Predicting Agricultural Crops from Soil Features in Chitradurga Area

Raghavendra M. Y.
Department of CSE, Sri Siddhartha Academy of Higher Education, Tumkur, India
raghavendramy09@gmail.com (corresponding author)

H. S. Annapurna
Department of ISE, Sri Siddhartha Institute of Technology, Tumkur, India
annapoornahs@ssit.edu.in

Received: 3 June 2025 | Revised: 2 July 2025 and 12 July 2025 | Accepted: 16 July 2025

 DOI: https://doi.org/10.48084/etasr.12537

ABSTRACT

Using soil parameters to predict crops can greatly improve farming results, as they are important factors that affect productivity. This study used machine learning models to predict which crops will grow best in the Chitradurga District based on the type of soil. Characteristics such as pH, macronutrients (N, P, K), and some micronutrients were examined using a large dataset of soil samples that were collected by hand from the Agriculture Department and APMC soil testing laboratories in six Chitradurga taluks. This public dataset, which is specific to this area, is the basis for a new crop prediction system based only on soil properties. Data preprocessing involved cleaning, normalizing, and addressing class imbalance using ADASYN and SMOTE. ANOVA F-score-based feature selection aimed to determine the most important soil characteristics. Four machine learning models, namely XGBoost, Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN), were examined to determine the best for predicting crop suitability. The XGBoost model achieved the best results, with an accuracy of 95%. The results show that soil characteristics can be used to make reliable crop recommendations and that data-driven methods can greatly improve decision-making in agriculture. This study shows that machine learning can be used in precision agriculture in Chitradurga, laying the groundwork for future improvements, such as hybrid modelling and the use of remote sensing data to make crop predictions that are even more accurate and specific to the region.

Keywords-Chitradurga; PH; macronutrients; precision agriculture; ADASYN; SMOTE; XGBoost; random forest; Support Vector Machine (SVM); Artificial Neural Network (ANN)

I. INTRODUCTION

Agriculture is the main way most people in Chitradurga district, Karnataka, make a living, so it is an important part of the local economy. The district has a semi-arid climate and a variety of soil types, both having a great effect on how productive farms are. Soil properties, such as pH, nitrogen, phosphorus, potassium, and micronutrients, are very important to determine crop growth and yield, as different crops need different types of soil [1]. A study of 854 soil samples from all six Chitradurga taluks found that soil pH ranges from 6.0 to 8.0, nitrogen levels range from 46 to 400 kg/ha, potassium levels range from 94 to 885 kg/ha, and phosphorus levels are usually medium to high [2]. Most soils had enough zinc, but there were signs that they were lacking micronutrients such as copper, manganese, and boron [2]. These differences show how important it is to carefully analyse the soil when choosing crops and making farming plans.

Farmers' past planting habits and experience have long been the main sources of information for crop planning in the area. These methods may be culturally relevant, but they often do not take into account how soil fertility, crop physiology, and environmental variability affect each other [1]. Thus, many farmers continue to grow crops that they know, even when the soil is not suitable for them, lowering their yields. In addition, traditional yield prediction models that look at rainfall correlations or trends have been shown not to be good enough to capture the many different parts of agricultural systems, especially when the weather is unpredictable [1].

Machine Learning (ML) and Artificial Intelligence (AI) are used to solve problems in agriculture. ML methods, such as Support Vector Machines (SVM), decision trees, neural networks, and ensemble learning models, such as Random Forest (RF) and XGBoost, can model complicated nonlinear relationships between soil, weather, and crop performance [3-5]. When used with datasets that are specific to a region, these tools are very useful to recommend crops, analyze soil fertility, and predict yields [6-8]. However, these technologies are still not widely used in places such as Chitradurga. Previous studies have shown high accuracy (99.51%) in controlled settings [10-

11]. However, they cannot be used in the real world because they do not cover enough areas, the data is not good enough, and farmers do not know about them [9-16]. Many ML models also have class imbalance issues or lack good feature selection, which makes them less generalizable and harder to understand [17-19].

