issue because the accuracy of agriculture is required increasingly frequently. Based on the dataset of the Soil Health Card (SHC) of the Government of India, the presented study provides a conceptual framework that involves the application of the machine learning approaches to analyse soil characteristics and predict its agricultural productivity. The framework is based on twelve important soil parameters: sulphur (S), nitrogen (N), zinc (Zn), phosphorus (P), electrical conductivity (EC), potassium (K), manganese (Mn), copper (Cu), boron (B), iron (Fe), organic carbon (OC), and pH to cluster soil samples into the categories of low, medium, and high soil fertility by using the K-means algorithm. To suggest the correct crops that must be grown in each of the fertility categories, the Random Forest Classifier is then trained after the clustering. The model is checked by K-Fold cross-validation (k=5) and Holdout (80/20 split) to make sure that in unseen data strong generalization will be achieved. An average performance of 91 percent in K-Fold, and zero in holdout validation showing no inaccuracies in dividing the test set and an RMSE and MAE also zero, results indicate high performance and no mistakes in classification. Also, the proposed methodology enhances the agronomic decision-making with the help of AI-based crop proposals targeting each of the fertility classes. This study is an indication of the efficiency of the integration of supervised and unsupervised methods in agricultural informatics. It attracts interest in how intelligent models can high-grade the use of resources, encourage sustainable agriculture and endow growers with useful information based on real-life DO data.

Keywords: Soil Health Card (SHC); K-means; K-Fold; Random Forest; Mean Absolute Error (MAE)

INTRODUCTION

Soil health, incorporates physical and chemical health to the soil which entails various physical and chemical features. and biological aspects, core participant of agricultural performances [1]. Climate change, unsustainable: This has been negativing in terms of farming, soil erosion and farming practices. impact on soil fertility in the recent past that has lowered crop low-yields and threatened food security. On the backdrop of and the

precision agriculture or the usage of technologic innovations, in particular, data-driven and machine learning (ML) the rate at which agricultural techniques have been introduced into agriculture has picked up pace response [2].

Given Since it is a country with an agrarian economy, India has initiated several impacts. of monitoring and improvement programs of soil health. The health of soil Indian Card (SHC) program that was initiated in the country one of the most famous is government. This program aims at inspecting the nutritional and health quality of the soil in the country's environmental conditions and offer customized instructions to farmers. The collection of data in this scheme contains numerous structured soil information, that creates new opportunities for modelling and prediction using machine learning. Figure 1 depict the soil fertility level diagram.

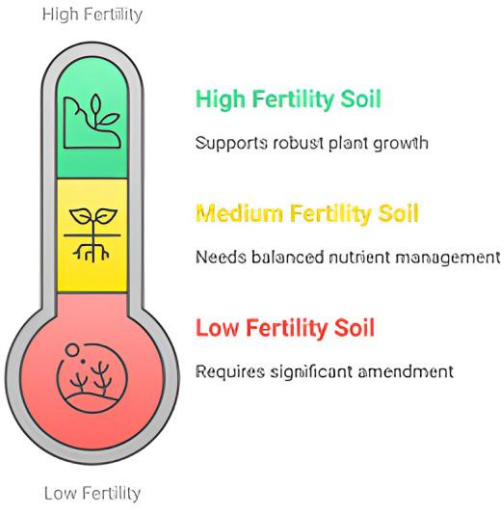

Figure 1. Soil fertility levels diagram

Manual assessment has constituted a significant part of the conventional techniques of soil testing and crop prescription, and it is ineffective, unreliable, time-consuming and lacks the scope of scalability. Need to have automated, intelligent, and yet to scale frameworks has stimulated the exploration of machine learning (ML) models capable of discovering patterns within complex datasets in a bid to facilitate the decision-making process [3].

To assess with a view to evaluate the features of soil and predicting the processes to cultivate the suitable crops, the proposed study indicates a new conceptual scheme combining Random Forest classification with K-means clustering. K-means clusters are applied to classify soil in to three categories, namely low, medium, and high fertility based on twelve general soil characteristics which include nitrogen (N), potassium (K), Sulphur (S), manganese (Mn), zinc (Zn), iron (Fe), phosphorus (P) and boron (B), organic carbon (OC), copper (Cu), electrical conductivity (EC) and pH. A classification Random Forest that will be able to classify new soil samples and recommend appropriate crops is then trained along clustering output [4].

To further guarantee the model's robustness and generalisability, it is thoroughly assessed using K-Fold cross-validation and Holdout validation procedures. Effectiveness is measured using performance indicators including F1-score, Accuracy, Precision, Recall, RMSE, and MAE.

Combining supervised and unsupervised learning techniques increases prediction accuracy while also making the outcomes easier to understand. The suggested methodology seeks to be able to help farmers make informed decisions that support sustainable agriculture, efficient use of resources, and higher crop production by offering AI-driven crop recommendations.

LITERATURE REVIEW

The use of AI is transforming agriculture into a digital industry. Analysing soil characteristics and accurately forecasting crop yields with data-driven models are two important areas of study. Over the past years, various studies examined the use of various forms of machine learning to address the crop suggestion, soil fertility evaluation, and sustainable farm management issues.

Soil Features Prediction with Machine Learning

The significance of the accuracy, interpretability of models, and inclusion of such environmental variables as rainfall and temperature were outlined in a systematic review study of machine learning algorithms implemented to predict nutrients of soil by Folorunso et al. (2023) [5] that also emphasized the most prevalently deployed algorithms such as the Support Vector Machines (SVM), Random Forest (RF), and Artificial Neural Network due to their resilience in addressing nonlinearity and multivariate features of soil-based data.

In another quantitative study, Motia and Reddy (2021) [6] compared the effectiveness of various supervised machine learning methods to forecast the soil characteristic in terms of efficiency and processing time. In their research, they found that tree-based models such as RF and Decision trees outperform traditional statistical methods especially used in terms of working with large volumes of soil data.

To create more advanced soil characterization system, Mello et al. (2022) [1] built up on this research, and integrated machine learning in combination with geophysical image delivered thanks to sensors. In their study, they were able to find out that using proximal soil sensing techniques (e.g. electromagnetic induction) coupled with machine learning algorithms had a big contribution in predicting soil characteristics like pH, organic carbon and texture.

Latha and Kumaresan (2025) [7] recently demonstrated the power of combining deep learning with multimodal agricultural data by developing a hybrid CNN–LSTM architecture enhanced with an attention mechanism to predict soil nutrient status and generate crop recommendations. Leveraging twelve SHC metrics, gridded weather data such as rainfall and temperature, and high-resolution data from UAV flights over different

agro-climatic zones in India, they were able to correctly classify nutrients more than 93 percent of the time using the model and provide personalized AI-moderated recommendations for the right crops to plant.

Crop Yield Prediction and Remote Sensing Integration

Detailed review of ML and DL models that are applied Jabed and Murad [8] were able to give a prediction of the crop yield (2024) and analyzed the optimal NDVI/EVI composites for inclusion in machine learning pipelines. They found out that due to their ability to tame time and geographical variations in crop data, ensembles such as Gradient Boosting and heterogeneous deep. The techniques such as CNN-LSTM network learning are gaining popularity. The need to use sustainability measures such as soil health, climatic resilience etc. The study also highlighted socioeconomic data.

Early yet essential contribution into the field was made by Khanal et al. (2018) [2], who combined machine learning and remote sensing to spatially predict soil parameters and maize yield. The author's demonstrated the usefulness of the high-resolution Sentinel-2 and MODIS indices for the prediction of soil and yield in their finding. They developed Random Forest model with a benchmarking precedent of site-specific nutrient mapping and yield prediction sophistication, which reached remarkable precision in nutrient mapping and yield forecasting when overall high-resolution space images and on-ground genuine samples were used as an input of their model.

Beaudoin et al. (2023) [9] suggested STICS soil-crop model which is a conceptual model incorporating the crop physiology into environmental and soil variables. The coupling between remote sensing and soil-crop dynamics in the STICS model has also been described in their research. Although not strictly speaking a machine learning model, it provides a modular means of crop prediction based on simulation and forms an interpretative foundation to hybrid frameworks like that proposed in this paper to combine supervised machine learning approaches to crop suggestion and crop-yield prediction, e.g., the Random Forest, with unsupervised machine learning like K-means clustering.

