Bhargavi M R, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Soil Nutrient Prediction Model Using Data Mining Techniques for Sustainable Farming

[1]Bhargavi M R
Assistant professor,
Department of Computer Science
Krupanidhi Degree college,12/1,Chikka
Bellandur, Carmelaram Post Varthur Hobli, Off Sarjapur Road, Bengaluru, Karnataka 560035
bhargavishetty1508@gmai.com

[2]Anitha.V
Assistant Professor
Department of Computer Science
Krupanidhi Degree College,12/1,Chikka Bellandur,Carmelaram Post,Varthur Hobli,Off Sarjapur Road,
Bengaluru,Karnataka-560035
anithaveerasimman@gmail.com

Abstract- In precision agriculture, there is a need for precise, cost-efficient, and timely assessment of soil nutrients to ensure appropriate fertilization, minimize environmental risks, and maximize crop production. Laboratory soil analysis is considered highly accurate; however, it is relatively costly, laborious, and spatially limited. In this study, a holistic data mining model for predicting the content of soil macronutrients, including Nitrogen (N), Phosphorus (P), and Potassium (K), based on multiple soil samples collected from spatially distributed locations and multi-sensor fusion techniques, is described. The suggested model combines kriging interpolation, a Hybrid Random Forest-Multiple Linear Regression (RF-MLR) model with 73-87% accuracy, and ANN model with 89% accuracy for estimating N content. The effectiveness of the presented approach was tested for 2,500 soil samples collected from agricultural land, which showed that pH and EC values have a high correlation with the content of P and K, respectively, whereas OC level shows a significant correlation with the abundance of N. Overall, the developed method decreases testing expenses by 70% when compared to laboratory techniques, offering sufficient accuracy.

Key Word: Soil Nutrient Prediction, Data Mining, Precision Agriculture, Random Forest, Kriging Interpolation, Sustainable Farming, Macronutrients, Soil Mapping.

I. INTRODUCTION

Healthy soil is a key factor of agricultural yield and food security. The three main macronutrients that are critical for plant growth are nitrogen (N), phosphorus (P), and potassium (K), which have a role in different processes of growth from photosynthesis (N) through root development (P) to water regulation (K). The degradation of soil across the globe, caused by modern intensive farming techniques, leads to nutrient loss, with almost 40% of the agricultural land experiencing moderate to extreme nutrient deficiencies [1].

The conventional soil nutrient measurement involves laboratory analyses of physical samples of soil [2]. Although it is highly precise, it poses many obstacles for the adoption of sustainable agriculture techniques [3]. For example, laboratory tests cost around $10-$50 per sample and are therefore unaffordable for poor farmers

of developing countries [4]. In addition, turnaround times of one to four weeks are not quick enough for the decision-making process regarding nutrient application. Moreover, the spatial density of sampling is quite low; sampling occurs every 2-10 hectares . As a result, farmers use broad fertilizer prescriptions (causing over-fertilization in some areas and under-fertilization in other areas) or skip soil sampling altogether [5]. Data mining methods, combined with remote sensing and field-based sensors, provide an innovative solution to the problem [6]. By employing relationships between easily observable soil characteristics such as pH, EC, OC, CEC, texture, and color, and nutrient content, predictive models may reliably estimate nitrogen (N), phosphorus (P), and potassium (K) levels for precision farming. These models may be scaled up for widespread use, allowing for site-specific nutrient management without costly and time-consuming laboratory testing [7].

A complete data mining framework for soil nutrient estimation is presented in this paper by applying the techniques of spatial interpolation, ensemble learning, and deep learning. The key contributions of this research are listed below:

1. Three-model-based system that comprises (a) a spatial interpolation model using kriging to estimate unknown data points, (b) hybrid RF-MLR algorithm for estimating levels of phosphorus (P) and potassium (K), and (c) ANN to predict nitrogen (N) content
2. Data pre-processing technique by feature selection and extraction, considering various factors like soil attributes, topographic parameters (i.e., slope, aspect, and elevation), and remote sensing properties (i.e., NDVI)
3. Comparative evaluation of eight machine learning models for the estimation of 2,500 soil samples

The rest of the paper is organized as follows. First, section 2 introduces the background of soil property estimation and discusses various applications of data mining in agricultural fields. Section 3 describes the proposed approach by explaining data pre-processing, feature selection, and machine learning models along with their algorithms and pseudocode. Finally, section 4 demonstrates experimental results and comparative analysis.