Due to these problems, this study aimed to develop a crop recommendation model based on soil and data for the Chitradurga district. The main goals are (i) to use manually collected real-world soil data to build a crop classification system, (ii) to test and compare the performance of different machine learning models, such as XGBoost, RF, SVM, and ANN; and (iii) to determine important soil features using ANOVA F-score–based feature selection. To improve the reliability of the predictions, the study also uses SMOTE to deal with class imbalance. This work distinguishes itself by using a manually curated region-specific dataset that was collected from APMC soil testing laboratories and checked by agricultural experts, which has been made public. The proposed system is a clear and useful solution for precision agriculture because it combines feature selection, ensemble models, and class balancing methods. This method is important not only for Chitradurga, but can help plan crops in other areas with similar climates and soils [20-21].

II. PROPOSED METHOD

Figure 1 shows the main steps of the proposed architecture.

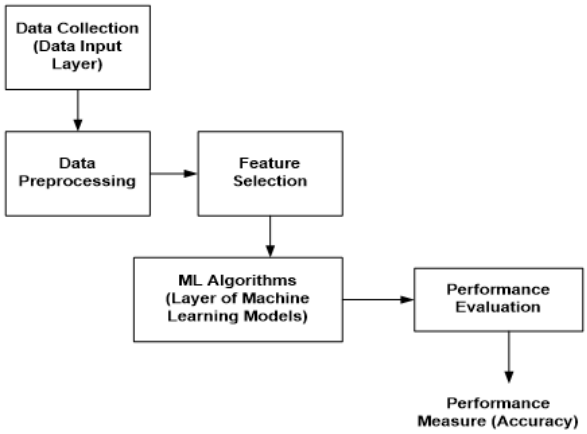

Fig. 1. Proposed architecture.

A. Data Input Layer

Soil data from the Chitradurga district is the main input to the data entry layer. An example of this is a table of soil samples, each containing some characteristics and a label that lists the best crop to grow there in the past or the yield that was harvested. The soil testing laboratories or records from the agriculture department provide the data. Analysts and farmers can use the interface to input updated soil test results and obtain predictions.

TABLE I. DATASET COMPOSITION BY CROP (AFTER DATA COLLECTION, BEFORE BALANCING)

Crop	No. of samples	Average yield (q/ha)
Finger Millet (Ragi)	280	18.5
Maize	220	35.0
Groundnut	180	15.2
Sunflower	150	12.8
Pigeonpea	120	10.4
Other minor crops	50	--

B. Data Collection and Preprocessing

The dataset was manually compiled from soil test reports collected by the Agriculture Department and APMC Chitradurga across all six taluks. The compilation was handled by the FDA, and crop labeling was verified by experts. The complete dataset and associated documentation are available [22]. The dataset has differences in data collection rates across several crop kinds, with finger millet and maize fields being sampled more often than pigeonpea fields, as shown in Table I.

C. Feature Selection

Having a balanced and clean dataset, feature selection aimed to find the soil parameters that best predicted crop success using the ANOVA F-test. For each characteristic, this statistical test calculates the variance between crop groups compared to the variance within each group. In essence, it examines to see if there is a significant difference in the mean values of a soil property across the different types of crops. If a feature has a higher F-score, it means that it can discriminate more effectively. Soil pH, for instance, had a high F-score, which means that some crops performed better in slightly acidic or slightly alkaline soils. In contrast, a characteristic having a low F-score (such as copper - Cu) means that its levels are rather constant across all crops.

TABLE II. FEATURE SELECTION (ANOVA F-TEST) RESULTS FOR KEY SOIL FEATURES

Feature	F-score	p-value	Selected?
pH	15.4	0.0003	Yes
Nitrogen (N)	12.7	0.0009	Yes
Phosphorus (P)	8.1	0.0052	Yes
Potassium (K)	3.5	0.047	Yes
Organic Carbon (OC)	2.8	0.062	No
Zinc (Zn)	1.1	0.356	No

The impact of pH, N, P, and K on crop performance was analyzed, focusing on the availability of nutrients and NPK. The study found that pH, N, P, and K were significant, while OC and Zn were marginal. Micronutrients such as Zn were not robust predictors due to the homogeneous use of micronutrient supplementation by farmers.

