INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Check for updates

OPEN ACCESS

Received: 19/06/2025

Accepted: 23/07/2025

Published: 20/08/2025

Citation: Sadasivan M, Krithika KV, Shinty PK, Deepa A (2025) Predictive Crop Yield Modeling and Soil Quality Classification using Machine Learning. Indian Journal of Science and Technology 18(32): 2594-2607. https://doi.org/10.17485/IJST/v18i32.1135

* **Corresponding author**.

varshini.krithika@cmr.edu.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Predictive Crop Yield Modeling and Soil Quality Classification using Machine Learning

Manju Sadasivan[1], K Varshini Krithika[1]*, P K Shinty[1], A Deepa[1]

1 School of Science and Computer Studies, CMR University, Bengaluru, Karnataka, India

Abstract

Background: Soil condition and yield are essential practices in contemporary farming, having a direct impact on food production efficiency and balanced resources consumption. Inability to measure these parameters with traditional methods usually takes time, is expensive and beyond the reach of many farmers. Such constraints demonstrate the necessity of data-based alternatives that are effective. **Objectives:** The study aims to aid in the development of a comprehensive machine learning system, which will categorize the soil quality grade in some classes, recommend possibly cultivation of crops against a particular soil quality, and an estimation of the crop yield given a few vital soil parameters. The approach integrates classifiers, recommenders and regressors into a single decision tool aimed at delivering sound farmer-friendly decision support in farming and advisory networks of farmers and agronomists. The long-term aim would be to enhance agricultural productivity, optimize the resources and also make data driven decisions easy for farmers. **Method:** Two Kaggle real world datasets were considered– one for crop recommendation and other for yield prediction. A custom Soil Quality Index (SQI) was developed based on important parameters such as nitrogen, phosphorus, potassium and pH.In both classification and regression, three classifications of machine learning models were deployed; Random Forest, XGBoost, and support vector machines (SVM/SVR). The interpretation of the models and understanding of the data were also achieved with the help of advanced visual analytics (heatmaps, scatter plots, feature importance graphs). **Findings:** Out of all three models, models with Random Forest Classifier produced the best performance with soil classification accuracy of 97.73% and this was significantly better compared to XGBoost and SVM. Regarding the estimation of crop yield, the ensemble and the kernel based regressors were better in the low value error and high prediction performance. **Novelty:** The study presents a broad and understandable machine learning model where soil quality classification, crop suggestion, and crop yield forecast are implemented together with real-world data. One of the innovations is that it has developed a custom Soil Quality Index (SQI) which combines four factors: nitrogen, phosphorus, potassium, and pH into one with a single, significant value, and, thereby, improving model results and model interpretability. Compared to the usage of many ML models (Random Forest, XGBoost, and SVM/SVR) for classification and regression. The task covered in this work is the completely based on agriculture with the use of real and publically accessible data

Keywords: Crop yield; Soil quality; Random Forest; Support Vector Machine; XGBoost; Nitrogen-Phosphorus-Potassium (NPK) Analysis; Precision Farming

1 Introduction

Looking at crops, soil quality can be fundamental when it comes to yield estimation since it determines health and nutrient uptake for plants, aspects such as water capacity for retention, aeration and the rates at which microbes perform their duties in the field[(1)]. However, the evaluation method of soil is an intricate process as it is determined by several interrelated factors such as nutrient status, moisture status, organic matter and soil pH. Farmers have for long depended on conventional methods of laboratory analysis, visual analysis, and past experience which are expensive and time-consuming respectively[(2)]. The results of their study were that Random Forest reached the most accurate result (R squared = 0.963) and thus able to effectively support agriculture planning based on data and help farmers enhance their yield prediction[(3)]. The study by Iniyan et al. introduces machine learning in determining the crop yield amount through the utilization of the district-wise environmental and soil data in Maharashtra compared to the relationship models using multiple regressions. The LSTM with feature engineering was identified as the best model with the accuracy rate of 86.3, and it is capable of helping farmers make better decisions due to yield forecasting[(4)]. The work of Elbasi et al. is an extensive crop prediction framework based on fifteen machine learning algorithms utilizing data collected by the IoT sensors in order to make more informed choices in the farm. Their findings demonstrate that algorithms like Bayes Net and Random Forest provided high accuracy in classification (up to 99.59 percent), facilitated the smart farming approach towards the optimal crop yield that would result in reduced wastage and sustainability[(5)]. Bai et al. develop a new GNN-RNN based crop yield predictor which combines geospatial and temporal data discovered on more than 2,000 US Countries investigated throughout 39 years. Their model is very successful, as they incorporate county-level spatial relationships and historical patterns, and because of this, it vastly out-predicts other traditional ML strategies, proving the need in geographic context helping to increase the accuracy of predictions[(6)]. In their paper, Jones et al.introduce interpretive machine learning (IML) and show how SHAP values can be used to diagnose yield constraints in furrow-irrigated cotton fields. The method combines digital soil mapping and gradient boosted (XGBoost) models that can diagnose the presence of yield-limiting factors, such as sodicity, salinity and terrain in a spatially explicit fashion, thereby providing prescriptive information that can be implemented into precision agriculture[(7)]. A complete smart farming system was implemented by Durai and Shamili, which involved machine learning and deep learning in crop recommendation, pest and weed detection, pesticide and herbicide recommendation and cost estimation. Their web-based framework yielded an accuracy of prediction as high as 95.45 per cent using such models as Random Forest and ResNet152V2, providing a realistic and data-driven answer to precision agriculture[(8)]. In their study, Mishra et al. highlights how AI/ML can contribute to boosting productivity, sustainability, and decision support in agricultural systems along with technical, ethical, and uptake issues[(9)]. Gosai et al. proposed Intelligent crop recommendation system by use of different machine learning models such as XGBoost, Random Forest, Naive Bayes to provide an optimal crop suggestion, which depends on the parameters used in the soil such as NPK, pH, temperature, humidity and rainfall. A system of theirs, XGBoost, was the most accurate to say the least (99.31), and it holds an opportunity to provide data-driven contributors to the increase in the productivity and profitability of agriculture[(10)]. Nyeki and Nemenyi summarize the recent progress in crop yield prediction in precision agriculture with particular attention to the introduction of AI and big data, IoT, and data fusion methodologies. They survey a wide range of approaches, including sensor-based monitoring, neural networks, and crop models, with far more focus given to how each improves the accuracy of yield forecasting and enables sustainable, site-specific management of farms in the long-term[(11)]. In their paper, Satya et al. identifies the crop recommendation of machine learning algorithms that include Gradient Boosting, XGBoost, and LightGBM based on soil indices and environmental factors such as humidity, temperature, and rainfall. The predictive models are supposed to complement and help in increasing agricultural productivity by prescribing the right kinds of crops to be produced as per the climatic and eco-spatial datasets[(12)]. By using seven machine learning algorithms, Dahiphale et al. design a powerful crop recommendation system proposing the best crops to be produced depending on the available information on the environmental and soil parameters. They measured their accuracy at up to 99.5%, resolving some of the agricultural issues, such as climate variability or resource or efficiency, and providing their sustainable farming with scalable and data-driven solutions[(13)]. Ángel Luis Perales Gómez et al. integrated Random Forest, SVM, and LSTM in the FARMIT platform, achieving a low error rate (6.59%) with better pest diagnosis and stress detection but faced challenges with high computational costs and scalability in practical settings[(14)]. Iqbal et al. revealed that the ensemble machine learning models, especially the Random Forest and XGBoost, performed much better than the traditional value of regression in predicting the wheat-yield under the climate-change conditions with an R^2 of 0.953 and RMSE of 0.107 using climatic and GCM data[(15)]. Hakkal and Lahcen proved that XGBoost is capable of dramatically improving predictive performance on various real-world data sets compared to the existing traditional logistic regression models and especially when high-dimensional and nonlinear feature representations are present[(16)]. In their study, Michelucci et.al. suggested new formulas that provide a more pessimistic estimate of the metrics as compared to the corresponding classical ones without considering measuring errors. Manipulation of the formulas will provide a more realistic evaluation of the model performance because it considers measurement errors present in the target variables[(17)]. The paper by Prity et al introduces a machine learning-based crop recommendation system comprising nine ML algorithms with the highest accuracy of 99.31%, randomly designed forest, which shows the relevance of ensemble learning in order to achieve successful data-based agricultural planning[18]. Afzal et al suggested one more ensemble model, RFXG, a combination of Random Forest, XGBoost to make a crop recommendation depending on the nutrients in the soil and the weather conditions that could be expected to reach a high accuracy rate of 98 per cent and outperform other classical machine learning methods in all 22 significant classes of crops[19]. Ajith et al. introduced an extensive review of the approaches based on AI, such as Random Forest, SVM, and XGBoost, and explained how they had the high predictive power in predicting crop yields and estimating soil nutrients, and Random Forest and the GBRT models regularly outperformed in the multi-nutrient estimation[20]. Sheeba et al. utilize an Internet of Things (IoT)-based automated system using Extreme learning machinery (ELM) for the micronutrient classification as well as soil fertility prediction results in the effective estimation of pH and nutrient mapping throughout multiple districts[21].

