Cureus Journal of Computer Science

Open Access Original Article

Monitoring Soil Health Using Machine Learning Techniques for Improved Crop Yield Prediction

Anmol Chaudhary [1] , Prabh Deep Singh [1] , Rohan Verma [1] , Manmohan Sharma [2]

1. Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, IND 2. School of Computer Applications, Lovely Professional University, Punjab, Jalandhar, IND

Corresponding author: Anmol Chaudhary, anmolc563@gmail.com

Received 04/07/2025
Review began 04/20/2025
Review ended 07/22/2025
Published 07/22/2025

DOI: https://doi.org/10.7759/s44389-025-04577-2

Abstract

Prediction of soil nutrients is one of the important primary input management systems to enhance crop yield. Soil is a natural resource that can be found everywhere on Earth's surface. It is composed of minerals, organic matter, air, water, and other elements. Soil nutrients are chemical elements that plants need for growth and survival. There are various nutrients and factors on which soil health directly or indirectly depends, like nitrogen, phosphorus, potassium, pH, microbial activity, soil structure, moisture retention, iron, zinc, temperature, soil salinity, soil compaction, organic matter, etc. Also, farmers tend to have incomplete knowledge and information about proper fertiliser usage and their impact on the soil. As a result, sometimes they apply fertilisers in greater quantity than necessary, leading to an imbalance in soil nutrients. Hence, the work aims to monitor the health of soil based on various factors using machine learning techniques.

In this work, 13 macronutrients that are needed for healthy plant growth and healthy soil are focused on. As a classification algorithm, light gradient boosting machine is used for the health prediction of soil. The proposed model gives an accuracy of 95.56%. Such a high level of accuracy demonstrates the model's ability to identify complex relationships between soil nutrients and health indicators. Furthermore, the insights gained from the proposed work can empower farmers to make smart and sustainable soil health management decisions. By accurately assessing soil health, farmers can optimise fertiliser use, enhance crop yield, and promote sustainable agricultural practices. The given approach not only contributes to increased productivity but also helps in preserving soil integrity for future generations.

Categories: Recommender Systems, Deep Learning, Machine Learning (ML)
Keywords: soil health monitoring, soil fertility assessment, soil nutrient analysis, soil fertility recommendation, precision agriculture, suistanable farming, machine learning in agriculture, soil quality evaluation, smart farming

Introduction

Agriculture has played a crucial role in shaping human evolutionary history and has significantly contributed to the development of civilisation. It provides a foundation for economic growth and development, sustaining populations and driving trade throughout history [1]. Approximately 37% of the world's land is utilised for agricultural purposes, with around 11% designated specifically for crop production. The remaining land is primarily used for livestock grazing. As of 2024, about 2.5 billion people worldwide rely on farming for their livelihoods, which includes farmers, agricultural labourers, and others involved in the agriculture sector.

Agriculture is an important part of India's economy and population, as a large percentage of the population is directly or indirectly related to farming and lives in rural areas. Around 72% of India's population lives in rural areas, and around 70% of people depend on agriculture for their livelihood. Agriculture also contributes a major part to India's economy by providing a livelihood to about 42.3% of the total population. It also contributes 18.2% to the country's GDP [1]. Today, agricultural produce demand has skyrocketed because of the rapid rise over the last several decades. At the same time as urbanisation expands, there is a noticeable drop in arable land for farming [2].

Countries are facing strain to confront the paired difficulty of rising food demand and a dip in available agricultural land. In the given context, it is essential to implement efficient and optimised methods to augment agricultural yields from the existing cultivable land without causing harm to the environment. The expansion of cultivation has resulted in nutrient loss, leading to an overall drop in soil fertility [3]. Cutting-edge computer-based optimised techniques can help avoid undesirable scenarios in agriculture. Various computational methods are used to find out the fertility characteristics of soil samples. However, these methods often encounter issues such as insufficient clarity and high capital costs. The proposed approach aims to overcome these limitations by providing farmers with the tools to effectively predict and interpret soil health.

Soil fertility

Soil fertility is crucial for agricultural production and plays a vital role in protecting global food systems by supporting the production of food needed to feed the population [4]. Soil fertility reflects the capacity of soil in a specific area to provide essential physical, chemical, and biological components that support plant growth. Fertile soil can greatly benefit the environment by enhancing vegetation restoration and offering a chance to create a zero-emission ecosystem [5]. It regulates the soil's capability to provide important nutrients needed for the overall plant's growth.

Soil fertility and the factors affecting it

There are several interrelated factors that directly or indirectly impact soil fertility. One main factor that impacts soil fertility is soil anatomy, i.e., the proportions of silt, sand, and clay. Also, another factor with the ideal combination of proper drainage, nutrient availability, pH level [6], microbial activity, soil structure, soil compaction, organic matter, and water retention, along with nutrients such as nitrogen (N), phosphorus (P), potassium (K), iron (Fe), and zinc (Zn), plays a vital role in modifying soil fertility positively [7]. These nutrients can also be restored through the degradation of organic matter, the use of fertilisers, and the erosion of minerals [8].

