

Integrating LightGBM with SHapley Additive exPlanations (SHAP) for Enhanced Soil Analysis and Interpretability

Dismas Tendet Kitaria

dtendet@must.ac.ke

Meru University of Science and Technology

Amos Chege Kirongo

Meru University of Science and Technology

Stephen Makau Mutua

Meru University of Science and Technology

Research Article

Keywords: SHAP, LightGBM, Mahalanobis Distance, Smart Agriculture, Soil Quality

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11040583/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Integrating LightGBM with SHapley Additive exPlanations (SHAP) for Enhanced Soil Analysis and Interpretability

Dismas Tendet Kitaria^{1,*}, Amos Chege Kirongo², Stephen Makau Mutua³

^{1,2,3} *Department Computer Science, Meru University of Science and Technology, P.O. Box 972-60200, Meru, Kenya*

**Corresponding author: dtendet@must.ac.ke*

Abstract

Agricultural production has been a cornerstone for food security worldwide. However, determining crops suitable for a particular soil type has been a challenge to most farmers due to varying soil properties. This has led to need of adoption of technology specifically in soil quality analysis towards boosting production. As a result, this paper presents a comprehensive approach that integrates machine learning and explainable artificial intelligence using SHapley Additive explanations (SHAP) in predicting the most suitable crop to be grown on a particular soil sample. The solution utilized data acquired from Kaggle dataset that comprised of soil quality parameters of Nitrogen (N), Phosphorus (P), potassium (K), pH, Temperature, and Humidity. A machine learning model was then built using Light Gradient Boosting Machine (LightGBM) onto a multi output regressor that accepted the input of the soil quality parameters to train the model for predicting crop suitability based on a soil sample. Grid search was then utilized for parameter tuning and then SHAP employed to provide explanation for the contribution of each soil parameter to prediction of a particular crop to be grown. Results validated the robustness and predictive accuracy of the model with a model performance of 0.9991 for R^2 score, 0.0047 of RMSE, and MAPE of less than 0.5%. The SHAP interpretability revealed that Phosphorous (P) and Potassium (K) are the most influential parameters across different crops, while humidity and Nitrogen (N) affected only specific crops suitability. The results support data driven interventions by farmers such as optimized fertilizer application, irrigation, and other resources utilization towards improving productivity.

Keywords: SHAP, LightGBM, Mahalanobis Distance, Smart Agriculture, Soil Quality

1.1 Introduction

Agricultural production has stood out over years as the main foundational and fundamental component towards global food security. The crucial component behind food security, sustainable Agricultural practices, carbon cycling and sustainable eco-system health is the soil. Consequently, actual soil analysis involving quantification of biological, physical and chemical properties remains a vital requirement. Effective realization of these leads to advancement in precision agriculture, climate change mitigation and appropriate land remediation strategies. As a result, soil analysis plays a pivotal role in this chain value of food security and various laboratory analysis techniques have been applied over time in determining soil health. However, whereas conventional laboratory approaches are accurate; they are not only time-consuming and costly, but also laborious, spatially sparse and beyond the reach of many farmers [1].

To address these concerns, researchers and academicians have preferred the adoption of technological techniques that are not only fast and reliable but also efficient and accurate. As a result, machine learning (ML) techniques coupled with chemometric and proximal sensing approaches, have provided scalable, data driven solutions for crop recommendation, yield production and soil quality assessments[2]. Machine learning models have demonstrated remarkable capabilities in predicting crop yields, diagnosing plant diseases, and optimizing resource allocation [3], [4]. Motivated by these strides, this work seeks to explore the utilization of ML techniques and explainable artificial intelligence (XAI) in determining crop suitability through soil analysis.

As earlier noted, soil health is a fundamental aspect that governs crop selection and productivity by mediating a balance between the soil's physical, chemical, and biological characteristics for nutrient bioavailability, root architectural development, and abiotic stress resilience [5]. Based on these soil health characteristics, it is possible to determine the suitability for a particular crop in a given area towards optimization of agricultural land and determination of land plans for current and future use [6].

Traditional land suitability assessments relied on multi-criteria evaluation frameworks, such as FAO land capability classification and empirical fertility indices, which integrated extensive pedological and climatic datasets [7]. While interpretable, these non-machine learning approaches suffer from subjective parameter weighting, static thresholding, and poor generalizability across heterogeneous agro-ecological zones, motivating the transition toward data-driven suitability modeling [8]; [9]; [10]. However, data driven approaches are highly reliant on the quality of the data and its appropriateness. This often leaves an open question – what is the desired soil metrics composition for model training and validation?

To address these limitations, it is imperative to establish the most critical data that can be collected and analyzed to make informed decision about soil suitability for a particular crop. Notably, soil is made up of seventeen (17) essential elements. These include calcium, carbon, hydrogen, oxygen, nitrogen, phosphorous, magnesium, sulfur and potassium that facilitates various aspects of plant growth; and the micronutrients include iron, manganese, zinc, copper, boron, molybdenum, chlorine and nickel. However, of these seventeen, carbon, hydrogen and oxygen are absorbed directly from the air and water thus leaving fourteen (14) essential elements [11]. From these, calcium, nitrogen, phosphorous, magnesium, sulfur and potassium form the *macro nutrients*. Whereas the micronutrients are important, they are often considered to be tracer elements, implying their availability is only desired in small quantities and thus their availabilities are strongly influenced by the soil pH and the macro nutrient balance. Subsequently, this makes their direct measurement less critical in building models for soil suitability just as water holding capacity which is indirectly represented by soil texture and moisture levels thus not prioritized in most suitability frameworks [12], [13].

Premised on these, [14] avers that nitrogen (N), phosphorus (P), and potassium (K) are considered as the most essential nutrients whose presence determines soil quality and crop recommendation. Additionally, [15] notes that pH directly influences soil health for determining crop production by balancing mineral particles, organic matter and water retention while influencing other processes such as mineral breakdown, nutrient solubility and microbial activities. Similarly, soil temperature is a crucial factor that has significant effect on soil physical, chemical as well as biological priorities. Temperature may change soil particle structure to increase its porosity, resulting in better soil aeration and improving permeability while at the same time decomposition rate of organic matter is also directly affected by changes in temperatures. These changes directly affect soil fertility and water holding capacity which in turn affects plant growth and development as well as microbial activities [16], [17], [18]. Anchored on these works, this study adopts a data composition of the macro nutrients N, P, K, pH, Temperature (T) and the humidity (H) as the appropriate data set to train a model for determination of crop suitability. We believe that it is critical to have in mind the other probable factors that would affect the productivity of a crop despite the appropriateness of the macro-nutrients and thus answering the question on the essential parameters for model building.

Having established the appropriate parameters, this study seeks to explore another concern. Whereas different approaches have been presented to address the soil suitability, we aver that the integral non-linearities and interactions amid the soil components calls for more expressive approaches. Due to this nature of this non-linearity, it is outright that regression models and simple ML models fail to be appropriate for this type of a problem. Nevertheless, ensemble methods, specifically gradient boosting machines (GBMs) have demonstrated a unique ability to perform well in heterogeneous data types, handling missing values, and non-linear interactions [19], [20] [21] [20]. Of these GBMs, the LightGBM (Light Gradient Boosting Machine) presents unique merits such as lower memory utilization, faster training and an Exclusive Feature Bundling (EFB) for heterogeneous data sets [22]. Apart from these merits, it also doesn't compromise on its accuracy making it a preferred approach over XGBoost and Random Forest in agronomic and environmental tasks like the one herein. However, despite its predictive excellence, just like other ML techniques, LightGBM lacks interpretability hence affecting its absorption. This is due to the black-box nature of most ML techniques.