1.1 Research Gap

Even with the fast development of precision agriculture, several small-sized and middle-sized farmers still use traditional, experience-driven approaches to the evaluation of soil and crop planning. Current mechanisms of crop yield prediction and soil quality assessment are prone to being distributed, too expensive rendering them unusable in large scales. In addition, the majority of the previous publications divided the issues of soil classification, crop recommendation, and yield forecasting into individual problems and did not provide efforts to estimate their combination in an end-to-end decision-support chain. Besides, black-box models, including deep learning and artificial neural networks (ANN), have been associated with a good predictive capability to an extent that sometimes they are not transparent and interpretable. Thus, there is an urgent necessity to have a simple, interpretable, and data-driven machine learning platform that can easily integrate all these, soil quality classification, crop recommendation, and yield prediction, thereby providing real-time solutions to farmers about what they can do sustainably and optimally to farm.

1.2 Contribution

The reason behind carrying out this study is to offer farmers credible, data-based resources in achieving sustainable crop farming, and this is based on the shortcomings of the conventional assessment and estimate techniques used in the evaluation of soil and the estimation of outputs which are both time consuming, expensive, and not available to small and medium scale farmers. The goal will be to establish an integrated machine learning system that will sort soil quality, recommendations on which crops to plant and expected crop yield based on major parameters of soil nitrogen, phosphorus, potassium and pH. This study also introduces a new innovative contribution of a specifically designed Soil Quality Index (SQI) that is a formulation that condenses several of the soil attributes into one simple comprehensible index to perform better classification. Influence of soil parameters on crop yield is depicted in figure 1.

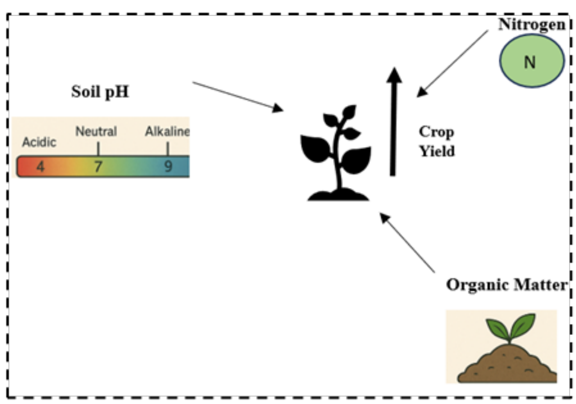

Fig 1. Influence of soil parameters in crop yield

In this study a machine learning model is generated using Random Forest to predict crop yield and classify soil quality. The prediction of the model indicates the quality of crop yield and popularization of categories of soil as using a unique Soil Quality Index (SQI) and analyzing the importance of features, the model demonstrates decent performance. The findings determine the strength of Random Forest models, which determines significant characteristics that influence productivity and soil quality.

This paper explains major activities like:

1. Predicting crop yield using regression techniques: Random Forest Regressor, XGBoost Regressor and SVR models
2. Classifying soil quality based via SQI
3. Identifying key soil parameters through feature importance analysis
4. Crop recommendation using Random Forest Classifier, SVM and XGBoost Classifier.

Future work involves semantic web technologies for more effective decision-making.

2 Methodology

2.1 Methodological Workflow

The various phases of the proposed methodology are illustrated below.

2.1.1. Data collection

There are two datasets used, and these datasets are collected from Kaggle website. The two datasets are Crop_Recommendation and Crop_yield_data. The parameters in Crop_Recommendation dataset involve some essential features like: Nitrogen, Phosphorus, Humidity, Potassium, Temperature, pH_value, Rainfall, Crop. The parameters in Crop_yield_data dataset involve some essential features like rainfall, Soil Quality, Farm_size, Sunlight, Fertilizer, Crop_yield.

2.1.2. Data Pre processing

Before implementing any machine learning models, any missing or invalid data was removed. The New Formula evaluated the soil's nitrogen (N), phosphorus (P), potassium (K), organic matter and pH and then used the results to come up with the Soil Quality Index (SQI). The StandardScaler technique was used to make sure all the features gave equal weight during model training.

2.1.3. Exploratory Data Analysis (EDA)

Initially, a correlation heatmap was produced to see the relationships and associations between different aspects and crop yield. Scatter plots were chosen to view how the Soil Quality Index (SQI) is linked to the crop yield, thus revealing patterns and single out any unusual values. Also, histograms and SQI distribution plots helped discover the range of soil quality seen in different samples, offering a clear overview of the grouped data.

2.1.4. Model Development

The training of the ML model takes place through using the cleaned dataset during this phase. This step involves dividing the dataset into training and testing subsets, and also to separate the feature from the target variable. The model is divided into two parts. In both parts of the process, a number of algorithms were applied to review performance. In Phase 1 (Classification), Random Forest, XGBoost and SVM classifiers were chosen since they can handle non-linear aspects, are efficient and work well with many input variables. Phase 2 (Regression) used Random Forest, XGBoost and SVR to model complicated setups, provide better results with tabular data and ensure the prediction model could be applied reliably.

2.1.5. Model Evaluation

In this phase, performance of the regression model is measured based on the training provided. In Phase 1 (Classification), the performance of the models was determined by Accuracy, Precision and F1-score and Random Forest and XGBoost demonstrated good results with balanced predictions. During Phase 2 (Regression), the R^2 Score and RMSE helped evaluate the model and plots of actual versus predicted yields were used to check its accuracy.

2.1.6. Feature Importance

The most important input features that have an impact with the target variables are identified in this phase. The models were used to determine that nitrogen, pH and organic matter were the major deciding factors among soil features. Bar charts demonstrated the effects of soil parameters on crop forecasting in Phase 1 and on yield forecasting in Phase 2 which helps farmers know what should be given priority to improve the fertility of their land.

2.1.7. Classification Evaluation

Soils were divided into Poor, Average, Good and Excellent groups based on their pH readings and three classifier algorithms—Random Forest, SVM and XGBoost—were each used to evaluate their other soil properties and give a rating. Models were evaluated with confusion matrix, accuracy and precision and it was determined that their classification results were reliable and consistent among the different algorithms.

2.1.8. Data Visualization

Heatmap of feature correlation and bar graphs are used to gain information about patterns and model behavior.

