International Research Journal of Multidisciplinary Scope (IRJMS), 2026;7(3):423-441

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2026.v07i03.011474

Data-driven Approaches in Soil and Crop Prediction: A Conceptual Framework Using Machine Learning

Vimla Dangi*, Kamal Kant Hiran, Manish Tiwari

Faculty of Computing and Informatics, Sir Padampat Singhania University (SPSU), Udaipur-Chittorgarh, Rajasthan, India. *Corresponding Author's Email: vimla.phd2023@spsu.ac.in

Abstract

This study presents a data-driven and explainable hybrid machine learning framework for soil fertility assessment and crop recommendation using the rds_crop_soil_dataset_2020_2025, a structured dataset aligned with India's Soil Health Card (SHC) scheme. The dataset comprises 1500 samples and 19 agronomic and climatic features, including macronutrients (N, P, K), micronutrients (Zn, Fe, Cu, Mn, B, S), and soil properties (pH, EC, OC), along with environmental variables such as temperature and rainfall. The proposed approach integrates unsupervised K-Means for soil fertility zoning with a soft-voting ensemble model combining Random Forest and XGBoost for crop prediction. Model interpretability is ensured using SHAP-based feature importance and dependency analysis, enabling transparent decision-making aligned with agronomic knowledge. The hybrid model achieves 97% accuracy, outperforming individual baseline models across precision, recall, F1-score, and ROC-AUC metrics. Visual analyses, including pair plots, correlation heatmaps, and cluster projections, validate the relationships among soil parameters and model robustness. Although the dataset is synthetically structured, it closely reflects real-world SHC distributions, making the framework suitable for precision agriculture applications, policy planning, and intelligent decision support systems. This study contributes toward building scalable, interpretable, and region-aware AI solutions for sustainable agriculture in India.

Keywords: Crop Prediction; PCA; Precision Agriculture; Soil Fertility Classification; Soil Health Card (SHC); t-distributed Stochastic Neighbor Embedding (t-SNE).

Introduction

The agricultural sector continues to be the backbone of the economies of developing nations, particularly the state of India, where the preservation of agricultural output plays a crucial role in sustaining the social and economic development (1,2). The many national programs have been effectively utilizing the information collected from the large number of established repositories, encompassing the soil profile of various macronutrient and micronutrient elements (3–5). Therefore, the farmers are unable to effectively allocate resources or receive data-driven information to optimize crop yield growth while maintaining SHC (6, 7). The introduction of artificial intelligence (AI) and machine learning (ML) in the past few years has been rapidly applied as key enablers of innovative precision agricultural solutions (8). Applications of techniques from traditional ensemble models to more complex models and techniques that utilize advanced learning architectures are being successfully deployed to address challenges and obstacles that confront agronomists (9). In addition, hybrid and ensemble learning methods have been shown to perform significantly better than individual models at classifying data as they can combine the strengths of the various algorithms used to predict the outcome (10–13).

A significant gap in research is the reliance on non-transparent models that lack XAI, leaving agronomists and end-users unable to understand how the models' logic generates crop recommendations (11, 14). The lack of a hybrid model integrating unsupervised spatial clustering followed by supervised classification is absent within the data. To address and bridge these gaps, a new explainable hybrid ML framework is proposed to predict combined soil fertility assessments and multiple crop recommendations. This framework establishes an architecture that combines unsupervised K-Means clustering with a supervised Random Forest and XGBoost algorithms. With a large dataset with complete soil and climate data, the proposed framework can predict which of the three recommended crops (Wheat, Rice, and Maize) would grow best in

(Received 28th March 2026; Accepted 02nd June 2026; Published 08th July 2026)

their respective regions. The system achieves that by using SHAP to create precise feature attributions, resulting in an algorithmic interpretation. To view the many variables affecting the predictive outcome of each region, it was necessary to employ both PCA and t-SNE to visualize the high-dimensional datasets. The outcomes indicate that it provides a scalable and interpretable AI solution that can facilitate the development of sustainable and climate-resilient agriculture policies with improved decision-making abilities based on accurate recommendations.

The research objectives of this study are:

(a) Build a hybrid machine learning architecture that combinesboth an unsupervised K-Means spatial clustering step with a supervised soft-voting ensemble to optimize crop recommendations.

(b) Compare the proposed hybrid framework against stand-alone baseline models using structured SHC parameters and evaluate the predictive validity of predictions produced by the hybrid framework.

(c) Validate the underlying soil nutrient distributions and model strength through high-dimensional visualizations.

The goal of the research is to answer the following questions:

RQ1: How does integrating unsupervised (via K-Means clustering) with a supervised approach improve the multi-crop classification accuracy of a hybrid machine-learning?

RQ2: What specific soil and land-based variables will be recommended of Wheat, Rice, and Maize, according to the SHAP analysis method of interpretability?

RQ3: To what extent will the use of district-provided Soil Health Cards eliminate the geographic limitations of precision agriculture systems?

An AI-driven predictive modeling framework aimed at estimating crop yields under varying climatic conditions to address vulnerabilities in global food security. The empirical results demonstrated that the Gradient Boosting model significantly outperformed the other approaches, successfully capturing the complex, nonlinear relationships between climate extremes and agricultural output. However, the authors noted a relatively low explanatory variance (R^2) across all models, identifying a substantial research gap: the models lack crucial unmeasured factors such as soil characteristics, crop phenology, and farm management practices (15). Various data, such as Nitrogen (N), Phosphorus (P), Potassium (K), pH level, temperature, humidity, and rainfall, are utilized by soil and environmental factors along with ML models to predict suitable crops (16). The inclusion of site-specific environmental variables results in more relevant and practical crop recommendations. To the best of our knowledge, the proposed paper provides a sound base for merging domain-specific knowledge with state-of-the-art ML techniques in the agriculture sector and confirms our assumption that soil health data can greatly improve crop production while decisions are made on precision agriculture systems (17). An extensive systematic literature review of the machine learning techniques applied to enhance agricultural productivity and, more particularly, for crop yield prediction. Research in turn follows PRISMA guidelines which reduces the studies from 776 to the 25 most relevant studies (18). Decision Trees, Random Forest, Support Vector Machines (SVM), K-nearest Neighbors, and Gradient Boosting are among the most used algorithms (19). A smart farming solution based on AI and IoT technologies for crop recommendations and nutrient management was performed by scanning 93 scientific publications (20). The research classified the existing methods into soil nutrient prediction, environmental condition prediction and analyzing the satellite imagery for the prediction of crop yield. It covers different machine learning models including Logistic Regression, Decision Trees, Random Forest, XGBoost and deep learning methodologies like CNN and LSTM networks (21–23). A synopsis of the other related studies is shown in Table 1. The research focuses on cotton, maize, and wheat crops and uses convolutional neural networks (CNNs), bi-directional long short-term memory (Bi-LSTM) networks, and ensemble regression models. The work motivates an openness to human-interpretable and scalable deep learning models. The previous research has confirmed the effectiveness of various ML models for predicting and recommending crop yields.

Additionally, few studies have effectively joined government-issued soil data with hybrid classification and ensemble methodologies (24, 25).

Table 1: Comparison with Related Studies

Key Focus	Techniques Used	Data Type	Contribution	Crop Type	Soil Name	Country	Dataset	Limitations or Research Gaps	References
Crop recommendation using climate + soil data	Random Forest (RF), Gradient Boosting, Logistic Regression + Bagging, SVM + Bagging; SMOTE	Climate (daily), soil types, productivity	Large-scale real data models; localized vs. general performance	95 crop types	32 soil types	Brazil (3961 cities)	Brazilian Ministry of Agriculture + INMET	Lower F1 in generalized models; class imbalance; generalization challenges	(26)
Optimizing irrigation and fertilization scheduling in arid sandy soils to maximize resource efficiency.	Random Forest, LSTM, Rule-based expert system (AI-DSS).	Real-time sensor data (soil moisture, nutrients, weather) and historical crop performance records.	Improved crop yields up to 13.1%, WUE by 15.5%, and reduced nitrate leaching by ~45% compared to conventional practices.	Wheat, Maize, Sugar Beet.	Sandy soils (Coarse-textured).	Egypt	10 years of historical soil, crop, and climate data + real-time sensor data.	High initial sensor/infrastructure costs; requires reliable rural connectivity; does not currently integrate remote sensing data or pest/disease forecasting.	(27)
Reviewing the adoption of precision farming technologies, data governance (Code of Conduct), and farm data spaces for agriculture.	GNSS, FMIS, optical/EC sensors, satellite imagery (Landsat/Copernicus), AI, PWM-nozzles, robotics, DSS.	Geospatial field data, satellite imagery (multispectral), in-situ ground data, and machine operational data.	Highlighted the necessity of secure data governance (CoC) and federated platforms (Farmmaps) to overcome interoperability barriers and drive sustainable crop management.	Potato (primary), Onion, Sugar Beet.	Not specified.	The Netherlands	Survey/field lab data from the Netherlands precision farming fieldlab (NPPL) comprising >30 farms.	Slow technology adoption due to uncertain ROI, high complexity, and poor data interoperability; yield mapping accuracy for tuber crops remains limited.	(28)
IoT-driven real-time data collection combined with machine learning for accurate crop recommendations.	IoT (LoRa gateway, sensors), XGBoost, Cloud computing.	Real-time sensor data (weather, soil parameters).	Achieved 99% accuracy with an XGBoost-IoT integrated system that updates soil and weather conditions	15+ crops including Cotton, Maize, Apple, Coffee, Mango.	Sandy loam, Silty loam, Clay loam.	India	Kaggle Crop Recommendation Dataset + Real-time simulated sensor data.	Simulated with laboratory sensors instead of actual field deployment; ignores pest infestations, crop rotation, and irrigation frequency.	(29)
Multi-year crop rotation forecasting and proactive soil health management.	ConvLSTM (Conv1D + LSTM), GRU, LSTM, Random Forest, XGBoost, SVM, KNN.	Multivariate time-series data (production, yield, cultivated area).	Developed a hybrid ConvLSTM model for crop rotation prediction.	Rice, Wheat, Pearl Millet, Cotton, Maize, Sorghum, Soyabean.	Not explicitly specified (General soil health).	India	ICRISAT District Level Database (1978-2017).	High computational cost, sensitivity to data imbalance, and lacks integration with real-time.	(30)
Systematic review of ML + IoT for soil parameters in smart farming	Supervised (RF, SVM, GBM, CNN, ensembles); IoT sensors	Soil parameters (pH, nutrients, moisture,	PRISMA review of 77 studies; identifies best	Various	Multiple (fertility, pH, quality)	Global	Various public & sensor datasets (e.g., Soil Fertility Dataset)	Fragmented research; need for more real-time integration	(31)

