Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 14, Issue. 7, July 2025, pg.78 – 89

Machine Learning-Driven Soil Health Analysis for Precision Agriculture: Sensor Based Fertilizer Recommendation

Emmanuel Chinembiri[1]; Rachel Chikoore[2]; Brian Mupini[3]

[1,2,3]Information Science & Technology, Harare Institute of Technology, Harare, Zimbabwe
[1] chinembiriemma@gmail.com; [2] rchikoore@hit.ac.zw; [3] bmupini@hit.ac.zw

DOI: https://doi.org/10.47760/ijcsmc.2025.v14i07.008

Abstract: Optimal application of fertilizer is key in maximizing crop yields with minimal damage to the soil health, yet conventional soil health analysis remain labour intensive, time consuming and inaccessible to many farmers. This study proposes a machine learning-based fertilizer recommendation system utilizing soil nutrient status to improve soil health analysis. A hybrid machine-learning model synthesizing Extreme Gradient Boosting and Random Forest was trained on a dataset containing Agronomical fertilizer recommendations based on Nitrogen, Phosphorous, Potassium, Soil type and crop type. The study evaluated the performance of an RS485-based digital NPK soil sensor in a sub-study by correlating sensor readings with laboratory test results. In 10 samples per soil type, sensor readings were found to be within ±10% of laboratory values, proving their reliability in real-time field measurements. The findings identify the potential for combining ensemble machine learning algorithms with low-cost sensors to offer scalable, real-time, and accurate fertilizer recommendations for smallholder and commercial farmers. The integrated approach supports data-driven agricultural decision-making and opens the way for smart, sustainable nutrient management.

Keywords: Machine Learning, Extreme Gradient Boosting, Random Forest, RS485 Sensor, NPK

I. Introduction

Agriculture is one of the main sources of livelihood in Zimbabwe, and it is central to the economy. About 70% of the population depends directly or indirectly on it for their survival and well-being and is crucial for food security [1]. Smallholder farmers dominate the sector, growing the country's staple grains, especially maize, sorghum and groundnuts, which occupy about 75% of their cultivated area. In this way, smallholder agriculture is central to national food security. In healthy soils, plants have adequate nutrients, moisture and physical support, and a diverse microbial community drives nutrient and energy cycling [2]. Many smallholder farmers in Zimbabwe are confronted with degraded soils because of factors acidification, low organic matter and multiple nutrient deficiencies. Field surveys confirm this: soils with very little topsoil or organic carbon yield only sparse, short-lived crops [3].

Studies show that soil nutrient deficiency is widespread in Zimbabwe, and many smallholder farmers do not even know the nutrient status of their soils. Conventional farming is known for over-fertilization in some areas at the cost of under-fertilization, in other cases farmers often use blanket applications to increase yield. This imprecise application of nutrients leads to lower yields and creates environmental problems such as pollution of surface and groundwater resources, and soil degradation [4].

In Zimbabwe soil health analysis relies on Conventional lab soil-testing methods to determine soil nutrients such as Nitrogen, Potassium and phosphorous(NPK) to recommend appropriate fertilizers to farmers. Most soil testing centers are predominately in Harare making it difficult for farmers in rural areas to access the facility. The conventional soil testing is often expensive, time-consuming, and unavailable, resulting in less-than-ideal application techniques by farmers. It takes more than a week for a farmer to get their soil sample results and an additional week to get crop recommendation. Recent advances in soil analysis technology open up opportunities for addressing these challenges. NPK Sensor, which measure nitrogen (N), phosphorus (P), and potassium (K) concentrations in soil, provide for rapid and cost-effective assessment of soil health . They allow obtaining real-time information about nutrient status, making it easier for farmers to apply fertilizers at the appropriate rate. Therefore, there is a greater need to explore modern technologies that can leverage on advancement in artificial intelligence and IOT to provide custom fertilizer recommendations based on real-time soil health data.

II. Review of Related Literature

A. *Importance of Soil Health Assessments*

Soil health assessment is crucial in understanding the fertility status of soils which directly influence agricultural productivity and environmental sustainability [4]. Healthy soils have diverse biological communities, water-holding capacity, and nutrient supply that contribute to plant growth. There must be effective soil testing methods to measure these attributes. Research has shown that soil health is linked with the physical, chemical, and biological properties of the soil that need an integrated method of soil testing. Quantifying soil health is dominated by chemical indicators such as the NPK values.

Soil health profiling is commonly prioritised as the foundation in bringing about environmental sustainability in agriculture. Chemical pesticide and fertilizer misuse and excessive application can damage soil in the long term and lead to undesirable characteristics such as topsoil erosion, nutrient leaching, water contamination, and devastation of beneficial soil organisms [5].

Soil analysis offers an immediate path to increased productivity, food security, and economic sustainability. A majority of rural agriculture is based on outdated methods of operation that are not adequate to the specific nutritional demands of their soils [6]. Without access to scientific information, they will under-fertilize, lower yields, or over-fertilize, which damages the environment and wastes resources. In contrast, site-specific and calibrated soil data allows farmers to make effective and precise input decisions [7].

Quantification of soil health is the future of sustainable agriculture. It enables more intelligent agronomic decisions to be made, improves environmental performance, and facilitates provision for developing farming systems that are resilient and can adapt to climate uncertainty. As global food systems get on with providing more people with more food while keeping the natural resource base intact in the presence of such challenges, investment in soil health is one of the most strategic options available today.

B. *Advances in Soil testing technologies*

New technologies in soil health analysis, particularly the advent of Inter of Things(IoT) soil sensors, offers effective solutions to the deficiencies of traditional soil testing [8]. The sensors deliver real-time, field-based estimations of soil health through live measurement of vital nutrient levels such as the NPK values. Changsha Zoko Link Technology Co., 2024 examined RS485 soil sensor's cost-effectiveness in Sub-Saharan Africa, where limitations in infrastructure demand low-cost, cabled systems [9]. Their study confirmed that RS485-based soil monitoring systems reduced deployment costs by 40% compared to wireless configurations because they eliminated repeaters and complex mesh networks. Farmers using RS485-connected NPK sensors reported 30% faster decision-making due to real-time data access. The RS485 sensors can detect nitrogen, phosphorus, and potassium concentrations, giving farmers timely feedback to maximize fertilizer use. Even though these possess their advantages, they come with limitations as well. Integration is complex with a lack of sensor register mapping standardization, requirements for custom firmware or manual configuration [10]. RS-485 devices are usually in need of an external power source, which is not always readily available within remote agricultural sites unless supported by solar or battery systems.