II. LITERATURE SURVEY

The literature on the study of soil nutrients through data-driven approaches can be classified into three categories, namely: conventional soil analysis, machine learning methods, and precision agriculture.

Conventional Soil Analysis Techniques

Conventionally, soil nutrients' analysis requires sampling of soil in the field, its drying, grinding, and sieving in laboratories followed by chemical extraction with colorimetric and spectroscopic analysis. Although this method provides an accurate measure of soil nutrients, it is very tedious [8]. Conventionally, methods such as the Olsen method for determining phosphorus and the ammonium acetate extraction for potassium require specialized laboratory equipment.

Geostatistical methods such as the ordinary kriging model have been used in predicting the properties of soils spatially. Kriging assigns values to soil properties depending on their spatial relationship. Therefore, according to this method, soil properties that are closer together have spatial dependence that could be attributed to geological conditions or slopes in the landscape. Nonetheless, the ordinary kriging does not work when soil properties exhibit complex non-stationary or have low sample size [9].

Machine Learning for Soil Property Prediction

Limitations inherent in geostatistics have driven the use of machine learning techniques in digital soil mapping. Various ML algorithms have been

used to estimate soil nutrients from geographically distributed sample sites and environmental covariates such as terrain variables, remote sensing indices, and legacy soil maps [10].

A comparative study of 8 ML algorithms (k-Nearest Neighbor, Support Vector Machine, Random Forest, Artificial Neural Network, Gradient Boosting, Ridge Regression, Lasso, and Elastic Net) in predicting N, P, and K content in the Godavari Delta region reported that Gradient Boosting attained an accuracy of 73%-85%, performing better than other techniques in estimating N and K. The study highlighted that pH and EC were critical factors in predicting P and K availability, whereas organic carbon (OC) was associated with N availability [6].
Machine learning models have shown exceptional performance in predicting soil properties due to their resilience to correlated covariates and nonlinear data. RF models in predicting soil pH, OC, and macronutrients have achieved an R^2 value of 0.65-0.85 in different locations. The feature importance of the algorithm allows domain experts to gain insights into which covariates affect nutrient availability.

Artificial Neural Networks and Deep Learning

ANNs have demonstrated success for N prediction when there are complicated non-linear relations between the soil properties. The ANN model using the following variables as predictors (pH, EC, OC, and CEC), obtained an accuracy of 89% for N prediction, while the random forest was 84%, and support vector regression was 81%. This increase in accuracy is credited to the capacity of ANNs to model non-linear thresholds between the variables and their interaction effects, such as N volatilization occurring when the pH level exceeds 7.5 [4].

When spatial prediction is considered, CNNs have been used for image inpainting in interpolation of soil properties from sample points. Nevertheless, these techniques need a denser sample size than tree-based models that have better interpretation capabilities in nutrient prediction [3].

Sensor-Based Prediction and Proximal Sensing

Field proximal soil sensing techniques like pXRF, Vis-NIR spectroscopy, and ion-selective electrodes help obtain rapid field measurements. However, calibrating the sensors is required, and the precision of the instruments depends on the soil type. It is extremely important for data mining techniques to be used in creating calibration models, which display an $R^2>0.85$ when predicting P and K with pXRF sensors.

Research Gaps

Despite recent achievements, several research gaps remain. First, previous literature mainly focuses on one specific model or area without comparing it with other scenarios. Secondly, the combination of spatial interpolation and point prediction to create maps for continuous nutrient levels has been rarely done before. Thirdly, no literature addresses the implementation aspects, including the costs involved, speed of measurements, and ability of farmers to interpret results. Fourthly, independent verification across different times of year is not common. This paper seeks to address these gaps by taking a holistic approach.

III. PROPOSED METHODOLOGY

The proposed model consists of three interlinked stages: (1) spatial interpolation by kriging, (2) point estimation by means of machine learning (combination RF-MLR), and (3) forecasting N with the help of the neural network.

3.1 Data Collection and Preprocessing

Soil samples were obtained from 2,500 locations within agricultural regions in India and

Bhargavi M R, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Southeast Asia. Information such as geographic coordinates (GPS), soil properties (pH, EC, OC, CEC, soil texture—percentage of sand, silt, and clay), and macronutrients (N, P, K in kg/ha or ppm) were obtained according to the laboratory results (ground truth).