D. Experimental Setup

Model training and testing were conducted in a Python environment with 16 GB RAM and an Intel Core i7 CPU. For crop prediction classification, total accuracy, precision, recall, and F1-score were calculated. To simulate real-world use, the experimental setup included historical data and unknown data points. The train-test split was run several times with different random seeds to prevent data loss.

Algorithm: Soil-Based Crop Prediction and Yield Estimation
Input: Soil properties $X = \{x_1, x_2, x_3, \ldots\ldots x_n\}$
where each x_i represents a soil attribute (e.g., pH, Nitrogen, Phosphorus, etc.), crop label y_c,
training dataset $_{i=1}^{m}\{(X_i, y_{c,i}, y_{r,i})\}$.
Output: Predicted crop $\hat{y_c}$
Step 1: Data Preprocessing
Handle Missing Values:
For each feature x_i, impute missing values using mean imputation:

$$x_i = \frac{\sum_{j=1}^{m} x_{i,j}}{m} \quad (1)$$

if $x_{i,j}$ is missing
Alternatively, use median imputation for skewed distributions.
Outlier Detection and Removal:
Define upper and lower bounds using Interquartile Range (IQR):
Q1 = 25^{th} percentile of x_i,
Q3 = 75^{th} percentile of x_i
IQR = Q3-Q1 (2)
Outlier thresholds:
x_i< Q1-1.5 * IQR or x_i> Q3+1.5 * IQR
Remove outlier data points.
Feature Scaling:
Normalize numerical features using min-max scaling:

$$x_i^{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (3)$$

Alternatively, apply Z-score normalization:

$$x_i^{std} = \frac{x_i - \mu}{\sigma} \quad (4)$$

where μ is the mean and σ is the standard deviation
Step 2: Feature Selection
Statistical Feature Selection (ANOVA F-score for Classification):
Compute ANOVA F-score for each feature:

$$F = \frac{\sum_{c=1}^{k} n_c (\bar{x}_c - \bar{x})^2}{\sum_{c=1}^{k} \sum_{j=1}^{n_c} (x_{c,j} - \bar{x}_c)^2} \quad (5)$$

Select top k features with the highest F-scores.
Correlation Analysis (for Regression):
Compute Pearson Correlation Coefficient r_{xy} between soil feature x_i and yield y_r:

$$r_{xy} = \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_r - \bar{y}_r)}{\sqrt{\sum_{i=1}^{m} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{m} (y_r - \bar{y}_r)^2}} \quad (6)$$

Remove features with $|r_{xy}|$<0.1 (weak correlation)
Step 3: Handling Class Imbalance
Apply SMOTE):
Given a minority class sample x_i, generate synthetic points:

$$x_{new} = x_i + \lambda * (x_{nearest} - x_i), \lambda \sim U(0,1) \quad (7)$$

Balance class distribution in the training set.
Step 4: Model raining
Train machine learning classifiers to predict the best crop given soil properties:
i) RF classifier
ii) XGBoost classifier
iii) SVM classifier
iv) ANN classifier
Final Prediction:
Given new soil properties X_{new}, predict:
Crop predict: $\bar{y}_c = argmax\, f(X_{new})$ (8)

III. RESULTS AND DISCUSSION

A. Dataset Description and Exploratory Analysis

Examining 950 samples of soil characteristics and target crops, the study found 6 significant soil features and 5 classes. Soil parameters differed between crops. Finger millet (ragi) required lower pH and P and K levels, sunflower fields required higher sulfur levels, and maize fields required higher nitrogen and phosphorus levels. Focusing on the advantages of choosing the appropriate crop for the soil, this study gave crop selection first priority above production prediction.