		fertility); IoT sensor data	models & frameworks						
Ensemble crop prediction from soil physico-chemical data	Grasshopper Optimization Algorithm, RF, XGB	pH, electrical conductivity, moisture, nutrient levels	Ensemble model (RF+XGB) guided by GOA outperforms baseline methods	Wheat, Rice, Maize	Varied physico-chemical soil profiles	India	Laboratory soil samples from Maharashtra	Laboratory-scale data; lacks dynamic IoT-sensor integration and large-scale validation	(32)
AI for sustainable agriculture and crop production	ML, DL, remote sensing; IoT; robotics; UAV analytics	Satellite imagery; sensor time-series; weather;	Comprehensive taxonomy of AI methods	Multiple crops (cereals, vegetables, fruits)	Review study	Global	Various public and proprietary datasets cited in reviewed studies	Limited field-scale validation; dataset heterogeneity	(6)
Integrated soil quality classification and crop yield forecasting	Random Forest, XGBoost, SVM/SVR	Custom Soil Quality Index (N, P, K, pH), historical yield records	Unified pipeline for soil grading, crop recommendation,	Multi-crop	Mixed (based on SQI classes)	India	Two Kaggle public datasets (soil parameters + yield)	Relies on publicly available datasets; needs local calibration and farmer-centric UI	(33)
Time-series crop yield forecasting	CNN-LSTM with attention and skip connections	Weather time series, soil moisture, temperature	Achieved RMSE 0.017 and R^2 0.967 for wheat and rice yield forecasts	Wheat, Rice	Not specified	India	State-level historical yield and weather data	Limited to two crops; lacks remote-sensing inputs and multi-crop generalization	(34)

Note: DSS- Decision Support Systems, FMIS- Farm Management Information Systems, GNSS- Global Navigation Satellite Systems, GRU- Gated Recurrent Unit.

The approach not only addresses AI technologies with reliable national data but also highlights the importance of preprocessing and dimensionality reduction to enhance model interpretability. This adds to the growing body of work promoting explainable, scalable and regionally adoptable AI solutions in agriculture, potentially informing precision farming deployment, driving policy making and supporting climate-resilient agricultural practices in India and elsewhere (24, 25). This study fills a critical gap by integrating unsupervised clustering with supervised crop recommendation, offering a scalable machine learning framework that advances soil health assessment and supports sustainable agricultural policy.

Methodology

The dataset is taken from India's Soil Health Card (SHC) scheme, via Cycle I, II and III, i.e., 2015–2017, 2017–2019 and 2020–2025. The data has an extensive geographical coverage, spanning 33 states and Union Territories (UTs) in India. This is a systematic and representative sampling of soil from a 2.5 (hectare irrigated)/10-hectare (rain-fed) grid as shown in Figure 1 for recommended crops across different Indian states (35). The domain comprises a six-year time frame, from 2020 to 2025, with 1500 samples that mimic the structure. One such record is for each soil analysis report for a specific region and it represents India's varied agro-climates like arid areas, subtropical plains and hilly areas.

Features and Variables

Each sample consists of 19 features, divided into the following sub-categories:

(a) Soil Nutrient Content:

Macronutrients: N (kg/ha), P (kg/ha), K (kg/ha);

Micronutrients: S, Mn, Cu, Zn, Fe;

Soil Chemistry: Organic Carbon (OC), pH, Electrical Conductivity (EC);

(b) Climatic Conditions:

Rainfall_mm: Annual average rainfall (500–2000 mm);

Temperature_C: Mean temperature during the crop cycle (10–35°C),

(c) Agricultural Outputs:

Yield_pred_ton_ha: Estimated yield potential in tons per hectare;

(d) Target Variables:

Fertility_Class: Categorical label (Low, Medium, High) derived from combined soil properties;

Recommended_Crop: AI-driven crop recommendation among 3 major Indian crops (Wheat, Rice, Maize).

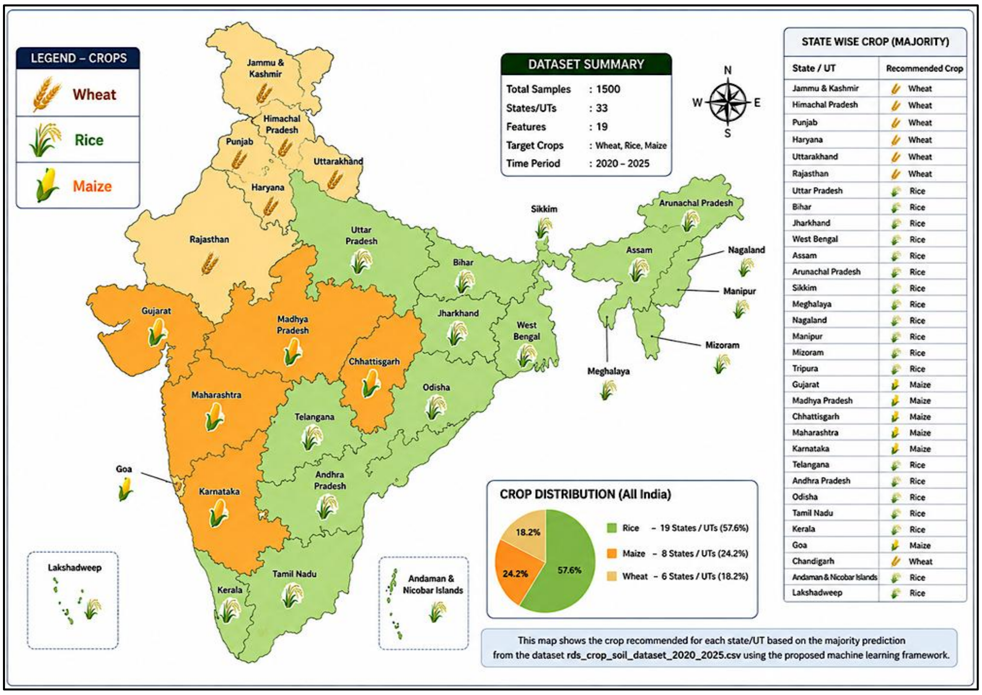

Figure 1: Recommended Crops Across Different Indian States

Data Preparation and Labeling

Soil Fertility Classification: The Soil Fertility Classification was based on the cutoff values detailed according to Indian agronomy standards, which separate the soil into Low, Medium and High fertility classes based on nutrient content. Crop Recommendation: A knowledge model with a logic that matches known crops suitability based on nutrient profile and environment, and which can perform both classification (supervised learning) and clustering (unsupervised learning) (36, 37). All input data is normalized and preprocessed to allow successful use of ML algorithms, and feature distributions also follow realistic shapes observed in government databases.

Quality and Reliability

The dataset maintains:

(a) There are no missing values or null entries.

(b) Reasonable representation of all target classes (to prevent the model from being biased).

(c) A realistic distribution of climatic and soil conditions in models like observed Indian regions (38).

(d) Potential to be a benchmark database for intelligent agronomy applications.

This data set is developed to meet the following research requirements: Comparison of classification models for predicting soil fertility and crop type; Comparing the performance of ensemble models – XGBoost + Random Forest; Applying K-means clustering for unsupervised grouping of soil types; Use of cross-validation (Holdout, K-Fold, Stratification) for reliable model performance assessment. This data-carving approach is largely consistent with national initiatives in India, such as the Digital India initiative, Soil Health Card Mission and Precision Agriculture, and promotes sustainable agriculture and improved decision-making among Indian farmers (39, 40).

Proposed Methodology

To see farmers' soil fertility conditions and predict their best crops, the research method applied in this study tries to take advantage of supervised and unsupervised machine learning. This method consists of a series of structured methods including dataset construction, feature engineering and preprocessing, unsupervised clustering, hybrid model construction, and interpretation through exploratory visualizations for the sake of interpretability to make it easy to modulate and understand the whole process.

