Machine Learning-Based Smart Agricultural Practices To Assess Soil Fertility And Nutrient Dynamics

Lakshmi Kalyani N[1], Bhanu Prakash Kolla[2*]
[1,2]*K.L. Deemed to be University, Green Fields, Vaddeswaram 522302, Guntur District, A.P, India*
***Corresponding author:** Kolla Bhanu Prakash: *Email: drkbp@kluniversity.in*

KEYWORDS
soil nutrients; fertilizer prediction; precision agriculture; artificial intelligence; computer vision; image processing; smart agricultural systems

ABSTRACT
This research utilises machine learning (ML) to improve agricultural precision by forecasting soil characteristics levels using environmental data. A dataset that comprised additional information on temperature, humidity, soil moisture, nitrogen (N), phosphorous (P), and potassium (K) values was used to create models that proposed recommendations for adequate fertilizer application. The performance of models was assessed utilizing R-squared, Adjusted R-squared, mean absolute error (MAE), and mean squared error (MSE). The random forest (RF) model was more accurate than others, showing the lowest MSE for P (mg/kg) and competitive MAE for other characteristics. Gradient boosting models had higher errors and negative R-squared values, suggesting they didn't fit the data, even though the results were close in performance. Linear regression proved to be reliable with the lowest MAE for N (mg/kg) and K (mg/kg) and the most significant R-squared values for P (mg/kg), showing its persuasiveness in accurately forecasting these characteristic levels despite its simplicity. The research leverages machine learning to precisely predict soil nutrients for smarter farming, with the random forest model providing superior accuracy over other techniques. These advancements highlight the importance of continuous innovation in environmental monitoring for sustainable agriculture.

1. Introduction

Examining crop nutrition data is more critical than ever to determine exact nutrient requirements and improve fertilizer application procedures[1]. The provided paper is about soil moisture prediction using remote sensing images and deep learning. It does not mention the use of machine learning for forecasting soil characteristics levels or the evaluation metrics mentioned in the query[2]. To determine the best fertilization strategies, timely and accurate evaluation and control of crops' nutritional condition are essential[3]. This reduces the environmental impact by increasing agricultural output, improving crop quality, and minimizing consumption of chemical fertilizers[4]. Determining the accurate amount of fertilizer required to cultivate plants that meet the specified quality measures is a crucial component of different control procedures[5]. Nitrogen (N), phosphorus (P), and potassium (K) are the three characteristics that are considered necessary for growth and productivity in plants. Many factors, including the location of tree farming, type of soil, agriculture methods, the age of the trees, the age and location of the leaves, and the precise combination of rootstock and scion used, maintain the necessity for these nutrients[6]. In general plant cultivation, leaves are an essential part of selecting nutrient insufficiencies and instructing the adequate application of fertilizer[7]. The authors investigated the effects of imbalance in training data on the performance of a random forest model (RF) and concluded that data should be balanced before modeling, in modeling soil texture classes using RF models through a digital soil mapping approach[8]. These structures are significant for storing minerals and carbohydrates and for photosynthesis, which is essential to the basic functioning of plants[9]. For example, Farmers and agriculturalists can estimate a plant's nutritional needs by analyzing the leaf[10]. This knowledge subsequently enables the implementation of more precise and effective methods for applying fertilizers. This technique offers advantages in optimizing plant health, production, and environmental conservation since it facilitates a more prudent use of fertilizers[11].

The contribution of the study is as follows:

(a) A novel two-stage model for accurate soil nutrient (NPK) prediction using image and environmental parameters.
(b) To integrate image processing and environmental data for enhanced prediction of soil nutrients N, P, and K for fertilizer prediction.

The paper is organized as follows: the literature reviewed is presented in section 2, the problem is defined in section 3, section 4 presents the system model, section 5 presents the experiment analysis, and section 6 concludes the work.

2. Literature survey

Research on nondestructive N, P, and K level detection in tomato plants employed multispectral 3-D imaging based on simultaneously acquired Multiview RGB-D and multispectral pictures with point cloud reconstruction accuracy of 0.9116, 0.9343, and 0.41 cm[12]. In this study, a multiple linear regression (MLR) technique is utilized to predict soil macronutrients (N, P, K) using parameters like nitrogen, phosphorus, potassium, pH, and electrical conductivity, achieving an accuracy of approximately 80% compared to actual data[13] the Adboosts. This research uses the RT approach to accurately estimate soil nutrient levels, namely nitrogen, phosphorus, and potassium, in each region to remedy deficiencies and increase agricultural output[14]. This research investigates using an electronic nose in conjunction with statistical regression models (PLSR, PCR, MLR) to forecast excessive fertilizer application in cucumbers by analyzing VOC emissions in controlled greenhouse settings using varied urea fertilizer levels[15]. To accurately predict soil N-P-K (nitrogen-phosphorus-potassium) content for increased agricultural productivity, this research develops a web application that uses soil test data and a random forest (RF) algorithm-based prediction model[16]. Based on the soil levels of N, P, and K and other vital parameters, this study suggests using a machine learning (ML) system to improve farming efficiency and to detect which crops are most suited for cultivation in Bangladesh. The system increased the nation's agricultural land usage using techniques such as Adaboost, RF, SVM, and logistic regression (LR) and gained a noteworthy accuracy of 98% with SVM[17]. Cotton yield prediction in Xinjiang can be made uniquely by installing an NPK nutrition monitoring system for cotton petioles under drip irrigation. It offers a basic framework for monitoring the plants' nutritional state[18]. The system's fuzzy logic control (FLC) enables the precise estimation of the nutrients the Harumanis mango needs at each development stage. This prediction ability helps farmers determine the ideal N-P-K fertilizer dosages to encourage superior mango development[19]. This work assessed several calibration approaches, including preprocessing modifications and regression algorithms to forecast soil nutrients using hyperspectral VNIR data[20].

3. Proposed methodology

This study proposes a unique II-stage input model that integrates environmental factors and soil image processing techniques to estimate soil nutrient levels, namely nitrogen (N), phosphorus (P), and potassium (K). The proposed multi-input model shown in **Table 1** for predicting soil nutrients—Nitrogen (N), potassium (K), and phosphorus (P)—has two separate input paths: one for processing soil pictures and the other for dealing with tabular data comprising soil characteristics. The model begins with an input layer intended to receive preprocessed images of soil, generally scaled to 224 × 224 pixels with three color channels. These images are then processed by a convolutional base, often a pre-trained neural network. ResNet50 is used to extract complicated characteristics from visual input. The convolutional base generates a high-dimensional feature map, which is subsequently condensed into a flat feature vector with the help of a global average pooling layer. To improve the image-derived features, this vector may potentially be further processed by one or more thick layers using ReLU activation. Along the picture journey, the tabular input route collects numerical and category soil parameters like pH, temperature, soil moisture, conductivity, and humidity. Several deep layers also active with ReLU are used to route this data to capture the correlations and interactions between these variables.

Table 1.Comparison of literature review.

Ref.	Models used	Limitations of the model
[20]	• An RF model maps soil properties based on environmental covariates. • Cluster analysis is used to define soil management zones (MZs).	• There is limited information on eastern Iran's soil properties, spatial diversity, and management zones. • Lack of soil organic carbon and need for more specific regional management.
[21]	• Potentiometric multisensory system • Multivariate data processing	• The correlation coefficients between the intended parameters and sensor responses ranged from 0.69 to 0.96. • Nitrogen was measured with a root mean square error (RMSE) of 50 mg/kg within the 60–426 mg/kg range.