Climate and weather conditions also have an impact on soil conditions, as they affect nutrient circulation and microbial activities [1,9]. Other factors like moisture, sunlight, and temperature also dominate how quickly organic matter is renewed, which impacts the rate of nutrient release. Additionally, the pH level of soil indicates its acidity or alkalinity, affecting both microbial activity and nutrient availability [6]. Human practices, like inclusive land management and agricultural methods, can either increase or decrease soil health. Favourable practices such as no-till or reduced till, cover cropping, organic farming, and crop rotation aid soil health, whereas poor irrigation and excessive fertiliser use can lead to soil health degradation [10].

Understanding soil parameters

In different geographical conditions, there are different types of soil, which are often characterised by slope, elevation, and some other specific soil types that require climatic conditions. So, there should be some standard key indicators for evaluating soil fertility for different soil samples. Sakin et al. [11] describe different parameters that are required for soil fertility prediction in a traditional setup and how to gather them using lab-based techniques. Table *1* [11] summarises these parameters and the techniques used to extract them.

Each metric in Table *1* plays a crucial role in analysing the soil samples. Organic matter refers to the decomposed remains of plants and animals in the soil, which enhances fertility, water retention, and microbial activity. Soil Organic Matter (SOM) contributes nutrients through recycling processes such as getting inputs from ash deposits and the mineralisation of plant materials [12]. Additionally, living organisms play a role in nutrient provision. SOM is essential for stabilising soil structure, facilitating the ability to retain plant nutrients, and maintaining water-holding ability [13]. Organic carbon (OC), a SOM component, is derived from plant residues, microorganisms, and decomposed organic materials [14]. Nitrogen is also important for plant growth, as it is a key component of chlorophyll and amino acids, influencing leaf development [15]. Phosphorus plays a vital role in enhancing root development, flowering, and energy transfer in plants, being essential for DNA and RNA synthesis [15]. Potassium improves plant stress resistance, water regulation, and enzyme activation, which helps plants promote strong stems and roots [16].

The pH level of soil reflects soil acidity or alkalinity, which affects nutrient availability and microbial activity, with an optimal range of 6.0-7.5 for most crops [17]. Microbial activity measures the presence of beneficial microbes that decompose organic matter and support nutrient cycling [13]. Soil Structure (1-5) describes soil aggregation, porosity, and compaction; good structure improves aeration and root penetration [18]. Moisture retention determines the soil’s ability to hold water, affecting plant hydration and drought resistance [18]. Iron and zinc are essential micronutrients required for chlorophyll synthesis, enzyme function, and plant metabolism [7]. Soil salinity refers to salt concentration, affecting plant water uptake. High salinity can lead to reduced crop yields [19]. Soil compaction indicates soil density, and high compaction reduces root growth, aeration, and water infiltration [18]. Boron contributes as a vital nutrient, as it is important for structural integration and cell wall synthesis. It also contributes to nitrogen and carbohydrate metabolism and is responsible for sugar transport within the plant [20]. Sulphur serves as a key signalling molecule that regulates stress management and supports essential metabolic processes [21].

Soil Parameters	Extracting Methodology
Soil Organic Matter (%)	Measured using loss on ignition, which involves oxidation of organic carbon.
Available N	Determined using the Kjeldahl method for total nitrogen, via acid digestion, while ion-selective electrodes measure nitrate nitrogen through ion activity detection.
Available P	Extracted using Olsen's method using sodium bicarbonate, and spectrophotometry quantifies its concentration.
Available K	Ammonium acetate extraction, followed by flame photometer analysis.
pH (1:2)	Collected using a pH meter after stirring soil with distilled water in a 1:2 ratio.
Microbial Activity	Assessed through colony-forming unit counts using serial dilution and plating on nutrient agar.
Soil Structure (1-5 rating)	Evaluated visually based on aggregation, porosity, and texture using standard soil structure assessment charts.
Moisture Retention (%)	Measured using the Gravimetric Method, time-domain reflectometry sensors, or pressure plate apparatus.
Plant Available Micronutrients (Zn, Fe)	Extracted using diethylene triamine pentaacetic acid method and analyzed via atomic absorption spectroscopy.
Temperature (°C)	Measured using soil thermometers or infrared sensors.
Soil Salinity (dS/m)	Found using electrical conductivity meters in a 1:5 soil-water extract.
Soil Compaction (g/cm^3)	Evaluated using a penetrometer or bulk density test.
Available Boron	The nutrient was extracted via hot water using a dilute CaCl solution and quantified colorimetrically with azomethine H.
Available Sulphur	Generally, sulphur extraction is carried out using CaCl, and the resulting opacity analysis is carried out at 440 nm using a UV-vis spectrophotometer.