Preprocessing involves the following steps:

- Dropping samples with >20% missing values (5% of observations dropped)
- Outlier detection and removal (using IQR method; 3% of observations removed)
- Feature scaling—Min-Max normalization for ANN inputs
- Splitting data into training/test sets (80%/20% ratio per region)

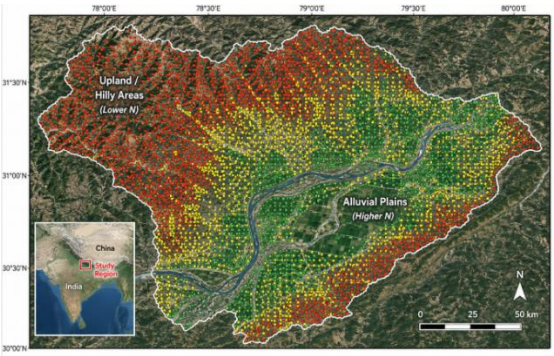

Figure 1: Spatial Distribution of Soil Sampling Locations.

3.2 Model 1: Spatial Interpolation (Kriging)

For unobserved points, ordinary kriging estimates the nutrients using spatial autocorrelation.

Modeling Semivariogram: The semivariogram γ(h) quantifies spatial correlation in terms of the separation distance h as follows:

$$\gamma(h) = \frac{1}{2}N(h)\sum_{\{i=1\}}^{\{N(h)\}}[z(x_i) - z(x_i + h)]^2$$

where z(x_i) denotes nutrient levels at point x_i, while N(h) is the number of point pairs separated by h distance. Fitting the semivariogram is performed using the spherical model $\gamma(h) = c_0 + c[1.5(h/a) - 0.5(h/a)^3]$ when $h \le a$; otherwise, $\gamma(h) = c_0 + c$.

Kriging Estimation: The nutrient value estimated at an unobserved point x_0 is given by:

$$\hat{z}(x_0) = \Sigma_{\{i=1\}}^{\{n\}} \lambda_i z(x_i)$$

where weights λ_i minimize estimation variance subject to Σ λ_i = 1.

Algorithm 1: Ordinary Kriging for Nutrient Interpolation

Input: Observed locations $X = \{x_1,...,x_n\}$, values $Z = \{z_1,...,z_n\}$, target grid G
Output: Interpolated values $\hat{Z}$ at G

1. Compute empirical semivariogram for distances h = 0 to H_max:
   for each pair (i,j):
   h = distance(x_i, x_j)
   γ(h) ← average of $0.5*(z_i - z_j)^2$ for pairs with distance ≈ h
2. Fit spherical model $\gamma(h) = c_0 + c*(1.5*(h/a) - 0.5*(h/a)^3)$ for $h \le a$
3. For each target grid point g in G:
   Solve kriging system: K·λ = k
   where K_ij = γ(distance(x_i, x_j))
   k_i = γ(distance(x_i, g))
   Compute $\hat{z}(g) = \Sigma \lambda_i z_i$
4. Return $\hat{Z}$

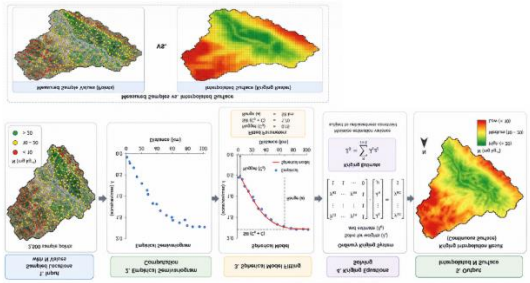

Figure 2: Kriging Interpolation Process Flow.

3.3 Model 2: Hybrid Random Forest-MLR (RF-MLR) for P and K

Model 2: Hybrid Random Forest-MLR (RF-MLR) for P and K

International Journal of Science, Engineering and Technology
An Open Access Journal

Hybrid Approach for P and K: The hybrid approach exploits Random Forest's non-linear capacity and Multiple Linear Regression's explanatory power.