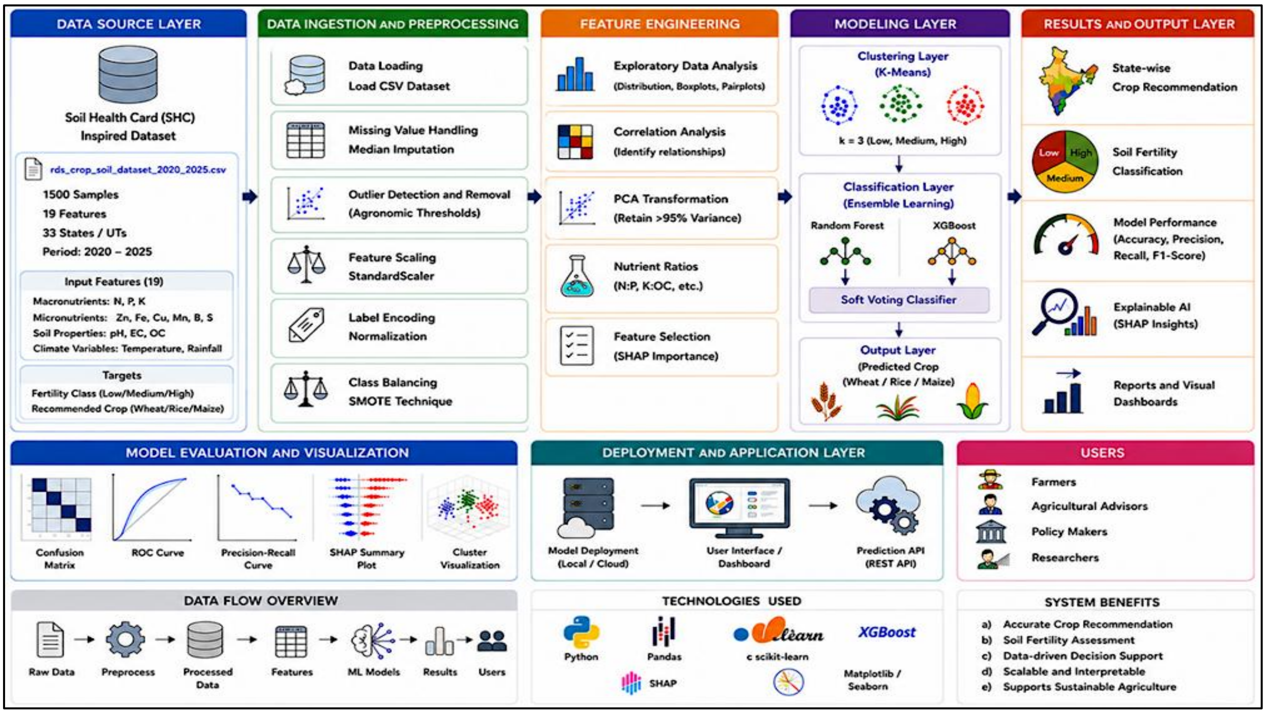

Figure 2: Flowchart of Proposed Methodology

Figure 2 outlines a machine learning workflow designed for crop recommendation and soil fertility classification. The data were generated from cycle I (2015–2017), cycle II (2017–2019) and Cycle III (2020–2025) and pilot projects as part of the Indian Government sponsored soil-testing program Soil Health Card. It includes major components such as nitrate (NO^{3-}), phosphorus (PO_4^{3-}) and potassium (K) and minor components such as pH, electrical conductivity (EC) and organic carbon (OC). It also contains micronutrients, which are divided into sufficient and deficient levels, including zinc (Zn), iron (Fe), manganese (Mn), copper (Cu), boron (B), and sulfur (S). The original dataset had to do with categorical columns that had to be parsed, multiple duplicate origination information columns, and percentage data fields in string form. This multistage cleaning, imputation, and aggregation has converted the data into a machine learning friendly data set consisting of 61 features for each of 33 state level records.

A few key preprocessing steps were taken to enhance the quality of the data and simplify model confirmation. Percentages were cast as floats and state name columns that are extraneous were added. Missing values, particularly in micronutrient data, were treated with forward fill and median imputation. Categorical variables for different crops (Cotton, Paddy, Wheat) were encoded using the label encoding technique and the one-hot encoding technique for ROC-AUC analysis, while continuous variables were normalized for analysis using StandardScaler.

The orange points indicate noisy data or outliers that were removed during the preprocessing stage. A crucial part of the preprocessing pipeline was the data cleaning and noise/outliers' treatment. Real agricultural data such as the SHC dataset may have results with incorrect entries due to measurement errors or inputting errors. To tune and strengthen our model we used the method of outlier detection, for which we compared the distribution of agronomic parameters of our features to known agronomic thresholds. This process is illustrated in Figure 3 which gives us scatter plots of unordered data for the main ingredients: Nitrogen (Figure 3A), Phosphorus (Figure 3B), Potassium (Figure 3C) and Rainfall (Figure 3D) with predicted yield. Orange points shown are noisy data or outlier points, which were detected and removed from the dataset. This cleaning process is required to avoid the model from learning from corrupted data which would increase its accuracy and its capacity to generalize to new real-world examples.

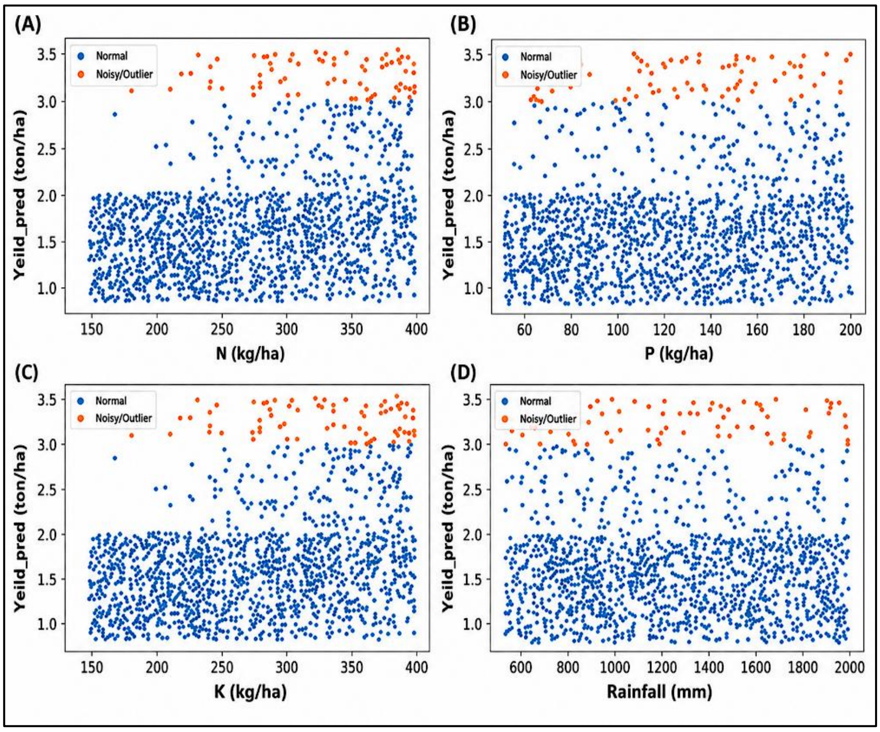

Figure 3: Visualization of Outlier Detection in Key Features. (A) Nitrogen, (B) Phosphorus, (C) Potassium, (D) Rainfall

We hand-crafted relevant representations during the feature engineering step. Overall nutrient sufficiency scores were generated by ranking the proportions of high, medium, and low concentrations for each macronutrient and micronutrient. The ratios of nutrients, N:P, K: OC were calculated. The high dimensional data was then compressed using PCA, capturing > 95% of the variance. The variation in soil fertility character and type among the states is evident from the PCA projection plot in Figure 4, which shows discrete groups of areas with comparable nutrient profiles. Feature distributions were analyzed for skewness and outliers before scaling. PCA was applied only after confirming high multicollinearity and to retain over 95% variance. Although KMeans was chosen because of its effectiveness and interpretability in delineating initial soil zones, this work did not compare it to other clustering algorithms. Due to the non-linear relation between these clusters and the cluster features, future work could use algorithms other than K-means such as DBSCAN or hierarchical clustering algorithms for nested soil profile relationships (41, 42). A comparison of this nature has the potential to define a basis for advancing the fertility zoning and thus enhancing the classification accuracy of the hybrid model. KMeans clustering with k=3 was used to find underlying patterns in the data, and th40 results showed a weak correlation with the established soil fertility categories of low, medium, and high.