B. Model Performance Analysis

The study evaluated four classification models on the soil dataset: XGBoost, RF, SVM, and ANN. Each model was chosen based on its advantages for classification tasks, particularly with tabular and imbalanced datasets. XGBoost is an optimized gradient-boosting framework that builds sequential decision trees using gradient descent, offering excellent performance on structured/tabular data. RF is an ensemble of decision trees using bootstrapped datasets and feature bagging, which is robust to noise and overfitting and well-suited for high-dimensional data. SVM with RBF kernel is effective in high-dimensional spaces and good for generalization on smaller datasets. ANN (MLP) is a feedforward neural network with one or more hidden layers trained through backpropagation, learning complex non-linear relationships and adapting to both classification and regression. The model's tuning includes hidden layer size, learning rate, number of epochs, and activation function.

TABLE III. PERFORMANCE OF DIFFERENT ML MODELS ON CROP PREDICTION (TEST SET)

Model	Accuracy	Precision (macro)	Recall (macro)	F1-score (macro)
XGBoost	0.95 (95%)	0.95	0.94	0.94
RF	0.92 (92%)	0.92	0.91	0.91
SVM (RBF)	0.88 (88%)	0.89	0.87	0.88
ANN (MLP)	0.90 (90%)	0.90	0.90	0.90

XGBoost achieved the highest accuracy at 95%, followed by RF at 92%, the ANN at 90%, and the SVM at 88%. XGBoost had a recall of 0.94 for all classes, with per-class accuracy and recall ranging from 0.92 to 0.97. Given their soil

profiles with unusually high P, the model's bias toward groundnut and millet appeared reasonable. Although oversampling and a well-balanced training set reduced this effect, the ANN model's performance indicated a little bias toward larger classes.

C. Visualization of Results

Figure 2 shows a bar chart of accuracy for each model. This visualization highlights the performance gap, with XGBoost at ~95%, RF ~92%, and so on.

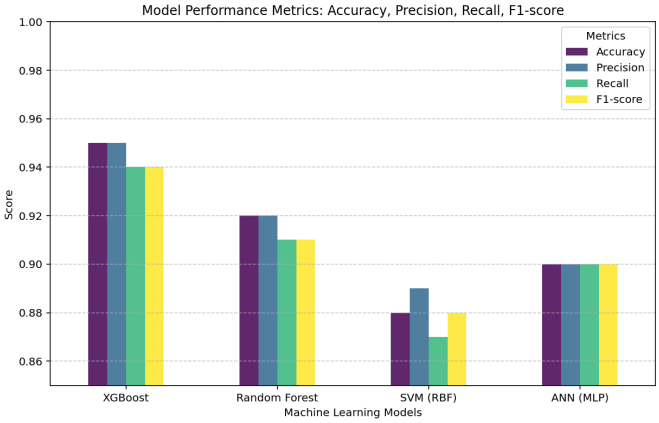

Fig. 2. Model comparison.

Figure 3 highlights a horizontal bar graph for the XGBoost listing of pH, N, P, K, S, etc., with their importance scores. This figure emphasizes pH and macronutrients as key factors in the model's decisions, consistent with domain expectations.

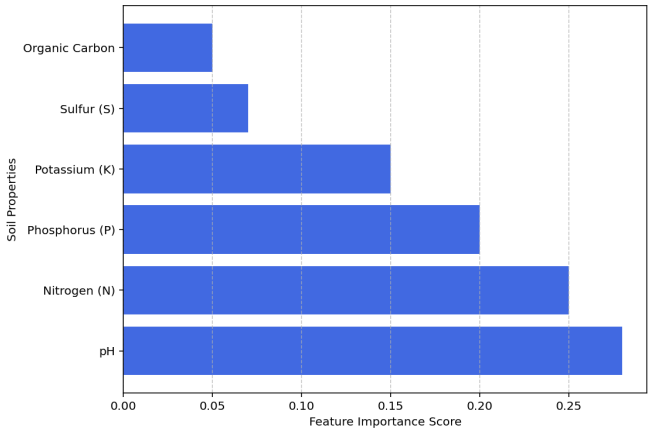

Fig. 3. Feature importance in the XGBoost model.