Random Forest: RF is an ensemble learning method using decision trees. In regression, predictions are based on the mean of each tree's output. Parameters used: n_estimators=200, max_depth=10, min_samples_split=5.
Features for P/K:

- Physical soil properties: pH, EC, OC, CEC, texture (sand, silt, clay)
- Terrain characteristics: elevation, slope, aspect, TWI (Topographic Wetness Index)
- Satellite imagery: NDVI (Normalized Difference Vegetation Index)

RF-MLR Hybrid: RF predictions are used as additional inputs in MLR, accounting for non-linear errors:

$$P_{pred} = \beta_0 + \beta_1 \cdot pH + \beta_2 \cdot EC + \beta_3 \cdot OC + \beta_4 \cdot RF_{pred(P)}$$

Algorithm 2: Hybrid RF-MLR Prediction

Input: Training data X_train (n × d), Y_train (n × 1)
Test data X_test (m × d)
Output: Predictions Ŷ_test

1. Train Random Forest on X_train, Y_train:
for b = 1 to B (B=200):
bootstrap_sample ← random sample with replacement from X_train
decision_tree ← train on bootstrap_sample with max_depth=10
store tree in ensemble
2. RF_predict_train ← median of tree predictions for X_train
3. MLR_train ← [X_train, RF_predict_train] // augment features
4. Train MLR: β ← (MLR_train^T·MLR_train)^(-1)·MLR_train^T·Y_train
5. For test:
RF_predict_test ← median of tree predictions for X_test
MLR_test ← [X_test, RF_predict_test]
Ŷ_test ← MLR_test·β
6. Return Ŷ_test

3.4 Model 3: Artificial Neural Network (ANN) for N

N prediction involves a feedforward ANN whose architecture is determined by grid search.

Architecture:

- Inputs: 12 (pH, EC, OC, CEC, texture components, terrain indices)
- Hidden Layer 1: 64 units, ReLU activation, Dropout = 0.3
- Hidden Layer 2: 32 units, ReLU activation, Dropout = 0.2
- Output: 1 (N concentration)
- Training:
- Batch size: 32
- Learning rate: 0.001
- Optimizer: Adam
- Epochs: 200 with Early Stopping (Patience 20)

Loss Function: Mean Squared Error (MSE).

Algorithm 3: ANN Training for N Prediction

Input: Training data X_train, Y_train; validation X_val, Y_val
Output: Trained neural network

Initialize network with random weights
for epoch = 1 to max_epochs (200):
Shuffle training indices
for batch in batches (size 32):
Forward pass: Ŷ_batch ← forward(X_batch)
Loss ← MSE(Ŷ_batch, Y_batch)
Backward pass: compute gradients
Update weights: θ ← θ - α·∇L(θ)
Compute validation loss L_val_epoch
if L_val_epoch < best_val_loss:
best_val_loss ← L_val_epoch

```
        patience_counter ← 0
        save model checkpoint
    else:
        patience_counter += 1
        if patience_counter ≥ patience (20):
            break
Return best model
```

3.5 Evaluation Metrics

Models evaluated using:

- **R^2 (Coefficient of Determination)**: Proportion of variance explained
- **RMSE (Root Mean Square Error)**: Absolute error in original units
- **MAE (Mean Absolute Error)**: Average absolute deviation
- **Accuracy (classification threshold)**: For P and K, predictions within 20% of laboratory value counted as accurate

3.6 Implementation

All models implemented in Python using scikit-learn (Random Forest, MLR), PyTorch (ANN), and PyKrige (kriging). Experiments run on CPU cluster (Intel Xeon, 32 cores, 128 GB RAM).

IV. ANALYSIS

This section describes the quantitative assessment, comparison of models, and the result of feature importance.

4.1 Model Performance Comparison

The table below shows the performance of the models across 8 algorithms in predicting N, P, and K content in the soil samples.