Overall, RS-485 soil sensors offer an efficient, robust, and technology appropriate way of creating precision agriculture [9]. Their adoption, however, is dependent on adequate infrastructure, training of the users, and calibration that is context-specific to ensure accuracy of sensor measurements and reliability of the system.

C. *Machine Learning in Soil Nutrient Prediction*

The recent years have seen a surge in the use of machine learning (ML) in agriculture, namely in soil data analysis for fertilizer optimization and nutrient prediction. A variety of model types ranging from ensemble decision trees to self-supervised models and neural networks has been used. The application of ML here has been driven by the large and growing volume of satellite data, in-situ soil sensor measurements, and high-dimensional agro ecological data. These models, besides enhancing the accuracy of nutrient estimation, enhance spatial generalizability and decision making in actual farm conditions.

Ensemble tree-based methods, notably Random Forest (RF) and Extreme Gradient Boosting (XGBoost), have been widely used in current studies as they can be utilized with heterogeneous data and identify nonlinear interaction among soil parameters. Kammerlander et al. (2025) contrasted RF and XGBoost with deep neural networks on a dense and diverse dataset of satellite images, weather, and ground-based soil measurements on various continents [11]. RF provided satisfactory interpretability and baseline performance, while XGBoost performed better than RF for cases with complex feature interaction in all but one scenario and provided higher accuracy when it came to the prediction of the soil nitrogen and phosphorus content. For instance, when estimating total nitrogen in the soil, XGBoost provided an R^2 of 0.63 compared to RF's 0.48, a testament to its higher capacity to detect complex relationships among variables.

In hyperspectral soil analysis, XGBoost in combination with advanced feature selection methods (like Continuous Wavelet Transform and Random Frog) significantly enhanced model accuracy. [26] demonstrated that the XGBoost model, following optimization with key hyperspectral bands, yielded better prediction of soil organic carbon than conventional decision trees or support vector regression. This indicates the potential of XGBoost in high-dimensionality spectral data situations, which are increasingly relevant in the midst of remote sensing technology expansion in precision agriculture.

III. METHODOLOGY AND MATERIALS

This research used a comparative and experimental research design to evaluate effectiveness of a machine learning model to recommend fertilizers; the model is benchmarked against agronomical decisions dataset. A key pillar of this research was also the empirical validation of a low cost RS485 soil sensor against results obtained from NPK nutrient chemical extraction. The research was conducted at an anonymized Soil testing facility in Zimbabwe which is specialized in agricultural innovation and soil fertility measurement. The institution provided a large-scale database of soil testing results, laboratory equipment, and local agronomic expertise, ensuring an intricate and relevant environment for experimental study. The research design allowed for the achievement of a comprehensive understanding of how machine learning algorithms and modern soil sensors can improve soil health monitoring [12].

A. *Comparative Design*

The comparative aspect of the research design was focused on comparing and evaluating some of the machine learning models to determine which model makes the best prediction using a dataset of past fertilizer recommendations. The study specifically compared the performance of five algorithms which are XGBoost, Multilayer Perceptron (MLP), Random Forest, Support Vector Machine, and a new hybrid model based on the literature from previous studies. These models were trained using a soil history data set with variables such as soil type, crop type, and macronutrient levels (Nitrogen, Phosphorus, and Potassium), in which the target variable is the recommended fertilizer. All algorithms were evaluated against equal performance metrics (accuracy, precision, F1-score, and Confusion matrix) to have a comparative framework. The result of this comparative analysis helped to decide the most accurate and consistent model for real-world application within fertilizer advisory systems. In addition to model comparison, comparative design was also used for the impact of data balancing techniques (e.g., SMOTE) on classification performance, particularly for minority fertilizer classes.

B. *Experimental Design*

The experimental was used on evaluating and comparing the efficacy, precision, and feasibility of soil nutrient (N, P, K) measurements using RS485 sensors against conventional laboratory soil testing. 50 Samples of soils were collected ensuring representative coverage for a variety of soil types along with varied levels of nutrients. NPK values were directly measured by the sensor and two measurements per sample were taken to account for variation, the average being recorded. The outcomes from both methods were next statistically contrasted to compare the consistency between the sensor-based and laboratory-based measurements.

C. *Key materials components of the research*

At the heart of the study was the RS485-based NPK soil sensor, selected for its robustness, cost-effectiveness, and serial communication compatibility for agricultural applications. The sensor measures the concentration of the three primary macronutrients which are Nitrogen (N), Phosphorus (P), and Potassium (K) in real time. RS485 is long-range communication compatible (up to 1.2 km) and not affected by electromagnetic interference, thus an ideal choice for field applications. The sensor is an ion-selective electrode-based sensor and is inserted into moist soil to detect concentrations of nutrient ions. Data is communicated in the form of Modbus RTU protocol, an industrial standard with broad recognition that facilitates easy communication with microcontrollers, data loggers, or edge computing platforms. USB-to-RS485 converter module was used in communication between the sensor and the laptop, converting differential signals from the sensor into data which the USB could understand. RS485-to-USB adapter is a communication interface converter that enables a computer to communicate with sensors or devices that use the RS485 serial communication protocol. RS485 is used in industrial and environmental application, such as soil NPK sensors, because it can sustain long-distance, noisy, and multi-point communication

[13]. Modern laptops and PCs lack RS485 ports, the converter bridges this gap by translating RS485 signals into USB signals compatible with the. On the RS485 side, the device is connected to the differential signal wires of the sensor (A+ and B−), while on the USB side, it is connected to the laptop as a virtual COM port. This allows standard serial software to interact with the sensor as if it were connected to a native serial port. The RS485-to-USB converter is, however, powered from the USB port, while the sensor on it typically requires its own external power supply, e.g., 12V DC. The converter is usually plug-and-play, compact, with high baud rate range, and there are transmission and receiving indicator LEDs. Suitable configuration of the communication parameters (baud rate, parity, stop bits) and application of suitable shielding and termination on the cables needs to be performed in order to offer signal integrity. This unit was chosen because of its ability to continually link the RS485-based soil sensor to the computer for efficient data capture and analysis towards fertilizer suggestion.