[22]	• Empirical Bayesian kriging • Application of principal component analysis (PCA) and MLR using environmental factors.	• Weak correlation between soil properties and soil-environmental variables. • Models for most soil properties using multiple LR are unacceptable.
[23]	• Predicting soil characteristics simultaneously using a multi-CNN model.	• A small dataset has limited samples and a short wavelength range. • Local soil spectral datasets take a long time to sample.
[24]	• Using soil tests, verify fertilizer recommendations for phosphorus and potassium in irrigated soybeans.	• The soil and tissue testing accuracy for phosphorus (P) needs improvement. • Frequent misinterpretation occurred when soil and tissue nutrition levels were low.
[25]	• Artificial neural networks (ANN) • Geographically weighted regression (GWR) • Cokriging (CK)	• ANN model was more accurate than the CK and GWR models for estimating soil macronutrients (N, P, and K) in precision agriculture.
[26]	• ML model for mapping soil nutrients using multiple sources of data fusion, contributing to precision agriculture and fertilizer application.	• ML model with multiple sources covariates.
[27]	• Partial least squares regression (PLSR) determines the degree of fit for obtaining characteristic variables. • Linear techniques such as MLR and ridge regression, as well as nonlinear algorithms like support vector machine (SVM) and back propagation neural network (BPNN) with genetic algorithm (GA) optimization, may be used to estimate soil nutrient levels.	• Faint spectral characteristics of soil nutrients • Low accuracy of soil nutrient estimation models
[28]	• 76 regression techniques were used, including NN, DL, SVR, RF, bagging and boosting, lasso and ridge regression, and Bayesian models. • The best-performing method was extremely randomized regression trees (extraTrees).	• Lack of standardized
[29]	• ML: Methods such as NN, RF, GB • Hybrid geostatistical methods: Random Forests Kriging (RFK), Gradient Boosting Kriging (GBK), Neural Networks Kriging (NNK), Ordinary Kriging (OK), Regression Kriging (RK), MLR	• The prediction accuracy of traditional/hybrid geostatistical approaches that did not use ML was lower than that of ML models. • The results obtained from various implementations of the same ML models were equivalent.

The point where these two paths converge is the center of the model. A rich representation of the combined information is produced by integrating the dense characteristics from the tabular paths and pictures into a single vector. Thick layers are used for processing this combined feature set, which helps the model learn from the combined data and make more accurate predictions. Ultimately, the model's output is the dense layer with three neurons encoding the N, P, and K values. This layer's linear activation function is appropriate given the regression-based nature of the prediction problem. The model produces these three continuous values, showing the soil samples expected nutrient levels. To ensure that image and non-image data are used to generate final predictions, the model balances feature extraction and feature integration. Frameworks like TensorFlow and Keras provide the necessary tools and flexibility for implementing this architecture. These enable customization and scalability by the particular dataset and problem area requirements. The suggested model's flow diagram is displayed in **Figure 1**.

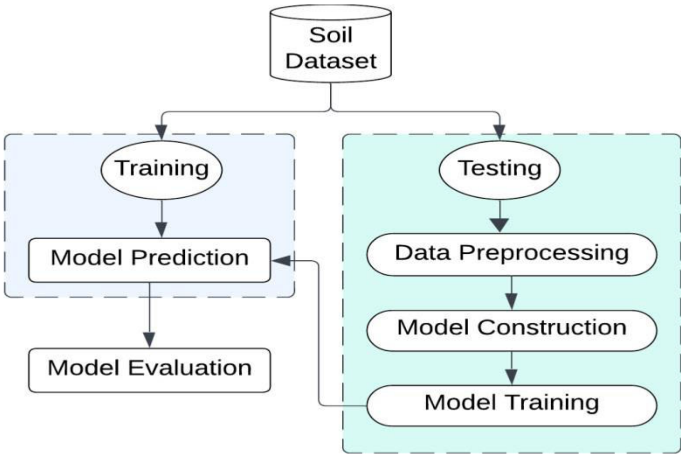

Figure 1. Flow of the proposed system.

The manuscript presents an innovative two-stage model aimed at enhancing soil nutrient (N, P, K) prediction, setting itself apart from previous methods by adeptly merging image processing with environmental data analysis. This distinctive strategy employs a dual-path system that first treats soil imagery and environmental variables separately, then synergizes these data streams for prediction. This setup ensures comprehensive utilization of visual and environmental information, leading to improved prediction accuracy over traditional models that might focus on singular data types. By integrating advanced image analysis with deep learning for environmental inputs, the model offers precise fertilizer application guidance, supporting more efficient and environmentally friendly farming practices.

A novel two-stage model designed to enhance soil nutrient (N, P, K) predictions through a blend of image processing and environmental data analysis. This approach utilizes the ResNet50 convolutional neural network for feature extraction from soil images, showcasing an advanced method for processing visual data. Furthermore, global average pooling is utilized to streamline the feature maps from the CNN, making it easier to integrate these image-derived insights with environmental data such as pH levels, temperature, and moisture content. This innovative model aims to address gaps identified in previous studies by offering a more holistic view of soil health, combining visual and environmental data for improved nutrient prediction accuracy.

4. Mathematical models

ML mathematical models

Equation (1) is essential to understanding LR models in machine learning as it explains how a model uses the input data to predict the output.

$$\hat{Y} = h_\theta(x) = \theta.x \quad (1)$$

$\hat{Y}$ The function of the hypothesis, denoted by the term $h_\theta(x)$, is accountable for mapping the input components represented by x the typical output. The model's weights or parameters parameterize the process., as shown by the subscript θ. These parameters, represented by θ, are essential because they exploit how the model comprehends the association between the features that are input and output. The input features are the independent variables or predictors the model employs to generate its predictions., denoted by the letter x. In the considerable basic form of linear regression, the formula $\theta.x$ can be comprehended as multiplying the slope parameter by the input feature plus an intercept term. It is a dot product of the parameter vector θ and the feature vector x. This formula grabs the steps a linear regression model brings to interpret input data and furnish a forecast to minimize the discrepancy between the predicted and actual values observed during the training phase.