Figure 4 highlights a heatmap confusion matrix illustrating how often each predicted crop coincided with the actual crop. Most weight is on the diagonal (correct predictions), with only a few off-diagonals. For example, a small confusion can be noted between the Groundnut and Pigeonpea classes. This offers insight into which crops were occasionally misidentified.

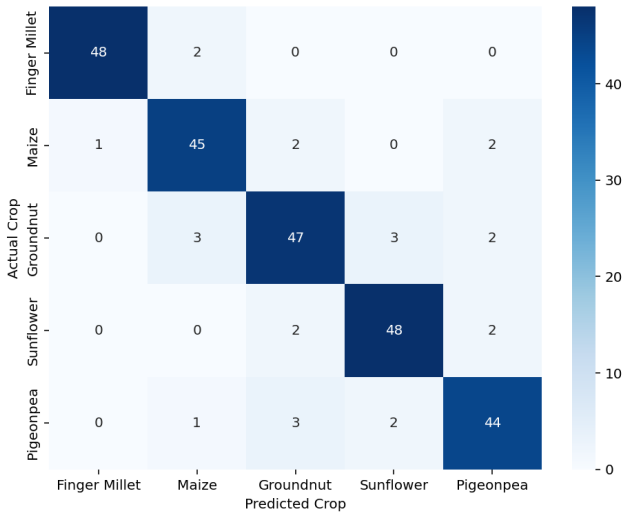

Fig. 4. Confusion matrix for XGBoost.

D. Comparison with Previous Studies

The study presents a novel approach to crop prediction using soil data, achieving an accuracy of 95%. The model's robustness is attributed to several factors, including ensemble methods (XGBoost/RF), data balancing, and feature selection. The focus on a specific region's data (Chitradurga) and tailoring the model to it could yield better performance than a more generic model. Comparing these results with recent approaches such as [1, 6, 19], among others, the XGBoost model shows promising results in terms of accuracy. Using self-taken or region-specific data allows for more accurate predictions in this context. Data specifically from the Chitradurga soils capture region-specific factors such as typical nutrient ranges and local crop varieties' performances. This highlights the value of local data collection for informed decisions. The integration of soil characteristics reaffirms previous research that ML can successfully leverage soil data to make better farming decisions. The results support ensemble tree methods (RF, XGBoost) as top performers in agricultural ML literature.

The use of SMOTE effectively addressed class imbalance, particularly improving the accuracy of prediction for minority crops such as Pigeonpea. Among all models, XGBoost yielded the best performance, highlighting its ability to handle tabular data and non-linear relationships in soil parameters. The novelty of the proposed method lies in combining region-specific data with robust ML strategies to provide actionable insights in precision agriculture.

IV. CONCLUSION

This study shows that it is possible and very useful to predict crops in the Chitradurga district only based on soil properties and identify the best crops for certain types of soils. XGBoost had the highest classification accuracy (95%) of all the models tested, outperforming RF, SVM, and ANN. The distinctiveness of this study lies in an integrated method that includes ANOVA F-test-based feature selection, class balancing with SMOTE, and advanced ensemble models on a

real-world dataset that was manually curated. The model's predictions fit well with what agronomists expect, so local farmers can trust its advice and understand it. This framework can improve traditional knowledge by giving farmers a tool for choosing crops based on evidence and helping them make better decisions, obtain better yields, and use resources more effectively. The study does, however, admit that there are some limitations, such as the fact that the dataset is only moderately large, it only looks at static soil properties without any changes over time, and it does not take into account any outside environmental or economic factors. The neural network model also looks good, but it could use more hyperparameter optimization.