To address this, Explainable Artificial Intelligence (XAI) frameworks came into being to seek to demystify complex models and make their decisions and judgement accessible to non-technical stakeholders [23], [24], [25]. Different researchers have since used XAI to explain their models. For instance, [26] applied SHAP to deep learning models for irrigation water quality assessment, revealing the influence of individual chemical features on predictions. [25] extended SHAP to workforce analytics, underscoring its value in human-centric agricultural decision-making. [27] integrated SHAP

and LIME into crop yield prediction models, identifying temperature and rainfall as the dominant factors influencing productivity. [28] demonstrated how SHAP could enhance transparency in crop recommendation systems, improving farmer's trust and compliance with GDPR mandates. [29] applied SHAP to weed detection using CNNs, achieving high accuracy while maintaining interpretability. However, SHAP cannot provide the desired interpretability of the AI models without having the appropriate parameterization and tuning. To address this, grid search was employed to enhance the framework by systematically evaluating a predefined combination of soil quality parameters towards recommending parameter combinations that would maximize suitability score that is tailored for a specific crop [30]. This was achieved by transparent manner in which grid search reproduces optimization strategy that exhaustively explores the specified parameter spaces towards identifying combinations that yield the best performance [31].

Motivated by these, this study therefore presents an approach that integrates LightGBM model with SHAP-based interpretability and grid search for parameterization. The model was trained, tested and validated based on a Kaggle soil dataset. Additionally, an independent Crop Recommendation using Soil Properties and Weather Prediction Dataset (Crop Recommendation using Soil Properties and Weather Prediction Dataset - Mendeley Data) was used to further validate the validity and reliability of the model as outlined in details under methods and materials. Our empirical results show that the platform facilitates well-informed, data-driven agricultural decisions and closes the gap between technical modeling and real-world field deployment by fusing high-performance machine learning with clear explanations; while providing for potential scalability and transfer learning.

2.0 Methods and Materials

2.1 Materials

The study utilized two data sets, a crop recommendation dataset from Kaggle (crop-recommendation-dataset) and Ethiopian Agricultural Transformation Agency (ATA) dataset (<https://data.mendeley.com/datasets/8v757rr4st/1>). The latter was purely an independent dataset which was not used to train or test the model but was re-organized to mimic the training data for the sole purpose of validating how the model performs on an entirely new dataset, its robustness and transfer learning capabilities. The former is the dataset (Kaggle dataset) that was used for training, testing and validation.

The Kaggle dataset is a result of augmentation of datasets of rainfall, climate and fertilizer in India and comprises of different fields namely; ratio of Nitrogen content in soil (N), ratio of Phosphorous content in soil (P), the ratio of Potassium content in soil (K) and the soil temperature in degree Celsius. Additionally, it also has fields of the relative humidity in percentage, Ph value of the soil (pH) and the amount of rainfall in millimeters. In the dataset, twenty-two (22) crops were used each with 100 rows per crop, resulting in a total of 22,000 samples. Each row included the presented field of Nitrogen N (mg/kg), Phosphorus P (mg/kg), Potassium K (mg/kg), Temperature (°C), Relative Humidity (%), pH, and Rainfall (mm).

To understand the distribution of these features and their variations across different crops, a Histograms with Kernel Density Estimation (KDE) overlay was used to provide insight of the range of soil quality parameters over time as shown in Figure 1 for each of the parameters that were in focus for this study. This was critical since it informed the study on the structure of the data thus informing the data pre-processing approaches and the algorithms to use in modeling to ensure reliable and accurate modeling.

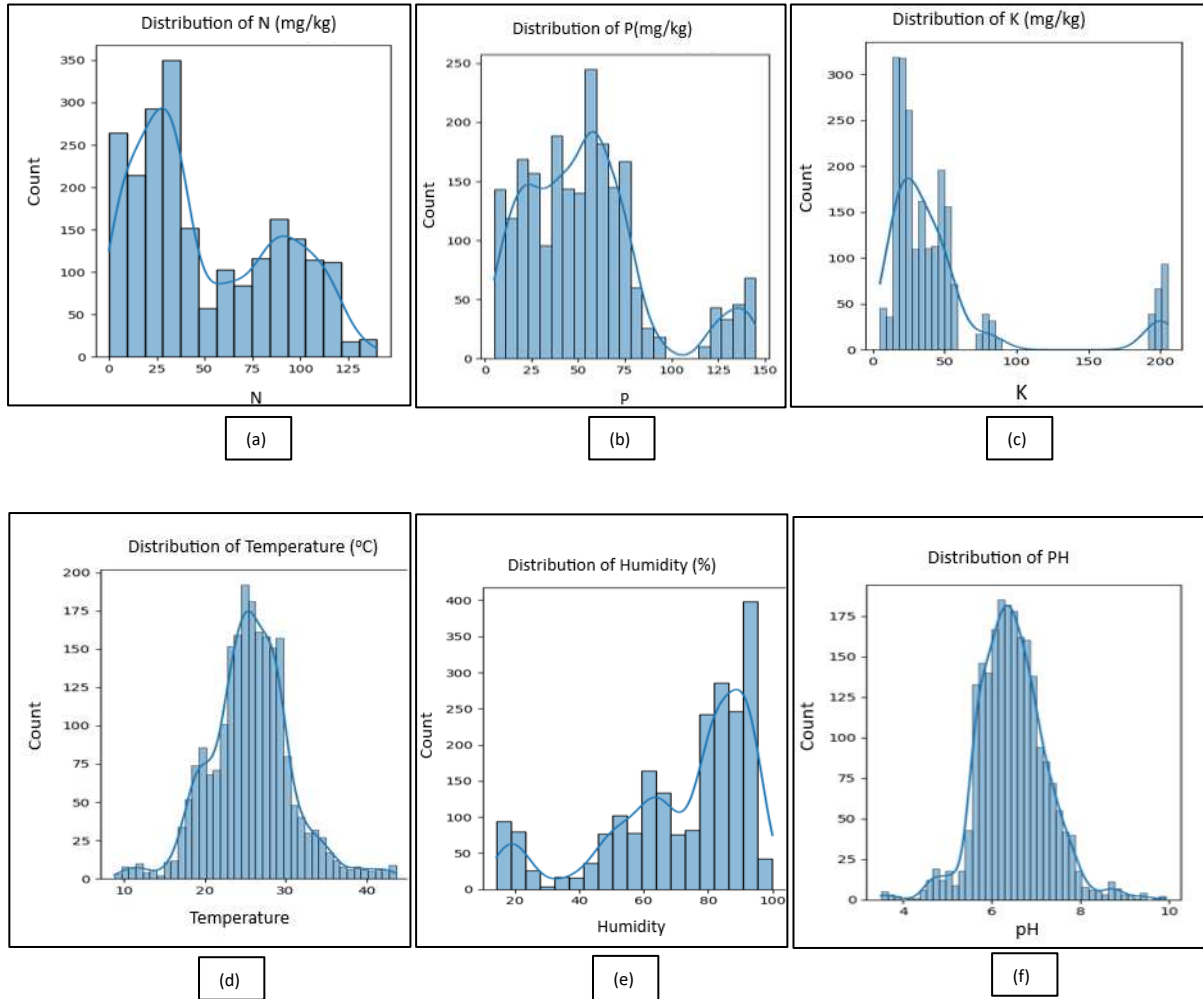


Figure 1 Illustration of the original data (Kaggle Dataset) Soil Quality Feature Distributions with (a) showing the Nitrogen distribution levels, (b) Phosphorous distribution levels, (c) showing Potassium distribution levels, (d) Temperature distribution levels, (e) Humidity distribution levels, and (f) showing pH distribution levels.

During modeling, the dataset was split into 80% training 20% testing; the training sample was later split into 80% training and 20% validation. Thus, the training set was 64%, the testing set was 16% and validation set was 20% accordingly.

2.2 Methods

This study adopted the quantitative research design. The respective methods for data pre-processing, feature engineering, model development and evaluation are outlined in the following sub-sections;

2.2.1 Data Preprocessing

The original data had seven features which are; rainfall, Nitrogen (N), Phosphorus (P), Potassium (K), temperature, humidity and the pH values. However, it was noted that for the rainfall feature, the input sensing device lacked capability to measure rainfall instantly implying that the values were manually inputted and thus it was dropped. Consequently, the other six features were employed. Nevertheless, the data had notable issues that needed redress.