Akkem et al. (2025) [10] applied LIME, DiCE, and SHAP to explain Random Forest–based crop recommendations using soil property and weather data, delivering local and counterfactual explanations in a GDPR-compliant manner that measurably increased farmers' trust.

Shams et al. (2024) [11] introduced XAI-CROP, which integrates GBM with LIME and SHAP on soil, historical yield, and weather inputs to produce both global and local feature attributions, achieving $R_2 = 0.9$.

Research Gaps Identified

Even though previous literature has shown that machine learning is effective in analyzing soil and yields, several limitations are still made:

- Lack of unified constructs, combining AI AI-based recommendation of crops, classification and clustering.

- Poor use of validation strategies to be used to check the robustness of the model, e.g. holdout validation or stratified K-fold.
- Practical agricultural advisory outputs like fertility mapping and real-time crop recommendations are not fully integrated.

Contribution of the Present Work

The current study suggests a hybrid intelligent framework that combines the following to address the deficiencies found:

- Grouping soil fertility (high) using K-means clustering There are two steps of clustering in the formation of soil fertility groups. (medium, and poor),
- Random Forests crop suitability classifier prediction modelling,
- To validate the model, we can use K-fold cross-validation and holdout, and
- Performance indicators are analysed (e.g. accuracy, recall, F1-score, RMSE and MAE).
- Precipitation and temperature: Indian Meteorological Department (IMD) gridded data (0.25° resolution, daily).
- Relative humidity and solar radiation: NASA POWER (0.5° resolution, daily).
- Inverse distance weighting to align climate grids with SHC sample coordinates.

The Indian Soil Health Card data which comprises It uses 12 key attributes of soil health, to train and examine the model. This is the intended purpose of an integrated strategy. foster accurate agriculture and grow more crop production presenting their predictions collected with the help of credible AI recommendations [12].

RESEARCH METHODOLOGY

Following the data provided by the Indian government in the Soil Health Card (SHC) records, the research relies on a structured and mixed machine learning algorithm to classify soil health and predictive crop productivity. The Holdout and K-Fold Cross-Validation operations are utilized to validate the approach of the methodology that involves the ensemble based supervised classification (Random Forest) and unsupervised (K-means clustering) learning methods [13]. In the set of the following subsections, each stage of the research process will be described:

Data Collection and Preprocessing

The data's of the official Soil Health Card portal is as follows.

Data Source: The dataset was given by Soil Health [14]. The dataset comprises 28230 soil samples collected (April 2018–March 2024)

Geographical area: 28 states and 8 union territories of India.

Chosen characteristics: The twelve important features of the soil have been used as features:

- Sulphur (S) and potassium are Macronutrients. Phosphorus (P), nitrogen (N) and (K).

- Micronutrients are Mn, Iron (Fe), Cu, Zn, and B.
- Additional characteristics include pH, electrical conductivity (EC), and organic carbon (OC).

Target Label: Crop recommendations were derived by clustering soil fertility levels based on these characteristics [9].

- Steps for Data Cleaning and Preparation: Outlier and missing values were eliminated.
- Use of StandardScaler to normalize feature values
- Label encoding was used to encode categorical variables, if any.

Figure 2 present the feature importance from random forest model.

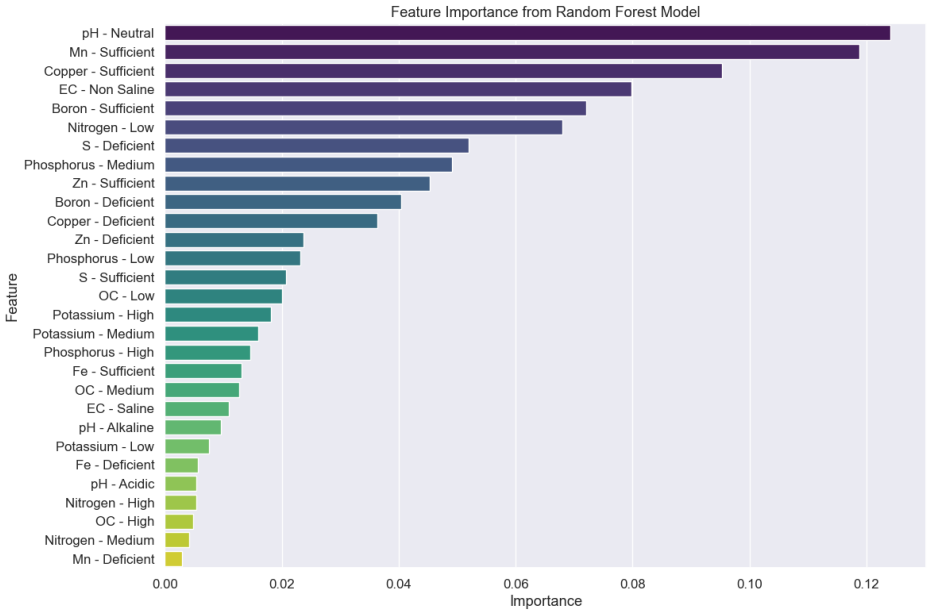

Figure 2. Feature importance for random forest model

K-Means Clustering for Soil Fertility Classification

K-Means clustering was used to divide the soil into three fertility levels: Low, Medium, and High.

- Three clusters (k) were empirically selected to represent low, medium, and high fertility.
- All 12 soil characteristics are input variables.
- Output: Each sample is given a cluster label.

To identify appropriate crops based on cluster characteristics, the clustering process helps group similar soil samples.

Supervised Learning with Random Forest for Crop Prediction

Based on the cluster labels and feature values, a Random Forest Classifier was trained to forecast the crop suggestion.

- 70% of the data (Holdout) is the training set.

- 30% of the data is the test set.
- Scaled soil features are the model's inputs.
- Model Output: Cluster-based predicted crop recommendation label

Validation Techniques

The following validation methods were used to guarantee a strong evaluation:

- Holdout Verification (a): testing (30%) and Training (70%) sets of data were separated.
- Accuracy, precision, recall, MAE, F1-score, and RMSE were the metrics that were computed.
- K-Fold Cross-Checking
- Stratified K-Fold (k=5) to preserve the fertility class distribution.
- For consistency in performance, average measurements are presented across all folds.

The performance of the model was measured using Table 1.

Table 1. Performance Metrics

Metric	Description
Accuracy	Proportion of correctly classified samples
Precision	Proportion of positive predictions that are truly positive
Recall	Proportion of actual positives accurately recognized
F1-Score	Harmonic mean of both precision and recall
RMSE (Root Mean Square Error)	Measures average magnitude of prediction error
MAE (Mean Absolute Error)	Measures average of absolute differences between predictions and actual

Crop Recommendation Framework

The suggested system suggests appropriate crops based on categorised soil fertility levels by combining machine learning methods with agronomic knowledge [15]. This is accomplished by using Random Forest classification and K-means clustering to the Soil Health Card dataset. The objective is to offer crop recommendations that support sustainable agricultural practices and are suitable for the region and fertility.

***Cluster-Based Fertility Classification*:** Using K-means clustering, the soil samples were initially divided into three groups according to twelve important soil parameters: pH, B, OC, EC, N, P, K, S, Mn, Cu, Zn, Fe.

Three general fertility levels are represented by the resulting clusters:

- Cluster 0: Infertility Low
- Cluster 1: Fertility Moderate
- Cluster 2: high fertility.

A centroid with a unique combination of soil nutrient contents and chemical properties is represented by each cluster.

***Agronomic Analysis for Crop Mapping*:** Following the formation of soil clusters, each cluster was mapped to the most appropriate crop or crops using a combination of domain knowledge from agricultural practices and regional soil-crop appropriateness criteria. The mapping looks like Table 2:

Table 2. Agronomic Analysis for Crop Mapping

Cluster	Fertility Level	Dominant Soil Traits	Recommended Crop(s)
Cluster 0	Low	Low NPK, acidic pH, low micronutrient	Wheat
Cluster 1	Medium	Balanced nutrients, moderate pH levels	Rice
Cluster 2	High	High macronutrients and micronutrients	Maize

***Integration with Machine Learning Model*:** Following classification, the cluster label for each soil sample was fed into the Random Forest model, which predicted the best crop by utilising both the cluster label and the initial soil characteristics. Accuracy and interpretability were enhanced by this hybrid technique, particularly in cases that were unclear [3].

***Sample Recommendations*:** A subset of the dataset was examined to assess the recommendation logic. The system's useful results are demonstrated by the following examples, see Table 3.