In the future, this framework can be improved and expanded by adding remote sensing data, real-time weather data, and assessments of how economically viable it is. Future studies can also examine hybrid models along with dynamic learning systems that use farmer feedback, and combine them with fertiliser and crop management recommendation systems. Working with experts in the field can help make the system even better by adding rules and data to make it work better in rare or edge-case soil conditions.

ACKNOWLEDGEMENT

The authors express sincere gratitude to Sri. Kalilsab (District Deputy Director, APMC) and Sadhasivanna (APMC Secretary) for facilitating access to the soil test data from six taluks in the Chitradurga District. The authors also express special thanks to Pradeep (First Division Assistant, APMC) for compiling the soil test reports and to Dr. S. Onkarappa, Krishi Vigyan Kendra (KVK), for crop suitability insights and validation support. Their contributions were crucial in ensuring the quality and regional relevance of the dataset used.

REFERENCES

[1] G. Manju, S. Thomas, and V. A. Binson, "Enhancing Agricultural Productivity: Predicting Crop Yields from Soil Properties with Machine Learning," *African Journal of Biological Sciences*, vol. 6, no. 12, pp. 294–403, 2024.

[2] A. H. K. Naik, B. M. Madhu, H. G. Sannathimmappa, G. Madhu, K. N. Kumar, and N. G. Hanumantha, "Soil fertility status in Taluks of Chitradurga district under zero budget natural farming of Karnataka," *International Journal of Chemical Studies*, vol. 8, no. 3, pp. 2555–2558, May 2020, https://doi.org/10.22271/chemi.2020.v8.i3ak.9595.

[3] D. ManendraSai, S. Dekka, M. Rafi, M. R. D. Apparao, T. Suryam, and G. Ravindranath, "Machine Learning Techniques Based Prediction for Crops in Agriculture," *Journal of Survey in Fisheries Sciences*, pp. 3710–3717, Mar. 2023, https://doi.org/10.53555/sfs.v10i1s.814.

[4] L. Kouadio, R. C. Deo, V. Byrareddy, J. F. Adamowski, S. Mushtaq, and V. P. Nguyen, "Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties," *Computers and Electronics in Agriculture*, vol. 155, pp. 324–338, Dec. 2018, https://doi.org/10.1016/j.compag.2018.10.014.

[5] M. A. Ghorbani, R. C. Deo, M. H. Kashani, M. Shahabi, and S. Ghorbani, "Artificial intelligence-based fast and efficient hybrid approach for spatial modelling of soil electrical conductivity," *Soil and Tillage Research*, vol. 186, pp. 152–164, Mar. 2019, https://doi.org/10.1016/j.still.2018.09.012.

[6] A. Suruliandi, G. Mariammal, and S. P. Raja, "Crop prediction based on soil and environmental characteristics using feature selection techniques," *Mathematical and Computer Modelling of Dynamical Systems*, vol. 27, no. 1, pp. 117–140, Jan. 2021, https://doi.org/10.1080/13873954.2021.1882505.

[7] Y. Akkem, S. K. Biswas, and A. Varanasi, "Smart farming using artificial intelligence: A review," *Engineering Applications of Artificial Intelligence*, vol. 120, Apr. 2023, Art. no. 105899, https://doi.org/10.1016/j.engappai.2023.105899.

[8] S. A. Bhat, I. Hussain, and N.-F. Huang, "Soil suitability classification for crop selection in precision agriculture using GBRT-based hybrid DNN surrogate models," *Ecological Informatics*, vol. 75, Jul. 2023, Art. no. 102109, https://doi.org/10.1016/j.ecoinf.2023.102109.

[9] Q. Chen, L. Li, C. Chong, and X. Wang, "AI-enhanced soil management and smart farming," *Soil Use and Management*, vol. 38, no. 1, pp. 7–13, Jan. 2022, https://doi.org/10.1111/sum.12771.