D. Data Source

The primary data source for this study was an anonymized Soil testing facility in Zimbabwe where a historical dataset of soil recommendations was collected. The data contains records collected between 2021 to 2023. This served as the foundation for training and testing the fertilizer recommendation model. The institution has standardized soil sampling and testing protocols that render the data reliable. The soil samples used for building the dataset were directly collected from farmers and tested in the laboratories using standard chemical analysis techniques. The data included some of the key soil parameters such as nitrogen, phosphorus, potassium (NPK), crop type and soil type.

In addition to the main dataset, a sub-study was also conducted to assess the efficiency and accuracy of the RS485-based soil sensors. For this purpose, data were collected directly from the sensors in a lab controlled environment, measuring real-time soil NPK values. Sensor readings were collected systematically and then compared with laboratory test results for the same soil samples. 50 samples were used covering five main soil types. This comparison exercise enabled the determination of the level of agreement between the sensor readings and the laboratory standards, and thus determining the feasibility of the implementation of sensor technology in everyday soil testing routines. This sub-study was not only necessary in verifying the accuracy of the sensors but also in understanding potential discrepancies between sensor readings and controlled laboratory tests.

The dataset had 26479 records that was collected between 2021 and 2024. The initial dataset before cleaning had six main features, the serial number, nitrogen (N), phosphorus (P), potassium (K), crop type, and soil type. The soil type has five categories: sandy, loamy, clayey, black, and red. Five crop categories were prioritized for this research, Maize, Millet, Tobacco, groundnuts, wheat.

E. Ethical Consideration

This research was compliant to Zimbabwe’s Data Protection Act, upholding integrity, transparency, and ethical use of resources and information. Under this study, all the soil samples and data were obtained through institutional sources through farmers’ submissions and their personal details were discarded. the study was conducted under the umbrella of the Harare Institute of Technology (HIT) Research Ethics Policy, under which integrity, transparency, and accountability in research are demanded. Approval to use the dataset was obtained before conducting the study, and all analyses were conducted solely for academic purposes in compliance with HIT's student and researcher code of ethics. Care was taken to present the findings and recommendations objectively, not exaggerated or misrepresented, and that the research provides a valuable contribution to Zimbabwean sustainable agriculture. Data employed in this study were provided generously by an anonymous institution that collected the soil data as part of its research and extension work. Gratitude was extended for their generosity while honoring their anonymity requirement. The author did not conduct the first data gathering and only received raw anonymized dataset.

F. Handling Missing Values

The dataset contained 99 entries with missing Soil Type data, 108 entries with missing Crop Type, 20 with missing Nitrogen, 10 with missing Potassium, 5 with missing Phosphorous and 96 missing Fertilizer Name. It can be suggested that the missing data might not be random but could be influenced by other variables. Then there is need for more statistical analysis to determine if the data is Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR)**Invalid source specified.**. The results can then inform suitable methods such as imputation, deletion, or adding indicators for missing entries to effectively manage the incomplete data without skewing the model.

As illustrated in the figure above, a Chi-squared test was conducted to explore whether the missing data in the Crop Type variable is linked to the observed values of Soil Type. The test helps us to conclude that missing data in Crop Type is influenced by another observed variable (which would indicate Missing at Random MAR) or that it occurs completely independent (Missing Completely at Random MCAR).

The null hypothesis (H_0) means that the missing entries in Crop Type are independent of Soil Type, supporting the idea that the missing entries are MCAR. On the other hand, the alternative hypothesis (H_1) suggests that the missing entries in Crop Type is related to Soil Type, which could imply MAR or even MNAR if it’s tied to unobserved factors. The p-value calculated was 0.7465, which is significantly higher than the usual significance threshold of 0.05. Since the p-value is greater than 0.05, we do not reject the null hypothesis. There’s no statistically significant evidence to suggest that the missing entries in Crop Type is connected to Soil Type. This indicates that the missing entries in Crop Type might reasonably be viewed as Missing Completely at Random (MCAR), at least in relation to Soil Type.

The test conducted proved no significant link between the missing Crop Type data and the observed Soil Type values ($p = 0.7465$). This indicates that the missing Crop Type entries is probably Missing Completely at Random (MCAR) in relation to Soil Type. Given the missing data under the MCAR assumption doesn't skew the analysis, it's recommended to drop records with missing Crop Type without affecting the model's validity. The same goes for Soil Type, which also had some missing values, but we found no evidence that these were systematically tied to other observed variables. Both Soil Type and Crop Type are crucial for making fertilizer recommendations, and since the missing data seems random, keeping incomplete records could actually lower the quality of the input features and add unnecessary noise. So, to maintain the dataset's integrity and ensure the model is trained on complete and trustworthy information, we decided to exclude records with missing Soil Type and Crop Type from further analysis.

Missing values in Fertilizer name is replaced with mode values. To cover this space, missing values in the Fertilizer Name column were imputed using mode imputation. Mode is the most frequent value in a dataset and is especially well-suited to categorical features. Since Fertilizer Name is a nominal categorical feature, applying mode for replacing missing values is appropriate to ensure that imputed values are meaningful and valid in context. This approach assumes missing values to be random (Missing Completely at Random - MCAR or Missing at Random - MAR) and that the most common fertilizer type will prove to be a suitable substitute in case of unavailability of any other knowledge. Mode imputation maintains the distribution integrity with no possible bias that may occur if random or outlier values are imputed.

The process of imputation was done through the pandas library of Python, the mode was calculated using non-missing values and then utilized to replace all missing values. This brought consistency to the data set as well as facilitated the process of training for the machine learning algorithms by removing any possibility of error because of null values.