TABLE 1: Traditional methods for extracting soil parameters

Table 1 shows the general steps used in laboratories and fieldwork for extracting soil parameters and analysing them. The processes might differ a little bit due to the different instruments, goals, and surroundings, but they seem to follow the common procedures of agricultural experts regarding the evaluation of soil conditions.

When analysed together, these parameters help determine how fertile a soil sample is. Table 2 shows the recorded experimental values and classifies the concentration of each parameter as low, medium, or high. It represents general guideline values based on standard agronomic references. However, actual thresholds may vary depending on soil type, crop requirements, and regional standards.

Element	Low	Medium	High
Organic Matter (%)	< 1.0	1.0–3.0	> 3.0
Nitrogen (ppm)	< 20	20–50	> 50
Phosphorus (ppm)	< 10	10–25	> 25
Potassium (ppm)	< 100	100–250	> 250
pH Level	< 5.5 (acidic)	5.5–7.5 (neutral)	> 7.5 (alkaline)
Microbial Activity (CFU) (g)	< 10^4	10^4–10^6	> 10^6
Soil Structure (1-5)	1–2 (poor)	3 (moderate)	4–5 (good)
Moisture Retention (%)	< 10	10–25	> 25
Iron (ppm)	< 2.0	2.0–4.5	> 4.5
Zinc (ppm)	< 0.5	0.5–1.5	> 1.5
Temperature (°C)	< 15	15–30	> 30
Soil Salinity (dS/m)	< 1.0	1.0–4.0	> 4.0
Soil Compaction (g/cm^3)	< 1.3	1.3–1.6	> 1.6
Boron (ppm)	< 0.5	0.5–1.0	> 1.0
Sulfur (ppm)	< 10	10–20	> 20

TABLE 2: Soil nutrient level classification by concentration ranges

CFU, Colony Forming Unit; PPM, Parts Per Million

So, understanding and overseeing these complex factors is important for sustainable agriculture practices. Evaluating soil fertility helps farmers learn new strategies that can maintain healthy microbial activity, optimise nutrient levels, and ensure long-term agricultural production. Traditional soil analysis methods can be labour-intensive and slow, pointing to the need for modern techniques that streamline the process [22]. Machine learning tools are emerging as effective means to analyse complex datasets and derive valuable insights for classifying soil fertility [6,23]. Also, many farmers lack a thorough understanding of the complex factors on which soil fertility relies, which is crucial for agricultural productivity and food security. Although farmers may be less aware of basic soil management principles, they may tend to fail to understand the relationships between soil composition, nutrient availability, microbial activity, and environmental conditions [24]. The existing knowledge gap or lack of understanding can adversely affect crop yields and overall agricultural production.

To address this issue, a machine learning-based model to evaluate soil fertility is required, which can provide applicable insights to farmers. By analysing key soil attributes, the proposed model aims to provide farmers with the basic information to help them make better decisions for good farming practices. As a result, this will enhance soil fertility and increase crop yields [25]. The main contributions of the work are as follows. First, a theoretical framework was developed and implemented for a soil health monitoring system using a light gradient boosting machine (LightGBM) classifier as a machine learning approach [26]. As part of the effort, key attributes for classifying soil health have been identified and analysed, including organic matter, nitrogen, phosphorus, potassium, pH level, microbial activity, soil structure, moisture retention, iron, zinc, temperature, soil salinity, and soil compaction [7,24]. The proposed work demonstrated that the LightGBM classifier is an effective tool for predicting soil health outcomes based on these attributes. Next, extensive validation is performed of the model through the cross-validation technique to assess its robustness and accuracy in classifying soil health. The results demonstrate that the system can effectively differentiate between healthy and unhealthy soil conditions, providing actionable insights for improved soil management. Ultimately, the findings indicate that the LightGBM classifier outperforms traditional classification methods, especially in scenarios with complex interactions among multiple soil attributes [26,27].

The data used here involves a range of soil samples from more than 10 Indian states, such as Uttar Pradesh, Bihar, Gujarat, Karnataka, Tamil Nadu, Kerala, Maharashtra, West Bengal, Rajasthan, and Punjab, covering a wide range of agro-climatic zones and cropping patterns. These are quite varied in soil structure, climate, pH, and nutrient status, offering a heterogeneous and real-world foundation for model training and validation. The availability of samples across humid, arid, semi-arid, and sub-tropical regions increases the model's robustness and geographic generalisability in India. Although this suggests high national-scale adaptability, validation on international region or highly localised cropping system datasets is needed in the future to establish cross-regional scalability and project its applicability worldwide. Unlike several previous ML-based soil health predictors that prioritise accuracy alone, the suggested LightGBM model places equal importance on interpretability as well as predictive performance. The feature importance analysis reveals significant soil parameters like microbial activity, organic matter content, and nutrient content consistent with agronomic theory, which provide useful insights for farmers and soil scientists alike. This not only facilitates classification but also informs tailored interventions, including nutrient balancing or personalised fertilisation. In addition, the model is shown to generalise well across a range of soils with fewer input features, thus making it scalable and affordable for real-world application, particularly in resource-constrained agricultural environments.