To address outliers in the data, the Interquartile Range (IQR) and Mahalanobis Distance was used. For each feature, the outliers were detected using the IQR as shown in equation 3.1.

$$IQR = Q_3 - Q_1 \quad \text{Eqn. 3.1}$$

Where Q_1 and Q_3 are the 25th and 75th percentiles, respectively.

Based on this, the data points outside the ranges of $Q_1-1.5(IQR)$ and $Q_3+1.5(IQR)$ were considered outliers and winsorized into the nearest boundary point.

Similarly, to address the multivariate outliers among the variables, the Mahalanobis distance was computed for each data point to identify multivariate outliers using the formula in equation 3.2 and normalized using the inverse covariance matrix.

$$D_M(x) = \sqrt{[(x - \mu)^T S^{-1} (x - \mu)]} \quad \text{Eqn. 3.2}$$

where x is the data point, μ is the mean vector of the features, and S is the covariance matrix.

Since the data was complete with no missing values, no imputation methods were employed.

2.2.2 Feature Generation and Extraction

Determination of Crop Standard Conditions

Each crop in the data set had varying values for the six features under study. Since the data was normalized and with no outliers beyond the standard deviation, the study computed the likely common standard condition for each of the feature's values that were considered critical for the optimal performance of the crop. To achieve this, the mean was computed across all the features creating an Ideal Matrix, I_M as illustrated in equation 3.3;

$$I_M = \begin{bmatrix} N & P & K & T & H & P_H \\ \mu_{1,N} & \mu_{1,P} & \mu_{1,K} & \mu_{1,T} & \mu_{1,H} & \mu_{1,P_H} \\ \mu_{2,N} & \mu_{2,P} & \mu_{2,K} & \mu_{2,T} & \mu_{2,H} & \mu_{2,P_H} \\ \mu_{3,N} & \mu_{3,P} & \mu_{3,K} & \mu_{3,T} & \mu_{3,H} & \mu_{3,P_H} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \mu_{22,N} & \mu_{22,P} & \mu_{22,K} & \mu_{22,T} & \mu_{22,H} & \mu_{22,P_H} \end{bmatrix} \quad \text{Eqn. 3.3}$$

where $\mu_{i,j}$ is the ideal condition of feature j for crop i ; and there are 22 crops in total each with six features

Suitability Score Computation

In order to retain the value proposition of each sample, the suitability score was computed using the Mahalanobis distance between the sample's feature values and the ideal conditions for each crop. The algorithm for this computation was as indicated below;

Algorithm: SuitabilityScorer

Step 1: For each sample, x and each crop, c , compute the Mahalanobis distance using equation 3.4:

$$D_M(x, c) = \sqrt{[(x - \mu_c)^T S_c^{-1} (x - \mu_c)]} \quad \text{Eqn. 3.4}$$

Where μ_c is the ideal condition vector for crop c , and S_c is the covariance matrix of the feature values for crop c .

Step 2: Convert the distance, D_M to a suitability score, $S(x,c)$ using an exponential decay function, equation 3.5:

$$S(x, c) = e^{\{-\alpha D_{M(x,c)}\}} \quad \text{Eqn. 3.5}$$

Where α is a scaling factor (set to 0.5 after experimentation to ensure scores range between 0 and 1).

Step 3: Generate a suitability matrix, S_M such that;

$$S_M = S(x, c) \quad \text{Eqn. 3.6}$$

Step 4: Repeat for all 22 crops

This suitability matrix served as the target variable for the multioutput regression model.

2.2.3 LightGBM Modeling

Light Gradient Boosting Machine (LightGBM) framework integrated with multi-output regression was used to address the multi-target prediction challenge in soil-based crop suitability. The approach was selected based on its efficiency, scalability and its ability to handle nonlinear relationships in high dimensional data.

Model Architecture and Ensemble Strategy

The modelling pipeline utilised the MultiOutputRegressor wrapper from scikit-learn to extend LightGBM's capabilities to multi-output regression. This wrapper trains one independent LightGBM model per target variable, allowing for parallelized learning and tailored optimization for each crop suitability output. Each base estimator was instantiated using LGBMRegressor, which inherently applies gradient boosting over decision trees, an ensemble technique that iteratively minimizes prediction error. The training process involved hyperparameter tuning using a grid search which was performed to find the optimal hyperparameters for the LightGBM model. The elastic Net (L1 and L2) were then used for regularization as defined herein;

Regularization and Objective Functions

Regularization parameters were used to optimize LightGBM model created to prevent any overfitting for robustness in prediction and making the model less complex. In the study, regularization parameters of *Reg_alpha* (L1 regularization) and *Reg_lambda* (L2 regularization) were applied. *Reg_alpha* was applied to penalize model complexity during training and encourage the model to minimize the absolute leaf values. Higher *Reg-alpha* value meant more in future importance with its optimal values being found during hyperparameters tuning and cross validation. *Reg_lambda* was applied to penalize the model as well to discourage overfitting by shrinking and smoothening the leaf weights of the features towards zero. In the LightGBM model, the regressor mapped all features from Matrix I_M onto a single column that represented the suitability score of each of the 22 crops. In the process, the *Reg_alpha* controlled how complex the model was when learning from each of the columns of matrix, I_M . For example, in a case where Phosphorus (P) and pH dominated prediction, they were regularized to reduce model overreliance on them across all the samples. For each model output, the predicted and the actual values were compared for each column outputs then the Mean Squared Error was computed over the entire column using the soil features from the input matrix I_M

Data Structure and Matrix I_m Alignment

The matrix I_M represents a multi-dimensional dataset where each row corresponds to individual soil samples and the columns represent critical soil parameters (pH, N, P, K, temperature, humidity).

Matrix I_M forms a Feature Matrix $I_m \in R^{n \times m}$ Eqn 3.7

The feature matrix is used in training the model where;

n = number of samples (rows)
 m = number of features (Columns)

The model predicts suitability scores for the 22 crops forming a Target Matrix

$$Y \in R^{n \times k} \quad \text{Eqn 3.8}$$

where;

$K = 22$ crops

Each column in Y corresponds to the predicted suitability score for one crop.

This alignment ensures that;

Each row in I_M maps to a corresponding row in Y

Each LightGBM regressor (via MultiOutputRegressor) learns to predict one column of Y using the same I_M .

Integration with SHAP and Interpretability

The Matrix I_M was integrated with the multi-output wrapper around LGBMRegressor. Then each of the regressors were trained independently on each input of I_M in predicting the suitability score of each of the 22 crops. From the suitability score, the SHAP values were then computed to determine the influence of each of the input parameters on prediction of each crop.

2.2.4 Model Evaluation

The model performance was compared across different crops in the dataset with varying yield ranges on the soil sample used. The metrics used for validation of the model were Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), R^2 Score, and Mean Squared Error (MSE) as defined in equations 3.9 to 3.13. These metrics collectively quantified the predictive accuracy and robustness of the crop suitability model. A summary of both validation and test performance of the model is provided in the section of results and discussion. More specifically, the R^2 measured the proportion of variance in the target variable that is explained by the model with its values ranging from 0 to 1. Values closer to 1 indicated better prediction performance. RMSE was used to measure the average of the magnitude of prediction errors, better model performance was indicated using a lower value of RMSE. On the other hand, MAE provided an average measure of prediction error by treating all errors equally without basing on the outliers as it's the case for RMSE. In the same regard, MAPE demonstrated prediction error inform of the actual value percentage that making the results easy for interpretation. MSE provided a squared average of prediction errors. While less interpretable than RMSE (due to squared units), it is mathematically convenient for optimization and is the basis for RMSE calculation. The formulae for the different metrics used for validation are as shown in eqn. 3.9 though to eqn. 3.13 below.

$$RMSE = \sqrt{\left(\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_i)^2\right)} \quad \text{Eqn. 3.9}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \bar{y}_i| \quad \text{Eqn. 3.10}$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \bar{y}_i}{y_i} \right| \quad \text{Eqn. 3.11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad \text{Eqn. 3.12}$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_i)^2 \quad \text{Eqn. 3.13}$$

Where:

n = number of samples

y_i = actual value for sample i

$\hat{y}_i$ = predicted value for sample i

2.2.5 SHAP (SHapley Additive exPlanations) Analysis

Sharpley Addictive exPlanation (SHAP) was integrated into LightGBM modelling pipeline to enhance transparency and gain trust among stakeholders in AI-driven crop suitability predictions. It provides a mathematical grounded approach in model interpretation by attributing contribution of each input feature to the final prediction. The concept of SHAP is based on a cooperative game theory that assigns a fair value to each player (feature) based on their marginal contribution to the outcome.