2.1.9. Dataset Description

Two primary datasets were used:

1. ***Crop Recommendation Dataset:*** Includes nutrients (NPK), pH, temperature, humidity, rainfall, and crop type.
2. ***Crop Yield Dataset:*** Includes soil quality index, rainfall, sunlight, fertilizer use, farm size, and yield.

2.2 Proposed model

2.2.1. SVM for crop recommendation

It performs a soil quality check by computing a Soil Quality Index (SQI) from the results of Nitrogen, Phosphorus, Potassium and pH which have all been normalized. With SQI, along with temperature, humidity and rainfall, an SVM model recommends the best crops to grow. The model is checked, and its importance of features is shown visually.

2.2.2. SVR for crop yield

Features from the crop_yield_data.csv dataset are used by the Support Vector Regression (SVR) model to estimate crop yield. It cuts the data into parts, scales the features and builds an SVR with the RBF kernel. A graph is created to compare the model's actual and predicted yields and the evaluation markers are R^2 value and RMSE.

2.2.3. XGBoost with Soil Quality Index (SQI)

The features of soil help determine a Soil Quality Index (SQI) and place it in one of the three categories: Good, Moderate or Poor. A classifier built using XGBoost advises on suitable crops by examining the features and this is evaluated by measuring how accurate and precise its recommendations are. Experts agree that soil quality categories provide another method to check and validate how accurate the SQI model is.

2.2.4. XGBoost Regressor for yield prediction

Crop yield is anticipated using rainfall, the quality of the soil, farm size, hours of sunlight and fertilizer as environmental and farming factors. Using the features provided, an XGBoost regressor is trained and how it does is measured with MAE, RMSE and the R^2 score. The visualizations displayed consist of scatter plots, feature correlation heatmaps and feature importance lists.

2.2.5. Random Forest with SQI

The model uses important soil components to get a Soil Quality Index, classifies the soil into grades and matches crops to soil and environment data by using Random Forest. There are charts and graphs of important features and how they relate to crop choice. The software is learned and assessed so that it can give precise predictions of crops, improving decision making in agriculture relying on soil data.

2.2.6. Random Forest Regressor for yield prediction

The model attempts to forecast crop yield based on characteristics from soil and their surroundings. The first step is to examine how the features in the data relate and to show the connection between soil quality and the amount of crops produced. After that, the Random Forest regressor is trained to predict yield and then checked with R^2 and RMSE as evaluation metrics. Lastly, the important features are highlighted to understand which ones affect the crop yield most.

The *feature_importances_* attribute of the trained Random Forest Regressor is used to extract the relative importance of each input feature. Random Forest is an ensemble learning method which works by building several decision trees and combining their output. In the process, it calculates how much each feature is able to minimize the impurity across all trees. The greater the contribution a feature makes in error reduction, the greater its importance score. Table 1 lists the features considered in the model.

Table 1. Feature description

Feature	Description	Agricultural Significance
pH	Measure of soil acidity or alkalinity	Affects nutrient availability; most crops grow best in pH 6.0–7.5
Nitrogen (N)	Concentration of nitrogen compounds in soil (g/kg)	Promotes leaf and stem growth; key element in chlorophyll and amino acid synthesis
Phosphorus (P)	Level of phosphorus nutrients in soil (mg/kg)	Essential for root development, energy transfer, and flowering
Potassium (K)	Concentration of potassium in the soil (mg/kg)	Enhances drought resistance, improves quality of fruits and grains, and regulates water and enzyme activity
Soil Quality Index (SQI)	Weighted composite of all soil features	A holistic score representing overall soil fertility and quality

2.3 Performance Evaluation Metrics

These models were evaluated using metrics such as Mean Absolute Error (MAE) for regression and Precision, F1 Score, and Accuracy for classification, providing insights into their performance and reliability.

2.3.1. Mean Absolute Error (MAE):

The common evaluation metric Mean Absolute Error (MAE) evaluates predictive model accuracy particularly in regression problems through measuring prediction errors in terms of their absolute values. This measurement technique computes the typical error size from model predictions but ignores their negative or positive value[17].

$$MAE = (1/n)\sum_{i=1}^{n}|y_i - x_i| \tag{1}$$

y_i = Prediction
x_i = True value
n= total number of data points

2.3.2. Accuracy:

Accuracy evaluates classification models through counting the correct predictions among all their outputs. The accuracy evaluation metric determines predictions' accuracy through the division between correct predictions and the complete set of predictions.

$$\text{Accuracy } = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

2.3.3. Recall:

Recall represents an evaluation method used to determine how well an algorithm detects all relevant positive results. The formula for recall calculation divides true positive occurrences by the combination of true positives and false negatives cases. The model suffers from decreased occurrence of omitted genuine positives when recall stands high.

$$\text{Recall } = \frac{TP}{TP+FN} \tag{3}$$

Where TP = True Positives,
TN = True Negatives,
FP =False Positives,
FN = False Negatives

2.3.4. Precision:

The accuracy level of positive predictions receives evaluation through the precision metric in classification tasks. The ratio between real positive outcomes and the addition of both correct and incorrect positive outcomes determines its calculation. The model shows high precision because it makes a limited number of incorrect positive predictions.

$$\text{Precision } = \frac{TP}{TP+FP} \tag{4}$$

2.3.5. F1 score:

When balancing metrics in one value the F1 Score provides the harmonic mean between precision and recall scores. A dataset with imbalanced distribution benefits from the application of this score. When applied to F1 Score we can identify excellent precision-recall harmony[18].

$$\text{F1 Score } = 2\, X \frac{Precision\ X\ Recall}{Precision + \ Recall} \tag{5}$$

2.3.6. Regressor Performance

The Random Forest Regressor achieved a Mean Absolute Error (MAE) of approximately 209.42 kg/ha, which indicates strong predictive performance.

2.3.7. Classification Performance

Classification Accuracy for soil quality labels was 97%. Precision and F1-score were both 97-98%, reflecting the classifier's balanced performance across multiple classes (Poor, Average, Good, Excellent).

2.3.8. Soil Quality Index (SQI):

A custom formula assigns weights to features based on agronomic influence. SQI serves as a comprehensive indicator of soil health. The parameters pH, Nitrogen, Organic Matter and Electrical Conductivity in the data set have strong feature importance values contributing to the SQI prediction. XGBoost Classifier achieved 91% accuracy, and Random Forest reached 88%.

$$SQI \ = ((0.3 * N) + (0.25 * P) + (0.2 * K) + (0.25 * pH))/4 \tag{6}$$

3 Results and Discussion

3.1 Random Forest

Random Forest Classifier is used for soil analysis and crop recommendation. The Correlation Heatmap of Soil Features in figure 2 represents how Nitrogen, Phosphorus, Potassium, pH_Value and Soil Quality Index (SQI) are linked. It shows that nutrients such as N, P and K often increase at the same time and play a major role in soil quality, but pH does not have a big or positive effect. With this, gardening can boost both healthy soil and greater crop productivity.

3.2 Random Forest Regressor for crop yield prediction

The heatmap in figure 3 helps to realize that most of the crop's yield increase comes from greater farm size, followed by a mostly weak effect from fertilizer use and rainfall on crop yield. Things like light and type of soil play a smaller role in plant growth.

3.3 XGBoost

XGBoost Classifier is used for soil analysis and crop recommendation. The heatmap in figure 4 helps us to understand the correlation between the Nitrogen, Phosphorus, Potassium, and pH value of soils. The intensity and direction of the correlation are shown by the blue for a negative relationship, the red for a positive relationship, and white for a situation with no relationship. The values 1 in the diagonal cells are there to show each feature is correlated with itself.

3.4 XGBoost Regressor for crop yield prediction

The heatmap in figure 5 presents the link between various aspects and the yield of crops. Varying blue to red colors mark the range from strong positive to strong negative correlations, and white stands for cases without any correlation. Every cell that is diagonally away from the main diagonal has a value of 1 since it is the correlation of a feature with itself. Farm size shows a strong positive correlation.