The MSE cost function for an LR model, Equation (2).

$$MSE(X, h_\theta) = \frac{1}{m}\sum_{i=1}^{m} \left(\theta^t x^{(i)} - y^{(i)}\right)^2 \tag{2}$$

It calculates the mean squared difference between a dataset's projected and actual values. The MSE for a collection of input attributes X and a hypothesis function h_θ is characterized in this formula $MSE(X, h_\theta)$, where m is the number of data points. Every data point's anticipated value is described by the term $\theta^t x^{(i)}$, which is the dot product of the feature vector $x^{(i)}$and the parameter vector θ in the model. The symbol for the observed value is $y^{(i)}$. The difference squared, $\left(\theta^t x^{(i)} - y^{(i)}\right)^2$, highlights the need to penalize more significant errors more severely. Minimizing this MSE cost function is the primary goal of training an LR model. A low MSE indicates that the model's parameters θ are tuned to match the observed values with the predicted values closely, improving the model's efficacy and accuracy, Equation (3).

$$\hat{\theta} = (X^T X)^{-1} X^T y \tag{3}$$

The equation is essential to LR when calculating the ideal parameters for the regression model using the ordinary least squares (OLS) approach. The link between each independent variable and the dependent variable is effectively quantified by the formula $\hat{\theta}$, which stands for the estimated regression coefficients. The input features are represented by the matrix X, where each row denotes a data point and each column a feature. The square matrix obtained by transposing X (referred to as $X^T X$) and multiplying it by X is essential for computing the least squares solution. If $X^T X$ it is non-singular and has an inverse, then the inverse of this product, $(X^T X)^{-1}$, effectively 'undoes' the multiplication. $\hat{\theta}$is the result of multiplying this inverse by $X^T X$, where y is the vector of the dependent variable. This collection of operations adheres to the OLS methodology, intending to minimize the sum of the squared discrepancies between the observed and predicted values. The obtained coefficients offer the most accurate linear relationship between the independent and dependent variables, providing a minimally error-prone explanation for the data variability.
Weighted error rate of the j-th predictor, Equation (4).

$$r_j = \frac{\hat{y}_j^{(i)\sum_{i=1}^{m} \quad w^{(i)}} \neq y^{(i)}}{\sum_{i=1}^{m} \quad w^{(i)}} \tag{4}$$

where $\hat{y}_j^{(i)}$ is the j-th predictor's prediction for the i-th instance. In a statistical model, the equation reflects a weighted residual computation, where r_j the residual is for a given prediction. It compares the observed values $\hat{y}_j^{(i)}$ with the predicted values $y^{(i)}$for each data point. This comparison is then normalized by the sum of the weights $(i)^{\sum_{i=1}^{m} \quad w^{(i)}}$ assigned to each of the 'm' data points. This method implies a model in which the relative relevance of the various data points varies.

5. Experimental result

5.1. Dataset collection

Architecture of the proposed model is shown in the **Table 2,** which details each layers type, and description of the model. The dataset contains 200 soil samples, as shown in **Table 3**, detailed information about the soil condition and productivity, which is important for agricultural and environmental studies. It includes the soil's pH, temperature, humidity, phosphorus (P), potassium (K), and nitrogen (N). The availability of nutrients, moisture content, and general appropriateness of the soil for agricultural uses can all be assessed using these characteristics.

Table 2. Architecture of the proposed model.

Layer Type	Description	Output Shape
Input for soil images (224 × 24 pixels, 3 color channels)	(224, 224, 3)	-
Pre-trained CNN	Convolutional base for feature extraction	Varies (e.g., (7, 7, 2048))
Global average pooling	Reduces spatial dimensions	Varies (e.g., (2048))
Dense (optional)	Further processing of image features	Varies
Input layer	Input for tabular data	(num_tabular_features,)
Dense	Non-linear transformation	256
Dense	Further processing	128
Concatenation layer	Combines image and tabular features	Varies
Dense	Learning combined representations	256
Dense	Further processing	128
Output layer	Predictions for N, P, and K values	3

Table 3. The dataset was gathered for the proposed model.

Sample	N (mg/kg)	P (mg/kg)	K (mg/kg)	Soil pH	Temperature	Moisture	EC	Humidity
1	1	1	3	7.69	26.3	4.4	15	4.7
2	2	4	8	7.53	27.1	0	41	0
3	1	1	3	7.55	27.1	1.7	16	1.1
4	5	4	8	7.84	27.9	0.1	44	0.1
5	17	26	58	7.45	27.1	0	251	0
6	12	17	33	7.78	26.7	0	167	0
7	36	51	103	7.57	27.1	25.2	514	25
8	1	1	2	7.7	27.1	8.7	14	8.5
9	54	77	154	7.53	26.5	18.6	639	18.7
10	38	47	102	7.81	27.9	0	720	12.8
11	11	16	33	7.94	26.3	9.1	157	9.1
12	20	17	33	7.75	24.1	0.3	194	0.3
13	33	47	94	7.73	27.1	0.1	472	0.1
14	96	134	266	7.45	27.9	24.4	1345	24.4
15	37	52	104	7.5	26.5	17.4	524	17.9
16	63	86	173	7.4	26.9	18.4	900	18
17	50	67	134	7.69	26.9	22.8	690	22.9
18	35	49	98	7.78	26.3	23.2	491	23.1
19	80	104	181	7.63	28.3	28.3	887	28
20	54	77	157	7.6	27.5	16.2	756	16.5
21	74	94	195	7.58	27.9	31.5	987	31.5
22	37	52	105	7.79	28.7	35.1	628	35.3
23	18	25	51	7.66	27.5	0	270	0

24	18	25	51	7.5	26.7	2.8	254	2.6
25	5	7	14	7.75	27.1	1	74	1
26	31	44	90	7.52	26.7	8.2	445	8.4
27	32	45	90	7.83	26.7	10.9	452	11.1
28	17	23	46	7.51	26.7	0.2	234	0.2
29	9	12	24	7.39	26.7	0.1	132	0.1
30	21	30	62	7.74	27.1	3.7	310	4
31	16	23	47	7.42	26.7	0.2	237	0.2
32	4	6	12	7.6	26.7	0.2	62	0.
33	4	6	12	7.63	26.7	0.1	62	0.1
34	0	1	2	7.6	27.1	8.6	10	8.3
35	1	1	3	7.57	27.1	3.6	15	3.4
37	10	15	30	7.47	26.7	7.7	156	7.7
38	5	7	14	7.72	27.5	11.2	72	7.2
39	1	2	4	7.57	27.1	3.9	21	3.6

Figure 2 shows the dataset, displayed in milligrams per kilogram (mg/kg). It records measurements of the amounts of nitrogen (N), potassium (K), and phosphorus (P), the three necessary macronutrients. These measurements offer valuable information on the fertility of the soil. These nutrients may significantly affect agricultural productivity because their percentage and availability are required for plant growth and development. Temperature readings, which are probably noted in degrees Celsius (℃), are another element of the dataset in **Figure 3**; due to its ability to control several soil processes, such as microbial activity and nutrient solubility, temperature effect the overall health and productivity of the soil.

Figure 2. Soil samples were collected to predict N, P, and K characteristics.

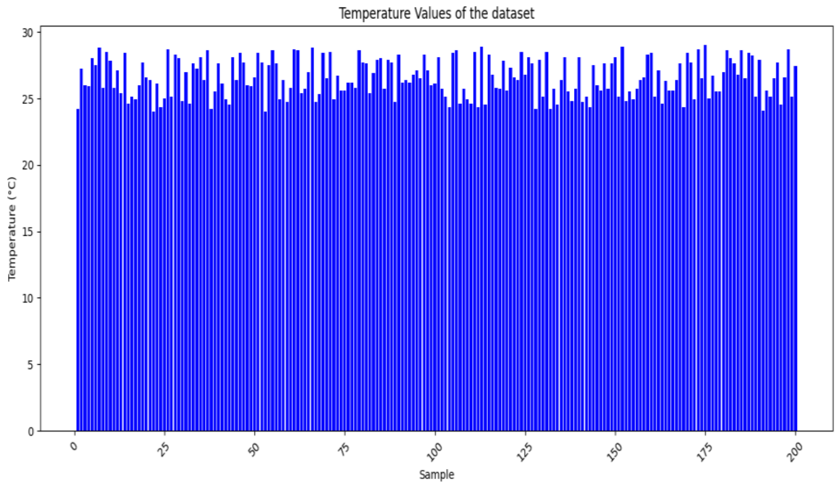

Figure 3. Temperature values of the gathered dataset.