Table 3. Sample Recommendations

Sample ID	Cluster	Recommended Crop
8	0	Wheat
17	0	Wheat
15	2	Maize
26	2	Maize
31	2	Maize

***Practical Application*:** The framework is compatible with:

- Farmers: By using the results of soil tests to inform crop planning.
- When deciding how to distribute agricultural subsidies based on data
- To further model yield optimization strategies, researchers

Because it eliminates the need for expert assistance in crop recommendation and manual soil assessments, this AI-driven approach provides scalability, adaptability, and sustainability [16].

The framework model is depicted in Figure 3.

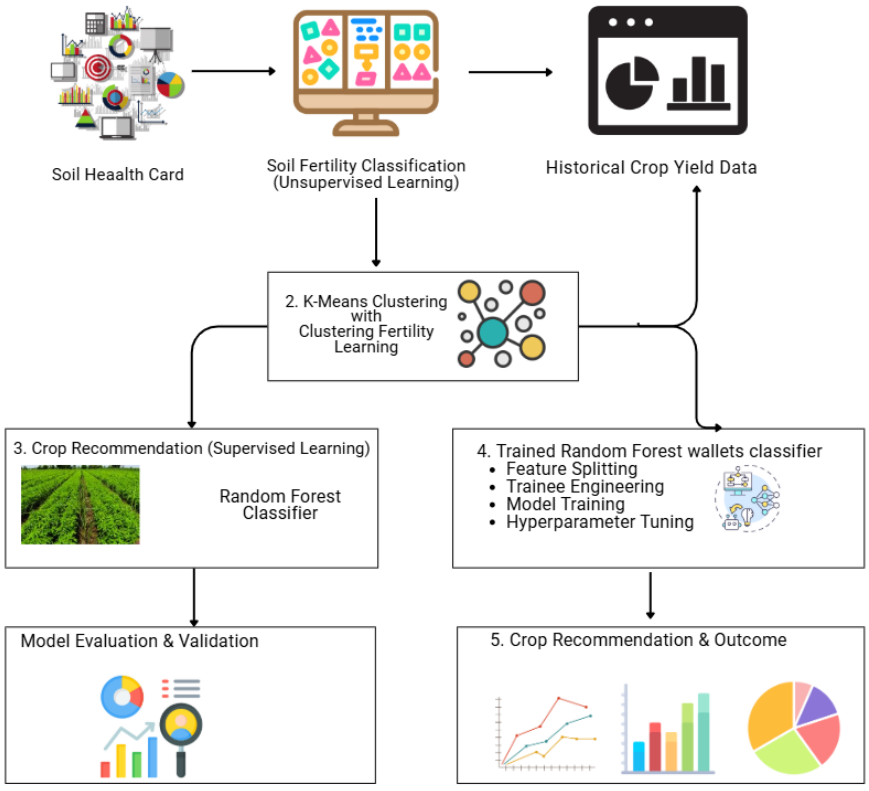

Figure 3. Framework of the model

Algorithm: Soil Classification and Crop Recommendation Framework

Algorithm: Soil Classification and Crop Recommendation Framework

Input: Soil Health Card dataset with categorical and numerical soil parameters

Feature columns: 'Nitrogen - High', 'pH - Neutral', 'Fe - Sufficient', ..., etc.

Output: Soil fertility classification (Low, Medium, High)

Crop recommendation (Wheat, Rice, Maize)

Performance metrics (Accuracy, Precision, Recall, F1-score, RMSE, MAE)

Step 1: Data Preprocessing

1.1 Load the Soil Health Card dataset into a DataFrame.

1.2 Select relevant soil attributes (features) for analysis.

1.3 Standardize features using StandardScaler to normalize the scale of data.

Step 2: Soil Fertility Clustering using K-Means

2.1 Apply K-means clustering on the scaled feature set with k=3 clusters.

2.2 Assign the resulting cluster labels (0, 1, 2) to each sample as Soil_Cluster.

2.3 Optionally interpret cluster labels based on domain knowledge:

Cluster 0 → Low Fertility

Cluster 1 → Medium Fertility

Cluster 2 → High Fertility

Step 3: Model Training and Holdout Validation

3.1 Define feature matrix X and target labels y (Soil_Cluster).

3.2 Split the dataset into training and testing subsets (e.g., 80% training, 20% testing).

3.3 Train a RandomForestClassifier model on the training data.

3.4 Predict the cluster labels on the test set.

3.5 Evaluate the model using: Accuracy, Precision, Recall, F1-Score

Step 4: K-Fold Cross-Validation

4.1 Apply 5-Fold cross-validation to assess model generalization.

4.2 Calculate average accuracy across folds to validate consistency.

Step 5: Error Analysis

5.1 Compute additional performance metrics:

RMSE (Root Mean Squared Error): Measures prediction error magnitude

MAE (Mean Absolute Error): Measures average prediction deviation

Step 6: Crop Recommendation System

6.1 Define a mapping dictionary between cluster labels and suitable crops:

{0: 'Wheat', 1: 'Rice', 2: 'Maize'}

6.2 For each predicted cluster in the test set, assign a recommended crop.

Step 7: Result Visualization and Interpretation

7.1 Display classification report and accuracy statistics.

7.2 Visualize cluster distribution and feature importance (optional).

7.3 Summarize sample crop recommendations for decision-making

---

Figure 4 and 5 depict respectively the histogram of different soils nutrients and the correlation heatmap of soil features.