3.5 SVM for crop recommendation

The study shows which aspects of an SVM model, with a linear kernel, have the biggest influence. Rainfall, Humidity, Temperature, pH Value, Potassium, Phosphorus, and Nitrogen are the features listed in terms of how important they are to agriculture.

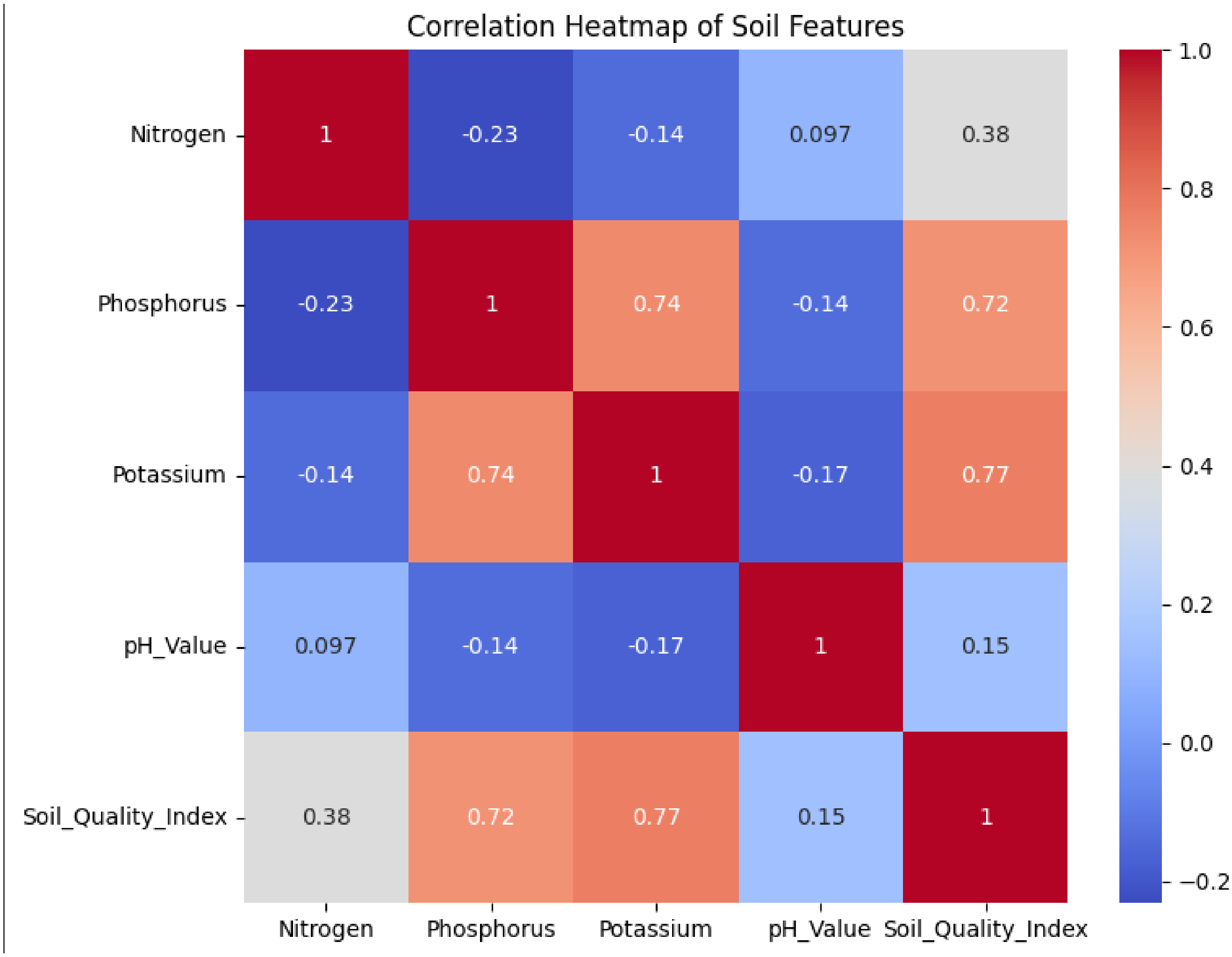

Fig 2. Correlation Heatmap of Soil Features

3.6 SVM for crop yield

In figure 6 the heatmap indicates the strength of relationships between various traits and the crop yield using different colors. Features that are looked at are mm of rainfall, soil quality index, hectares of farm, sunlight hours, use of fertilizer (kg), and yield of the crop.

The bar chart in figure 7 illustrates the level of importance of each feature for predicting outcomes using Random Forest. It points out Rainfall, Humidity, Potassium, Phosphorus, Nitrogen, Temperature and pH_Value, putting their importance values between 0.00 and 0.20. In this way, the model can show which parts of the environment or soil are the most important for the model's outcomes, assisting in agriculture. It means that Potassium and Phosphorus are essential for predicting outcomes in agriculture, so they should be the concentrated focus areas for soil quality analysis.

Random Forest delivers the best performance in terms of both R^2 Score and RMSE, indicating it captures the relationship between soil parameters and crop yield more effectively. XGBoost performs comparably well, but slightly less accurately than Random SVR (Support Vector Regression) shows a significantly lower R^2 Score and a much higher RMSE, indicating weaker predictive ability. Figure 8 shows the R^2 score and RMSE values for various regression models for crop yield prediction.

XGBoost consistently performs well across both regression and classification tasks, showcasing its robustness. While Random Forest remains a strong baseline, boosting techniques can further enhance classification precision. SVC may need better kernel tuning or feature engineering for better classification accuracy.