For each crop specific LightGBM regressor, SHAP computes the contribution of feature i to the model's prediction using the formulae;

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad \text{Eqn 3.14}$$

Where;

N = set of all features (All Columns in the matrix)

S = subset of features excluding i

$f(S)$ = model prediction using only features in subset S .

$f(S \cup \{i\})$ = prediction when feature i is added

Φ_i = Shapley value for feature i (Quantifies how much feature i changes prediction across all possible combinations).

This computation is done for each row (Soil sample) in the matrix to produce a SHAP value matrix $X \in R^{n \times m}$ where each entry Φ_{ij} represents the contribution of feature j to the prediction for sample i

SHAP explanation

SHAP approximates the model output using a linear explanation model;

$$f(x) \approx g(x') = \phi_0 + \sum_{i=1}^M \phi_i x'_i \quad \text{Eqn 3.15}$$

Where;

$F(x)$ = original LightGBM Model prediction for a given soil sample

$G(x')$ = SHAP explanation model

Φ_0 = Base value (average prediction across all samples)

Φ_i = SHAP value for feature i

X'_i = binary indicator of whether feature i is present (1 if feature i is present, 0 if not)

In the matrix I_M , Each row x is a soil sample, each column i is a soil parameter.

SHAP computes Φ_i for each feature in each row, linking the feature matrix X to the SHAP value Matrix Φ .

In the study, SHAP was applied to each LightGBM regressor withing the multi-output framework to interpret crop specific predictions. The soil features used in the model served as the player in the shapely framework.

3.0 Results and Interpretation

3.1 Data Preprocessing Results

Both univariate and multivariate outlier detection techniques were applied during preprocessing. For Univariate outlier detection, Interquartile Range (IQR) was computed for six soil parameters of Nitrogen (N), Phosphorus (P), Potassium (K), pH, temperature, and humidity. All the data points falling outside $(Q_1 - 1.5) \text{ IQR}$, $(Q_3 + 1.5) \text{ IQR}$ were considered to be outliers. However, in the dataset that was used, no values were found to breach the thresholds across all the features, indicating a clean distribution with no extreme deviations.

For multivariate outlier detection, Mahalanobis distances were computed for each observation using full features set against a chi-squared distribution at 95% confidence level. Similarly, just as was observed in the univariate pre-processing, no data values exceeded the critical values confirming absence of multivariate outliers. Consequently, no winsorization or imputation was done and the raw data was retained for modelling.

3.1.2 Collinearity Testing

The global correlation matrix across all the crops in the dataset shows how the six soil quality parameters relate with one another in the entire dataset. The matrix provides an understanding of the balance soil nutrients parameters in the soil with regards to particular crop be grown under the baseline conditions. As illustrated in figure 2, each cell represents a correlation coefficient that ranges from -1.00, indicating strong negative relationship to +1.00 that indicates strong positive relationship. The figure results indicate that Phosphorous (P) and potassium (K) have strong positive correlation of 0.74 showing that the soil sample analyzed had levels of both P and K due to possible similar soil fertility or soil management practices. On the other hand, Nitrogen (N) had a weaker correlation with most of the other soil quality parameters meaning that N variation is largely independent of other soil quality parameters. Similarly, Temperature and Humidity have moderate correlation of 0.21 while pH remains largely uncorrelated to any other soil parameter, meaning that soil acidity or alkalinity does not strongly depend on soil nutrients concentration on the dataset used. For instance, the correlation between N and P of -0.23 and K and pH of -0.17 shows inverse relationships, meaning when one parameter increases, the other tends to decrease however, these variations are weak and statistically dominant.

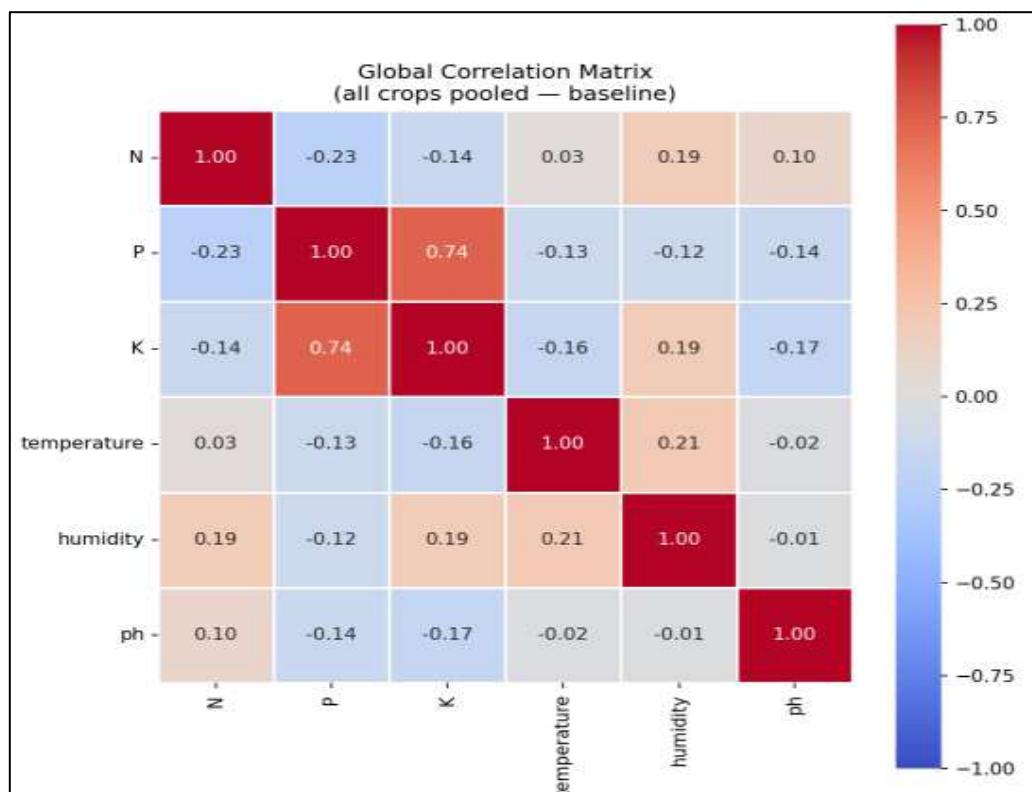


Figure 2 Global correlation matrix for all crops in the dataset showing relationship between the six soil quality parameters.

Given the multivariate nature of the data used in the model, collinearity issues could arise due to high correlation between P and K causing redundancy and instability in the regression model as shown in the correlation matrix. To solve the issues, LightGBM offered solution through its tree-based learning framework ability, by constructing decision trees that split data based on the most informative feature at each node progressively. In a case where two parameters are highly correlated, the algorithm is designed to make them competitive for each split, and the parameter the best maximizes on loss reduction is selected automatically.

3.1.2 Feature Generation and Extraction Results

To support multi-output regression modelling for crop suitability, ideal condition matrix (I_M) was computed for each of the 22 crops. This computed mean value for the six soil parameters (N, P, K, pH, temperature, humidity) represented optimal soil condition for each crop and the matrix served as a reference baseline for suitability scoring. Then each soil sample was then evaluated against the I_M using Mahalanobis distance to quantify how close the sample matches the ideal condition for each crop. This distance was transformed into a suitability score, that resulted to generation of a suitability matrix for all the 22 crops. This suitability matrix became the target variable for the multi-output regression model as shown in table 1.