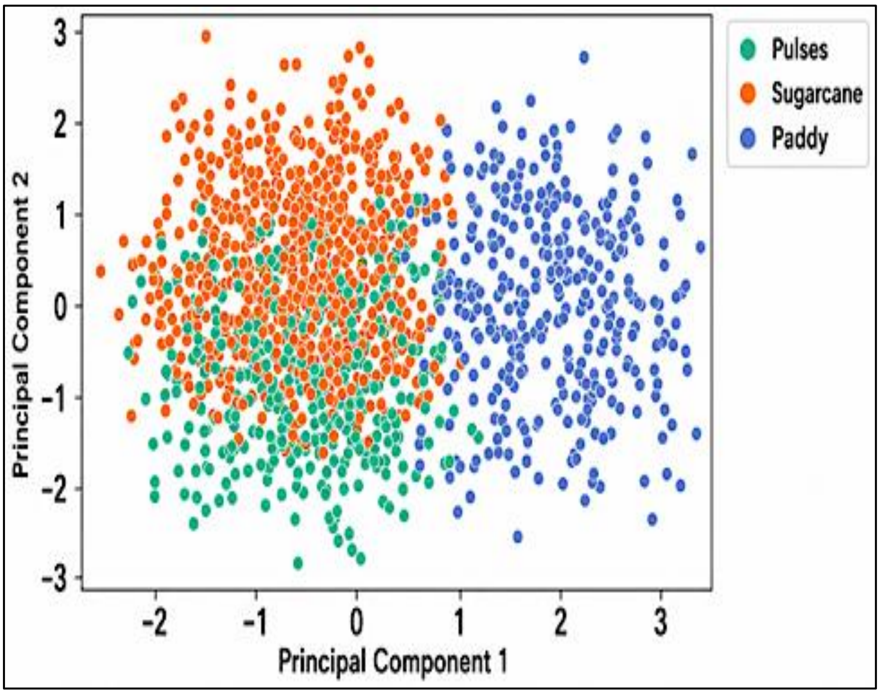

Figure 4: PCA Projection of K means Clustering

The clustering's biological and agricultural significance is highlighted by the visualisation of these KMeans clusters, which show distinct groupings that correlate to geospatial and agro-climatic zones. t-Distributed Stochastic Neighbour Embedding (t-SNE) was used to further investigate the separability of data points in a non-linear context. The result is a scatter plot in Figure 5 that displays distinct clusters that represent the latent structure in the high-dimensional dataset.

The spatial and structural analyses of the chemical traits of the soil starting by pH levels distribution. Figure 6 represents the streaming percentages of acidic, neutral and alkaline soils across the dataset with neutral soils clearly dominating. Understanding that pH affects where nutrients are and the overall health of plants throughout the model run, we can see this as a useful backdrop for our visualization.

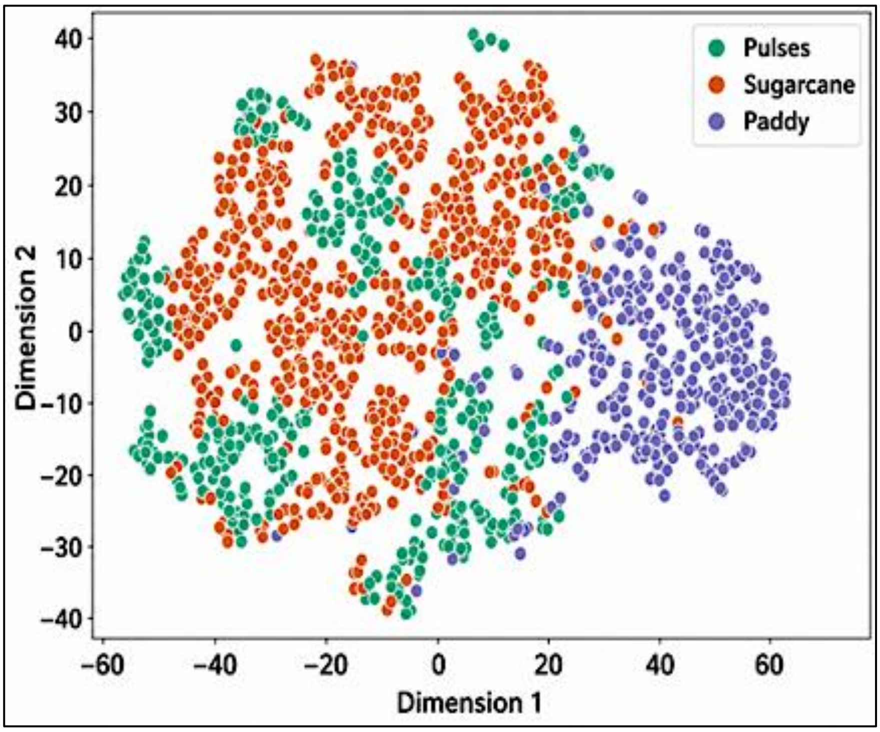

Figure 5: t-SNE Visualization of Soil Samples by Crop Type

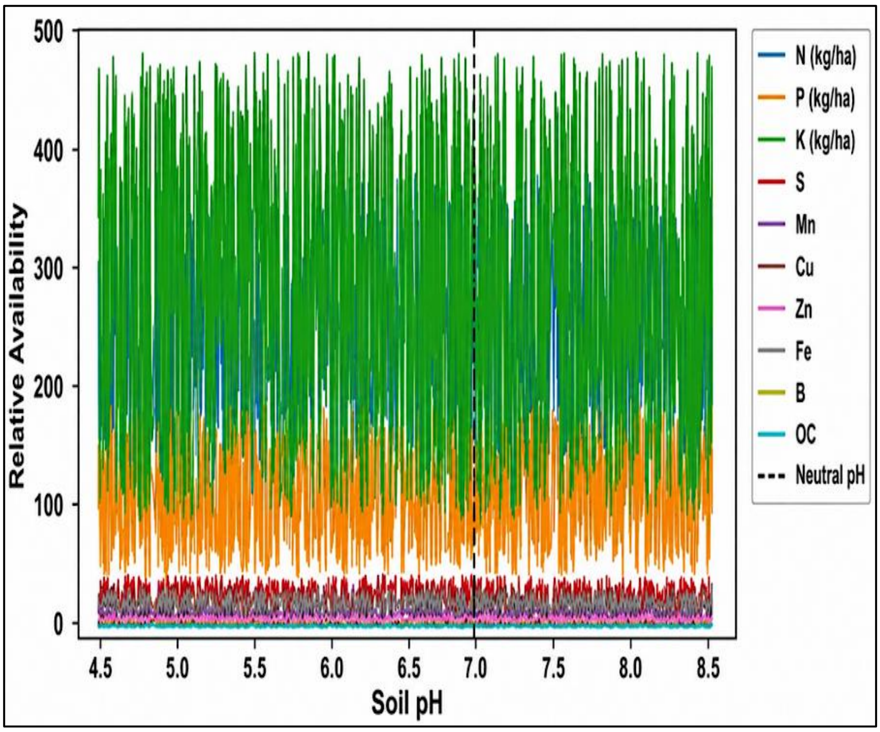

Figure 6: Soil pH Distribution Across India

Figure 7 was constructed to explore further the links among the main soil parameters. In the next plot, we have mapped Nitrogen (N), Phosphorus (P), Potassium (K), and pH against each other as well as showing how these are distributed for different fertility classes. The scatter plots display potential relations of various nutrients against each other and the density plots plotted in a diagonal show how their distributions change with levels of low, medium or high fertility soils. Such an exploratory analysis is necessary to reveal the nature of the data object to model.

.