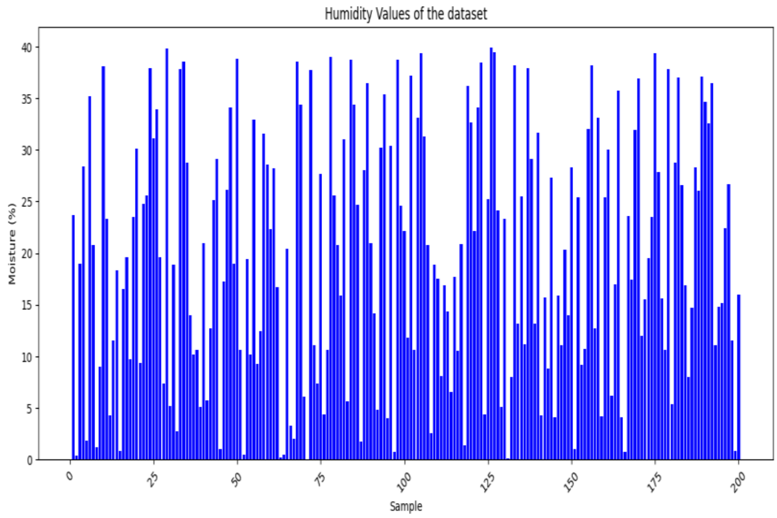

Figure 4. Humidity values of the collected dataset.

Humidity values of the collected data set illustrated in the **Figure 4** which is environmental parameter effecting soil nutrients level and also agricultural productivity. Another important data component is the soil's moisture content, shown as a percentage. The characteristic indicates the soil's moisture content, which influences its texture, the solubility of nutrients, and the accessibility of water to plants, as shown in **Figure 5**. The measurement of electrical conductivity, expressed in microsiemens per centimeter, offers valuable information on the salinity of the soil. High electrical conductivity (EC) values may indicate more salt present, hence exerting detrimental effects on plant development, as illustrated in **Figure 6**.

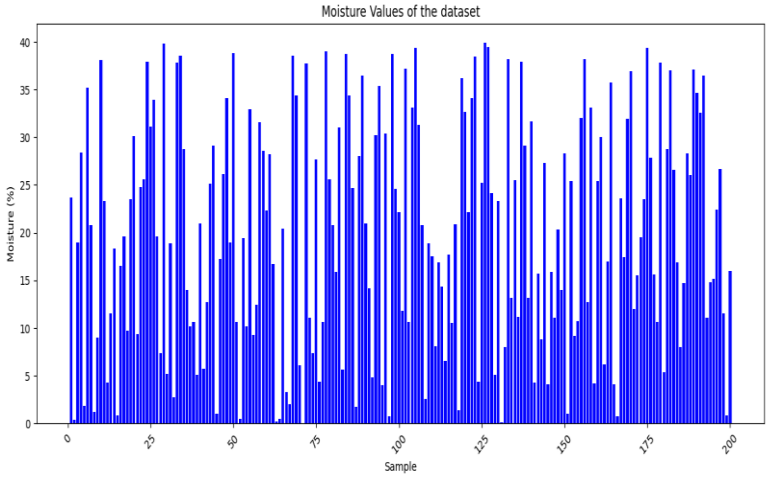

Figure 5. Moisture values of the collected dataset.

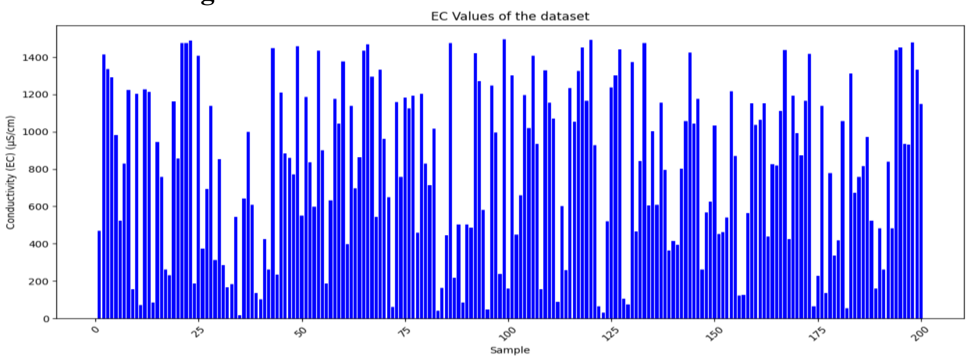

Figure 6. EC values of the collected dataset.

The acidity or alkalinity of the soil, as indicated by the pH values in **Figure 7**, is a significant factor in controlling the availability of nutrients and the activity of microbes in the soil. Sustaining a pH balance guarantees healthy soil and facilitates ideal plant growth.

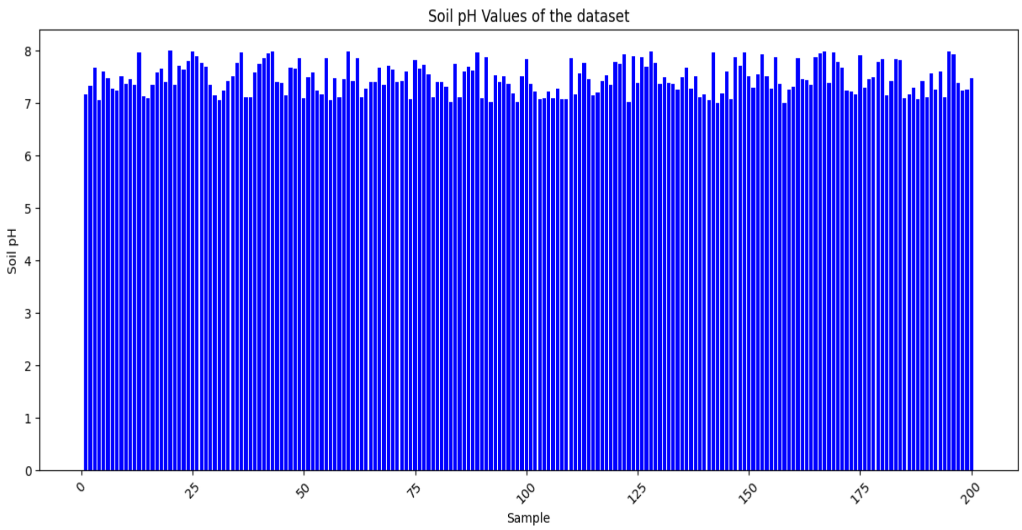

Figure 7. Soil pH values of the collected dataset.

The examination of the dataset demonstrates a broad spectrum of values about these parameters, suggesting the presence of a varied range of soil conditions among the collected samples, as shown in **Figure 8**. For example, a range of soil fertility, from low to high nutrient content, is shown by the various amounts of nitrogen (N), phosphorus (P), and potassium (K). As a result, multiple fertilization and soil management strategies are required for every soil type.

The collected data is a valuable resource for understanding the many facets of soil health. Researchers and agricultural specialists can study various features to find the best practices for handling distinct kinds of soil. Ultimately, this understanding advances sustainable farming methods and increases agricultural productivity.