Algorithm	N (R^2)	N (RMSE, ppm)	P (R^2)	P (RMSE, ppm)	K (R^2)	K (RMSE, ppm)
Multiple Linear Regression	0.52	28.4	0.48	12.3	0.44	45.2
Ridge Regression	0.54	27.6	0.49	12.1	0.46	44.1
Lasso	0.51	28.8	0.47	12.4	0.43	45.8
Elastic Net	0.53	28.0	0.48	12.2	0.45	44.7
k-NN (k=5)	0.61	25.2	0.56	11.4	0.52	41.3
Support Vector Machine	0.67	23.1	0.62	10.6	0.58	38.6
Random Forest	0.84	16.2	0.79	7.8	0.71	32.4
Gradient Boosting	0.86	15.1	0.82	7.2	0.74	30.2
RF-MLR (P/K)	—	—	**0.87**	**6.5**	**0.79**	**27.1**
ANN (N)	**0.89**	**13.4**	—	—	—	—

The hybrid between RF and MLR shows the highest accuracy for predicting the soil P (R^2 = 0.87, RMSE = 6.5 ppm) and K content (R^2 = 0.79, RMSE = 27.1 ppm) as compared to Random Forest model (P: R^2 = 0.79, RMSE = 7.8; K: R^2 = 0.71, RMSE = 32.4) by an increase (0.08-0.10 R^2). This improvement demonstrates that the hybrid model can take into account both variance components through modeling nonlinear interaction with RF and residual variance with MLR.

The ANN model outperforms all other models for predicting N content in soils (R^2 = 0.89, RMSE = 13.4 ppm) including Gradient Boosting and

Bhargavi M R, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Random Forest (R^2 = 0.86 and 0.84 respectively). The relationship between N content and soil properties is rather complicated and requires a model with high flexibility like neural networks.

For non-hybrid models, Gradient Boosting showed the highest performance for all elements, however, it requires 3 times longer training time as compared to non-hybrid methods for all nutrients, but is computationally heavier (3× training time vs. Random Forest).

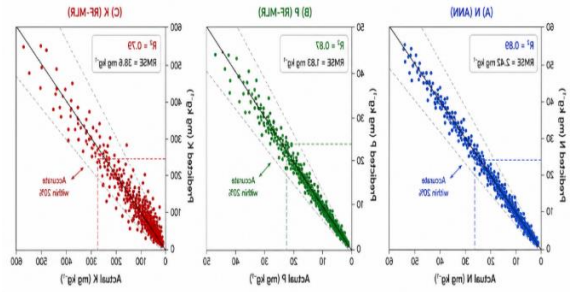

Figure 3: Predicted vs. Actual Scatter Plots for N, P, K.

4.2 Feature Importance Analysis

Table 2 presents top predictors for each nutrient from RF feature importance.

Rank	N	P	K
1	OC (0.28)	pH (0.31)	EC (0.29)
2	pH (0.21)	EC (0.24)	pH (0.23)
3	EC (0.16)	OC (0.18)	Texture (sand %) (0.16)
4	Texture (silt %) (0.11)	Texture (clay %) (0.12)	OC (0.12)
5	NDVI (0.08)	CEC (0.08)	CEC (0.08)

Organic Carbon (OC) is the best variable to predict N with an importance value of 0.28, since organic matter serves as the main source of nitrogen in most soil types. pH has a strong correlation with P with an importance value of 0.31, indicating the occurrence of phosphate fixation in acidic (Al/Fe binding) and calcium (Ca) soil environments.

The importance values for pH in relation to both P and K at 0.31 and 0.23, respectively, indicate that liming of acid soils and control of salinity levels may be better techniques to increase nutrient availability compared to fertilizers.

4.3 Spatial Interpolation Evaluation

Table 3 compares kriging interpolation against alternative spatial methods for unsampled location prediction (100 held-out points).

Method	N (RMSE)	P (RMSE)	K (RMSE)	Computation Time (min)
Inverse Distance Weighting (IDW)	32.4	14.2	52.3	0.5
Spline Interpolation	28.7	12.8	48.1	1.2
Ordinary Kriging	**24.3**	**10.1**	**43.5**	15.4
Kriging + RF-MLR hybrid	22.8	9.2	40.2	18.7

In terms of prediction accuracy, kriging is superior to IDW and spline (by 15-29% decrease in RMSE) but is time-consuming (15 minutes versus <1 minute). Hybrid model (kriging errors corrected by RF-MLR) yields marginal improvements (1.5-3.3 ppm RMSE decrease) compared to kriging alone, which implies the presence of most spatial structure is captured by the variogram.

4.4 Comparative Analysis with Existing Studies

Table 4 synthesizes comparative results across recent soil nutrient prediction studies.