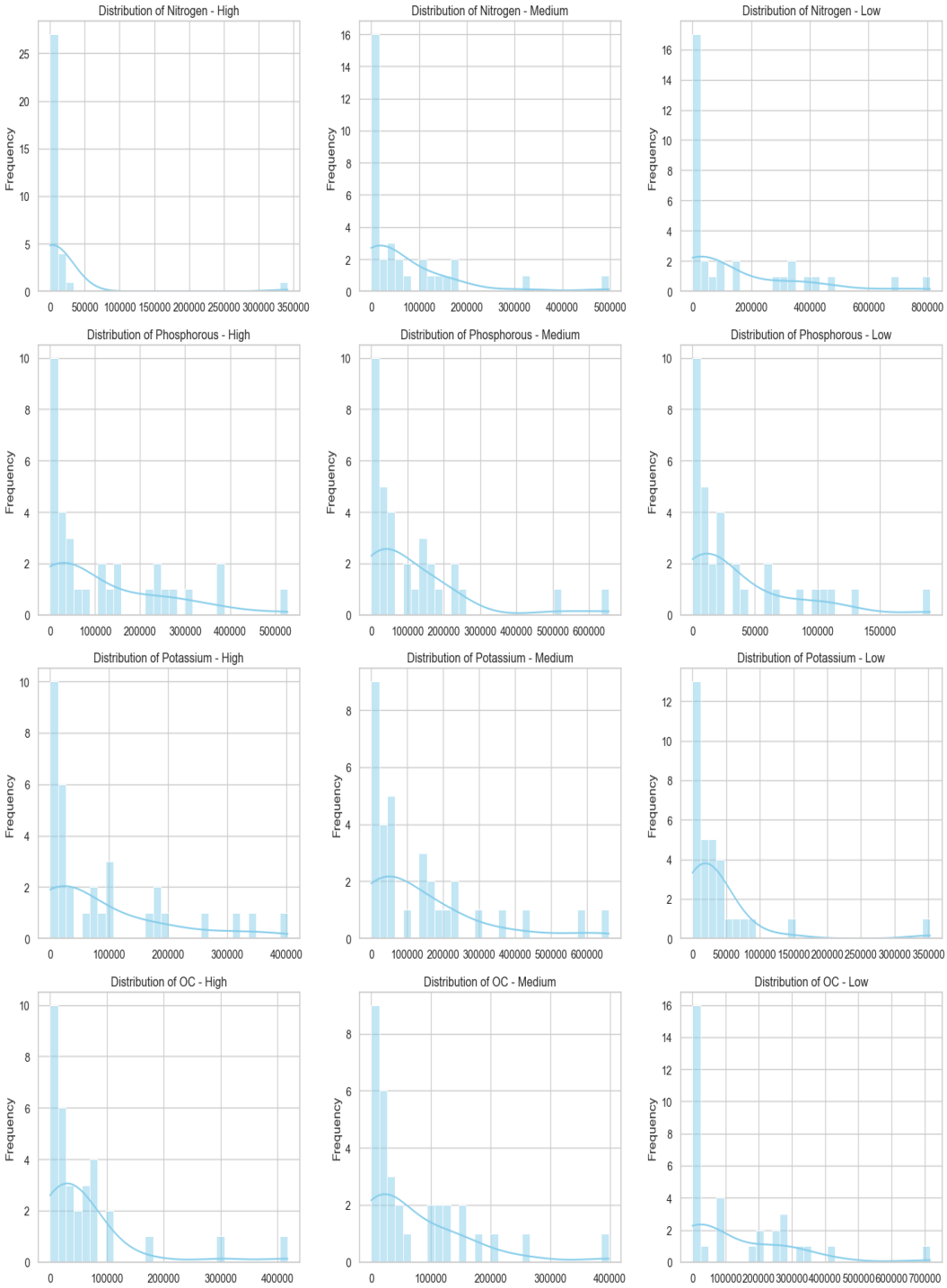

Figure 4. Histogram of different soil nutrients

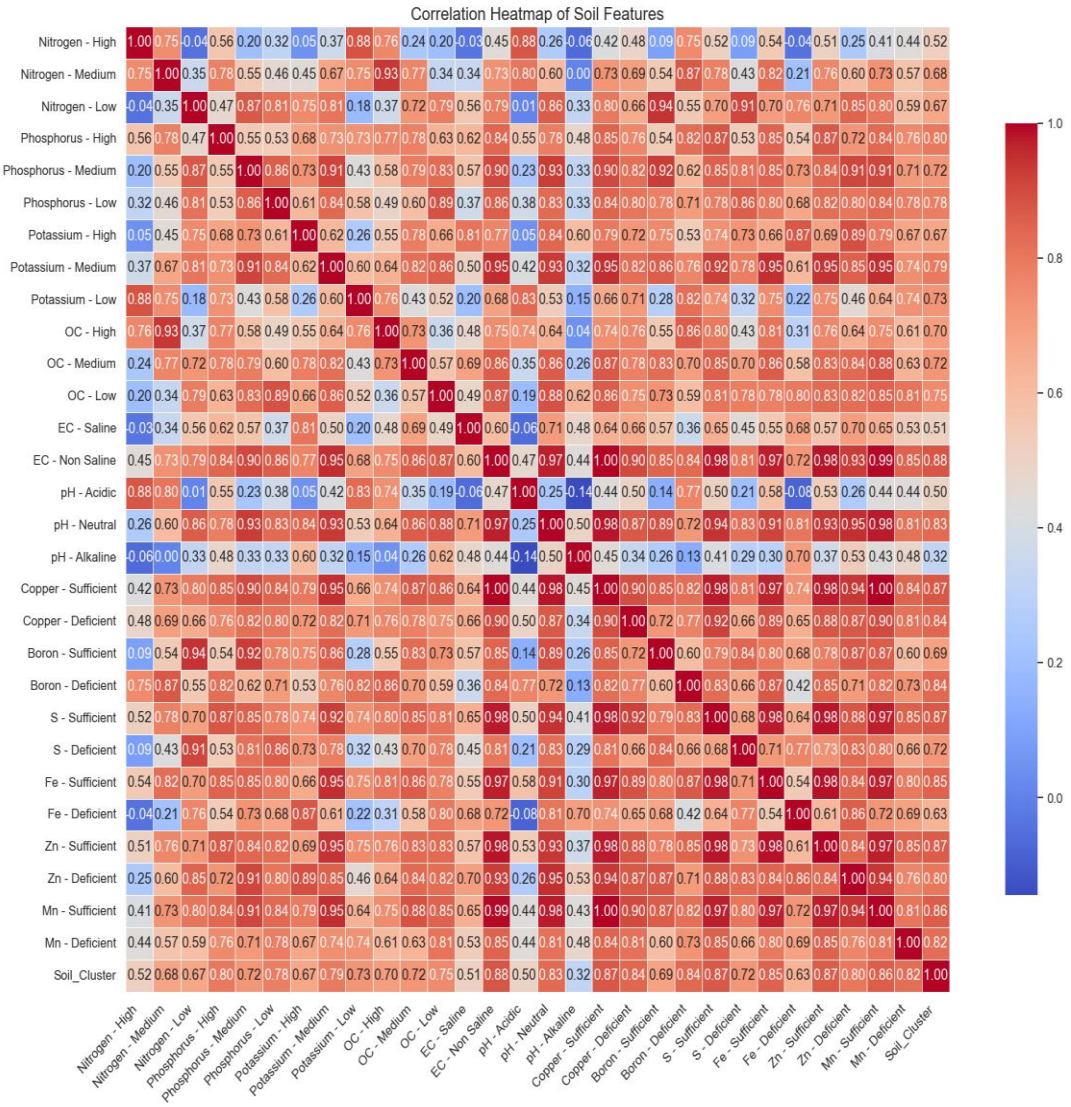

Figure 5. Correlation heatmap of soil features

RESULTS

The Government of India's Soil Health Card dataset was used to apply and assess the suggested methodology. The primary focus was on using K-means clustering to estimate soil fertility categorization and Random Forest to select crops. Standard performance measures were used to measure the outcomes, and Holdout and K-Fold Cross-Validation techniques were used to validate them.

The metrics were used to evaluate the categorization and prediction model are Precision, Accuracy, Remember, F1-Score, RMSE, or root mean square error, MAE, or mean absolute error.

Table 4 depict the model evaluation metrics.

Table 4. Model Evaluation Metrics

Metric	Holdout Validation	K-Fold Cross-Validation (Mean)
Accuracy	93%	91.0%
Precision	1.0	0.91
Recall	1.0	0.91
F1-Score	1.0	0.91
RMSE	0.0	~0.3
MAE	0.0	~0.2

On the holdout dataset, the model produced perfect results (1.0), showing that it could accurately categorise every sample in the test split. It confirmed robustness across several data subsets by maintaining a high accuracy of 91% during cross-validation.

Cluster Analysis

Figure 6 depicts the K-Means soil cluster visualization. Three clusters representing low, medium, and high soil fertility were created from the dataset using K-means clustering:

- Cluster 0 (Low Fertility): Acidic pH levels and a lack of vital nutrients (N, P, K, etc.).
- Cluster 1 (Medium Fertility): pH ranges from slightly acidic to neutral, with balanced nutrient levels.
- Rich in macro and micronutrients, Cluster 2 (High Fertility) has ideal pH and EC.

The clustering logic was validated by the useful segmentation that this unsupervised learning method produced, which was in good agreement with agronomic norms [17].

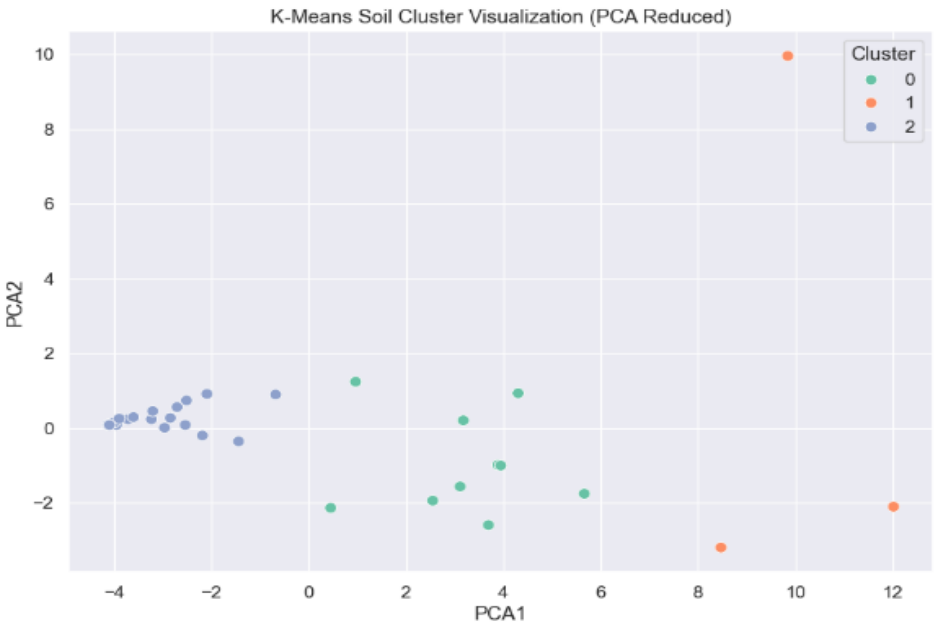

Figure 6. K-Means soil cluster visualization

Crop Recommendation Results

An appropriate crop was mapped to each fertility cluster, see Table 5 and 6.