G. Feature Selection and Feature Engineering

Two feature were created, the nutrient sum and the average sum. This is an important step in preprocessing to build a robust fertilizer recommendation model. Through the summation of the separate nutrient values (Nitrogen, Phosphorous, and Potassium) into composite measures, we can identify higher-order soil fertility patterns which single nutrient measurements cannot. The nutrient sum provides a general approximation of total macronutrient content, allowing the model to distinguish between severely depleted soils (in which extensive fertilization is required) and already dense soils (in which even selective supplementation may be sufficient). By contrast, nutrient average is a normalized score of overall soil nutrient balance which enables the model to detect cases where soils might have adequate total nutrients but extreme imbalances (e.g., high Nitrogen but low Potassium) which would prevent crop yield. Such computed features fare well to reduce dimensionality without sacrificing significant biological relationships, with simpler model structures being capable of achieving good performance. They also provide basic thresholds for business rules (e.g., "if nutrient average < X, recommend basal fertilization") that can be utilized to enhance predictions by machine learning**Invalid source specified.**. The version used currently utilizes efficient vectorized operations, though the typographical error within quotation marks must be addressed so that it runs properly. Ultimately, these characteristics will improve the model's capability for generalization across varied soil conditions while retaining interpretability - essential qualities for agricultural decision-support systems where transparency and trust are as valuable as predictive accuracy

The Column "serial number" was removed from the dataframe, as it is just an identifier and has no inherent relationship with the target variable "fertilizer name". In predictive modelling such identifiers would introduce complexity and without any contribution to learning**Invalid source specified.**. Given that machine learning models attempt to establish relationships, returning unique identifiers can lead to overfitting.

H. Hyper Parameter Tunning

Hyper parameter tuning was done to improve the performance of the machine learning algorithms and ensure a proper comparison. For the purpose of striking a balance between accuracy and computational efficiency, the tuning was focused on a limited, knowledgeable set of values for each model based on literature review as well as initial trial runs. For example, Support Vector Machine was tried out across different levels of C, gamma, and kernel types, while Multilayer Perceptron's architecture and activation function were optimized for better convergence. Similarly, for ensemble models such as Random Forest and XGBoost, estimators, depth, and learning rates were tuned to get the optimal settings without overfitting.

In the attempt to speed up the process, instead of full grid searches over enormous parameter spaces, reasonable defaults and small grid searches were employed, which provided acceptable results within manageable time. The hybrid model was built as a voting ensemble using Random Forest and XGBoost, base estimators with hyper parameter optimization were utilized, and the complete meta-classifier was a minimized Random Forest for performance optimization. SMOTE technic was used to ensure models were trained on balanced data, so the hyper parameter selection focused on improving precision and F1-score rather than accommodating majority classes alone.

Models	**Hyper parameter Tuning**		
	Hyper parameter(s) Tuned	***Tested Values / Ranges***	***Final Chosen Value(s)***
Support Vector Machine (SVM)	C, gamma, kernel	C: [0.1, 1, 10]; gamma: ['scale', 0.01, 0.001]; kernel: ['linear', 'rbf']	C=1, gamma='scale', kernel='rbf'
Multilayer Perceptron (MLP)	hidden_layer_sizes, activation, alpha	hidden_layer_sizes: [(50,), (100,), (50,50)]; activation: ['relu', 'tanh']; alpha: [0.0001, 0.001]	hidden_layer_sizes=(50,), activation='relu', alpha=0.0001, max_iter=300
Random Forest (RF)	n_estimators, max_depth	n_estimators: [50, 100, 200]; max_depth: [None, 10, 20]	n_estimators=50, max_depth=None
XGBoost	n_estimators, learning_rate	n_estimators: [50, 100]; learning_rate: [0.1, 0.01]	n_estimators=50, learning_rate=0.1
Hybrid (Voting RF + XGB)	Base: RF + XGB, Final estimator: RF	Final estimator n_estimators: [10, 30, 50]	Base: tuned as above; Final: n_estimators=30

Table 1: Hyper parameter tuning

I. *Model Training and Testing*

Performance of five supervised machine learning were evaluated on a historical Agronomical dataset for recommending fertilizers. The dataset was preprocessed in order to handle missing values, normalize the features, and encode categorical features. After cleaning, it was split into training (80%) and test (20%) sets in a manner such that the test set remained unseen while training the model and hyper parameter optimization. The models: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), XGBoost, Random Forest, and an ensemble combining XGBoost and Random Forest through soft voting.

Hyper parameter tuning was done to improve the performance of the machine learning algorithms and ensure a proper comparison. For the purpose of striking a balance between accuracy and computational efficiency, the tuning was focused on a limited, knowledgeable set of values for each model based on literature review as well as initial trial runs. For example, Support Vector Machine was tried out across different levels of C, gamma, and kernel types, while Multilayer Perceptron's architecture and activation function were optimized for better convergence. Similarly, for ensemble models such as Random Forest and XGBoost, estimators, depth, and learning rates were tuned to get the optimal settings without overfitting.

IV. RESULTS AND DISCUSSION

Under this study, a sub study was conducted to test the effectiveness of an RS485 NPK sensor against the conventional lab soil testing methods. 50 soil samples were tested on using chemical extraction methods and then tested using the RS485 soil sensor. Each soil class (Clay, sand, Black, Red, Loam) had 10 representatives. For each recorded values 2 readings were taken and then averaged.

A. *Comparative Algorithm Perfomance Results*

The comparative aspect of the research design was focused on comparing and evaluating some of the machine learning models to determine which model makes the best prediction using a dataset of past fertilizer recommendations. The study specifically compared the performance of five algorithms which are XGBoost, Multilayer Perceptron (MLP), Random Forest, Support Vector Machine, and a new hybrid model based on the literature from previous studies. These models were trained using a soil history data set with variables such as soil type, crop type, and macronutrient levels (Nitrogen, Phosphorus, and Potassium), in which the target variable is the recommended fertilizer. All algorithms were evaluated against equal performance metrics (accuracy, precision, F1-score, and Confusion matrix) to have a comparative framework. The result of this comparative analysis helped to decide the most accurate and consistent model for real-world application within fertilizer advisory systems. In addition to model comparison, comparative design was also used for the impact of data balancing techniques (e.g., SMOTE) on classification performance, particularly for minority fertilizer classes.

Models	**Metrics**			**Comment**
	Accuracy	***Precision***	***F1 Score***	
Support Vector Machine (SVM)	90.52%	90.68%	90.54%	Performs adequately, but struggles with complex patterns
Multilayer Perceptron (MLP)	92.18%	92.22%	92.18%	Learns non-linear relationships better than SVM
Random Forest (RF)	94.16%	94.22%	94.18%	Excellent at capturing feature interactions; strong performance
XGBoost	94.25%	94.11%	94.09%	Robust and reliable, slightly below XGBoost
Hybrid (Voting RF + XGB)	94.44%	94.38	94.33%	Best overall; combines strengths of both tree-based models

Table 2: Comparison of Machine learning algorithms

The results show that tree-based hybrid model (XGBoost, Random Forest) outperform both Support Vector Machine and Multi Layer Perceptron, which suggests that the soil feature space is highly nonlinear and benefits from the ability of tree ensembles to capture complex feature interactions. Among the single models, XGBoost achieves the highest accuracy (94.25%), slightly edging out Random Forest (94.16%). The hybrid model further improves performance, achieving an accuracy of **94.44%**, with high precision and F1-score as well.