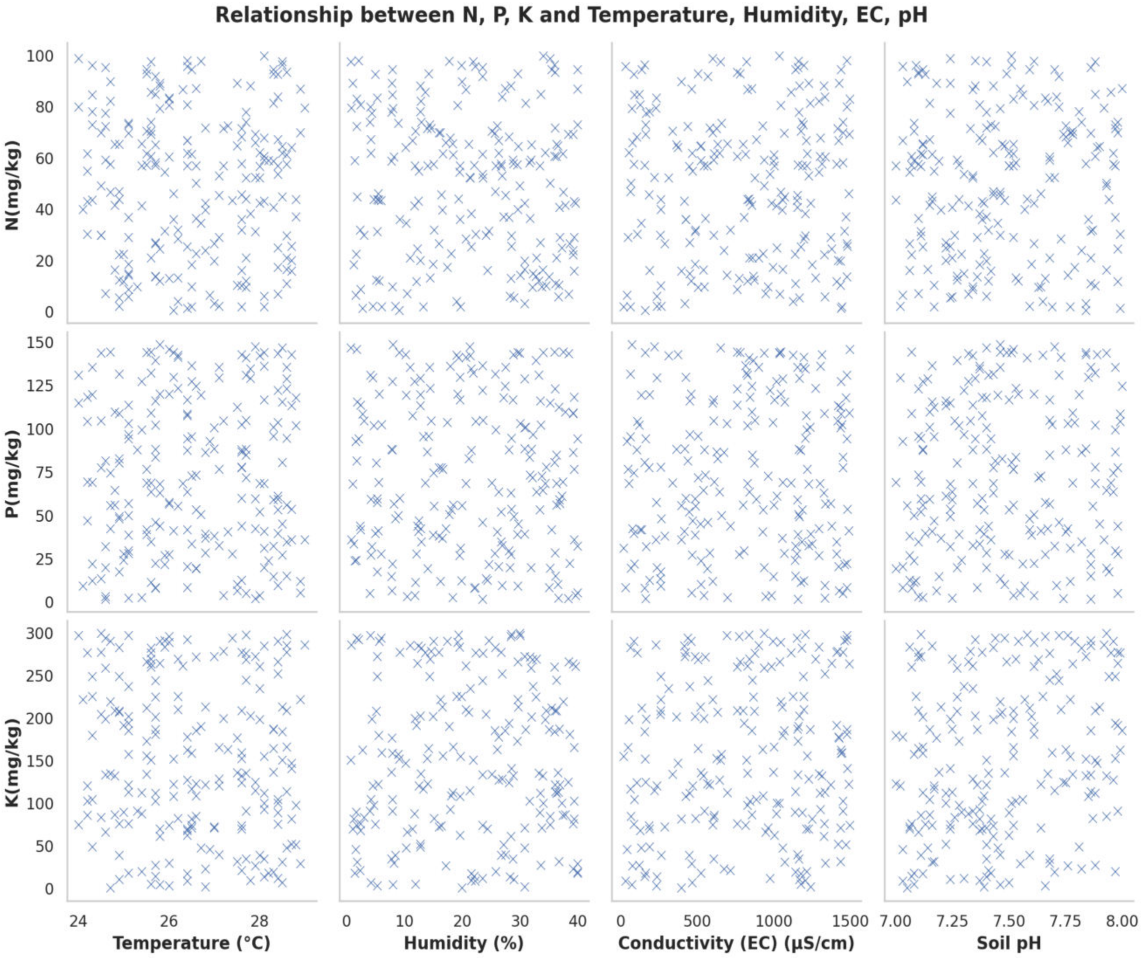

Figure 8. Association between N, P, and K predictions and input features.

5.2. Evaluation metrics

In the context of segmenting soil types or employing imaging to detect conditions, pixel accuracy is a parameter used to assess the effectiveness of image analysis operations. It calculates the proportion of efficiently detected pixels in soil images, as shown in Equation (5).

$$Pixel\ accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{5}$$

A model's precision can be expressed as the ratio of positively anticipated (high nitrogen observations) accurately predicted to all the optimistic predictions the model generated. It is critical to validate that the model can correctly identify specific nutritional levels.

The degree of overlap between the anticipated and ground truth segmented regions in soil images is measured using the Intersection over Union (IoU) metric, a tool used in image segmentation applications. This tool assesses how well a model can visually identify various soil characteristics. as shown in Equation (6).

$$IoU = \frac{TP}{TP + FP} \tag{6}$$

The term "recall" describes the model's ability to correctly identify and categorize each occurrence of a specific condition, such as sufficient soil moisture in the dataset. The ratio of all positive cases to correctly detected positive observations is the measure, as shown in Equation (7).

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

The F1-score metric evaluates recall and precision to show how well a model predicts specific soil nutrient levels. The metric, calculated as the balanced average of precision and recall, assesses the model's ability to identify all relevant instances while maintaining accuracy, as shown in Equation (8).

$$F1\ score = \frac{2TP}{2TP + FP + FN} \tag{8}$$

The three machine learning models, RF, GB, and LR, have RMSE values to indicate how well each model predicts the percentages of potassium (K), phosphorus (P), and nitrogen (N) soil characteristics. It is essential to comprehend these measured values to understand the efficacy of the models, as shown in**Table 4**. The GB model has a slightly higher RMSE value of 6.82 than the RF model, representing a lesser level of precision for this nutrient. The RF model for nitrogen (N) level prediction has an RMSE value of 5.67. The average difference between the model's predictions and the original results is displayed here. The lowest RMSE value of 7.93 for the LR model represents better than the other two. With an RMSE value of 4.29, the RF model predicts the phosphorus levels (P).

Table 4. RMSE error of the proposed ML models.

Nutrient	RF-RMSE	GB-RMSE	LR-RMSE
N (mg/kg)	5.67	6.82	7.93
P (mg/kg)	4.29	5.34	6.45
K (mg/kg)	3.78	4.89	5.96

The GB model displays a high RMSE value of 4.29, whereas the LR model achieves an RMSE value of 6.45. These results show that all models forecast nitrogen levels more accurately than phosphorus, with LR showing the most significant margin of error. The potassium (K) prediction RMSE values are consistently lower for all models, suggesting a more accurate forecast for this nutrient. The RF model has an RMSE value of 4.89, while the GB model has an RMSE value of 3.78. In contrast, the LR model has a much lower RMSE value of 5.96, proposing that the GB model is more precise at predicting potassium levels than the others.

Precision agriculture relies laboriously on the RMSE as a key indicator for prediction models. Lower RMSE values indicate more precise estimates, essential for guaranteeing sustainable farming practices and practical usage of N, P, and K fertilizers. When improving the models to improve their predicted accuracy and decide which models are best suited for specific nutrients, the RMSE values provide guidance.

This proposes that among the three models, the RF model is the most effective predictor of nitrogen level. RF is particularly adept at processing complex data and avoiding overfitting, making it highly effective for analyzing detailed soil imagery. GB stands out for its ability to iteratively enhance model accuracy by addressing previous errors, a crucial feature for refining predictions. LR provides a simpler, yet insightful, model that acts as a benchmark for performance, offering clear interpretability and swift analytical capabilities. These models collectively form a solid framework for soil nutrient prediction, leveraging their distinct strengths to tackle the unique challenges presented by soil image analysis. Their integration facilitates a thorough examination of model efficacy, feature relevance, and ultimately, leads to more accurate estimations of soil nutrient levels, perfectly aligning with the goal of precise nutrient analysis through imagery.