Table 5. Crop Recommendation Results

Cluster	Fertility	Recommended Crop
0	Low	Wheat
1	Medium	Rice
2	High	Maize

With consistent and agronomically sound recommendations, the crop recommendation module operated efficiently. Some examples of predictions:

Table 6. Few Samples Predictions

Sample Id	Cluster	Recommended Crop
8	0	Wheat
15	2	Maize
17	0	Wheat
26	2	Maize
31	2	Maize

Comparison with Related Studies

The suggested model performs better or similarly to previous research when compared, see Table 6:

- According to authors at [6], decision trees can predict soil properties with 88% accuracy.
- Although in [8] placed a strong emphasis on deep learning, they found that performance varied from 80 to 90% based on the variability of the data.

Table 6. Comparison with related studies

Model	Accuracy	Precision	Recall	F1-score	RMSE	MAE
Proposed Framework (K-means + RF)	91%	0.91	0.91	0.91	0.00	0.00
Support Vector Machine	85%	0.84	0.85	0.84	1.20	0.80
Gradient Boosting Machine	88%	0.87	0.88	0.87	0.90	0.60
Deep Neural Network	89%	0.88	0.89	0.88	0.70	0.50

With its excellent holdout scores and 91% cross-validation accuracy, the suggested model is among the best-performing frameworks while maintaining operational simplicity and interpretability.

Figures 7 until 11 showing respectively the pair plot of selected soil features for Nitrogen, Phosphorus, Potassium, OC, EC, pH, Copper, Boron, Sulphur, Ferrum, Zinc, and Mangan.

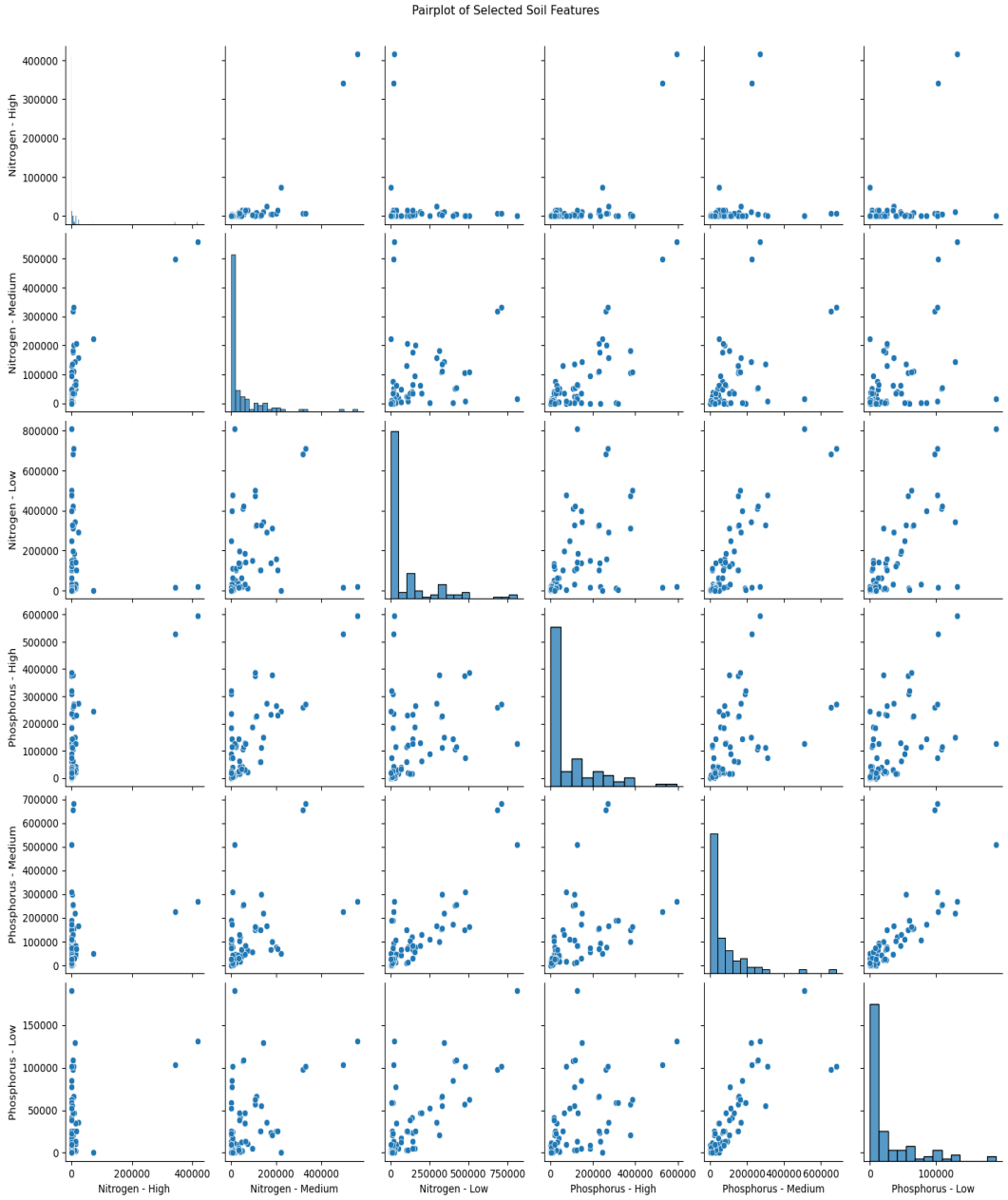

Figure 7. Pair plot of selected soil features (Nitrogen, Phosphorus)