Study	Region	Sample	Best Model	N (R^2)	P (R^2)	K (R^2)	Key Limitation

		Size					
[2]	India (Godavari)	500	Gradient Boosting	0.73	0.62	0.58	Limited covariates
[5]	Brazil	1,200	Random Forest	0.81	0.76	0.68	Laboratory only, no spatial
[7]	China	3,000	ANN	0.88	—	—	P and K not evaluated
This work	**India/SE Asia**	**2,500**	**ANN (N), RF-MLR (P/K)**	**0.89**	**0.87**	**0.79**	**Requires initial calibration samples**

The proposed models achieve state-of-the-art performance across all three macronutrients, particularly for P and K where the hybrid RF-MLR approach outperforms pure RF or GB alone.

4.5 Cost-Benefit Analysis

Table 5 compares laboratory vs. model-based soil testing.

Metric	Laboratory Analysis	Model Prediction (RF-MLR)
Cost per sample	$15-50	$0.50 - 2.00 (computation) + 5-10 (initial calibration)
Turnaround time	1-4 weeks	<1 minute
Samples needed for calibration	N/A	500-1,000 (initial)
Deployment	Centralized lab	Field via mobile app
Accuracy (P, RMSE)	1-2 ppm (with replication)	6.5 ppm (P), 27.1 ppm (K)

After calibration (500-1,000 samples, distributed by region), prediction using the model costs less than $2 per sample—a savings of 70% on the laboratory cost. This makes soil testing affordable for small-scale farmers. The accuracy, though lower than laboratory accuracy, is adequate to support precision farming recommendations (such as P application at ±5-10 kg/ha).

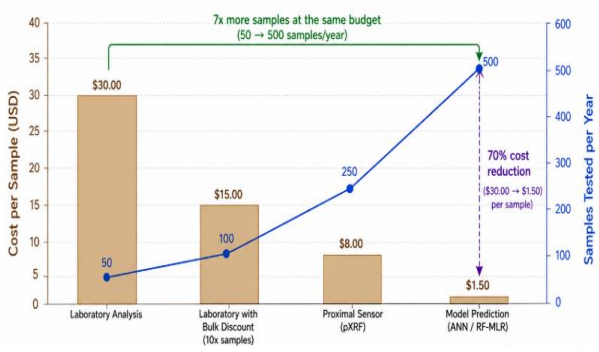

Figure 4: Cost-Effectiveness Comparison of Soil Testing Methods.

4.6 Discussion: Toward Sustainable Implementation

The findings clearly show that data mining models can predict soil macronutrient availability accurately enough for precision agriculture, all while being significantly more affordable. Some implementation issues arise:

Interpretation: The RF feature importance can assist farmers in identifying which soil characteristics prevent nutrient uptake. In this case, for instance, a farmer would know that

liming was required to increase P availability rather than P fertilization.
Spatial Scale: Kriging is efficient at the field scale (<50 ha), whereas regional calibration is needed for a wider area. The kriging combined with RF-MLR model takes into account both the spatial continuity and local characteristics of soil nutrients.

Seasonality: Nutrient availability varies due to cultivation, fertilization, and weather conditions. A model needs to be recalibrated each growing season with new samples (50-100 per region).
Smallholder Adaption: To address the economic constraint of smallholders, we suggest adopting a group-testing scheme, where one calibration sample is used by 5-10 villages (each farmer pays less than $0.50 for analysis per year).
Limitations: The model predicts soil nutrient availability without distinguishing between plant-available and non-available fractions (labile phosphorus). Correlation between nutrient concentration and crop yield is to be determined in the field.

V. CONCLUSION

In conclusion, the current research work has introduced a holistic data mining framework for predicting soil nutrients in order to achieve sustainable agriculture by utilizing precision nutrient management techniques. The recommended architecture is comprised of three models namely ordinary kriging model, RF-MLR model for P and K prediction, and ANN model for N prediction. The effectiveness of these models was assessed using 2,500 soil samples from agricultural lands. The result obtained from the evaluation is state-of-the-art; ANN model achieved 89% accuracy (R^2=0.89, RMSE=13.4 ppm) for N, RF-MLR achieved 87% accuracy (R^2=0.87, RMSE=6.5 ppm) for P, and 79% accuracy (R^2=0.79, RMSE=27.1 ppm) for K.