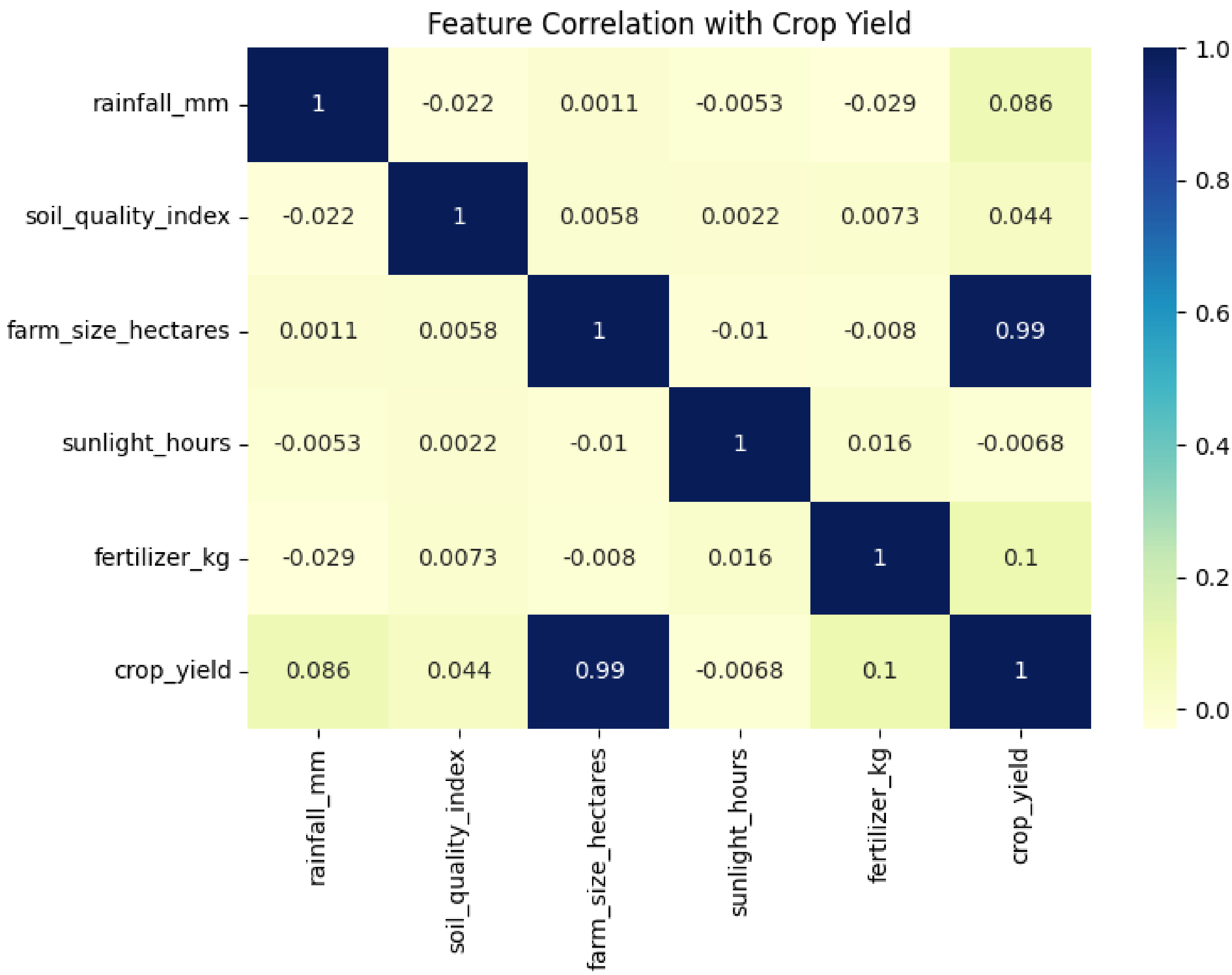

Fig 3. Feature Correlation with Crop Yield

3.7 Limitation and Potential biases

Despite the fact that the developed machine learning models, i.e., Random Forest, Support Vector Machine (SVM), and XGBoost demonstrated favorable outcomes in the domain of soil quality classification and crop yield forecasting, certain restrictions and biases are to be referred to form a more objective critique of them. Firstly, the data considered in the present research is rather specific and narrow in the focus and extent and remains limited to the chosen geographical and agro-climatic zones only. A regional bias can harm the model of such significance in terms of generalizing findings to some other ecological zones that have a different soil type, crop type, or climatic patterns. Future studies involving larger proportions of multi-regional and heterogeneous information should be incorporated to be stronger to work with. Secondly, the study is grounded to data on soil and agricultural products yield of the past time that may entail inconsistency, noise or even incompleteness. Such quality data issues can also cause a bias when training a model and during forecasting, particularly in case some classes of the soil or categories of the crop are underrepresented (class imbalance). Moreover, the models that belong to the ensemble model (Random Forest and XGBoost) are vulnerable to overfitting occasionally, and thus, where the model is implemented, it must be suitably regularized or tuned, particularly in circumstances of high dimensions. On the contrary, the sample may be a potential issue in the SVM model due to scalability and sensitivity to parameter selection in place of the non-linear kernels. On top of that, the scope of variables is limited since they are cited to physicochemical soil properties and climatic conditions. The other important Agro-Economic factors that could not be considered using the modeling pipeline include irrigation slant, pest

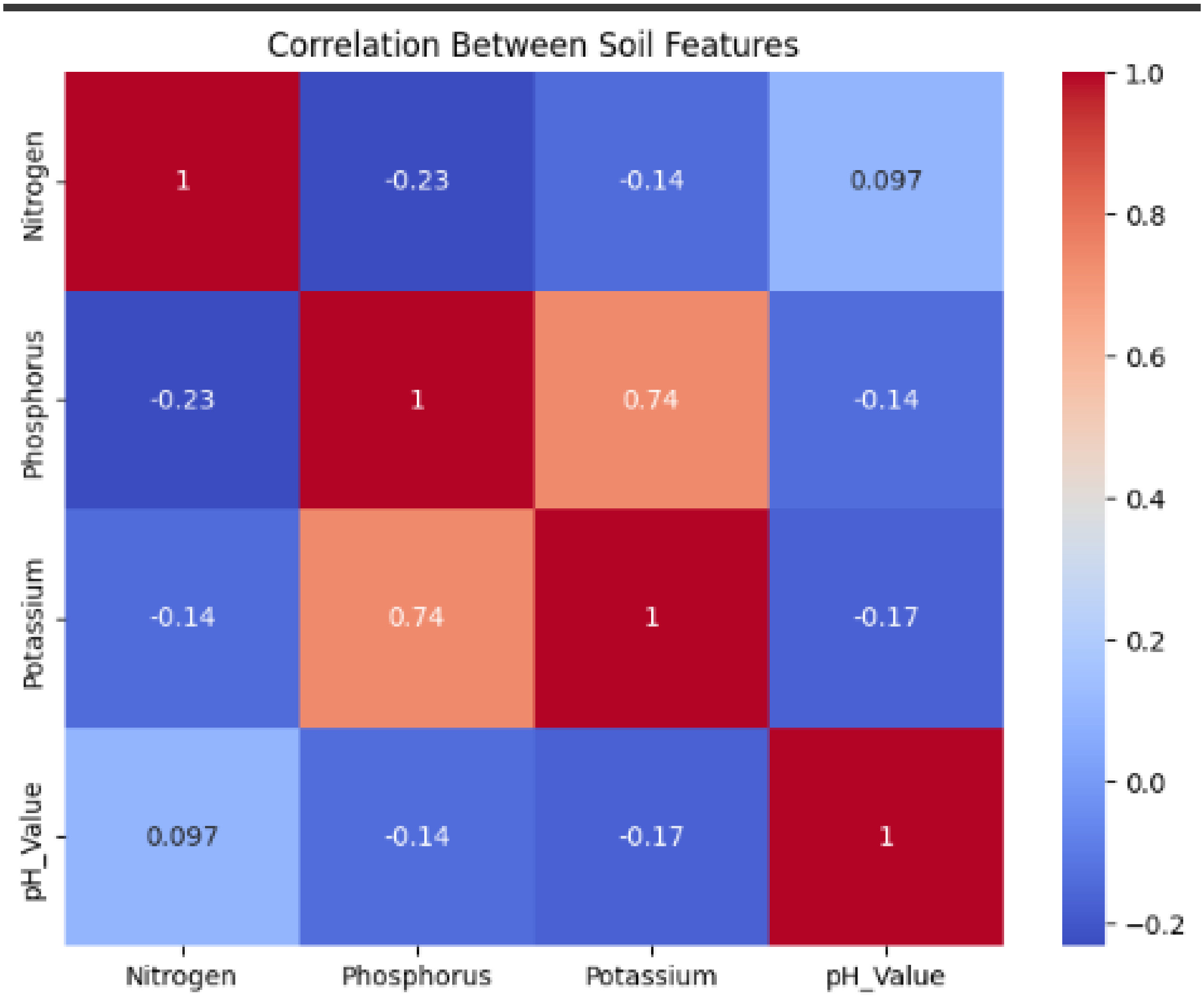

Fig 4. Correlation Between Soil Features

outbreaks, illness tension and fertilizer types and market access. Such non-accounted variables can be very serious as far as yield is concerned though they are very difficult to measure or standardize.

3.8 Comparative Model Evaluation

A comparison of various models given in Table 2.