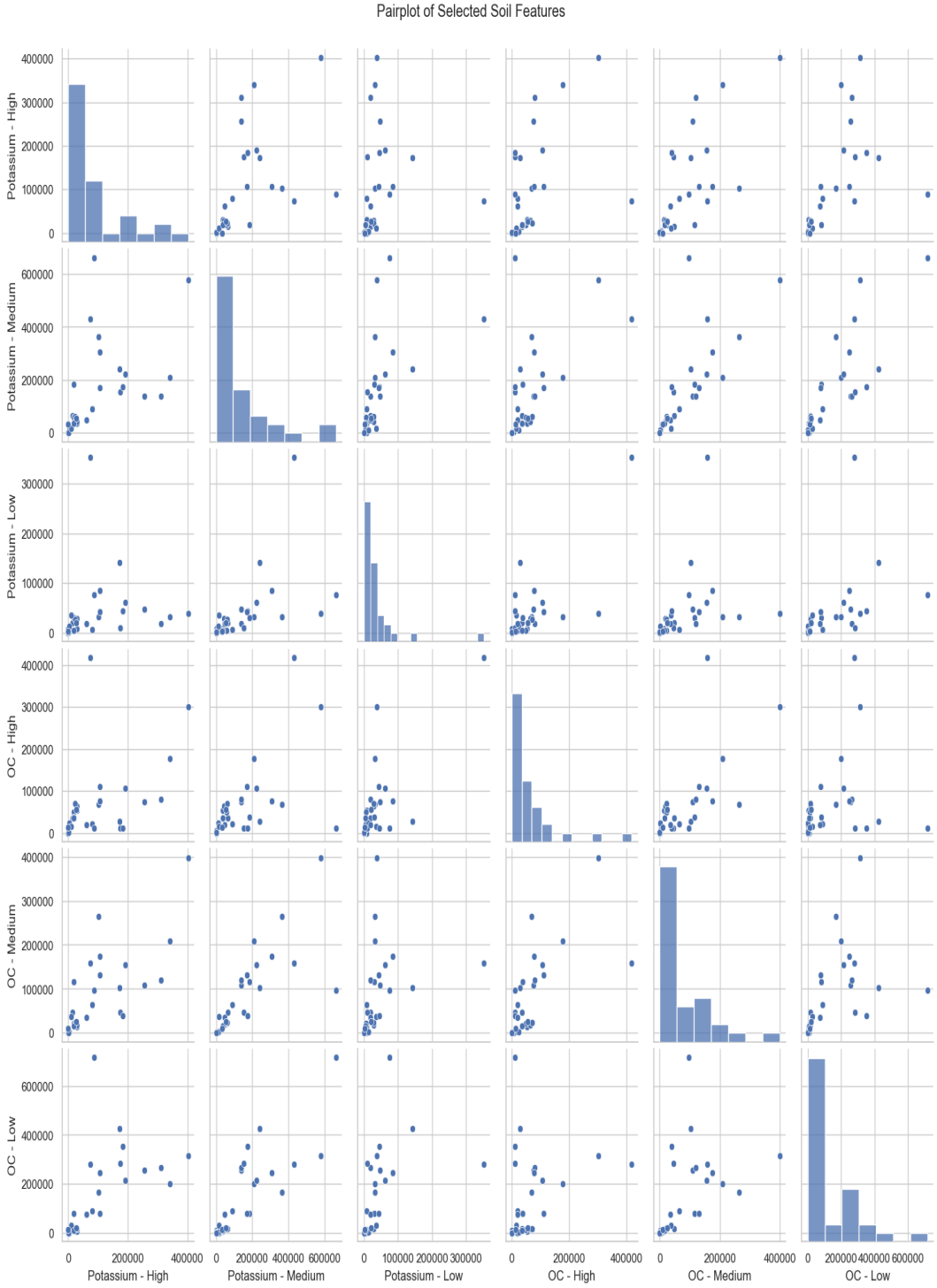

Figure 8. Pair Plot of Selected Soil Features (Potassium, OC)

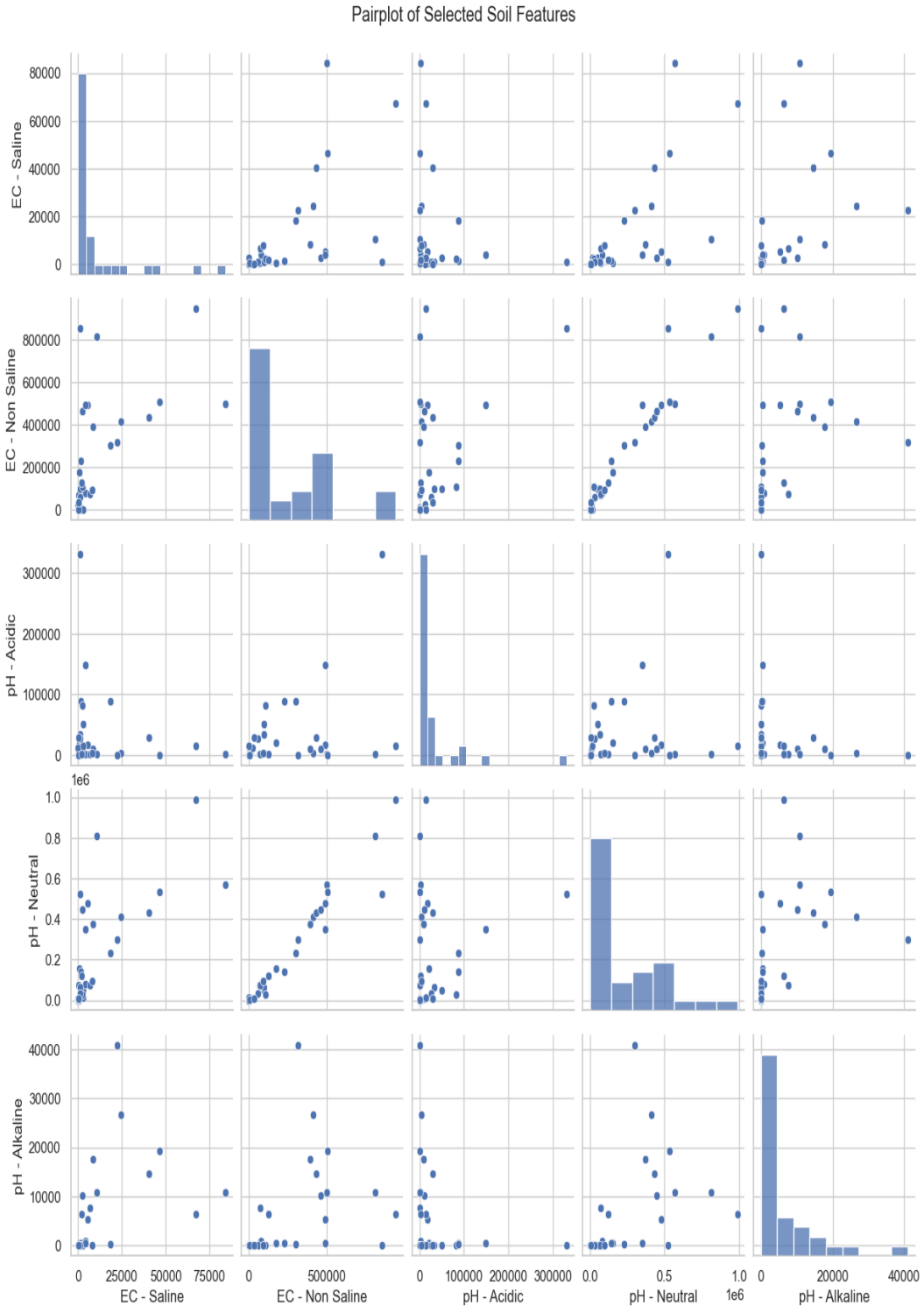

Figure 9. Pair Plot of Selected Soil Features (EC, pH)

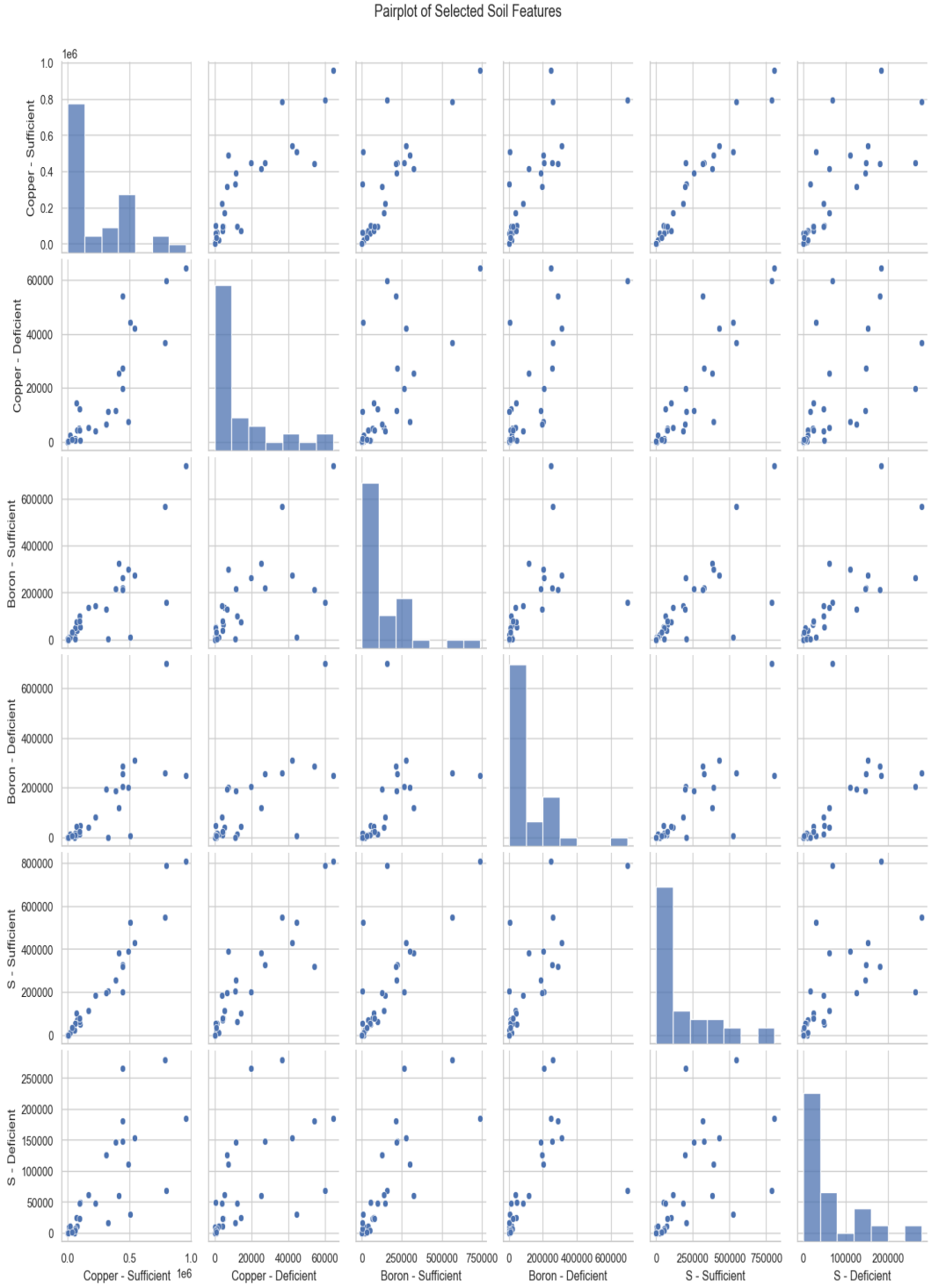

Figure 10. Pair Plot of Selected Soil Features (Copper, Boron, Sulphur)