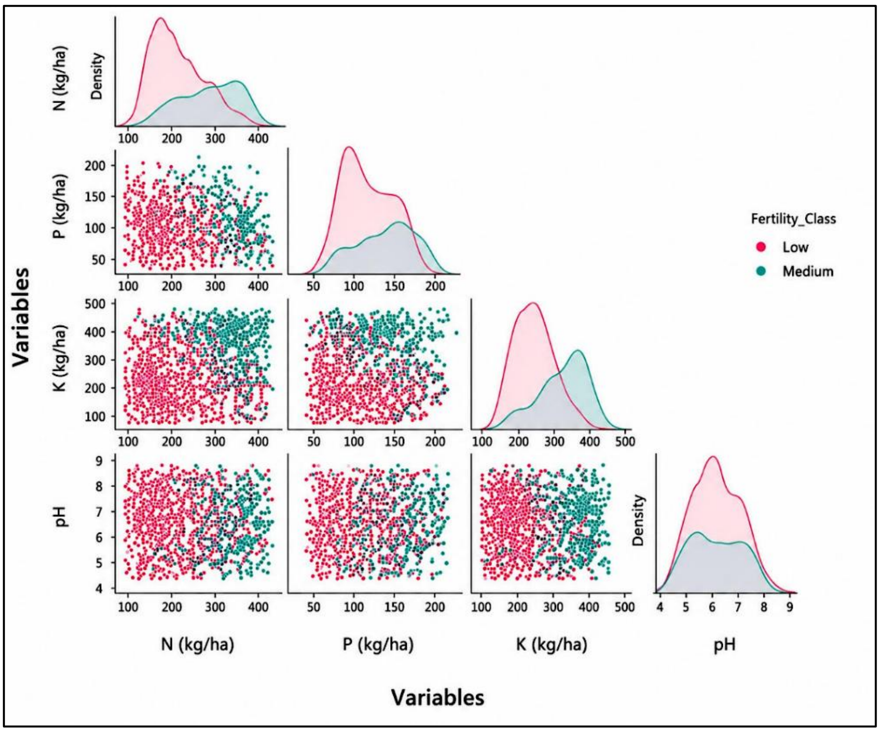

Figure 7: Pair Plot shows the Relation Between the Features

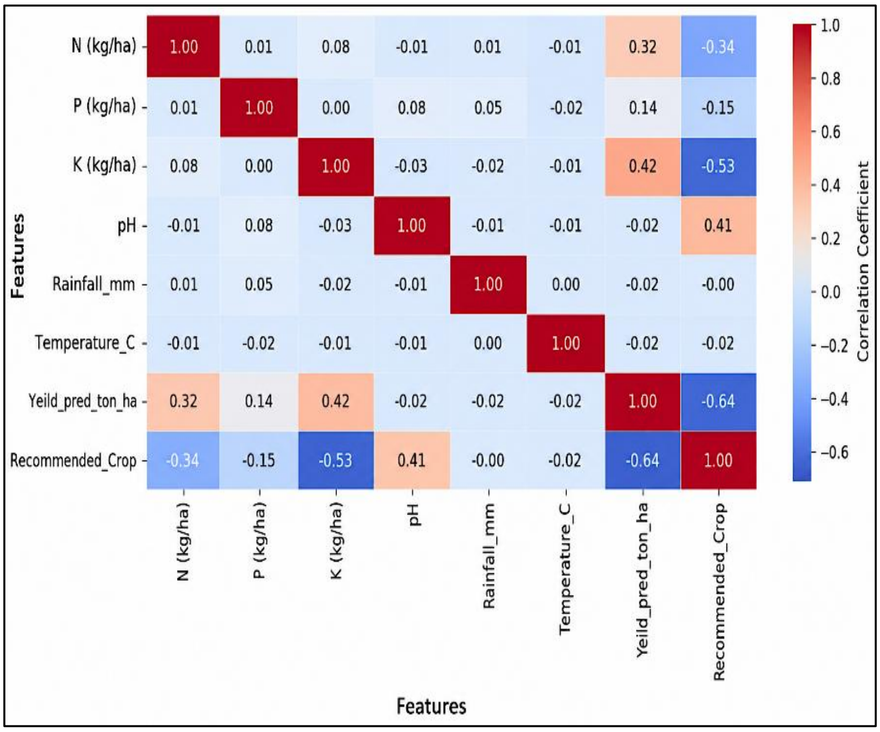

Figure 8: Correlation Heatmap Showing the Correlation Between all the Features

We created a correlation heatmap shown in Figure 8 to determine the strength of relationships between all the soil parameters, in addition to the pair plot. It helps to identify redundant features and potential multicollinearity issues. When we compared the heatmap of the dataset before and after PCA was used, it was seen that just around 5% variance had been loss, but major redundant features were removed. This step helped ensure the subsequent modeling is both computationally efficient and statistically rigorous, thereby enhancing the hybrid performance as well as its interpretability.

The SHAP analysis showed that only seven features were significant for crop prediction: N (kg/ha), P (kg/ha), K (kg/ha), pH, Rainfall (mm), Temperature (°C), and Yield_pred (ton_ha). The selected features are consistent with existing agronomic knowledge, as soil nutrients (N, P, K), soil reaction (pH), and climate factors (temperature, precipitation) are known to be important predictors for crop yield. We continued to step past traditional feature importance methods and utilized the SHAP explainer for better interpretability into how individual features impacted the model's predictions. Figure 9 illustrates this point with the resulting SHAP summary plot, which provides a more comprehensive explanation by both showing a feature's importance and its impact on different crop classes. This methodology reflects the principle of agronomic relevance, as it indicates that pH, Potassium (K), and the predicted Yield are the most important features.

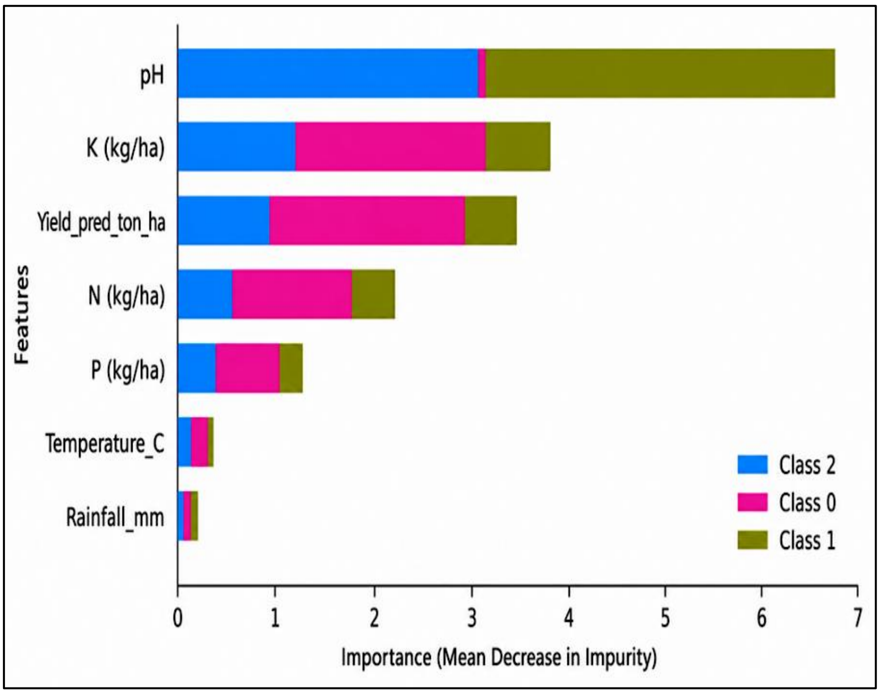

Figure 9: Feature of the Dataset Using SHAP Analysis

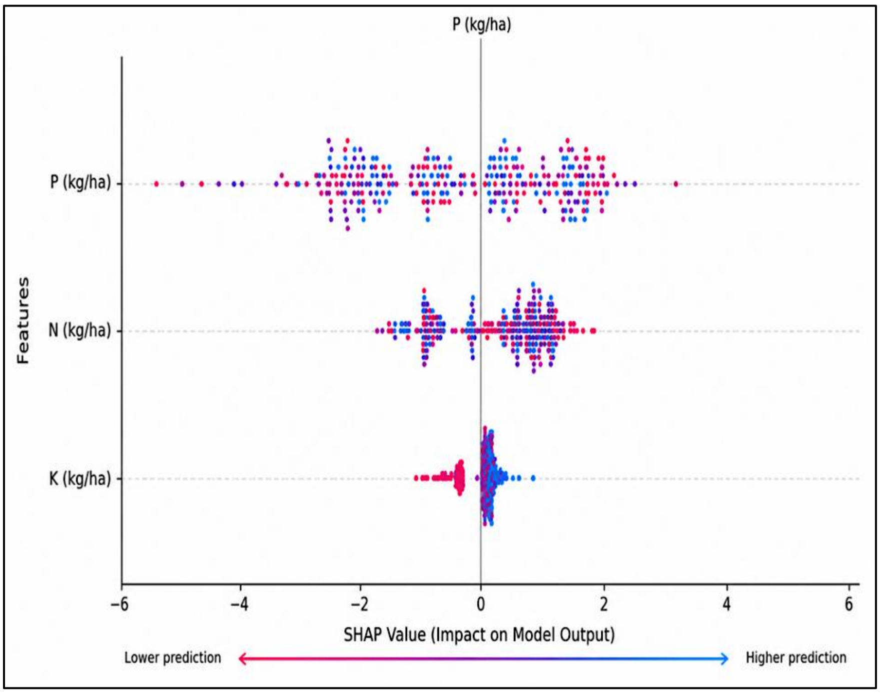

Figure 10: SHAP Dependency Plot

Finally, to better understand the model's internal mechanism, a SHAP dependency plot was shown in Figure 10. Here, the diagram is isolating a feature (Phosphorus - P) with its value on the x-axis and its SHAP value on y-axis (how much the feature affected the model output). Instead of an overview of feature importance as in the summary plot, the dependency plot shows you how a single attribute or even 2 features combined result in the model learning. This results in an extremely high-resolution view of how the recommended impact of phosphorus on the crop changes with its concentration for this specific model decision-making nature. Such a level of interpretability is fundamental to corroborating the model's behavior with the excellence in domain expertise and contributing to trust in its predictions.

All the previously mentioned elements had an impact on the creation of the final model architecture. The suggested model is a hybrid ensemble classifier that incorporates aspects of supervised and unsupervised learning. A soft-voting ensemble of Random Forest and XGBoost classifiers was used to perform supervised classification after KMeans cluster labels were added to the dataset as an additional feature. This hybrid approach improves performance and interpretability by using the knowledge gathered from clustering to fine-tune the classification decision boundaries. Here it should be mentioned that the current framework is deployed based on a database of soil and climate properties presented in tabular form using the SHC scheme. This is also a potential future direction for increasing the utilization of real-time remote sensing data in EWS which has been discussed in the Future Work section. Following is Algorithm 1 of the proposed hybrid model.