Table 2 provides the outline of the comparative study of the support vector machine model (SVM), XGBoost, Random Forest (RF), and Random Forest Regressor (RF Regressor) models. In our experiments, Random Forest classifier recorded the best results in soil quality ranking and crop suggestion (97.73%), whereas XGBoost logged the best results in crop yield scoring with R^2 score of 0.88. SVM classifier offered similar performance (94%) in classification of soil and selection of crops, which was not as good as the ensemble-based models. All in all, there was little difference between ensemble methods and kernel-based learning models in terms of predictive performance, especially in balancing predictive performance and interpretability in a variety of agricultural prediction problems. Compared to the earlier research, our findings meet or overtake the current standards. As an example, well-known methods such as XGBoost could achieve crop recommendation with an accuracy of 99.31%, and the Random Forest could provide an R^2 value of 0.963 of the yield prediction of the efforts of Gosai et al. and Jhajharia et al. based in Rajasthan, respectively[10],[3]. On the same note, Dahiphale et al. have shown that SVM and RF could obtain an accuracy of more than 97% in classification-based crop prediction tasks[13]. Iqbal et al. demonstrated some promising

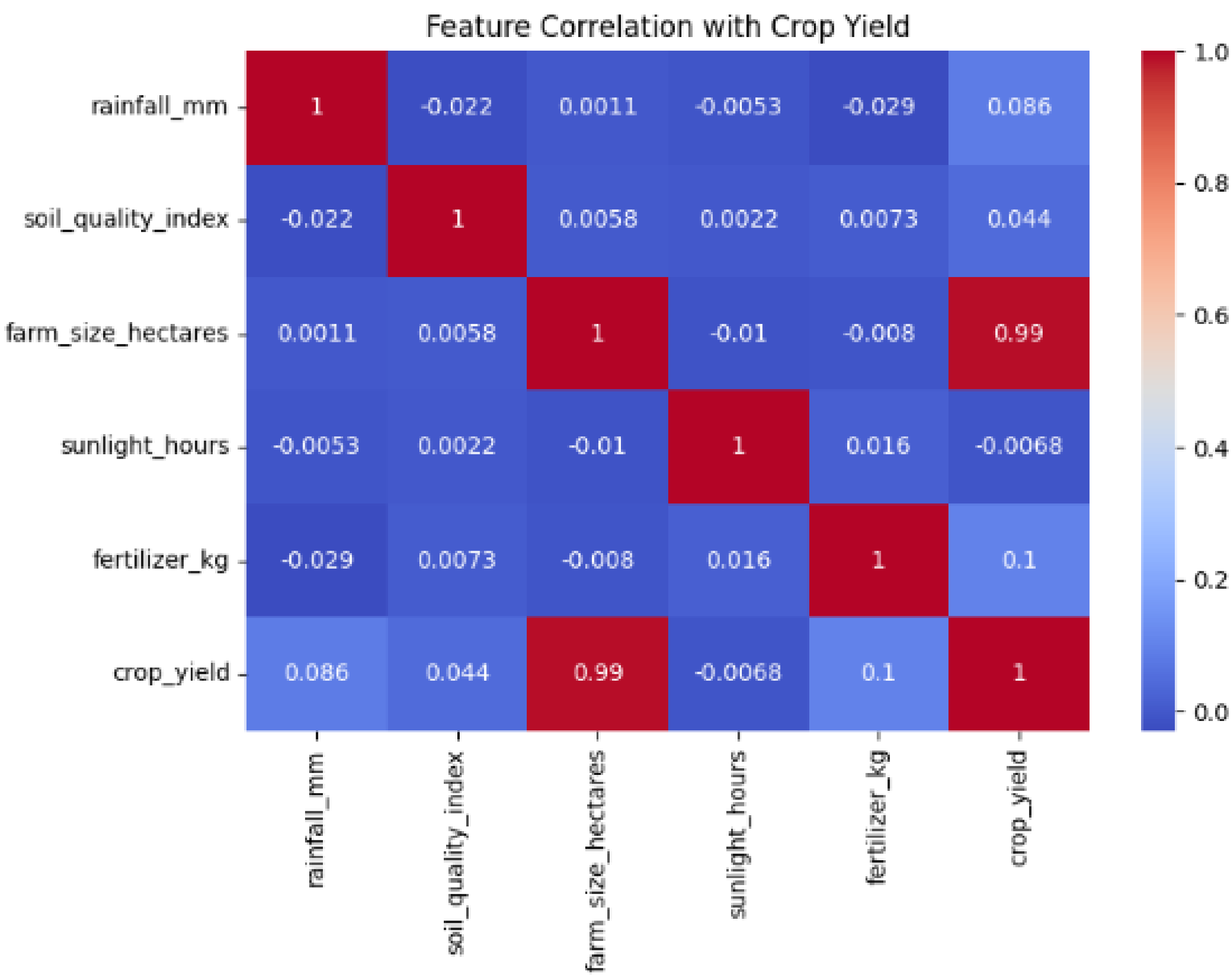

Fig 5. Feature Correlation with Crop Yield

Table 2. Model Comparison

Aspect	**SVM**	**XGBoost**	**Random Forest**	**RF Regression**
Task	Classification	Classification	Classification	Regression
Purpose	Crop Recommendation	Yield Prediction	Crop Recommendation	Crop Recommendation
SQI Used	No	Yes	Yes	Yes
Model Type	Linear Classifier	Boosted Trees	Bagged Trees	Regression Ensemble
Accuracy / R^2	94%	96%	97.73	$R^2 \approx 0.88$
Interpretability	Coefficients	Feature Importance	Feature Importance	Feature Importance
Strengths	Lightweight, clear	High performance	Balance of speed & accuracy	Non-linear yield learning
Limitations	Less robust to non-linearity	Longer training time	Needs feature scaling	May overfit on small data