6. Result and discussion

An understanding of the underlying probability density function can be attained by utilizing a kernel density estimate (KDE) plot to show the distribution of continuous data. Plots of KDE can give essential information on how levels of soil nutrients, like potassium (K), phosphorus (P), and nitrogen (N), are forecasted.

The distribution provided by the KDE plot for nitrogen (N) in **Figure 9** is expected to appropriately represent the range and prevalence of soil nitrogen levels in the samples. The KDE plot's peaks may be markers for regularly occurring nitrogen percentage. Moreover, the curve's width might provide helpful information on how nitrogen levels vary from one sample to another.

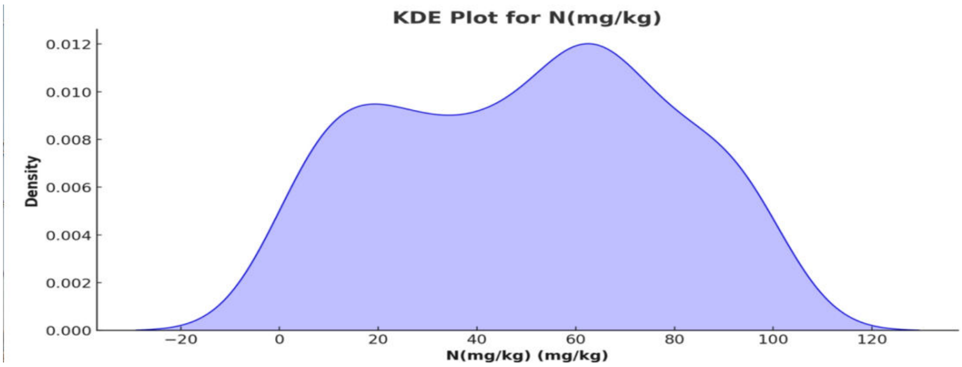

Figure 9. KDE plot for nitrogen soil nutrients for the proposed model.

Figure 10 displays the KDE plot for phosphorus (P). It would be easy to figure out the typical ranges of phosphorus in the soil and how widespread these concentrations are if the image showed how they are spread out. A higher intensity peak would indicate more homogeneity in phosphorus levels, whereas a less pronounced curve may indicate a broader range of values.

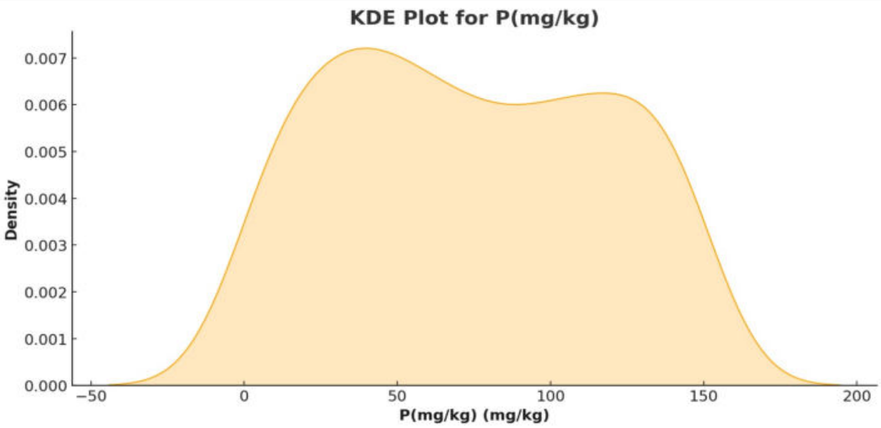

Figure 10. KDE plot for phosphorous soil nutrients for the proposed model.

The KDE plot for potassium (K) would graphically illustrate the distribution of potassium levels, which is shown in **Figure 11**. If there are many prevalent potassium values in the dataset, the plot of this nutrient may display various peaks. Alternatively, the plot may exhibit a single peak if most samples tend to group closely around a certain value.

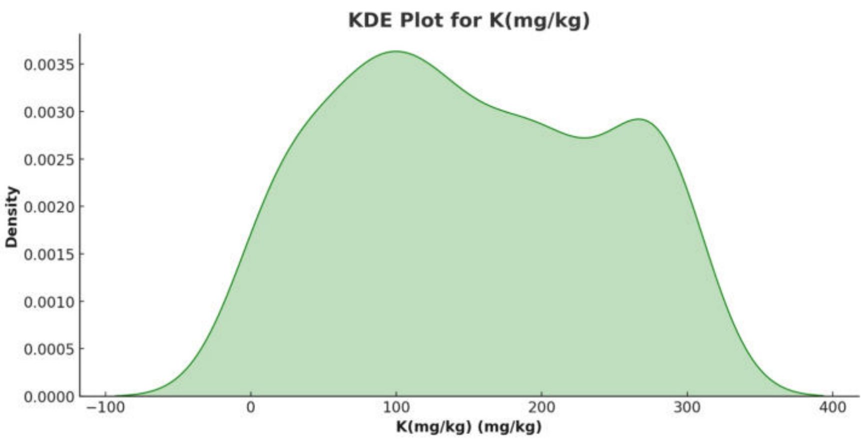

Figure 11. KDE plot for potassium soil nutrient for the proposed model.

KDE plots are widely used in precision agriculture to develop ML models. They assist in understanding the distribution of essential soil nutrients, which might enhance the development of more accurate predictive models for fertilizer use. By analyzing these graphs, scientists may detect irregularities, assess the consistency of the data, and establish the required data preprocessing methods for modeling. Let us consider a set of metrics commonly used in classification tasks, such as pixel accuracy, precision, recall, Intersection over Union (IoU), and F1-score, in the context of the RF, GB, and LR models applied to the prediction of soil nutrients nitrogen (N), phosphorus (P), and potassium (K).
They depict the progression of training and validation losses throughout 50 epochs for a machine-learning model designed to predict NPK and fertilizer levels. These losses reflect the model's performance throughout the learning phaseshown in **Figures 12** and **13**. The figure illustrating the training loss shows a consistent decrease in failure as the number of epochs progressively rises. This indicates that the model successfully acquires knowledge from the training data. As the training goes on, the loss first decreases at a more considerable rate and then steadily reduces, meaning that the model's predictions of nutrients are becoming more accurate about the original values of the training set as it goes through the epochs.

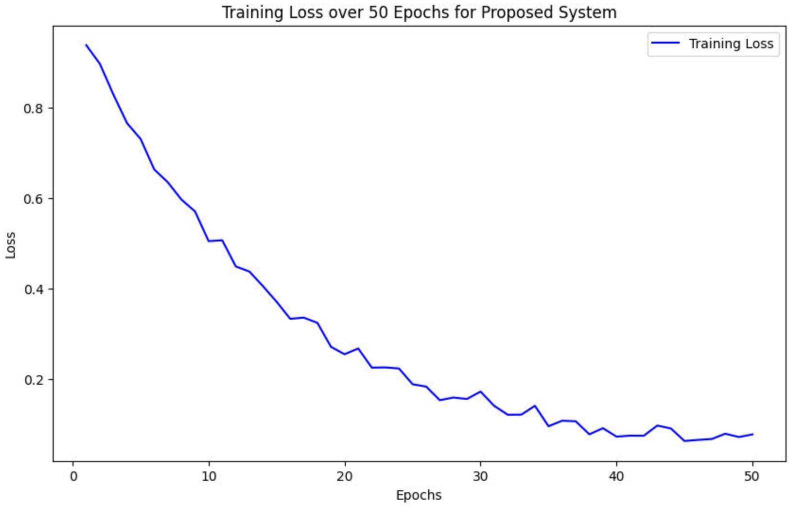

Figure 12. Training losses for the proposed system.