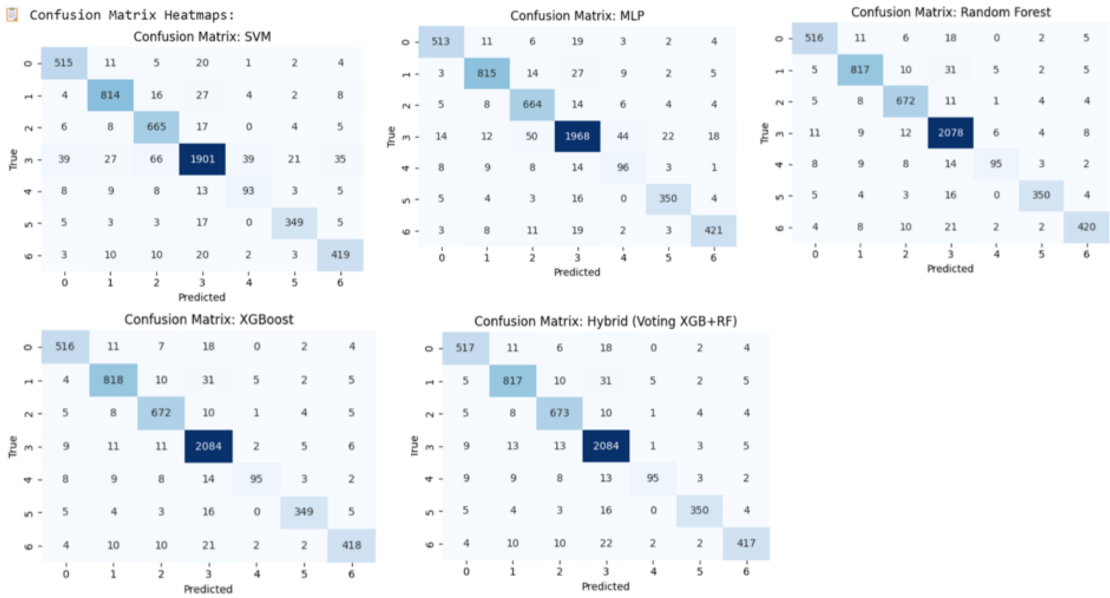

Figure 1: Confusion Matrix of five supervised algorithms

Figure 1 shows five heat maps of confusion matrices illustrate the classification accuracy of each of the five models: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest, XGBoost, and the Hybrid (XGBoost + Random Forest). Each comparison matrix graphs show predicted versus actual classes, with darker diagonal cells indicating more accurate classifications. The SVM performs the worst, with much lighter diagonals and an accumulation of misclassifications throughout the off-diagonal cells, particularly for classes 2, 4, and 5. This is reflective of its lower classification accuracy of around 90.5%, which shows that SVM does not find it easy to properly predict the complex patterns in the soil data.

The MLP performs better than the SVM, with heavier correct predictions along the diagonal and fewer errors. Random Forest and XGBoost both show significant improvements on SVM and MLP as their significantly darker diagonals and less off-diagonal noise reveal. In particular, class 3 consistently records the largest number of correct predictions across all models, suggesting that it is either the most common or most unique class in the data. Misclassifications for class 0, 4, and 6 are reduced in these tree models, which reflects the way they improve nonlinear relationships among the soil attributes.

Finally, the Hybrid model combining XGBoost and Random Forest by voting achieves the best performance among all models. Its confusion matrix contains the darkest and densest diagonal elements that denote few classification mistakes and maximum precision for the models. Misclassifications are negligible and less severe compared to the rest of the models, further confirming its excellence in high accuracy of about 94.44%. This clearly indicates that the ensemble approach leverages the complementary strengths of both tree algorithms to generate good and reliable fertilizer recommendations based on soil sensor measurements.