Table 1: Presenting the suitability score of the six soil quality parameters for different crops in the dataset to represent a suitability matrix for all the 22 crops

Crop	N	P	K	Temperature	Humidity	pH
Apple	20.8	134.2	199.9	22.6	92.3	5.93
Banana	100.2	82	50	27.4	80.4	5.98
Blackgram	40	67.5	19.2	30	65.1	7.13
Chickpea	40.1	67.8	79.9	18.9	16.9	7.34
Coconut	22	16.9	30.6	27.4	94.8	5.98
Coffee	101.2	28.7	29.9	25.5	58.9	6.79
Cotton	117.8	46.2	19.6	24	79.8	6.91
Grapes	23.2	132.5	200.1	23.8	81.9	6.03
Jute	78.4	46.9	40	25	79.6	6.73
kidneybeans	20.8	67.5	20	20.1	21.6	5.75
Lentil	18.8	68.4	19.4	24.5	64.8	6.93
Maize	77.8	48.4	19.8	22.4	65.1	6.25
Mango	20.1	27.2	29.9	31.2	50.2	5.77
mothbeans	21.4	48	20.2	28.2	53.2	6.83
mungbean	21	47.3	19.9	28.5	85.5	6.72
muskmelon	100.3	17.7	50.1	28.7	92.3	6.36
Orange	19.6	16.6	10	22.8	92.2	7.02
Papaya	49.9	59	50	33.7	92.4	6.74
pigeonpeas	20.7	67.7	20.3	27.7	48.1	5.79
pomegranate	18.9	18.8	40.2	21.8	90.1	6.43
Rice	79.9	47.6	39.9	23.7	82.3	6.43
watermelon	99.4	17	50.2	25.6	85.2	6.5

3.1.3 LightGBM Modeling

Model Training

The model was developed and trained (based on training dataset) using LightGBM with a Scikit-Learn MultiOutputRegressor wrapper to handle the 22 regression tasks, one per crop. In this phase, the model learnt general patterns of how each input feature affect suitability based on the soil sample provided.

To optimize on the predictive performance of the crop suitability model, a systematic grid search was performed across four key hyperparameters towards finding the optimal hyperparameters for the LightGBM model using the validation dataset. These hyperparameters were `n_estimator` (Number of boosting iterations (epochs)), `learning_rate` (Step size shrinkage to prevent overfitting), `max_depth` (Maximum depth of individual trees) and `num_leaves` (Number of leaves per tree, controlling model complexity). The best parameters were selected based on the lowest mean squared error (MSE) on a validation set as illustrated in Table 2. This indicated that the model's prediction was closer to the true suitability score, making it a reliable indicator for the overall accuracy of the model.

Table 2: A comparative performance of the four key performances tracked against Root Mean Squared Error (RMSE), Mean Squared Error (MSE), R^2 Score, and Mean Absolute Percentage Error (MAPE) performance matrix

n_estimators	learning_rate	max_depth	num_leaves	RMS E	MSE	R^2 Score	MAPE	Observation
50	0.01	3	15	0.0098	0.0096	0.9941	1.12%	Underfitting, slow convergence
100	0.1	5	31	0.0063	0.004	0.9981	0.67%	Balanced fit, moderate complexity
200	0.1	-1	15	0.0047	0.0022	0.9991	<0.5%	Best performance, selected model
200	0.2	3	63	0.0059	0.0035	0.9974	0.71%	Slight overfitting, unstable generalization
100	0.01	-1	15	0.0075	0.0056	0.9967	0.89%	Slow learning, suboptimal accuracy

Model Testing

On testing data, the model showed a good overall performance for unseen data with R^2 of 0. and MAPE of 0.481 as shown in table 3. Apple, Pigeonpeas, and kidneybeans showing a good R^2 of 0.9990, 0.9998, and 0.9971 respectively showing that the model's prediction is closer to the actual suitability scores. The lower level of MAPE also indicated that on average, prediction errors were less than half a percent of the true value that demonstrated high predictive precision.

Table 3: Per-crop testing metrics for the all the crops based on testing data from Kaggle dataset

Per-Crop Testing Metrics

R^2 , adjusted R^2 , explained variance, MSE, RMSE, MAE, median AE and MAPE per crop; bottom row pooled across all crops.

Crop	R^2	Adj. R^2	Explained Var	MSE	RMSE	MAE	MedAE	MAPE %
Apple	0.9990	0.9990	0.9990	0.000025	0.0050	0.0033	0.0023	0.788
Banana	0.9985	0.9985	0.9985	0.000030	0.0055	0.0039	0.0028	0.510
Blackgram	0.9994	0.9994	0.9994	0.000019	0.0043	0.0031	0.0023	0.395
Chickpea	0.9992	0.9992	0.9992	0.000020	0.0045	0.0034	0.0027	0.707
Coconut	0.9992	0.9992	0.9992	0.000026	0.0051	0.0038	0.0030	0.515
Coffee	0.9994	0.9994	0.9994	0.000019	0.0044	0.0031	0.0025	0.403
Cotton	0.9994	0.9994	0.9994	0.000019	0.0044	0.0032	0.0025	0.435
Grapes	0.9994	0.9994	0.9994	0.000015	0.0038	0.0022	0.0014	0.546
Jute	0.9992	0.9992	0.9992	0.000020	0.0044	0.0032	0.0024	0.399
Kidneybeans	0.9971	0.9970	0.9971	0.000072	0.0085	0.0043	0.0026	0.636
Lentil	0.9996	0.9995	0.9996	0.000015	0.0039	0.0029	0.0022	0.368
Maize	0.9997	0.9997	0.9997	0.000014	0.0037	0.0026	0.0018	0.362
Mango	0.9995	0.9995	0.9995	0.000018	0.0042	0.0029	0.0022	0.409
Mothbeans	0.9997	0.9997	0.9997	0.000011	0.0034	0.0024	0.0018	0.324
Mungbean	0.9987	0.9987	0.9987	0.000032	0.0056	0.0039	0.0027	0.498
Muskmelon	0.9985	0.9985	0.9985	0.000036	0.0060	0.0042	0.0029	0.568
Orange	0.9996	0.9996	0.9996	0.000016	0.0040	0.0029	0.0021	0.461
Papaya	0.9988	0.9988	0.9988	0.000030	0.0055	0.0038	0.0028	0.534
Pigeonpeas	0.9998	0.9998	0.9998	0.000010	0.0032	0.0023	0.0017	0.363
Pomegranate	0.9995	0.9995	0.9995	0.000018	0.0043	0.0033	0.0026	0.449
Rice	0.9994	0.9994	0.9994	0.000019	0.0044	0.0032	0.0024	0.423
Watermelon	0.9984	0.9984	0.9984	0.000028	0.0053	0.0038	0.0030	0.486
OVERALL	0.9994		0.9994	0.000023	0.0048	0.0033	0.0023	0.481

On the unseen test set the model attains $R^2 = 0.9994$ overall (per-crop R^2 from 0.9971 to 0.9998), with RMSE 0.0048 and MAE 0.0033 on the 0–1 suitability scale and a mean absolute percentage error of 0.48%. Test performance essentially matches validation (R^2 0.9995, RMSE 0.0048), indicating the model is well-fit and not over-fit — it generalises consistently within the Kaggle domain.

Tuned multi-output LightGBM (*multioutput_lgbm_tuned.pkl*) evaluated on the test split of *data_splits.npz*. Reproducible via *compute_kaggle_test_metrics.py*.

Model validation

The model was validated based on the data from the two datasets (2) from Kaggle and Ethiopian Agricultural Transformation Agency (ATA)). This evaluated how the trained model would perform on unseen data to ensure generalization of the results and avoid overfitting and to meet the intended purpose of the model before its deployed. Results showed a near perfect R^2 of 0.9995 and RMSE of 0.0048 on Kaggle validation while the unseen ATA validation data results showed an R^2 of 0.9201) with a small absolute error (RMSE) of 0.044, with an MAE of 0.028 and indicated in the Table 4. The drop confirms that the model generalises results on an independent dataset while performing best on its in-domain data.