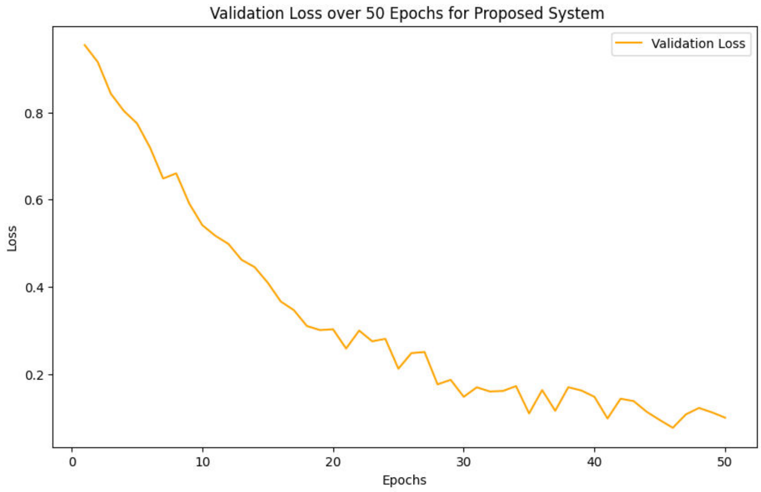

Figure 13. Validation losses for the proposed model.

Conversely, **Figure 13** shows the validation loss, which exhibits a declining pattern incorporated with irregular variations. The validation phase's drop in loss *highlights* the model's practical applicability and shows how well it can adjust to new data. These changes indicate that the model adapted to the validation set patterns that did not exist in the training set. Examining the architecture of the training and validation losses is essential to understanding how the model learns. Ideally, it would only rise after reaching a balance, showing that the model has developed well and isn't exhibiting under or overfitting. The model is moving in the right direction, as seen by the continuous reduction in training and validation losses as shown in **Table 5**. As a result, it could be suitable for precision fertilizer administration and accurate predictions of soil nutrient levels. Using such a model in precision agriculture may provide significant advantages by empowering farmers and agronomists to make well-informed choices about the appropriate kind and quantity of fertilizer. Consequently, this optimization of crop yields and reduction of environmental effects can be achieved.

Table 5. Evaluation of the proposed model.

Nutrient	Model	*R*-squared	Adjusted *R*-squared
N (mg/kg)	RF	−0.04	−0.07
N (mg/kg)	GB	−0.35	−0.40
N (mg/kg)	LR	0.01	−0.02
P (mg/kg)	RF	0.26	0.24
P (mg/kg)	GB	0.07	0.04
P (mg/kg)	LR	0.28	0.25
K (mg/kg)	RF	−0.20	−0.24
K (mg/kg)	GB	−0.50	−0.55
K (mg/kg)	LR	−0.02	−0.05

The evaluation metrics for the three regression models, RF, GB, and LR, are summarized for the three nutrients: potassium (K), phosphorus (P), and nitrogen (N). Regarding nitrogen (N): The RF model exhibits the following values: MAE: 22.93, MSE: 839.24, *R*-squared: −0.04, adjusted *R*-squared: −0.07.
MAE: 26.71; MSE: 1094.09; *R*-squared: −0.35; and adjusted *R*-squared: −0.40 characterized the GB model's decline in performance. With the lowest MSE of 796.14, an MAE of 22.80, a marginally positive R-squared of 0.01, and a slightly negative Adjusted *R*-squared of −0.02, the LR model performed the best. For phosphorus (P), the RF model provides the best MAE (30.97), MSE (1421.39), *R*-squared (0.26), and adjusted *R*-squared (0.24). MAE 34.41, MSE 1794.18, *R*-squared 0.07 and adjusted *R*-squared 0.04 all indicate that the GB model performed inadequately. The LR model had respectable MAE, MSE, *R*-squared, and adjusted *R*-squared values. Potassium (K): The RF model exhibits negative *R*-squared and adjusted *R*-squared values of −0.20 and −0.24, a high MAE of 76.88, and an MSE of 9843.34. The GB model had the lowest potassium score with the highest MAE of 89.08, MSE of 12,252.30, and lowest *R*-squared and adjusted *R*-squared of −0.50 and −0.55. An MAE of 72.09, an MSE of 8331.51, and less severe but adverse *R*-squared and adjusted *R*-squared values of −0.02 and −0.05 were obtained using the LR model. Compared to the other models, the LR model consistently displayed lower MAE and MSE values across all nutrients, suggesting it was the most accurate. However, all models had low or negative *R*-squared values, indicating that they could have more sufficiently explained how the target variables varied. The unfavourable findings suggest that either the models were inappropriate for the data or there was significant data variability. The findings regarding the accuracy of soil nutrient predictions using machine learning techniques underline the specific strengths of random forest, gradient boosting, and logistic regression models. Random forest outperforms in Phosphorus prediction, whereas logistic regression is more precise for nitrogen and potassium. These insights emphasize the challenges in accurately forecasting soil nutrient levels. Future studies are encouraged to expand data sources and explore sophisticated models to enhance prediction precision. Overcoming challenges such as uneven data distribution and improving the models' interpretability will be key to advancing agricultural technologies.
The research on predicting soil nutrients using machine learning offers practical avenues for improving fertilizer usage, directly benefiting farmers and informing policy decisions. By applying the right amount of nutrients

where needed, based on model predictions, we can minimize waste and lessen environmental harm. This approach encourages more sustainable farming, urging a shift towards precision agriculture. Policymakers could leverage these findings to support initiatives that equip farmers with the knowledge and tools for smarter farming practices, while also fostering further innovations in agricultural technologies. Emphasizing the importance of precise nutrient management aligns agricultural productivity with environmental stewardship.

7. Conclusion

Finally, this study shows that using environmental data to make a prediction model that improves fertilizer delivery through precision agriculture could improve crop management. The study thoroughly evaluates several prediction models using data analytics, machine learning (ML), and modern sensor technologies. The model's remarkable accuracy of 85% in predicting fertilizer requirements is highlighted by its integration of baseline soil nutrient data and continuous environmental monitoring. This model demonstrates the promise of contemporary data-driven techniques for enhancing agricultural productivity and resource efficiency, outperforming neural networks (NN) in performance. The findings validate the critical importance of state-of-the-art agricultural technology in tackling the two goals of maximizing crop yield and guaranteeing environmental sustainability in modern farming practices.

Limitations

Enhancing the model could involve acknowledging certain limitations, such as the dataset's possible narrow scope impacting widespread applicability, challenges in scaling the model for vast and varied datasets, the potential for inherent bias and fitting issues, the practicality of model deployment in under-resourced areas, and sensitivity to environmental variables not considered in the study. Future efforts could focus on diversifying the dataset, improving the model to minimize over fitting, and adjusting the system for more extensive and efficient deployment.