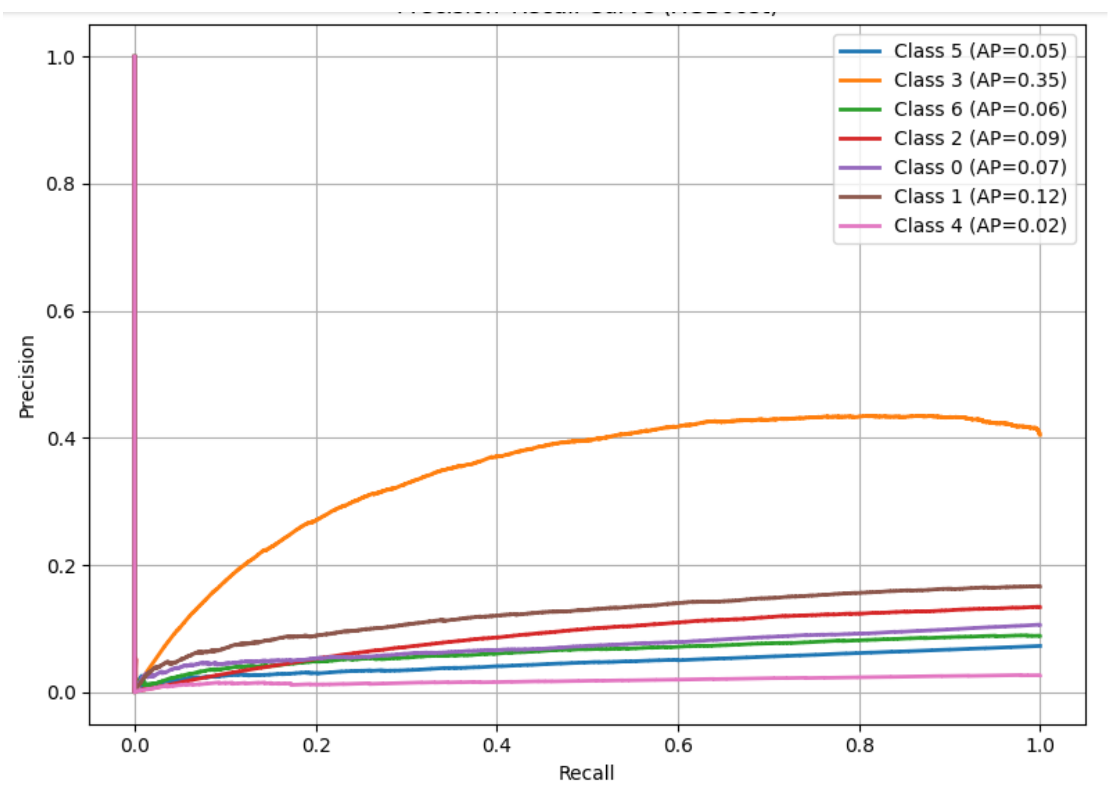

Figure 2: Precision Recall graph for the XGBoost + Random Forest hybrid model

The graph shows precision recall for the hybrid model (XGBoost + Random Forest) focusing on Average Precision (AP) per class. Class 3 has the highest AP score of 0.35, indicating relatively higher precision in prediction compared to other classes. Class 4 is lowest at a mere AP of 0.02, indicating severe difficulties in being able to successfully predict this class. The remaining classes (0, 1, 2, 5, and 6) have moderate precision with AP scores ranging from 0.05 to 0.12. The ordering of the classes from lowest to highest AP indicates a similar ordering of performance, with the only exception being Class 3, which is the top performer. Overall, the model does not do so well at making targeted predictions in the majority of classes, and this indicates a requirement for improved distinction among the classes.

B. *Experimental Sensor Validation Results*

Under this study, a sub study was conducted to test the effectiveness of an RS485 NPK sensor against the conventional lab soil testing methods. 50 soil samples were tested on using chemical extraction methods and then tested using the RS485 soil sensor. Each soil class(Clay, sand, Black, Red, Loam) had 10 representatives. The results of the experiment comparing the sensor data with the lab tests provide insights into the accuracy and reliability of the sensor technology in extracting soil health parameters. The experiment was conducted in a controlled laboratory environment to minimize extraneous effects and ensure consistency between tests. Two separate readings were recorded for all of the soil samples, and then averaged to eliminate random errors and reduce potential bias.

=== Descriptive Statistics ===

	Sensor_N	Lab_N	Sensor_P	Lab_P	Sensor_K	Lab_K
count	50.000000	50.00000	50.000000	50.000000	50.000000	50.00000
mean	17.200000	17.28000	33.560000	33.340000	49.820000	49.46000
std	11.405334	11.56003	22.923262	22.864829	32.655087	32.86721
min	0.000000	0.00000	0.000000	0.000000	3.000000	3.00000
25%	9.000000	9.00000	12.500000	13.000000	22.250000	21.00000
50%	15.000000	13.50000	32.000000	32.000000	39.500000	39.00000
75%	25.750000	26.00000	56.250000	54.000000	80.500000	81.75000
max	40.000000	42.00000	69.000000	72.000000	108.000000	112.00000

Figure 3: Descriptive statistics from Lab against sensor study

Figure shows measurements from both the sensor and the laboratory are in close agreement across all three nutrients. Each method recorded 50 observations. The mean nitrogen values were 17.20 (sensor) and 17.28 (lab), with similar standard deviations (~11.4–11.6), suggesting comparable central tendency and variability. For phosphorus, the mean values were 33.56 (sensor) and 33.34 (lab), with standard deviations around 22.9.

Potassium showed means of 49.82 (sensor) and 49.46 (lab), with higher variability (standard deviations ~32.7). Minimum values for nitrogen and phosphorus were zero for both methods, while potassium had a minimum of 3. The maximum observed values were slightly higher in the laboratory, particularly for potassium (112 vs. 108). Quartile values (25%, 50%, 75%) for all nutrients were also closely matched between the sensor and laboratory.

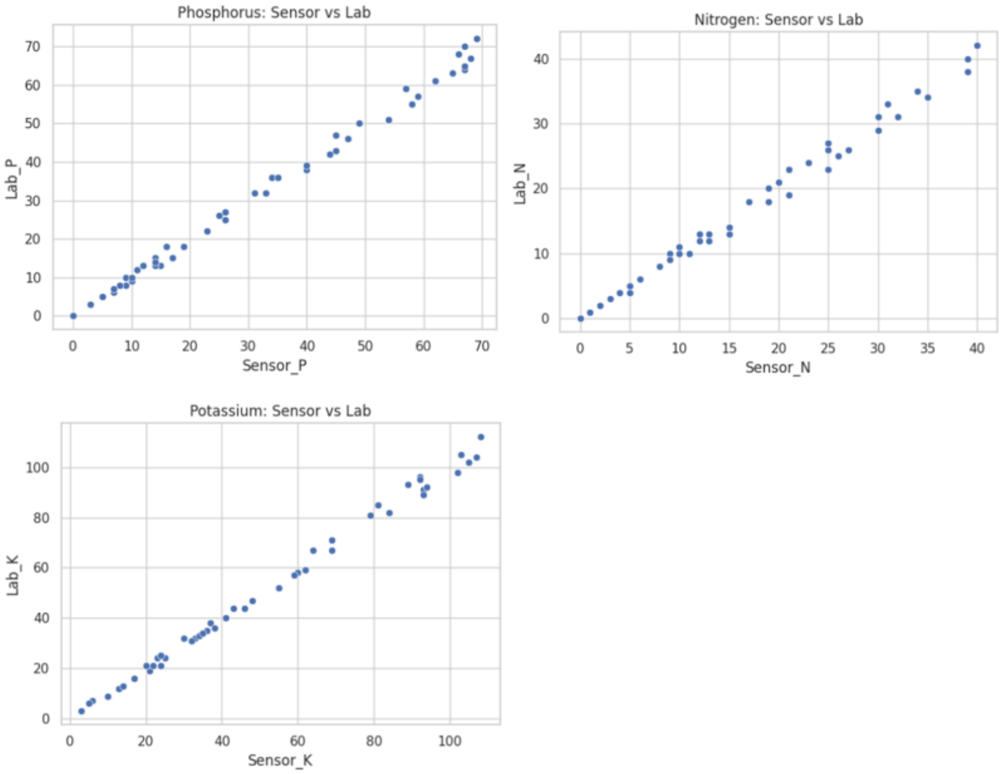

Figure 4: NPK Scatter plot of Sensor readings against Lar results

Figure 4 presents three scatter plots comparing sensor and laboratory measurements for phosphorus (P), nitrogen (N), and potassium (K). In all three plots, the points are tightly clustered around a diagonal line from the origin (0,0) upwards, indicating a strong positive linear relationship between the sensor and lab readings. This suggests that the sensor measurements closely match the laboratory measurements across all three nutrients. The alignment of points along the diagonal in each case indicates high agreement and consistency between the two methods, with little evidence of systematic bias or large outliers.