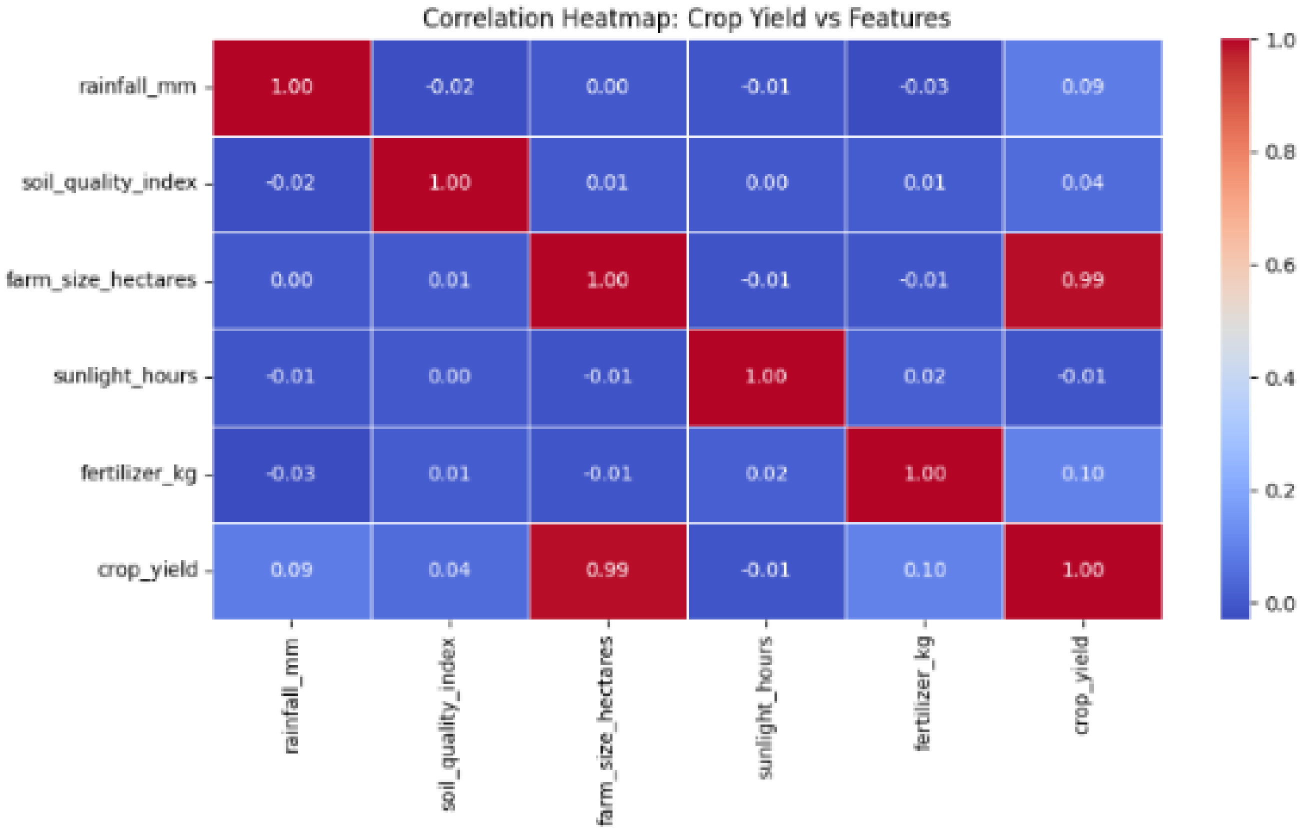

Fig 6. Correlation Heatmap : Crop yield VS Features

performance of XGBoost regression in yield prediction of wheat along with other climatic zones[15]. Moreover, Blesslin Sheeba et al. used yet have used ML models for soil classification, but not in combination with yield forecasting[21], and Afzal et al. discussed in their systematic review that the majority of agricultural ML studies consider prediction and classification separately, subdividing them, and thus they are not integrated[19]. On the contrary, our study demonstrates an integrated, end-to-end design so that the soil quality classification, hunting only suitable crops, and determining crop yield uses a uniform dataset pipeline. As compared to the previous solutions, which either use complex, uninterpretable deep learning models or have a limited set of features belonging to a narrow class of agronomic features, this integrated decision support framework is based on real-life field data. Added to the custom-built Soil Quality Index (SQI) and the interpretable models, such as Random Forest and XGBoost, our system achieves not only a high level of accuracy but also the explainable predictions easily displayed in the form of feature importance. This makes the decisions (data-driven, transparent, and actionable) more suited to small and medium-scale farmers and thus increases the quality of the real-life application and acceptance of smart farming systems.

4 Conclusion and Future Scope

This research is created to transform the scenario of agricultural decision making for farmers, as it develops an advanced integration of essential soil qualities, such as pH, phosphorus, potassium, and nitrogen levels, together with current meteorological conditions. Its purpose is to use important soil measurements—pH, phosphorus, potassium and nitrogen—together with crop yield prediction models to help in farmland management. With data-driven methods, the system allows farmers to choose options that help them get the best yields while being more sustainable. These models perform excellently and are more robust than traditional methods, The accuracy of Random Forest classifier was high (97.73%) on the soil quality classification categorization as compared to the other models because of the advantages it had over other models such as ensemble, resistance to overfitting, and explanation of feature importance. In comparison, XGBoost, which has an accuracy of 79.77% and macro precision of 78.80%. Unlike other models such as SVM and ANN, Random Forest is more interpretable