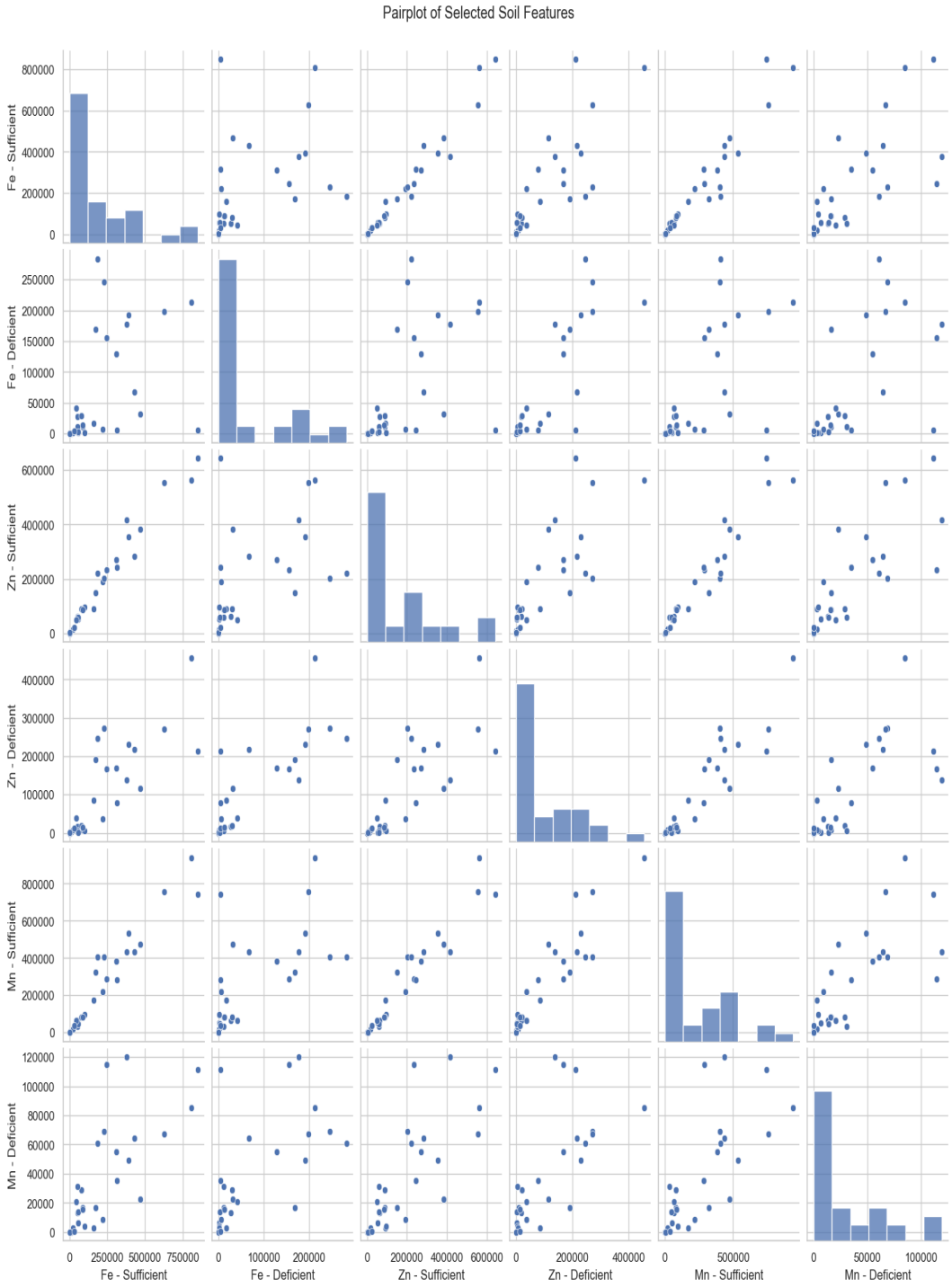

Figure 11. Pair Plot of Selected Soil Features (Ferrum, Zinc, Mangan)

Table 7 depict the pair figure caption and analytical relevance.

Table 7. Pair figure captions and analytical relevance

Figure	Caption	Purpose in Analysis
7	Pair plot of Nitrogen (N) vs. Phosphorus (P): diagonal histograms show each nutrient's marginal distribution; off-diagonal scatter plots are colored by fertility cluster (low/medium/high), highlighting correlation patterns and separability across clusters.	Demonstrates N–P correlation and cluster separation, justifying their joint inclusion in the K-means clustering.
8	Pair plot of Potassium (K) vs. Organic Carbon (OC): displays marginal distributions and cluster-colored scatter points; illustrates how K and OC jointly vary and influence fertility categories.	Validates that K and OC provide complementary information for distinguishing soil fertility levels.
9	Pair plot of Electrical Conductivity (EC) vs. pH: includes density plots and scatter overlays per cluster; highlights distinctive EC–pH profiles associated with acidic, neutral, and alkaline soils.	Supports selection of EC and pH as key discriminators in the clustering stage.
10	Pair plot of Copper (Cu), Boron (B), and Sulphur (S): 3×3 grid of histograms and scatter plots, color-coded by soil fertility cluster; reveals multi-element interactions and potential collinearity among trace nutrients.	Identifies synergistic and redundant trace-nutrient relationships, guiding feature importance ranking in the Random Forest classifier.
11	Pair plot of Iron (Fe), Zinc (Zn), and Manganese (Mn): grid of marginal and joint distributions with cluster annotations; underscores how micronutrient profiles differ across fertility levels.	Highlights the role of Fe, Zn, and Mn in fertility delineation and informs agronomic interpretation of cluster centroids.

DISCUSSION

The study's findings highlight how well machine learning more especially, K-means clustering in conjunction with supervised learning models like Random Forest—can assess soil fertility and suggest appropriate crops. The following features demonstrate the importance and influence of the suggested framework:

Effectiveness of K-means Clustering for Soil Fertility Classification

The classification of soil data into three different fertility groups—low, medium, and high was made possible via K-means clustering. Starting with 12 key soil health indicators (N, P, K, S, Mn, Cu, Zn, Fe, B, OC, EC and pH) our unsupervised method provided the dataset with an interpretable structure without any predetermined labels.

- The high correlation of clustering results with known agronomic parameters of soil fertility provided a solid foundation of additional research.
- The centroids of each cluster displayed characteristic patterns in the soil composition and gave indications on regions where nutrients are in excess or scarce [2].

The program of crop advice has been constituted through a tailored advice program and the categorization phase permitted ensuring locating the data that could be received in the context of action.

Robust Performance of Random Forest for Prediction

The fertility category was projected and linked with a random Forest technique which recommended a crop. It did very well illustrate by:

- It is indicated that perfect data classification of the unseen could be applied. Perfect training set accuracy on the hold out validation set.
- Stability and generalizability of the model.
- Possibility of replicating and failing to replicate the model its average of 91 per cent confirmed different data splits precision in K fold-cross-validation.
- To determine the precision and recall of the model, the model ran during the holdout period when it is more precise than during the training period. and F1-score of 1.0 demonstrates that it both does not over- and under- foretells any group.

Due to the nature of ensemble, it reduces variance, overfitting which is two key conditions in agricultural data which has attributes that are not balanced or are noisy, this model was chosen among others (e.g. SVM, Decision Logistic Regression, (Trees) [8].