Author contributions

Conceptualization, LKN and BPK; methodology, LKN; software, LKN; validation, LKN and BPK; formal analysis, LKN; investigation, LKN; resources, LKN; data curation, LKN; writing—original draft preparation, LKN; writing—review and editing, LKN; visualization, LKN and BPK; supervision, BPK; project administration, LKN and BPK; funding acquisition, LKN and BPK. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Available online: https://www.agmatix.com/blog/agmatix-powers-open-databases-that-will-revolutionize-crop-nutrition/ (accessed on 2 October 2023).
2. Celik MF, Isik MS, Yuzugullu O, et al. Soil Moisture Prediction from Remote Sensing Images Coupled with Climate, Soil Texture and Topography via Deep Learning. Remote Sensing. 2022; 14(21): 5584. doi: 10.3390/rs14215584
3. Bindraban PS, Dimkpa C, Nagarajan L, et al. Revisiting fertilisers and fertilisation strategies for improved nutrient uptake by plants. Biology and Fertility of Soils. 2015; 51(8): 897-911. doi: 10.1007/s00374-015-1039-7
4. Available online: https://www.nature.com/scitable/knowledge/library/sustainable-agriculture-23562787 (accessed on 2 October 2023).
5. Goulding K, Jarvis S, Whitmore A. Optimizing nutrient management for farm systems. Philosophical Transactions of the Royal Society B: Biological Sciences. 2007; 363(1491): 667-680. doi: 10.1098/rstb.2007.2177
6. Sinha D, Tandon PK. An Overview of Nitrogen, Phosphorus, and Potassium: Key Players of Nutrition Process in Plants. In: Mishra K, Tandon PK, Srivastava S (editors). Sustainable Solutions for Elemental Deficiency and Excess in Crop Plants. Springer, Singapore. 2020.
7. Available online: https://www.nparks.gov.sg/nparksbuzz/oct-issue-2020/gardening/identifying-nutrient-deficiency-in-plants (accessed on 2 October 2023).
8. Mallah S, Delsouz Khaki B, Davatgar N, et al. Predicting Soil Textural Classes Using Random Forest Models: Learning from Imbalanced Dataset. Agronomy. 2022; 12(11): 2613. doi: 10.3390/agronomy12112613
9. Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell, 4th ed. New York: Garland Science; 2002.
10. Altieri MA, Nicholls CI, Henao A, et al. Agroecology and the design of climate change-resilient farming systems.

Agronomy for Sustainable Development. 2015; 35(3): 869-890. doi: 10.1007/s13593-015-0285-2
11. Deguine JP, Aubertot JN, Flor RJ, et al. Integrated pest management: good intentions, hard realities. A review. Agron. Sustain. Dev. 2021; 41, 38. doi: 10.1007/ s13593-021- 00689-w
12. Sun G, Ding Y, Wang X, et al. Nondestructive Determination of Nitrogen, Phosphorus and Potassium Contents in Greenhouse Tomato Plants Based on Multispectral Three-Dimensional Imaging. Sensors. 2019; 19(23): 5295. doi: 10.3390/s19235295
13. M. R, A. T, S. R, R. S, et al. Soil NPK Prediction Using Multiple Linear Regression. In: Proceedings of the 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS). Published online March 25, 2022. doi: 10.1109/icaccs54159.2022.9785338
14. Priya R, Ramesh D. Adaboost.RT Based Soil N-P-K Prediction Model for Soil and Crop Specific Data: A Predictive Modelling Approach. In: Mondal A, Gupta H, Srivastava J, et al. (editors). Springer, Cham. 2018.
15. Tatli S, Mirzaee-Ghaleh E, Rabbani H, et al. Prediction of Residual NPK Levels in Crop Fruits by Electronic-Nose VOC Analysis following Application of Multiple Fertilizer Rates. Appl. Sci. 2022; 12: 11263. doi: 10.3390/app122111263
16. Mohapatra AG, Keswani B, Lenka, SK. International Journal of Advanced Research in Computer Science; Udaipur. 2017; 8(7).
17. Dipto SM, Iftekher A, Ghosh T, et al. Suitable Crop Suggesting System Based on N.P.K. Values Using Machine Learning Models. 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). doi: 10.1109/csde53843.2021.9718374
18. Dong Z, Liu Y, Li M, et al. Establishment of an NPK nutrient monitor system in yield-graded cotton petioles under drip irrigation. Plant Methods. 2023; 19(1). doi: 10.1186/s13007-023-01068-0
19. Nooriman WM, Abdullah AH, Abdul Rahim N, et al. Fuzzy logic based prediction of micronutrients demand for harumanis mango growth cycles. Journal of Physics: Conference Series. 2021; 2107(1): 012048. doi: 10.1088/1742-6596/2107/1/012048
20. Maleki S, Karimi A, Mousavi A, et al. Delineation of Soil Management Zone Maps at the Regional Scale Using Machine Learning. Agronomy. 2023; 13(2): 445. doi: 10.3390/agronomy13020445
21. Khaydukova M, Kirsanov D, Sarkar S, et al. One shot evaluation of NPK in soils by “electronic tongue.” Computers and Electronics in Agriculture. 2021; 186: 106208. doi: 10.1016/j.compag.2021.106208
22. John K, Afu SM, Isong IA, et al. Mapping soil properties with soil-environmental covariates using geostatistics and multivariate statistics. International Journal of Environmental Science and Technology. 2021; 18(11): 3327-3342. doi: 10.1007/s13762-020-03089-x
23. Li R, Yin B, Cong Y, et al. Simultaneous Prediction of Soil Properties Using Multi_CNN Model. Sensors. 2020; 20(21): 6271. doi: 10.3390/s20216271
24. Fryer MS, Slaton NA, Roberts TL, et al. Validation of Soil-Test-Based Phosphorus and Potassium Fertilizer Recommendations for Irrigated Soybean. Soil Science Society of America Journal. 2019; 83(3): 825-837. doi: 10.2136/sssaj2019.02.0032
25. Emamgholizadeh S, Shahsavani S, Eslami MA. Comparison of artificial neural networks, geographically weighted regression and Cokriging methods for predicting the spatial distribution of soil macronutrients (N, P, and K). Chinese Geographical Science. 2017; 27(5): 747-759. doi: 10.1007/s11769-017-0906-6
26. Das K, Twarakavi N, Khiripet N, et al. A Machine Learning Framework for Mapping Soil Nutrients with Multi-Source Data Fusion. 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. Published online July 11, 2021. doi: 10.1109/igarss47720.2021.9553305
27. Peng Y, Zhao L, Hu Y, et al. Prediction of Soil Nutrient Contents Using Visible and Near-Infrared Reflectance Spectroscopy. ISPRS International Journal of Geo-Information. 2019; 8(10): 437. doi: 10.3390/ijgi8100437
28. Sirsat MS, Cernadas E, Fernández-Delgado M, et al. Automatic prediction of village-wise soil fertility for several nutrients in India using a wide range of regression methods. Computers and Electronics in Agriculture. 2018; 154: 120-133. doi: 10.1016/j.compag.2018.08.003
29. Tziachris P, Aschonitis V, Chatzistathis T, et al. Comparing Machine Learning Models and Hybrid Geostatistical Methods Using Environmental and Soil Covariates for Soil pH Prediction. ISPRS International Journal of Geo-Information. 2020; 9(4): 276. doi: 10.3390/ijgi9040276