Table 4: Aggregate metrics across all 22 crops. Kaggle (in-domain) sets the ceiling; Mendeley (external) is lower, as expected under distribution shift.

Dataset	R^2	Explained Var	RMSE	MAE	MSE	MAPE %
Kaggle (Validation)	0.9995	0.9995	0.0048	0.0033	0.00000	0.48
Mendeley (External)	0.9201	0.9312	0.0443	0.0283	0.00200	4.48

Per crop validation metrics for both the data sets as indicated in Table 5 with the left block showing the Kaggle validation and left indicating the ATA data indicated as Mendeley external.

Table 5: Per-Crop Validation Metrics for both datasets. Left block indicates Kaggle validation while Right block shows Mendeley external with Bottom row shows overall pooled across crops.

Crop	KAGGLE (validation)				MENDELEY (external)			
	R ²	RMSE	MAE	MAPE%	R ²	RMSE	MAE	MAPE%
Apple	0.9980	0.0070	0.0039	0.85	0.6251	0.0458	0.0339	11.06
Banana	0.9988	0.0049	0.0037	0.48	-0.4697	0.0693	0.0601	8.90
Blackgram	0.9995	0.0041	0.0030	0.39	0.5273	0.0385	0.0318	4.16
Chickpea	0.9990	0.0047	0.0035	0.73	0.1990	0.0604	0.0330	4.71
Coconut	0.9994	0.0047	0.0034	0.48	0.7055	0.0396	0.0298	4.66
Coffee	0.9995	0.0040	0.0030	0.38	0.5724	0.0350	0.0150	2.20
Cotton	0.9994	0.0045	0.0032	0.45	0.4674	0.0404	0.0242	3.34
Grapes	0.9989	0.0052	0.0025	0.60	0.8084	0.0375	0.0235	6.56
Jute	0.9992	0.0044	0.0033	0.41	-0.0146	0.0400	0.0299	3.94
Kidneybeans	0.9986	0.0059	0.0040	0.61	0.5554	0.0411	0.0262	3.39
Lentil	0.9996	0.0037	0.0027	0.35	0.6518	0.0349	0.0272	3.51
Maize	0.9997	0.0034	0.0025	0.35	0.7825	0.0328	0.0175	2.50
Mango	0.9994	0.0043	0.0031	0.42	0.2026	0.0595	0.0511	6.38
Mothbeans	0.9997	0.0037	0.0025	0.32	0.6919	0.0370	0.0239	3.24
Mungbean	0.9989	0.0052	0.0037	0.49	0.2807	0.0441	0.0388	5.46
Muskmelon	0.9987	0.0057	0.0041	0.57	0.6504	0.0482	0.0313	5.10
Orange	0.9996	0.0041	0.0027	0.43	0.7625	0.0505	0.0205	3.75
Papaya	0.9988	0.0054	0.0038	0.52	0.7695	0.0424	0.0179	2.91
Pigeonpeas	0.9998	0.0032	0.0023	0.35	0.7897	0.0342	0.0145	2.47
Pomegranate	0.9994	0.0046	0.0034	0.48	0.8317	0.0281	0.0176	2.53
Rice	0.9993	0.0047	0.0033	0.44	0.7424	0.0496	0.0214	3.39
Watermelon	0.9981	0.0060	0.0041	0.55	0.2836	0.0415	0.0323	4.49
OVERALL	0.9995	0.0048	0.0033	0.48	0.9201	0.0443	0.0283	4.48

Overall validation results showed per crop R² is at 0.9995 while for ATA data is at 0.9201 with RMSE of 0.0048 and 0.0443 respectively. The results further showed that on ATA data, R² is well represented for pomegranate, pigeonpeas, and maize at 0.83, 0.79 and 0.78 respectively, however it was lower for banana and jute. The results showed that crop suitability has little variance on external soil and compared to the training data. It was also noted that based on MAPE, the model performed well on Coffee, pigeonpeas, maize and pomegranate at 2.20, 2.47, 2.50, and 2.53. while the percentage error was more on Apple, Banana, Grapes and Mango at 11.06, 8.90, 6.56, and 6.38 respectively. The lower percentage error in Maize and pigeonpeas is attribute to the fact that these crops were common in both the datasets while the high errors in the listed crops were attributed to their absence in the data that was used to train the model. These results showed that if the model was trained on the ATA dataset, the model would have worked better in predicting suitability of the crops listed in the dataset.

3.1.4 SHAP (SHapley Additive exPlanations) Analysis

To best understand the SHAP interpretations, consider an example for maize as a crop among the others. From the dataset, a soil sample with features as indicated in Table 6 is analyzed;

Table 6: Sample soil features for calculation of maize suitability score

Feature	Absolute Value	Generated SHAP Value
Nitrogen (N)	13.0 mg/kg	-0.012

Phosphorus (P)	23.0 mg/kg	0.062
Potassium (K)	6.0 mg/kg	0.023
Temperature	23.96 °C	0.002
Humidity	90.26%	- 0.022
pH	7.37	0.000

This sample formed a single row $x \in \mathbb{R}^{1 \times 6}$ in the feature matrix I_M , which is the input to the LightGBM regressor trained to predict maize suitability.

The SHAP explanation model approximated the LightGBM prediction for maize suitability as;

$$f(x) \approx g(x) = \phi_0 + \phi_i$$

Where ϕ_0 = Expected base value (average maize suitability score across all samples)

$\phi_i = \phi_N + \phi_P + \phi_K + \phi_{Temp} + \phi_{Humidity} + \phi_{PH}$ which represents the SHAP value of feature i
(Quantifies contribution of each feature to the prediction of this specific sample)

Expected base value (Mean across all samples) $\phi_0 = 0.8111$

The value of $f(x)$ as computed by the model is 0.8654

SHAP Contribution for the soil sample in the Table 6 thus becomes;

$$g(x) = \text{Expected Baseline value } (\phi_0) + (\phi_i) - \text{Sum of SHAP Values } (\phi_N + \phi_P + \phi_K + \phi_{Temp} + \phi_{Humidity} + \phi_{PH})$$

$$g(x) = 0.8111 + (-0.012 + 0.062 + 0.023 + 0.002 - 0.022 + 0.000) \\ = 0.8641$$

To confirm that SHAP values correctly explain the prediction, a difference of $f(x)$ and $g(x)$ was calculated.

$$f(x) - g(x) = 0.8654 - 0.8641 \\ = 0.0013$$

The value 0.0013 indicated that the prediction for maize suitability is correct with results analysis suggesting that Phosphorus (P) and Potassium (K) had the highest global importance across in maize while Nitrogen (N) and Humidity were of lower value comparatively.

This phenomenon was observed across most crops and the contribution importance of pH and temperature contributed variably, often interacting with other features dependent on the crop as indicated by the generated SHAP values in Table 7.

Table 7: Table summary showing SHAP values of each of the 22 crop in the dataset

Crop	Predicted Suitable Score $f(x)$	Expected Base Value (ϕ_0)	Sum of SHAP Values (sum ϕ_i)	Computed $g(x) = \phi_0 + \text{sum } \phi_i$	$f(x) - g(x)$
apple	0.3644	0.4623	-0.098	0.3643	0.0001

banana	0.7698	0.7835	-0.014	0.7695	0.0003
blackgram	0.8587	0.8067	0.051	0.8577	0.001
chickpea	0.3591	0.5108	-0.151	0.3598	-0.0007
coconut	0.8796	0.7741	0.106	0.8801	-0.0005
coffee	0.8397	0.8119	0.027	0.8389	0.0008
cotton	0.8652	0.8027	0.062	0.8647	0.0005
grapes	0.4008	0.4944	-0.094	0.4004	0.0004
jute	0.8454	0.8357	0.01	0.8457	-0.0003
kidneybeans	0.6262	0.6718	-0.046	0.6258	0.0004
lentil	0.869	0.8081	0.06	0.8681	0.0009
maize	0.8654	0.8111	0.053	0.8641	0.0013
mango	0.7587	0.7682	-0.01	0.7582	0.0005
mothbeans	0.9019	0.8142	0.087	0.9012	0.0007
mungbean	0.9078	0.8021	0.106	0.9081	-0.0003
muskmelon	0.8385	0.7697	0.068	0.8377	0.0008
orange	0.9744	0.7304	0.245	0.9754	-0.001
papaya	0.8374	0.7672	0.07	0.8372	0.0002
pigeonpeas	0.8729	0.8069	0.067	0.8739	-0.001
pomegranate	0.8498	0.7794	0.072	0.8514	-0.0016
rice	0.8349	0.7929	0.043	0.8359	-0.001
watermelon	0.8086	0.8062	0.002	0.8082	0.0004

However, these need to be compared to the standardized optimal ranges to ascertain the interpretability and applicability of the proposed model. Table 8 presents the optimal ranges of these metrics for maize suitability.