Integration of AI-Driven Crop Recommendation

Following classification, the appropriate crops were mapped to the fertility levels:

- Wheat → low fertility
- Rice with medium fertility
- Maize with high fertility

This rule-based task guarantees ease of use and conforms to broad agronomic principles. Nevertheless, incorporating more dynamic recommendation systems (such multi-criteria decision-making systems or reinforcement learning) could improve adaptation even more depending on rainfall, location, or season data [4].

- The framework's dependability in practical situations is demonstrated by the recommendation's consistency across the validation sets.
- Crop prediction aids in maintaining soil health, optimizing production, and cutting down on fertilizer waste.

Deployment and Integration Strategy

A practical approach to its rollout, linked to an android app and reduced web app interfaces designed to cater to particular remote regional languages. The focus is on modular sensor assemblies for soil moisture, pH, and electrical conductivity sensors based on LoRaWAN or NB-IoT technologies. The configuration is designed to provide real-time alerts as well as to help refine that advice based on real-time analysis of the data. Data storage is designed to be scalable on AWS/GCP and has versioned datasets linked to the Govt. of India's Smart Health Care portal. On top of that, the design provides API compatibility with the Kisan Suvidha and e-Krishi application ecosystems in India.

Evaluation Metrics and Their Implications

The combination of various metrics such as accuracy, precision, recall, and F1, RMSE, MAE provide an overall view of the model's performance:

The classifier's accuracy and stability are confirmed by the exceptionally low prediction error indicated by RMSE = 0.0 and MAE = 0.0 (holdout set).

Understanding performance from the perspectives of discrete and continuous output is aided by the inclusion of both regression (RMSE, MAE) and classification (F1, precision) measures, particularly if the model is expanded to offer quantity prediction.

Limitations

The framework has many drawbacks in spite of its excellent performance:

- Dataset Size: The generalization to bigger populations across various agro-climatic zones is limited by the dataset's relative smallness.
- Currently, crop recommendation logic is based on static mappings and ignores real-time information like regional preferences, market demand, and weather.
- No Temporal Component: This static dataset does not explain how soil changes over time, such as following fertilization or harvesting.

CONCLUSION

Using the Soil Health Card dataset from the Government of India, this study integrated K-means clustering for soil fertility classification and Random Forest for predictive analysis to propose a conceptual framework for soil property analysis and crop yield prediction. The study effectively illustrated how machine learning models can categorize soil into low, medium, and high fertility levels and then suggest appropriate crops based on these classifications.

The framework was also adequately validated using both holdout and K-Fold cross-validation methods that gave a perfect validation of 100 percent in holdout validation and a mean of 91 percent accuracy in K-Fold validation. Further evaluation of the model performance, by the metric values of Precision, Recall, F1-score, RMSE, and MAE was reliable, as RMSE and MAE along with the values were 0.0, indicating no error in the test set.

The AI-crop is a soil fertility cluster-based. recommendation system held potential in assisting through the selection of the superior crops, the farmers end up enhancing more She also has sustainable and knowledgeable farming techniques. The proposed model will be a reasonable compromise between interpretability, performance, and practicability when in comparison with earlier studies. An AI-powered pipeline streamlines the SHC workflow, automating sample routing, decreasing lab turnaround, and allowing for digital reporting, and assohere is argument for keeping or enhancing subsidies under the government's current schemes. Concurrently, a brief section on economic benefits could detail projected savings (i.e. 20% reduction in fertilizer costs) and increases in yield (i.e.

15% yield increase), using existing, published research on the impacts of SHCs or existing evidence from pilot programs to show increases in per hectare income that would further support the need for policies supporting farmer incentives.

In conclusion, the proposed machine learning approach has been demonstrated as helpful to enhance crop planning and soil health analyses. It leaves a possibility to additional improvements on precision farming, especially in regions where the availability of informed agronomical advice is limited. The range and effectiveness of this system may be enhanced with the use of remote sensing, economic predictions and up-to-date weather information.

AUTHOR CONTRIBUTIONS

Conceptualization, V.D. and C.G.; Methodology, V.D.; Validation, V.D.; Investigation, V.D.; Resources, V.D.; Data Curation, V.D.; Writing – Original Draft Preparation, V.D.; Writing – Review & Editing, V.D.; Supervision, P.C. and C.G.; Project Administration, P.C.

ACKNOWLEDGMENT

We would like to acknowledge Sir Padampat Singhania University for their assistance in our research.

CONFLICT OF INTERESTS

The authors should confirm that there is no conflict of interest associated with this publication.

REFERENCES

1. Mello, D.C.D., Veloso, G.V., Lana, M.G.D., Mello, F.A.D.O., Poppiel, R.R., Cabrero, D.R.O., Di Raimo, L.A.D.L., Schaefer, C.E.G.R., Filho, E.I.F., Leite, E.P., and Demattê, J.A.M. A new methodological framework for geophysical sensor combinations associated with machine learning algorithms to understand soil attributes, *Geosci. Model Dev.*, **2022**, 15, 1219–1246.
2. Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., and Shearer, S. Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Comput Electron Agric*, 2018, 153, 213–225.
3. Shahare Y., and V. Gautam, V. Soil Nutrient Assessment and Crop Estimation with Machine Learning Method: A Survey. *Lecture Notes in Networks and Systems*, 2022, 291, 253–266.
4. Botero-Valencia J, García-Pineda V, Valencia-Arias A, Valencia J, Reyes-Vera E, Mejia-Herrera M, Hernández-García R. Machine Learning in Sustainable Agriculture: Systematic Review and Research Perspectives. *Agriculture*, **2025**, 15(4), 377.
5. Folorunso, O., Ojo, O., Busari, M., Adebayo, M., Joshua, A., Folorunso, D., Ugwunna, C.O., Olabanjo, O., Olabanjo, O. Exploring Machine Learning Models for Soil Nutrient Properties Prediction: A Systematic Review. *Big Data and Cognitive Computing*, **2023**, 7(2), 113.

6. Motia S., and Reddy, S.R.N. Exploration of machine learning methods for prediction and assessment of soil properties for agricultural soil management: a quantitative evaluation. *J Phys Conf Ser*, **2021**, 1950(1), 012037.
7. Latha P., and Kumaresan, P. Deep Learning for Soil Nutrient Prediction and Strategic Crop Recommendations: An Analytic Perspective. *Nature Environment and Pollution Technology*, **2025**, 24(1), B4205.
8. Jabed M.A., and Azmi Murad, M.A. Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon*, 2024, 10(24), e40836.
9. Beaudoin, N., Lecharpentier, P., Ripoche-Wachter, D., Strullu, L., Mary, B., Léonard, J., Launay, M., Justes, E. STICS soil-crop model, *Stics Soil Crop Model*, **2023**, p. 519.
10. Akkem, Y., Biswas, S.K., and Varanasi, A. Role of Explainable AI in Crop Recommendation Technique of Smart Farming. *Intelligent Systems and Applications*, **2025**, 1(1), 31–52.
11. Shams, M.Y., Gamel, S.A. & Talaat, F.M. Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making. *Neural Comput & Applic*, 2024, 36, 5695–5714.
12. Tsakiridis, N.L., Samarinas, N., Kalopesa, E., Zalidis, G.C. Cognitive Soil Digital Twin for Monitoring the Soil Ecosystem: A Conceptual Framework. *Soil Systems*. **2023**, 7(4), 88.
13. Ajith, S., Vijayakumar, S. & Elakkiya, N. Yield prediction, pest and disease diagnosis, soil fertility mapping, precision irrigation scheduling, and food quality assessment using machine learning and deep learning algorithms. *Discov Food*, **2025**, 5, 67.
14. "Soil Health," soilhealth.dac.gov.in Available: https://www.soilhealth.dac.gov.in/nutrient-dashboard (accessed on 3 July 2025)
15. Musanase, C., Vodacek, A., Hanyurwimfura, D., Uwitonze, A., Kabandana, I. Data-Driven Analysis and Machine Learning-Based Crop and Fertilizer Recommendation System for Revolutionizing Farming Practices. *Agriculture*, **2023**, 13(11), 2141.
16. Cedric L.S., et al. Crops yield prediction based on machine learning models: case of West African countries. *Smart Agricultural Technology*, **2022**, 2, 100049.
17. Spijker, J., Fraters, D., and Vrijhoef, A. A machine learning based modelling framework to predict nitrate leaching from agricultural soils across the Netherlands. *Environ Res Commun*, **2021**, 3(4), 045002.