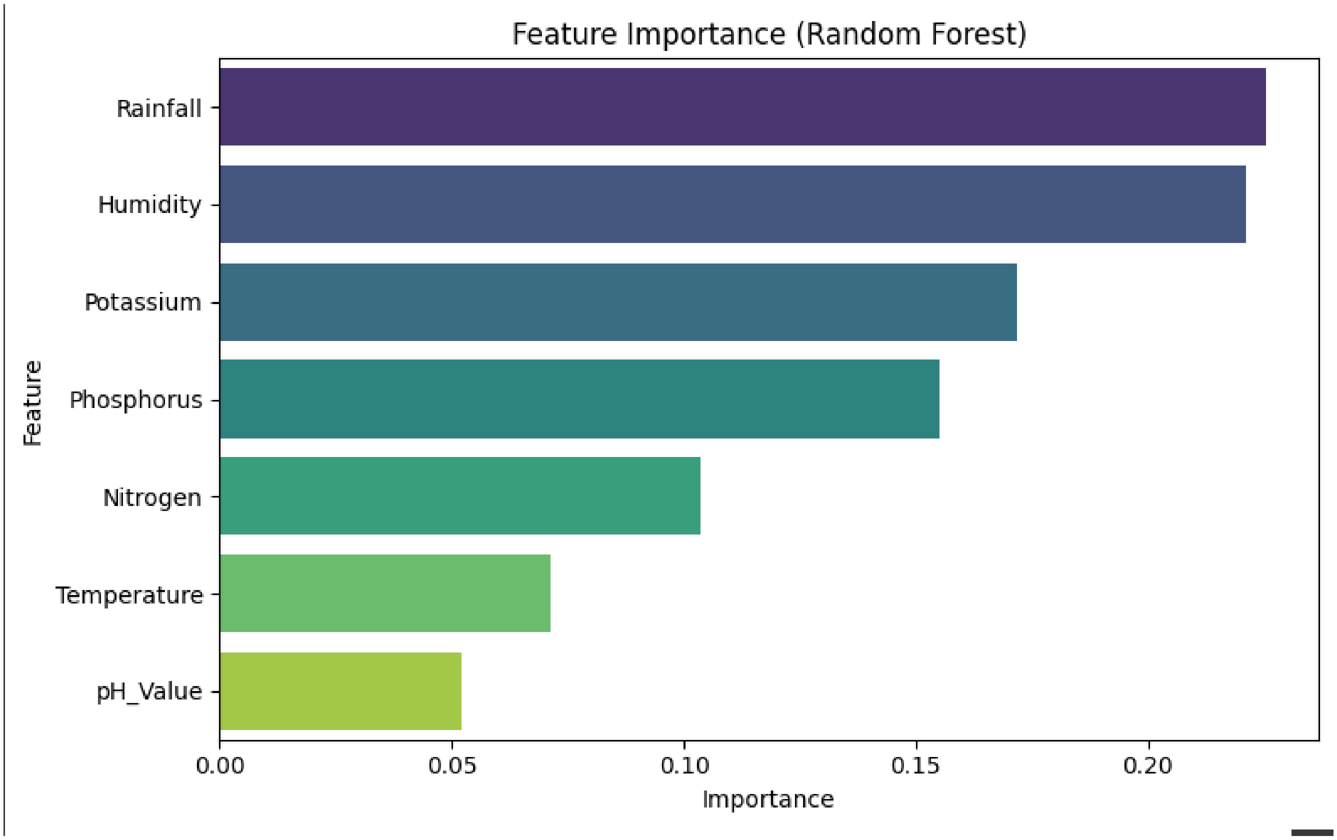

Fig 7. Feature Importance

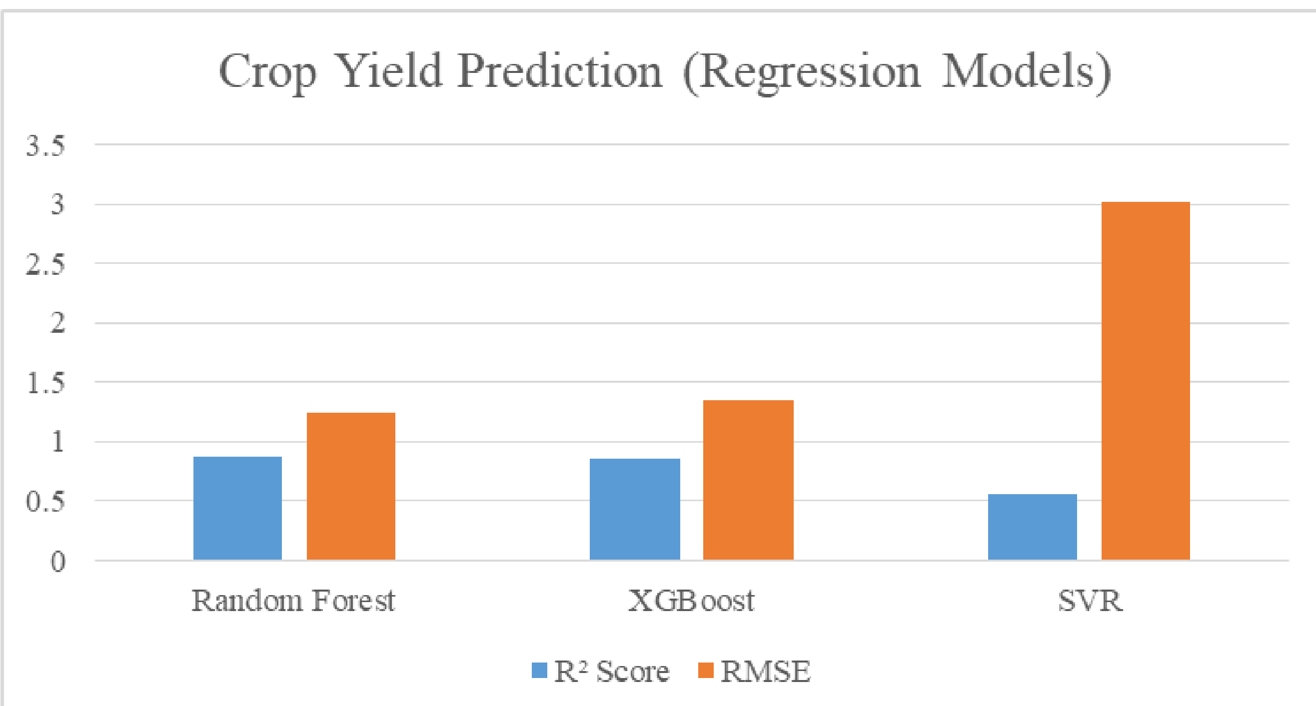

Fig 8. Crop yield prediction

and thus can be well utilized in real life on farms. The addition of the Soil Quality Index (SQI) unites several soil attributes into a single number to support the classification. With the help of ensemble and kernel-based models, the system can produce clear results. It enables the automation of crop suggestion, determination of soil type and prediction of the crop yield. The system could be expanded by introducing satellite images, Internet of Things instant soil sensors, an app to use in the field, systems for predicting upcoming pests and diseases and improved prediction of when and where crops will grow.

References

1) Das BS, Wani SP, Benbi DK, Sekhar M, Tapas B, Biswapati M, et al. Soil health and its relationship with food security and human health to meet the sustainable development goals in India. *Soil Security*. 2022;8:100071. https://doi.org/10.1016/j.soisec.2022.100071.
2) Chaudhry H, Vasava HB, Chen S, Saurette D, Beri A, Gillespie A, et al. Evaluating the Soil Quality Index Using Three Methods to Assess Soil Fertility. *Sensors*. 2024;24(3):864. https://doi.org/10.3390/s24030864.
3) Kavita J, Pratistha M, Sanchit J, Sukriti N. Crop Yield Prediction using Machine Learning and Deep Learning Techniques. *Procedia Computer Science*. 2023;218:406–417. Available from: https://www.sciencedirect.com/science/article/pii/S187705092300023. https://doi.org/10.1016/j.procs.2023.01.023.
4) Iniyan S, Varma VA, Naidu CT. Crop yield prediction using machine learning techniques. *Advances in Engineering Software*. 2023;175:103326. Available from: https://www.sciencedirect.com/science/article/pii/S0965997822002277. https://doi.org/10.1016/j.advengsoft.2022.103326.
5) Elbasi E, Zaki C, Topcu AE, Abdelbaki W, Zreikat AI, Cina E, et al. Crop Prediction Model Using Machine Learning Algorithms. *Applied Sciences*. 2023;13(16):9288. https://doi.org/10.3390/app13169288.
6) Fan J, Bai J, Li Z, Ortiz-Bobea A, Gomes C. A GNN-RNN Approach for Harnessing Geospatial and Temporal Information: Application to Crop Yield Prediction. 2021. https://doi.org/10.48550/arXiv.2111.08900.
7) Jones EJ, Bishop TFA, Malone BP, Hulme PJ, Whelan BM, Filippi P. Identifying causes of crop yield variability with interpretive machine learning. *Computers and Electronics in Agriculture*. 2022;192:106632. https://doi.org/10.1016/j.compag.2021.106632.
8) Kumar SDS, Divya SM. Smart Farming Using Machine Learning and Deep Learning techniques. *Decision Analytics Journal*. 2022;3:100041. https://doi.org/10.1016/j.dajour.2022.100041.
9) Harshit M, Divyanshi M. ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN AGRICULTURE: TRANSFORMING FARMING SYSTEMS. 2024. https://doi.org/10.5281/zenodo.15157536.
10) Dhruvi G, Chintal R, Rikin N, Hardik J, Axat P. Crop Recommendation System using Machine Learning. 2021;7(3):554–557. https://doi.org/10.32628/CSEIT2173129.
11) Anikó N, Miklós N. Crop Yield Prediction in Precision Agriculture. *Agronomy*. 2022;12(10):2460. https://doi.org/10.3390/agronomy12102460.
12) Sathya J, Asha V, Kalaivani A, Sangappa HS, Sarana F, Sachin KS. Soil Analysis for Crop Recommendation using Machine Learning Algorithms. 2025. https://doi.org/10.1109/IDCIOT64235.2025.10915185.
13) Devendra D, Pratik S, Koninika P, Vijay D. Smart Farming: Crop Recommendation using Machine Learning with Challenges and Future Ideas. *TechRxiv*. 2023. https://doi.org/10.36227/techrxiv.23504496.v1.
14) Ángel Luis PG, de Teruel Pedro E L, Alberto R, Ginés GM, Gregorio BG, J GCF. FARMIT: Continuous Assessment of Crop Quality Using Machine Learning and Deep Learning Techniques for IoT-Based Smart Farming. *Cluster Computing*. 2022;25:2163–2178. https://doi.org/10.1007/s10586-021-03489-9.
15) Nida I, Umair SM, M SES, Usman TM, Javed R, Tuan-Vinh L, et al. Analysis of Wheat-Yield Prediction Using Machine Learning Models under Climate Change Scenarios. *Sustainability*. 2024;16(16):6976. https://doi.org/10.3390/su16166976.
16) Soukaina H, Ayoub AL. XGBoost To Enhance Learner Performance Prediction. *Computers and Education: Artificial Intelligence*. 2024;7:100254. https://doi.org/10.1016/j.caeai.2024.100254.
17) Umberto M, Francesca V. New metric formulas that include measurement errors in machine learning for natural sciences. *Expert Systems with Applications*. 2023;224:120013. https://doi.org/10.1016/j.eswa.2023.120013.
18) Prity FS, Hasan MM, Saif SH. Enhancing Agricultural Productivity: A Machine Learning Approach to Crop Recommendations. *Hum-Cent Intell Syst*. 2024;4:497–510. https://doi.org/10.1007/s44230-024-00081-3.
19) Afzal H, Amjad M, Raza A. Incorporating soil information with machine learning for crop recommendation to improve agricultural output. *Sci Rep*. 2025;15:8560. https://doi.org/10.1038/s41598-025-88676-z.
20) 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. https://doi.org/10.1007/s44187-025-00338-1.
21) Sheeba TB, Anand LDV, Gunaselvi M, Saravana S, Wilfred CB, Muthukumar K, et al. Machine Learning Algorithm for Soil Analysis and Classification of Micronutrients in IoT-Enabled Automated Farms. *Journal of Nanomaterials*. 2022;p. 5343965–1–7. https://doi.org/10.1155/2022/5343965.