Table 8: Soil Quality Indicator Threshold ranges for maize suitability, source [32]

Parameter	Recommended Range
Soil Nitrogen (N)	80-100 mg/kg
Soil Phosphorus (P)	45-60 mg/kg
Soil Potassium (K)	35-50 mg/kg
Temperature	22-30 °C
Relative Humidity	55-75%
Soil pH	5.5-7.0

In this case of Maize as the sample crop, the results as shown in table 6 above, Phosphorous (P) had a SHAP value of 0.062 which formed the highest positive contributor. This indicated that, a P level of 23.0 mg/kg is suitable for maize growth, however, fertilizer in P content needs to be applied to raise the levels to the required optimum of 45-60 mg/kg to realize improved maize growth. Potassium (K) had a positive SHAP contribution of 0.023 for 6.0 mg/kg of k levels. This was below the optimum range of 35-50 mg/kg meaning for maize suitability, fertilizer that is more of K should be added to improve on maize suitability. Temperature as well had a positive SHAP contribution of 0.002 for a temperature value of 23.96°C. This meant that the temperature value was favorable for maize productivity that required a favorable temperature of between 23°C and 30°C. Negative SHAP values indicated detrimental contribution to the growth of a crop. The results showed that Humidity had the highest negative SHAP contribution of -0.022 for a humidity level of 90.26% which was not suitable for maize production as it prefers humidity levels of 55% to 75%. Similarly, Nitrogen (N) had a negative SHAP contribution of -0.012 for a nitrogen level at 13.0 mg/kg. Meaning fertilizer that is more in Nitrogen is required to boost its level to the optimum of 80 to 100 mg/kg for maize suitability. pH had SHAP contribution of 0.000 for the pH value of 7.37. This was closer to the range of optimal value of maize growth of 5.5-7.0 making the effect be suitable for maize growth.

From the results, the actionable insight towards improving suitability score for maize include farmers focusing on increasing N (e.g., via nitrogen fertilizers) and reducing humidity (e.g., through improved drainage). While P and K are positive contributors, they are still below optimal levels and could be adjusted further to improve maize productivity. Similarly, the optimal values all the other crops need to be determined to adequately provide the necessary insights to a farmer or the user of the model.

Whereas this example presents the feature importance of only maize as a crop, Table 9 presents a summary for all the 22 crops used during the study in the Kaggle Dataset. Table 5 presents the results for the validation dataset.

Table 9: A summary of feature importance of all the 22 crops in the dataset

<i>Crop</i>	<i>N</i>	<i>P</i>	<i>K</i>	<i>Temperature</i>	<i>Humidity</i>	<i>pH</i>	<i>Base Value</i>	<i>Predicted f(x)</i>
apple	0.0009	0.0558	0.0686	0.0007	0.028	0.0031	0.4623	0.3644
banana	0.0253	0.0047	0.0027	0.0031	0.0131	0.0067	0.7835	0.7698
blackgram	0.0027	0.0384	0.0253	-0.0006	-0.019	0.0053	0.8067	0.8587
chickpea	0.0169	0.0095	0.0335	0.0026	-0.1175	0.0042	0.5108	0.3591
coconut	0.0074	0.0168	0.044	0.0049	0.0418	0.0093	0.7741	0.8796
coffee	0.0256	0.0605	0.0051	0.0061	-0.0102	0.0021	0.8119	0.8397
cotton	0.0297	0.0538	0.0093	0.013	0.0153	0.0008	0.8027	0.8652
grapes	0.0009	-0.051	0.0637	-0.0007	0.0203	0.0007	0.4944	0.4008
jute	0.0207	0.0467	0.0346	0.008	0.0087	0.0017	0.8357	0.8454
kidneybeans	0.0038	0.032	0.009	0.0001	-0.0758	0.0147	0.6718	0.6262
lentil	0.0064	0.0443	0.0272	0.0007	-0.0201	0.0024	0.8081	0.869
maize	0.0115	0.0616	0.0234	0.0024	-0.022	0.0004	0.8111	0.8654
mango	0.0047	0.0605	0.0116	-0.0003	-0.0626	0.0002	0.7682	0.7587

mothbeans	0.0091	0.0662	0.0292	0.0002	-0.0179	0.001	0.8142	0.9019
mungbean	0.0079	0.015	0.0522	0.0009	0.0271	0.0027	0.8021	0.9078
muskmelon	- 0.0038	0.0038	0.0042	0.002	0.0633	- 0.0006	0.7697	0.8385
orange	0.0114	0.0066	0.1197	0.0008	0.1045	0.001	0.7304	0.9744
papaya	0.0022	- 0.0011	- 0.0018	-0.0029	0.0759	- 0.0021	0.7672	0.8374
pigeonpeas	0.0072	0.0446	0.0267	-0.0015	-0.0089	- 0.0022	0.8069	0.8729
pomegranate	0.0088	0.0545	- 0.0434	0.0039	0.0477	- 0.0012	0.7794	0.8498
rice	- 0.0164	0.0022	0.0136	0.0019	0.0409	- 0.0002	0.7929	0.8349
watermelon	- 0.0151	0.0361	- 0.0459	0.0084	0.0203	- 0.0014	0.8062	0.8086

4. Conclusions and Recommendations

This study successfully developed a robust and interpretable machine learning framework for precision agriculture by integrating IoT sensor data, LightGBM modeling, grid search, and SHapley Additive exPlanations (SHAP). The multi-output regression model demonstrated exceptional predictive performance, achieving an R^2 score of 0.9991, RMSE of 0.0047, and MAPE below 0.5%, validating its accuracy and reliability across 22 crop varieties. SHAP analysis provided critical insights into feature importance, revealing that Phosphorus (P) and Potassium (K) were the most influential soil parameters, while humidity and Nitrogen (N) played crop-specific roles. These findings empower stakeholders to make data-informed decisions on fertilizer application, irrigation, and crop selection. The integration of grid search further enhanced the model's prescriptive capabilities, enabling targeted soil adjustments. The modular dashboard design ensures accessibility for non-technical users, bridging the gap between complex analytics and practical agronomic action. The study recommends integration of sensors to enrich the dataset and provide real time monitoring and prediction of soil quality parameters as well as collaboration with agricultural policy makers for adoption of the model in supporting data driven decisions with regards to agricultural production.

References

- [1] K. N. Setty, "Modern Techniques and Applications in Soil Testing: A Comprehensive Study," JETIR, 2025. [Online]. Available: www.jetir.org/d115
- [2] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review," 2021, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2020.3048415.
- [3] Zhao, "477800620," 2021. [Online]. Available: <https://www.researchgate.net/publication/354802124>
- [4] M. A. Javed and M. A. Azmi Murad, "Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability," Dec. 30, 2024, *Elsevier Ltd.* doi: 10.1016/j.heliyon.2024.e40836.
- [5] D. C. Țopa, S. Căpșună, A. E. Calistru, and C. Ailincăi, "Sustainable Practices for Enhancing Soil Health and Crop Quality in Modern Agriculture: A Review," May 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/agriculture15090998.
- [6] G. Bilas *et al.*, "Land Suitability Analysis as a Tool for Evaluating Soil-Improving Cropping Systems," *Land (Basel)*, vol. 11, no. 12, Dec. 2022, doi: 10.3390/land11122200.