Algorithm 1 - Proposed Hybrid Crop Prediction Model

```
Input:
  Soil Health Dataset D
  Features F = {N, P, K, pH, EC, OC, Rainfall, Temperature, Micronutrients}
  Target Variables:
    Crop_Type = {Wheat, Rice, Maize}
    Fertility_Class = {Low, Medium}
Output:
  Predicted Crop Type
  Predicted Soil Fertility Class
  Model Performance Metrics
Begin
Step 1: Data Collection
  Load Soil Health Card (SHC) inspired dataset
  Import soil nutrient and climatic attributes
Step 2: Data Preprocessing
  Remove duplicate records
  Handle missing values using median imputation
  Detect and remove noisy outliers
  Convert categorical variables into numerical format
  Apply StandardScaler normalization
  Balance target classes using SMOTE
Step 3: Feature Engineering
  Perform correlation analysis
  Generate nutrient ratio features:
    N:P, N:K, K:OC
  Apply Principal Component Analysis (PCA)
  Retain components explaining >95% variance
  Select important features using SHAP analysis
Step 4: Soil Fertility Zonation
  Apply K-Means clustering on processed features
  Set number of clusters k = 3
  Assign samples into:
    Low Fertility
    Medium Fertility
```

```
      High Fertility Zones
  Step 5: Hybrid Model Training
    Initialize Random Forest classifier
    Initialize XGBoost classifier
    Train Random Forest on training dataset
    Train XGBoost on training dataset
    Combine predictions using Soft Voting Ensemble
  Step 6: Crop Type Prediction
    Predict Crop_Type:
      Wheat / Rice / Maize
  Step 7: Soil Fertility Classification
    Predict Fertility_Class:
      Low / Medium
  Step 8: Model Evaluation
    Compute:
      Accuracy
      Precision
      Recall
      F1-Score
      ROC-AUC Score
    Generate:
      Confusion Matrix
      ROC Curve
      SHAP Summary Plot
      PCA and t-SNE Visualization
  Step 9: Result Analysis
    Compare Hybrid Model with:
      Random Forest
      XGBoost
    Identify best-performing model
 End
```

Results and Discussion

The high predictive accuracy of the machine learning models employed in this study justifies the transition from conventional statistical analysis to computer intelligence and data mining technologies in agricultural science (43). Random forest (RF) and other ensemble learning models developed as part of this study achieved better performance and outperformed the linear baseline models, which is comparable with recent research, indicating that random forest can effectively address the non-uniformity and noise issues of agricultural datasets (44). The performance of the proposed hybrid model was verified through standard classification measures such as accuracy, precision, recall, and AUC. The model was tested in two different tasks: the prediction of soil fertility and the classification of the type of crop. Data separation into training and testing sets. To ensure that each crop class is well represented in the training and testing subsets, a stratified 80:20 train-test split was performed (keeping the having proportions of Wheat, Rice, and Maize, with Maize proportions of Wheat, Rice, and Maize, with Maize a relatively larger number of classes). Thus, it causes less sampling bias, and the model evaluation happens smoothly with the real-world class distribution. This model achieved very good predictive performance with soil fertility classification accuracy at 96% and crop type prediction at 97%, thus indicating great ability to generalize from the training data to novel ones.

Table 2: Accuracy Matrix of all the Proposed Hybrid Models

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
XGBoost	0.9253	0.9266	0.9253	0.9256	0.97
Random Forest	0.9173	0.9191	0.9173	0.9166	0.98
Hybrid Model (XgBoost+ Random Forest)	0.97	0.96	0.97	0.96	0.99

The different accuracy scores of the crop prediction task are presented in the classification report in Table 2 and Figure 11. The results reveal that the model can discern crops from each other regionally in terms of soil nutrient supply. All crops were classified perfectly. This is supported by looking at the confusion matrix shown in Figure 12, where most of the predictions are on the diagonal. The Wheat, Rice, and Maize class was accurately predicted on all the test instances. The high true positive rate, as well as the low false positive rate for all classes from the model demonstrate the model's robustness. Table 2 outlines the performance evaluation matrix comparing the standalone baseline models with the proposed hybrid approach. The results demonstrate that the proposed Hybrid Model (XGBoost and Random Forest) significantly outperforms the individual models, securing the highest classification accuracy of 0.97 and an exceptional ROC-AUC score of 0.99. The Hybrid Model consistently outperforms individual models across all metrics.