In order to get insights of reliability and accuracy of RS485 soil sensors against conventional chemical lab testing, three statistical methods were applied on the 50-paired samples: Mean Percentage Deviation (MPD), Paired T-Test, and Pearson Correlation Coefficient. The tests were applied to data for key soil nutrients—Nitrogen , Phosphorus , and Potassium. The purpose of such tests was to find out whether sensor measurements are in accordance with, and statistically equivalent to lab test values and hence validating their relevance in real-time, field-based application.

Paired T-Test was conducted to determine whether there were any statistically significant differences between the means of the two measurements. Since one sensor reading is equivalent to a laboratory measurement of the same sample, the paired structure gives a good comparison of paired observations.

Pearson Correlation Coefficient was calculated to verify the strength and direction of the linear relationship between sensor and laboratory measurements. A strong positive correlation would indicate that sensors accurately follow the direction and size of laboratory-derived nutrient measurements.

Together, these measures of statistics give an overall validation system by measuring the magnitude of difference (MPD), and statistical significance of differences (t-test), and measurement pattern consistency (correlation). [13]The results give important evidence about whether or not low-cost soil sensors can serve as good substitutes for laboratory analysis in precision agriculture applications.


```
=== Mean Percentage Deviation ===
Nitrogen: 4.40%
Phosphorus: 4.73%
Potassium: 4.87%

=== Paired t-test Results (Sensor vs Lab) ===
Nitrogen: t = -0.540, p = 0.5913
Phosphorus: t = 0.939, p = 0.3525
Potassium: t = 1.181, p = 0.2432

=== Correlation (Sensor vs Lab) ===
Nitrogen: 0.9959362462306791
Phosphorus: 0.9973832040974836
Potassium: 0.9978578458450175
```

Figure 5: statistical evaluations from sensor validation

Mean Percentage Deviation MPD was employed to calculate the mean relative difference between sensor and lab values in percentage. This test provides a direct quantitative measure of the practical precision of the sensors.

From the results that were obtained, below were the deviations

- Nitrogen 4.4% ,
- Phosphorous 4.73%
- Potassium 4.87%.

On average the sensor values are within 5% deviation which is acceptable in practical scenarios.

Paired T-Test: This test compares two related groups (Sensor Readings and Lab readings) to determine whether the difference between them is statistically significant.

Hypothesis

H0 – There is no difference between sensor values and Lab values

H1 – There is significant difference

If the p values > 0 we fail to reject H0, this indicate that there is no significant difference

- o Nitrogen: p = 0.59
- o Phosphorous: p = 0.35
- o Potassium: p = 0.24

Given that all values all values are great than 0.05, it means that sensor readings are can be used in place of the conventional chemical methods.

Correlation – This measures the strength and the direction of readings between the Sensor data and the lab results. +1 indicate positive relationship, 0 means there is no correlation -1 indicate negative relationship.

- o Nitrogen: 0.9959,
- o Phosphorous: 0.997
- o Potassium: 0.9978

The results were all close to 1 indicating a near perfect relationship.

C. *Critical Analysis of Findings*

The Under this study the interpretation of results is drawn from the performance of machine learning models. A hybrid machine learning (ML) model using XGBoost and Random Forest (RF) yielded potent results in the domain of fertilizer recommendation. Achieving a classification accuracy of 94.44%, satisfactory recall and F1-scores for most classes, the model is both technically efficient and practically relevant. The hybrid model (XGBoost + Random Forest) confirms its suitability in recommending fertilizers. The model was the best in handling class imbalances. This hybrid model also had the strongest recall across all the 7 classes, it will give accurate recommendations. The results confirms that the hybrid generalizes well and has high robustness than all the other models. The hybrid model leverages the ensemble synergy between XGBoost and Random Forest two decision tree based algorithms with differing yet complementary strengths. Random Forest adds robustness through bagging and bootstrapping, reducing variance and improving performance on noisy or absent data. XGBoost employs boosting with regularization, in which model generalization is promoted and overfitting is reduced through iterative correction of weak learners.

This fusion enabled the model to learn subtle patterns and interdependencies between crop type, soil type, and soil nutrients (N, P, K), with high classification accuracy of 94.44% and well-balanced performance across all five classes of fertilizer recommendation. Class 5, which previously had poorer performance, recorded improved recall and precision in the hybrid model—demonstrating its effectiveness in mitigating class imbalance and boundary overlap.

Results obtained from comparing sensor data and Lab results shows that the RS485 sensors can reliably be used to extract NPK values from the soil. The Mean Percentage Deviation values are less than 5%, although there is a slight variation, but the readings are still within acceptable range. These results validate the RS485 soil sensor as reliable, cost effective alternative for extracting the NPK values from the soil.

Results from both experiments indicate that the RS 485 sensors can be used to extract NPK values from the soil and the values can evaluated by the hybrid model and recommend fertilizers. The findings demonstrate the potential of transforming conventional agricultural practices into precision agriculture.

D. Comparison with Previous Academic Studies

The invent of Machine learning has presented opportunities in precision agriculture, particularly for the optimization of fertilizer recommendation systems [14]. Several academic research studies have demonstrated the potential of ML algorithms for soil fertility classification and recommending nutrient management practices based on NPK values, and crop type. These models have served to improve the precision of decision-making and reduce reliance on broad, region-wide fertilizer recommendations.

A review of literature, however, raises some limitations common to most such studies. To begin with, most of them focus on one algorithm, i.e., Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), or single Artificial Neural Networks (ANNs), without taking into account the complementarity of the strengths of hybrid approaches. While these individual models tend to achieve modest accuracy levels, they tend to suffer from performing badly in highly imbalanced datasets, such as those found in real-world agricultural setups where some fertilizer needs are much less frequent than others.