- [7] H. Mugiyo *et al.*, “Evaluation of land suitability methods with reference to neglected and underutilised crop species: A scoping review,” Feb. 01, 2021, *MDPI AG*. doi: 10.3390/land10020125.
- [8] M. Atzmueller, J. Fürnkranz, T. Kliegr, and U. Schmid, “Explainable and interpretable machine learning and data mining,” Sep. 01, 2024, *Springer*. doi: 10.1007/s10618-024-01041-y.
- [9] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable ai: A review of machine learning interpretability methods,” Jan. 01, 2021, *MDPI AG*. doi: 10.3390/e23010018.
- [10] R. Marcinkevičs and J. E. Vogt, “Interpretable and explainable machine learning: A methods-centric overview with concrete examples,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 13, no. 3, May 2023, doi: 10.1002/widm.1493.
- [11] R. P. Potdar, M. M. Shirolkar, A. J. Verma, P. S. More, and A. Kulkarni, “Determination of soil nutrients (NPK) using optical methods: a mini review,” 2021, *Bellwether Publishing, Ltd.* doi: 10.1080/01904167.2021.1884702.
- [12] D. M. Sobhy and A. Anandhi, “Soil Nutrient Monitoring Technologies for Sustainable Agriculture: A Systematic Review,” Sep. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/su17188477.
- [13] A. Sagwal, P. Wadhwa, Shubham, and S. Kaushal, “Essentiality of Micronutrients in Soil: A Review,” *Int. J. Plant Soil Sci.*, vol. 35, no. 24, pp. 56–65, Dec. 2023, doi: 10.9734/ijpss/2023/v35i244297.
- [14] S. Ameer, H. Ibrahim, F. N. U. Kulsoom, G. Ameer, and M. Sher, “Real-time detection and measurements of nitrogen, phosphorous & potassium from soil samples: a comprehensive review,” Jul. 01, 2024, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s11368-024-03827-5.
- [15] R. Yaulilahua-Huacho *et al.*, “Examining the adaptability of soil pH to soil dynamics using different methodologies: A concise review,” *Journal of Experimental Biology and Agricultural Sciences*, vol. 12, no. 4, pp. 573–587, Sep. 2024, doi: 10.18006/2024.12(4).573.587.
- [16] Z. Guo, J. Han, Y. Zhang, and H. Zhuang, “Effects of temperature and time treatment on soil characteristics and wheat growth,” *Front. Sustain. Food Syst.*, vol. 9, 2025, doi: 10.3389/fsufs.2025.1499224.
- [17] T. Grodek, “The Secret of Soil Humidity Comprehending, Assessing, and Resulting in Rural Farming,” 2023, doi: 10.4172/2325-9647.1000291.
- [18] D. Reisig, “Soil moisture influence on nutrient availability in the soil,” 2023. [Online]. Available: <http://www.internationalscholarsjournals.com/>
- [19] B. Dey, J. Ferdous, and R. Ahmed, “Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables,” *Heliyon*, vol. 10, no. 3, Feb. 2024, doi: 10.1016/j.heliyon.2024.e25112.
- [20] H. Sugeru, E. P. Wibowo, P. Musa, R. K. Harahap, and S. Harmanto, “Integrating IoT and Machine Learning for Real-Time Soil-Based Crop and Fertilizer Recommendations: The UG-AgroPlan System,” Sep. 24, 2025. doi: 10.20944/preprints202509.2044.v1.
- [21] B. Dey, J. Ferdous, and R. Ahmed, “Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables,” *Heliyon*, vol. 10, no. 3, Feb. 2024, doi: 10.1016/j.heliyon.2024.e25112.
- [22] M. Bugaj, K. Wrobel, and J. Iwaniec, “Model Explainability using SHAP Values for LightGBM Predictions,” in *International Conference on Perspective Technologies and Methods in MEMS*

Design, Institute of Electrical and Electronics Engineers Inc., May 2021, pp. 102–106. doi: 10.1109/MEMSTECH53091.2021.9468078.

- [23] S. M. Lundberg, P. G. Allen, and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” 2018. [Online]. Available: <https://github.com/slundberg/shap>
- [24] A. M. Salih *et al.*, “A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME,” *Advanced Intelligent Systems*, vol. 7, no. 1, Jan. 2025, doi: 10.1002/aisy.202400304.
- [25] S. Kawakura, M. Hirafuji, S. Ninomiya, and R. Shibasaki, “Analyses of Diverse Agricultural Worker Data with Explainable Artificial Intelligence: XAI based on SHAP, LIME, and LightGBM,” *European Journal of Agriculture and Food Sciences*, vol. 4, no. 6, pp. 11–19, Nov. 2022, doi: 10.24018/ejfood.2022.4.6.348.
- [26] E. E. Hussein *et al.*, “Harnessing Explainable AI for Sustainable Agriculture: SHAP-Based Feature Selection in Multi-Model Evaluation of Irrigation Water Quality Indices,” *Water (Switzerland)*, vol. 17, no. 1, Jan. 2025, doi: 10.3390/w17010059.
- [27] R. N. V. J. Mohan, P. S. Rayanoothala, and R. P. Sree, “Next-gen agriculture: integrating AI and XAI for precision crop yield predictions,” *Front. Plant Sci.*, vol. 15, 2024, doi: 10.3389/fpls.2024.1451607.
- [28] Y. Akkem, S. K. Biswas, and A. Varanasi, “Role of Explainable AI in Crop Recommendation Technique of Smart Farming,” *International Journal of Intelligent Systems and Applications*, vol. 17, no. 1, pp. 31–52, Feb. 2025, doi: 10.5815/ijisa.2025.01.03.
- [29] A. Mamatha and G. Santhosh Reddy, “Enhanced Weed Detection in Agricultural Fields Using Convolutional Neural Networks and SHAP Interpretability Techniques,” 2024. [Online]. Available: <https://www.jisem-journal.com/>
- [30] Y. A. Ali, E. M. Awwad, M. Al-Razgan, and A. Maarouf, “Hyperparameter Search for Machine Learning Algorithms for Optimizing the Computational Complexity,” *Processes*, vol. 11, no. 2, Feb. 2023, doi: 10.3390/pr11020349.
- [31] N. Subaşı, “Comprehensive Analysis of Grid and Randomized Search on Dataset Performance,” *European Journal of Engineering and Applied Sciences*, vol. 7, no. 2, pp. 77–83, Dec. 2024, doi: 10.55581/ejeas.1581494.
- [32] S. Nepali and R. Tamang, “Phosphorus and potassium interaction in maize: enhancing yield and soil health under Nepalese agroecosystems,” *Frontiers in Agronomy*, vol. 8, 2026, doi: 10.3389/fagro.2026.1795379.

Ethical statement

The study does not in any way involve experiment on human or animals, as a result, no ethical approval was required or applicable.

Clinical trial number

Not Applicable.

Consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Funding

Not applicable.

Code availability (software application or custom code)

Code will be made available on request.

Credit Authorship contribution statement

Kitaria D.: Conceptualization, Writing the original draft, IoT setup, machine Learning model development. Mutua S.: Supervision, Methodology, Review and editing, code review. Chege A.: Review and editing, related work review, formal analysis, IoT and machine learning model review.

Declaration of competing interest

No competing interests (Financial or personal) from the authors that could influence the work in this article.

Acknowledgement

The authors are thankful to Solomon Kamau and Morris Murigi for the role they played in setting up the IoT and Machine Learning Models and integrating them to the cloud servers. This was very essential for achieving the results of this research work. Special thanks to Isaac Nyagah Maina for his meticulous proofreading, which greatly enhanced the clarity and readability of the article for a wider audience.

Data availability

The Kaggle dataset was used for training, testing and validation of the model and can be found on crop-recommendation-dataset.

The Ethiopian Dataset was used purely for validation to assess the Models performance on new datasets and can be found on Crop Recommendation using Soil Properties and Weather Prediction Dataset - Mendeley Data