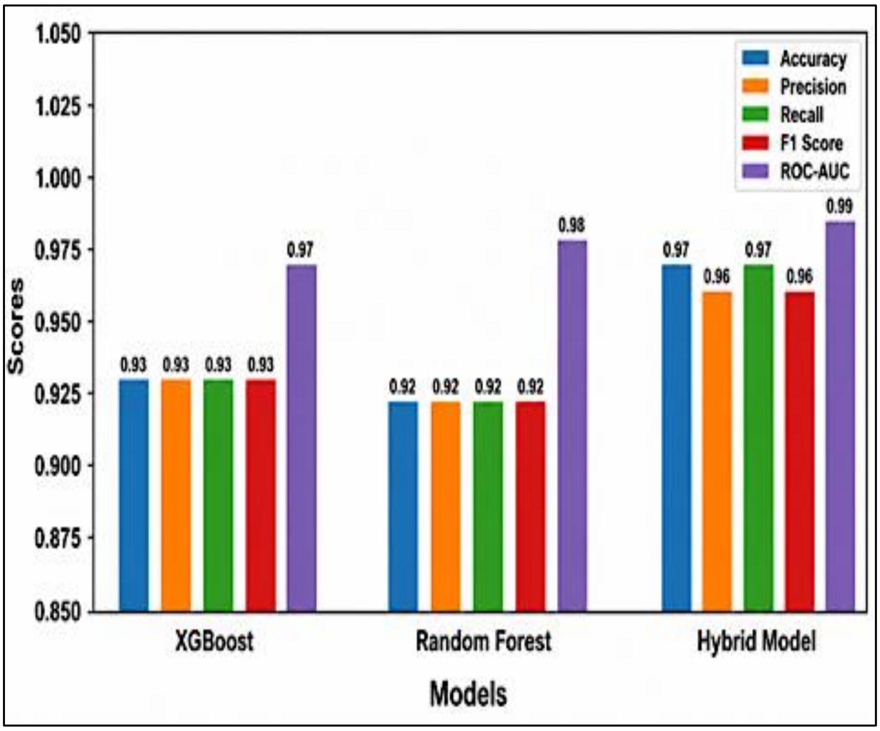

Figure 11: Comparison of Model Performance Metrics

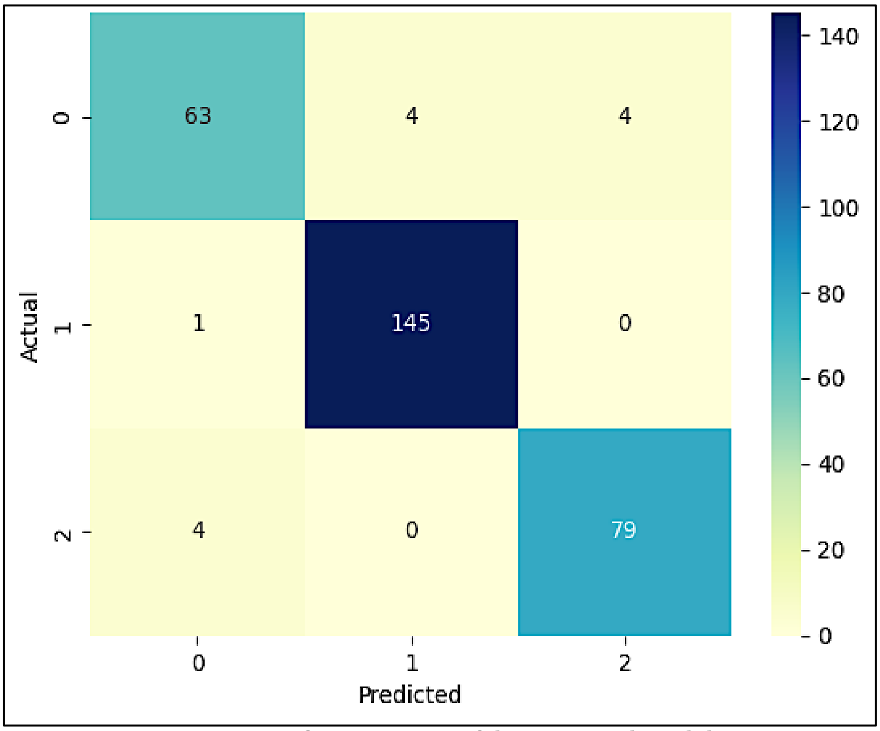

Figure 12: Confusion Matrix of the Proposed Model

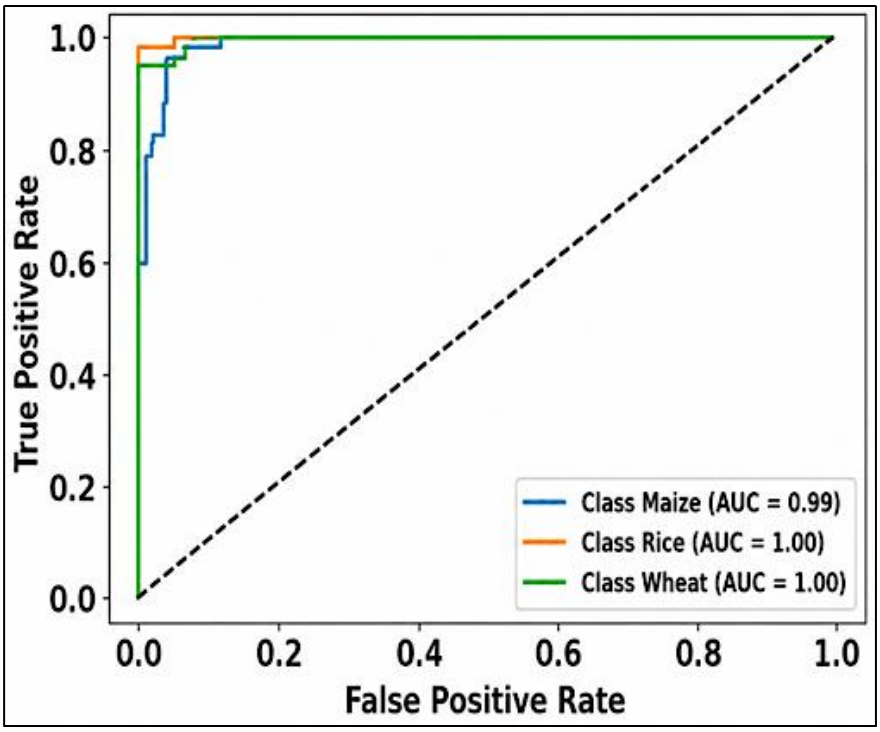

Figure 13: ROC Curve Analysis of the Proposed Model

In addition to the classification metrics, a Receiver Operating Characteristic (ROC) curve is shown in Figure 13 to evaluate the discriminative power of the model among crop categories. The heavy rotation of the curve towards the top left represents strong outcomes. AUC score: The macro-averaged Area Under the Curve (AUC) is 0.99, indicating that the model is overall sensitive and specific for all three crop classes and is very suitable for multi-class crop prediction.

Additionally, the graphical representation of feature importance helped interpret the logic of the model. Nitrogen (with greater importance in "medium" and "low" sufficiency), pH (all levels) and Potassium were some of the main variables in the analysis. In summary, the hybrid model-based K-Means clustering with RF and XGBoost exhibited exceptional accuracy and acceptable interpretability for data-driven agricultural advisory applications. This method overcomes the difficulties of big data analytics in precision agriculture. Experimental results validate the effectiveness of the proposed method, which is comparable with the state-of-the-art methods and further advances (16). Comparing with previous works also supports these results. Previous multi-crop or multi-region research showed R^2 ranging from 0.65 to 0.92, and our framework showed close or better results by conceptual verification, emphasizing feature space, temporal aspects, and diversified model ensemble (29, 30, 32–34).

Conclusion

The hybrid ensemble model with two base learners, K-Means++ clustering algorithm, and soft voting ensemble of Random Forest and XGB classifiers is observed to be excellent in its prediction potency with remarkable accuracy, precision, recall and F1-score of 97%. The strong performance also receives support through an AUC macro-averaged of 1, which adds to its robustness. One of the reasons behind such a remarkable performance of the model is its spectacular design, as evidenced by meticulous data pre-processing and the tactical utilization of unsupervised K-means clustering. The model can handle complex, non-linear relationships in the data. It also retains a higher interpretability, as such is shown in the feature importance plots expressed in accordance with the existing agronomic knowledge. Additional confirmations are given by visual aids, including ROC curves, confusion matrices, and t-SNE visualizations, which confirm the model's robustness and generalization for other soil types. Given that macro and micro levels of nutrients are comprehensively combined, the model represents the complex nature of the soil-crop system that is usually overlooked in most conventional assessments. In summary, this work shows that the hybrid AI method is successful and is useful for precision agriculture and it provides a manageable and intelligent decision support system for agronomists, policy makers and farmers. The framework offers an effective grounding for future studies and research in localized crop recommendations, soil treatment tactics, and scaling AI-enabled farming solutions to the commercial level.

Potential limitations of the proposed framework include risks of overfitting due to model complexity and possible sample imbalance within the Soil Health Card dataset. These factors may influence generalization across diverse agro-climatic regions and should be considered when interpreting the results. While the framework is highly accurate, it is not a definitive solution for growth in deploying the trained model. The steps set out our vision for turning the research into a usable tool for farmers i.e by using a mobile application for farmers, would democratize entry to data, bringing the value that may be gleaned from SHC data - actionable insight.

To address these restrictions and bring this research into use as a scalable, real-world precision agricultural tool, the future part of this work will concentrate on an implementation plan. The creation of a mobile GPS application to democratize predictive insights with specific crop recommendations and nutrient management for any farmer accessing the application. Secondly, the inclusion of IoT sensors located in the field, in order to provide updates on dynamic soil moisture, temperature and N-P-K fluctuations will surpass the restrictions of static data to supply continuous, dynamic measurements that allow for hyper-specific, situationally-prescribed modelling. Finally, the framework will be extended to include remote sensing data from satellites which will produce real-time vegetation indices (e.g., NDVI) and early monitoring of pest infestations, mapping of water-related stresses, and highly accurate estimates of yield. Automated retraining pipelines capable of synthesizing the repeated updates of the SHC database with the real-time IoT data and multispectral satellite data will enable the establishment of a real-time, responsive adaptive ecosystem for the management of resilient crops.

Abbreviations

AUC: Area Under the Curve, DBSCAN: Density-Based Spatial Clustering of Applications with Noise, DDD: Driver Drowsiness Dataset, IoT: Internet of Things, NDVI: Normalized Difference Vegetation Index, ROC: Receiver Operating

Characteristic, SHAP: SHapley Additive exPlanations, SHC: Soil Health Card, T-SNE: t-distributed Stochastic Neighbor Embedding, UTs: Union Territories.

Acknowledgement

We express our profound gratitude for the accessibility to scholarly resources and institutional facilities afforded by Sir Padampat Singhania University, which significantly facilitated the fulfillment of this research endeavor. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions

Vimla Dangi: Conceptualization, Methodology, Data Collection, Data Analysis, Original Draft Preparation, Kamal Kant Hiran: Supervision, Formal Analysis, Validation, Final Editing, Manish Tiwari: Supervision, Formal Analysis, Validation, Final Editing.

Conflict of Interest

The author declares no conflict during the study.

Data availability

The datasets used in this study are obtained from a publicly available government data portal.

Declaration of Artificial Intelligence (AI) Assistance

The authors used AI-assisted language editing tools solely for grammatical check and refinement. All research design, data analysis, interpretation and intellectual content were carried out by the authors themselves. The authors take full responsibility for the content's originality, interpretation and accuracy.

Ethics Approval

Not applicable.

Funding

No funds were received to conduct the study.

References

1. Kumar S, Siddiqui F, Perwej Y, Rizvi H, Akhtar N. Robust Crop Prediction Framework Using Soil Classification and Ensemble Methods. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2025;11 (6):248–60.
doi: 10.32628/cseit2511646
2. Bharadiya JP, Tzenios N, Reddy M. Forecasting of Crop Yield using Remote Sensing Data, Agrarian Factors and Machine Learning Approaches. Journal of Engineering Research and Reports. 2023;24(12):29–44.
doi: 10.9734/JERR/2023/v24i12858
3. Subramaniam LK, Marimuthu R. Crop yield prediction using effective deep learning and dimensionality reduction approaches for Indian regional crops. e-Prime - Advances in Electrical Engineering, Electronics and Energy. Elsevier. 2024; 8:100611.
doi: 10.1016/j.prime.2024.100611
4. Goel M, Pandey M. Crop Yield Prediction Using AI: A Review. 2nd International Conference on Disruptive Technologies, ICDT. 2024;1547–53.
doi: 10.1109/ICDT61202.2024.10489432
5. Pingali P, Boiteau J, Choudhry A, Hall A. Making meat and milk from plants: A review of plant-based food for human and planetary health? World Dev. 2023; 170:106316.
doi: 10.1016/j.worlddev.2023.106316
6. Ali Z, Muhammad A, Lee N, Waqar M, Lee SW. Artificial Intelligence for Sustainable Agriculture: A Comprehensive Review of AI-Driven Technologies in Crop Production. Sustainability. 2025;17(5):2281.
doi: 10.3390/su17052281
7. Paramesha V, Kumar P, Francaviglia R, *et al.* Evaluating land use and climate change effects on soil organic carbon. A simulation study in coconut and pineapple systems in west coast India. Catena. 2025;248:108587.
doi: 10.1016/J.CATENA.2024.108587
8. Latha P, Kumaresan P. Deep Learning for Soil Nutrient Prediction and Strategic Crop Recommendations: An Analytic Perspective. Nature Environment and Pollution Technology. 2025;24(1): 1-2.
doi: 10.46488/NEPT.2025.v24i01.B4205
9. Aijaz N, Lan H, Raza T, Yaqub M, Iqbal R, Pathan MS. Artificial intelligence in agriculture: Advancing crop productivity and sustainability. J Agric Food Res. Elsevier. 2025;20:101762.
https://doi.org/10.1016/j.jafr.2025.101762
10. Sowmya BJ, Meeradevi AK, Supreeth S, *et al.* Leveraging machine learning for intelligent agriculture. Discover Internet of Things. 2025;5(1): 33.
doi: 10.1007/s43926-025-00132-6
11. Tyagi A, Haritash AK. Climate-smart agriculture, enhanced agroproduction, and carbon sequestration potential of agroecosystems in India: a meta-analysis. J Environ Stud Sci. 2025;15(1):167–85.
doi: 10.1007/s13412-024-00917-1
12. Mukherjee B, Jha RK, Sattar A, *et al.* Harnessing the potential of millets for climate-resilient and sustainable agriculture. Front Plant Sci. 2025;16: 1574699.
doi: 10.3389/fpls.2025.1574699
13. Singh R, Singh S. A Review of Indian-Based Drones in the Agriculture Sector: Issues, Challenges, and Solutions. Sensors. 2025;25(15):4876.
doi: 10.3390/s25154876