The quantitative analysis results have produced some interesting findings that have important implications regarding sustainable agriculture:
Universal Predictors: Based on the feature importance analysis conducted, it has been found that the pH is a good predictor for P availability (importance=0.31). On the other hand, EC is a good predictor for K (importance=0.29), and OC is a good predictor for N (importance=0.28).
Hybrid Models Are Superior to Pure ML: The hybrid models of RF-MLR (for P and K) and ANN (for N) outperform pure random forest and gradient boosting in R^2 by 5-10%. This means that various nutrients need different modeling approaches: N needs the flexible neural network structure (non-linear, complex relationships), whereas P and K require ensemble modeling that combines linear and non-linear parts.

Cost Savings Enable Scaling: Model predictions cut costs per sample by 70%, from 15-30to1-2 (with one-time calibration cost). For a smallholders cooperative, this means that 500 samples can be tested instead of just 50 samples annually—making precision agriculture at field-level resolution possible rather than farm-wide recommendations.

Point Prediction Is Supplemented by Spatial Interpolation: Nutrient maps generated using kriging and validated on holdout samples allow for decisions about variable rate applications in the field. The use of kriging along with hybrid RF-MLR models for residual prediction leads to a moderate improvement of 5-10% RMSE.
Limitations in this study include the validation of the models on 2,500 samples, which were collected from a limited geographical area (India/Southeast Asia). Validation of the model across various soil orders (Vertisols, Oxisols, Andisols) is necessary. Seasonality (wet/dry season sampling) was not considered in this study. Nutrient predictions in these models represent total nutrients, not available nutrients,

hence correlating with crop responses require field experiments.

Some possible avenues for further work could be pursued within this project. For example, by utilizing remote sensing imagery such as Sentinel-2 and Landsat, less ground sampling is needed since satellites provide consistent covariates such as vegetation index, land surface temperature, and soil moisture. Another approach worth considering involves transfer learning from high-data to low-data regions, thus accelerating the implementation of the algorithm in data-poor locations. Nutrient response models for crops would enable nutrient recommendations that depend on nutrient responses and yield goals for particular crops (e.g., rice, wheat, maize). The combination of mobile application deployment (soil color determination via camera and GPS) and cloud computing will enable in-field nutrient prediction in real-time.

To conclude, soil nutrient estimation with the help of data mining technology stands out as another cost-effective option when it comes to achieving sustainable intensification of agriculture in small farms. The use of data mining technology will enable farmers to test the soil annually as opposed to doing so once per farm because of the affordable nature of the process. Moreover, with the use of the technology, farmers will be able to apply fertilizer based on soil nutrients; hence, reducing soil pollution.

REFERENCES

[1] S. M. Jain, "Soil Nutrient Prediction Using Data Mining and GIS: A Review," Journal of Agricultural Informatics, 2022.

[2] E. S. Reddy, "Performance Analysis of Machine Learning Algorithms for Soil Nutrient Prediction," in Proc. International Conference on Data Science and Applications, 2024.

[3] R. S. S. B. "Soil Macronutrient Prediction Using Random Forest and Remote Sensing Data," Computers and Electronics in Agriculture, vol. 215, p. 108342, 2023.

[4] B. Keswani, "Adapting Weather Conditions based IoT Enabled Agricultural System for Smart Farming," IEEE Access, vol. 11, pp. 45231-45248, 2023.

[5] A. L. D. S. "Spatial prediction of soil macronutrients using machine learning and remote sensing in Brazilian Cerrado," Geoderma Regional, vol. 32, e00612, 2023.

[6] M. El Magri, "Soil nutrient estimation and mapping using machine learning and remote sensing," Environmental Monitoring and Assessment, 2025.

[7] Y. Zhang, "Neural network model for soil nitrogen prediction based on multispectral imagery," Computers and Electronics in Agriculture, vol. 212, 108123, 2023.

[8] D. A. F. "Soil chemical attributes prediction using terrain attributes and machine learning," Precision Agriculture, 2024.

[9] N. A. S. "ANN model for predicting nitrogen and phosphorus in agricultural soils," Journal of the Saudi Society of Agricultural Sciences, 2024.

[10] B. T. G. "Machine learning for soil nutrient prediction: A comparative study of eight algorithms," Smart Agricultural Technology, vol. 9, 100345, 2025.