The majority of the studies are either crop- or geographically limited [15], focusing on single regions or single crops, which restricts the generalizability of their findings. Such models are not easily upscaled to heterogeneous agro-ecological zones, or real-time use by smallholder farmers in general, especially in low-resource settings like Zimbabwe, without external validation or sensor fusion.

On the other hand, this study introduces a hybrid machine learning approach (Random Forest and XGBoost fusion) that not only improves classification accuracy but also addresses real-world usability through the inclusion of real-time sensor data. By validating RS485 soil sensor readings against lab values and inputting these directly into the model, this study offers an operational, affordable decision support system that is both technically viable and also practicable to deploy in rural farming settings.

V. CONCLUSION

This research used a comparative and experimental research design to evaluate effectiveness of a machine learning model to recommend fertilizers, the model is benchmarked against agronomical decisions dataset. A key pillar of this research was also the empirical validation of a low cost RS485 soil sensor against results obtained from NPK nutrient chemical extraction. By merging data collected from an RS485-connected soil sensor with advanced machine learning models, the proposed integrated system produces timely, accurate, and crop-specific fertilizer recommendations in order to help farmers optimize soil fertility while minimizing environmental impact. A number of machine learning algorithms were used and experimented with, including Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), XGBoost, Random Forest, and a hybrid model combination of XGBoost and Random Forest. Experimental results show that tree-based models, particularly the ensemble hybrid, achieved the best accuracy, precision, and F1-score, with the hybrid model achieving a staggering accuracy of 94.40%, surpassing individual models. Confusion matrix analysis also validated the better prediction of the hybrid model, which continuously minimized misclassifications for all classes. This supports the value of making use of ensemble approaches in an attempt to leverage complementary strengths of models. The system proposed has immense potential to foster sustainable practice of agriculture through the offering of an automated data-driven decision support system that can help farmers apply the right amount and type of fertilizer [16]. Given that the system uses a low-cost RS485 sensor in the acquisition of soil data makes it cost-effective and scalable, particularly in environments where resources are limited. Future studies can include expanding the dataset with more diversity in soils, crops, and climatic conditions, incorporating other soil factors such as micronutrients and pH, and deploying the system on actual farm fields to assess performance in real scenarios. Overall, the work here highlights the transformative power of machine learning and IoT technologies to enable effective, efficient, and green agriculture.

REFERENCES

[1]. FAO, "Food and Agriculture Organization," 2021. [Online]. Available: https://www.fao.org/zimbabwe/fao-in-zimbabwe/zimbabwe-at-a-glance/en/. [Accessed 12 June 2025].

[2]. FAO and ITPS, "State of knowledge of soil biodiversity – Status, challenges and potentialities," 2021. [Online]. Available: https://www.fao.org/documents/card/en/c/cb1928en/. [Accessed 2025 June 12].

[3]. R. Chikowo, J. Mutegi, J. Nyamugara and P. Mapfumo , Soil fertility management in southern Africa: Balancing productivity, sustainability, and resilience. Frontiers in Sustainable Food Systems,, 2020.

[4]. P. Bindraban, ,. C. Dimkpa and ,. L. Nagarajan, "Improving fertilizer use efficiency: State of the art and prospects.," Argus Media - Fertilizer Focus, pp. 16-21, 2021.

[5]. J. Lehmann and M. Kleber, "The contentious nature of soil organic matter.," Nature Publishing Group, pp. 60-68, 2015.

[6]. FAO, "State of Knowledge of Soil Biodiversity.," Food and Agriculture Organization of the United Nations., 2020.

[7]. N. AP, "New technologies empower farmers with site-specific soil data," 2024. [Online]. Available: https://apnews.com/article/agriculture-soil-data-2024. [Accessed 18 April 2025].

[8]. IntechOpen, "Precision agriculture: Technology and economic perspectives," 2021. [Online]. Available: https://www.intechopen.com/chapters/77608. [Accessed 20 June 2025].

[9]. M. Mwangi, J. Wambua and J. Maina, "Cost-effective RS485 sensor networks for precision soil monitoring in Sub-Saharan Africa," Elsevier, 2022.

[10]. Changsha Zoko Link Technology Co., "Advanced RS485 Soil Sensors: User Manual.," Changsha Zoko Link Technology Co, 2024.

[11]. Niubol, "Soil NPK Sensors: Next-Generation Guidebook.," Shenzhen: Niubol Technology Co., 2024.

[12]. T. Kammerlander, R. Huber, R. Schulin and T. Keller, "Machine learning-based soil health monitoring under field conditions," Wiley, pp. 8-12, 2025.

[13]. L. Yang, Q. Zhou, Y. Wu and H. Zhang, "Real-time monitoring systems for precision agriculture: A review.," Frontiers Media, 2023.

[14]. H. Li and Y. &. Z. X. Wang, " Development of a portable RS485-USB interface module for environmental monitoring. Sensors and Materials," Myu Scientific Publishing, pp. 18-19, 2022.

[15]. D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BioMed Central, pp. 1-13, 2020.

[16]. A. Gupta, A. Jain, , S. Prakash, and S. Roy, , "Soil nutrient prediction and fertilizer recommendation using machine learning models," Elsevier : Computers and Electronics in Agriculture, pp. 106-134, 2022.

[17]. R. Koyande and M. Thakur, "Optimizing soil health and crop productivity using sensor-driven decision support systems," Agricultural Informatics Journal, , pp. 45-53, 2024.

[18]. D. Chikodzi and G. Mutowo, "Digital tools in climate-smart agriculture: Enhancing fertilizer efficiency among smallholders in Zimbabwe," Zimbabwe Ministry of Lands, Agriculture, Water and Rural Resettlement, pp. 105-117, 2022.

[19]. P. Bindraban, C. Dimkpa, and L. Nagarajan, , "Improving fertilizer use efficiency: State of the art and prospects.," Argus Media - Fertilizer Focus, pp. 16-21, 2021.

[20]. M. Zhou, W. Luo, Q. Xu and Y. Chen, "A soil sensor system based on RS485 protocol for real-time monitoring of NPK in precision agriculture," MDPI, 2021.

[21]. K. Kumar, S. R. D. Patel and V. Singh, "Accurate soil health profiles for sustainable agriculture," Springer: Journal of Soil Science and Plant Nutrition, p. 134–147, 2024.