14. Nikhil UV, Pandiyan AM, Raja SP, Stamenkovic Z. Machine Learning-Based Crop Yield Prediction in South India: Performance Analysis of Various Models. Computers. 2024;13(6):137. doi: 10.3390/computers13060137
15. Akwabeng PM, Awele O, Jegede OA, Odili J. A Data-Driven Assessment of Crop Yield Variability and Global Food Security under Climate Change. World Journal of Advanced Research and Reviews. 2026;29(1):295–304. doi: 10.30574/WJARR.2026.29.1.0028
16. Afzal H, Amjad M, Raza A, *et al.* Incorporating soil information with machine learning for crop recommendation to improve agricultural output. Sci Rep. 2025;15(1):8560. doi: 10.1038/s41598-025-88676-z
17. Pal K, Gupta N. Machine learning and deep learning approaches for soil classification: methods, challenges, and future prospects. Int J Remote Sens. 2025;46(22):8488–545. doi: 10.1080/01431161.2025.2570552
18. Screpnik CR, Zamudio E, Gimenez LI. Artificial Intelligence in Agriculture: A Systematic Review of Crop Yield Prediction and Optimization. IEEE Access. 2025;13:70691–7. doi: 10.1109/ACCESS.2025.3560631
19. Sirsat MS, Cernadas E, Fernández-Delgado M, Khan R. Classification of agricultural soil parameters in India. Comput Electron Agric. Elsevier. 2017;135:269–79. http://dx.doi.org/10.1016/j.compag.2017.01.019
20. Filippi P, Han SY, Bishop TFA. On crop yield modelling, predicting, and forecasting and addressing the common issues in published studies. Precis Agric. 2025;26(8). https://doi.org/10.1007/s11119-024-10212-2
21. Patil Y, Ramachandran H, Sundararajan S, Srideviponmalar P. Comparative Analysis of Machine Learning Models for Crop Yield Prediction Across Multiple Crop Types. SN Computer Science 2025; 6(1):64. https://doi.org/10.1007/s42979-024-03602-w
22. Elbasi E, Zaki C, Topcu AE, *et al.* Crop Prediction Model Using Machine Learning Algorithms. Applied Sciences. 2023;13(16):9288. doi: 10.3390/APP13169288
23. Zidan F, Febriyanti DE. Optimizing Agricultural Yields with Artificial Intelligence-Based Climate Adaptation Strategies. IAIC Transactions on Sustainable Digital Innovation (ITSDI). 2024;5(2):136–47. doi: 10.34306/ITSDI.V5I2.663
24. Dangi V, Goswami C, Chakrabarti P. Developing a Conceptual Framework for Soil Property Analysis and Crop Yield Prediction Using Machine Learning Techniques. International Journal of Innovative Technology and Interdisciplinary Sciences. 2025;8(3):513–36. doi: 10.15157/IJITIS.2025.8.3.513-536
25. Ahmad B, Alam A, Hamid, A, *et al.* Modern tools for sustainable agriculture: a review of intelligent crop protection technologies. Discover Agriculture. 2026;4(1):19. doi: 10.1007/s44279-025-00467-2
26. Gomes da Costa Cavalcanti JH, de Brito Souto MC, de Moura Oliveira Filho E, *et al*. Machine learning for optimized crop selection: a data-driven approach considering climate and soil type data. International Journal of Data Science and Analytics 2025 22:1. Springer. 2026;22(1):12. doi: 10.1007/s41060-025-00939-0
27. Emara EIR, Al-Saeed AM, Hamed LMM. AI-driven decision support system enhances productivity, water use efficiency, and soil sustainability of strategic crops in sandy soils. Agric Water Manag. 2026;325:110198. doi: 10.1016/J.AGWAT.2026.110198
28. Kempenaar CC, Sijbrandij FFD, Been TTH, *et al.* Towards Data-Driven Precision Crop Management of Potato. Potato Res. 2026;69(1):10. doi: 10.1007/S11540-025-09996-1
29. Balaji V, Balamurugan KS, Thanarajan T, Selvi Sundarapandi AM. Machine Learning-Based Crop Recommendation System Integrating Soil Properties and Weather Conditions with IoT-Driven Data Collection. Agrociencia. 2026;60(1):1–17. doi: 10.47163/AGROCIENCIA.V60I1.3428
30. Sabeena AA, Islam MA, Rahman R, *et al.* AI-Driven Predictive Analytics for Crop Rotation and Soil Health Management. In Proceedings of the International Conference on Sustainable Economics and Finance in the Digital Business Transformation (INCOSEF 2025). 2026.254–65. https://www.researchgate.net/publication/403461993_AI-Driven_Predictive_Analytics_for_Crop_Rotation_and_Soil_Health_Management
31. Chaudhary P, Gulia P, Gill NS, Alduaiji N, Shukla PK, Vishnoi S. An evaluation of machine learning for soil analysis in internet of things-enabled smart farming. Scientific Reports. 2026;16(1):10318. doi: 10.1038/s41598-026-36017-z
32. Nanwal J, Sethi P. An ensemble framework for crop prediction based on soil's physical and chemical properties using grasshopper optimization algorithm (GOA). International Journal of Information Technology. 2025;17(5):2841–6. doi: 10.1007/S41870-025-02485-1
33. Sadasivan M, Krithika KV, Shinty PK, Deepa A. Predictive Crop Yield Modeling and Soil Quality Classification using Machine Learning. Indian J Sci Technol. 2025;18(32):2594–607. doi: 10.17485/IJST/v18i32.1135
34. Dharwadkar NV, Kalmani VH, Thapa V. Crop Yield Prediction using Deep Learning Algorithm based on CNN-LSTM with Attention Layer and Skip Connection. Indian J Agric Res. 2024;59 (8):1303. doi: 10.18805/ijare.a-6300
35. Soilhealth.dac.gov.in. Soil Health. 2025. https://www.soilhealth.dac.gov.in/nutrient-dashboard
36. Akhatov A, Fayzulla N, Eshmurodov M. Ensemble Machine Learning Algorithms for Crop Types Classification Based on Intelligent Detection of Soil Properties. 2025 International Russian Automation Conference (RusAutoCon). 2025:691–6. doi: 10.1109/RUSAUTOCON65989.2025.11177338
37. Malik SJ, Hiran KK, Barua T, Doshi R. AI-Powered Precision Agriculture: Integrating Smart Systems for Sustainable Yield Forecasting and Climate Resilience. 2025 8th International Conference on Emerging Technologies in Computer Engineering: Advances in

Computing, Healthcare and Smart Systems (ICETCE). IEEE. 2025;139–46.
doi: 10.1109/ICETCE66848.2025.11389084
38. Kadu AV, Reddy KTV. Agriculture Yield Forecasting via Regression and Deep Learning with Machine Learning Techniques. InInternational Conference on Communication and Intelligent Systems. 2023;219–33.
doi: 10.1007/978-981-97-2079-8_18
39. Garg S, Sagtani A, Kothari A, Golakia J, Olaniyi RO. Climate Change Analytics Using Artificial Intelligence: Focusing on Legal and Ethical Considerations. Community Resilience and Climate Change Challenges: Pursuit of Sustainable Development Goals (SDGs). United States: IGI Global Scientific Publishing. 2025;13-28.
doi: 10.4018/979-8-3693-6522-9.CH002
40. Nishant PS, Sai Venkat P, Avinash BL, Jabber B. Crop Yield Prediction Based on Indian Agriculture Using Machine Learning. International Conference for Emerging Technology, INCET. Institute of Electrical and Electronics Engineers Inc. 2020.1-4
doi: 10.1109/INCET49848.2020.9154036
41. Gupta R, Sharma AK, Garg O, *et al.* WB-CPI: Weather Based Crop Prediction in India Using Big Data Analytics. IEEE Access. 2021;9:137869–85.
doi: 10.1109/ACCESS.2021.3117247
42. Liu B, Li X, Yang J, Liang Z, Zheng X, Zou X. Cotton field population phenotyping analysis based on 3D Gaussian reconstruction and dynamic spatial constraints. Front Plant Sci. 2026;17:1790984.
doi: 10.3389/FPLS.2026.1790984
43. Cravero A, Sepúlveda S, Gutiérrez F, Muñoz L. From Precision Agriculture to Intelligent Agricultural Ecosystems: A Systematic Review of Machine Learning and Big Data Applications. Agronomy. 2026;16(5):516.
doi: 10.3390/agronomy16050516
44. Kindra K, Bhuvaneswari Amma NG, Nageswari Amma NG. Towards enhancing the performance of crop prediction system for precision agriculture using feature correlation square-based nearest neighbor classifier. Scientific Reports. 2026;16(1):8807.
doi: 10.1038/s41598-026-35212-2

How to Cite: Dangi V, Hiran KK, Tiwari M. Data-driven Approaches in Soil and Crop Prediction: A Conceptual Framework Using Machine Learning. Int Res J Multidiscip Scope. 2026;7(3):423-441.
DOI: 10.47857/irjms.2026.v07i03.011474.