[10] S. K. Apat, J. Mishra, K. S. Raju, and N. Padhy, "An Artificial Intelligence-based Crop Recommendation System using Machine Learning," *Journal of Scientific & Industrial Research*, vol. 82, no. 05, May 2023, https://doi.org/10.56042/jsir.v82i05.1092.

[11] M. Madhumitha and R. Ambikapathy "Soil Analysis and Crop Recommendation Using Deep Learning," *International Research Journal of Modernization in Engineering Technology and Science*, Aug. 2024, https://doi.org/10.56726/irjmets60753.

[12] S. Patil, D. Hajare, S. Gavhane, N. Panchal, S. Shelke, and A. Nikam, "AI-Driven Approach for Optimal Soil-Based Crop Recommendations," in *Advances in Intelligent Systems for Sustainable Agriculture*, 2025, pp. 221–237, https://doi.org/10.1007/978-981-97-9839-1_14.

[13] M. Chandraprabha and R. K. Dhanaraj, "Soil Based Prediction for Crop Yield using Predictive Analytics," in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, Greater Noida, India, Dec. 2021, pp. 265–270, https://doi.org/10.1109/icac3n53548.2021.9725758.

[14] E. Elbasi *et al.*, "Crop Prediction Model Using Machine Learning Algorithms," *Applied Sciences*, vol. 13, no. 16, Aug. 2023, Art. no. 9288, https://doi.org/10.3390/app13169288.

[15] M. Awais *et al.*, "AI and machine learning for soil analysis: an assessment of sustainable agricultural practices," *Bioresources and Bioprocessing*, vol. 10, no. 1, Dec. 2023, https://doi.org/10.1186/s40643-023-00710-y.

[16] V. A. J. Mahjenabadi *et al.*, "Digital mapping of soil biological properties and wheat yield using remotely sensed, soil chemical data and machine learning approaches," *Computers and Electronics in Agriculture*, vol. 197, Jun. 2022, Art. no. 106978, https://doi.org/10.1016/j.compag.2022.106978.

[17] S. Naimi, S. Ayoubi, J. A. M. Demattê, M. Zeraatpisheh, M. T. A. Amorim, and F. A. D. O. Mello, "Spatial prediction of soil surface properties in an arid region using synthetic soil image and machine learning," *Geocarto International*, vol. 37, no. 25, pp. 8230–8253, Dec. 2022, https://doi.org/10.1080/10106049.2021.1996639.

[18] S. Jain, D. Sethia, and K. C. Tiwari, "A critical systematic review on spectral-based soil nutrient prediction using machine learning," *Environmental Monitoring and Assessment*, vol. 196, no. 8, Aug. 2024, https://doi.org/10.1007/s10661-024-12817-6.

[19] S. R. Gopi and M. Karthikeyan, "Effectiveness of Crop Recommendation and Yield Prediction using Hybrid Moth Flame Optimization with Machine Learning," *Engineering, Technology & Applied Science Research*, vol. 13, no. 4, pp. 11360–11365, Aug. 2023, https://doi.org/10.48084/etasr.6092.

[20] F. Kaya, A. Keshavarzi, R. Francaviglia, G. Kaplan, L. Başayiğit, and M. Dedeoğlu, "Assessing Machine Learning-Based Prediction under Different Agricultural Practices for Digital Mapping of Soil Organic Carbon and Available Phosphorus," *Agriculture*, vol. 12, no. 7, Jul. 2022, Art. no. 1062, https://doi.org/10.3390/agriculture12071062.

[21] D. A. Reddy, B. Dadore, and A. Watekar, "Crop Recommendation System to Maximize Crop Yield in Ramtek region using Machine Learning," *International Journal of Scientific Research in Science and Technology*, pp. 485–489, Feb. 2019, https://doi.org/10.32628/ijsrst196172.

[22] M. Y. Raghavendra, "Raghavendramy09/crop-recommendation." Jul. 02, 2025, [Online]. Available: https://github.com/Raghavendramy09/crop-recommendation.