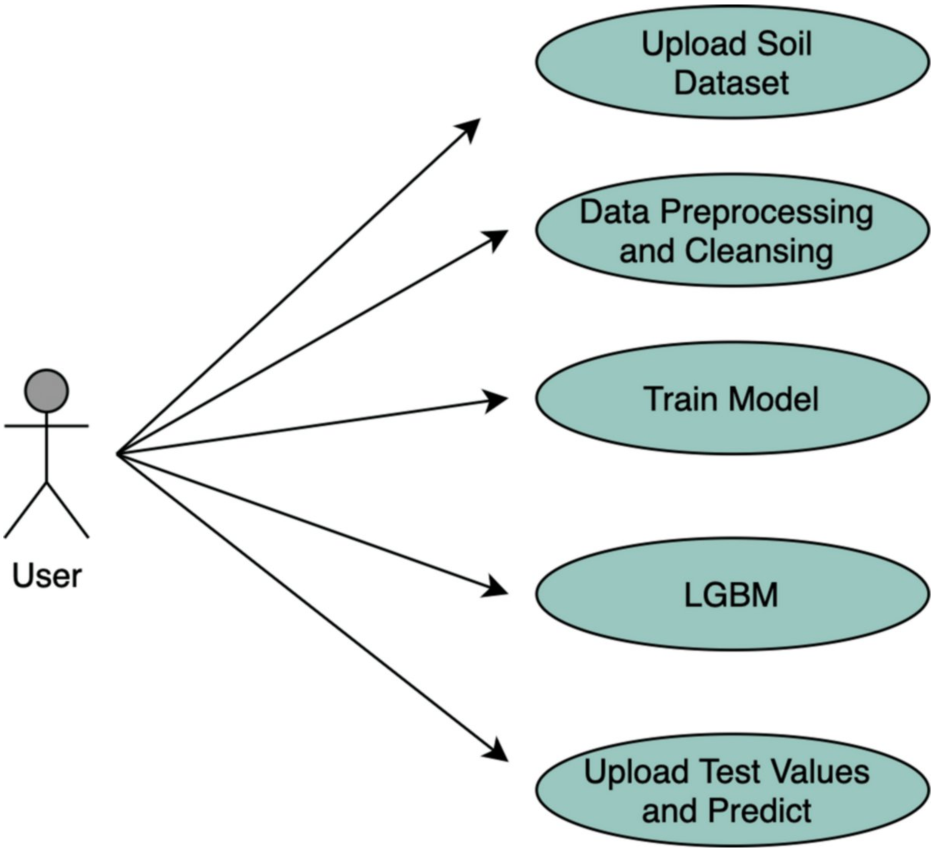

FIGURE 1: Use case diagram of soil health monitoring system

LGBM, Light Gradient Boosting Machine

Figure *1* clarifies the interaction between the soil health monitoring system and the user. Here, the user requests an analysis of soil health from the system and inputs soil parameters as the first use case. The monitoring system then uses these inputs using the trained LightGBM model and updates the model based on user feedback. Finally, the system returns the provided soil health status along with the user-provided input values.

The paper has the following structure. The literature review under Material and Methods section analyses the previous work in soil health monitoring and the application of machine learning techniques in agriculture [28]. This is followed by the formulation of the problem of a soil health monitoring system, focusing on the development of a machine learning-based approach to assess and predict soil health [27-29]. The system aims to analyse key soil parameters such as pH, moisture content, nutrient levels, etc., while addressing challenges related to data availability, feature selection, and model optimisation [6]. Finally, it proposes a machine learning-based soil health monitoring system that integrates data collection of key soil parameters and environmental factors to accurately assess soil health [30]. The model aims to empower farmers with the basic information to help them make better decisions for good farming practices. As a result, this will enhance soil fertility and increase crop yields [25]. Results and Discussion sections describe the expected outcomes of the proposed model, which highlights its potential to improve soil health evaluation, which can enhance crop yields and positively impact sustainable agricultural practices. Finally, Conclusions section concludes by summarising the significance of the proposed soil health monitoring system and its potential impact on agriculture.

Materials And Methods

Literature review

The selected papers focus on soil fertility, soil health, and machine learning algorithms. Meanwhile, Keerthan Kumar et al. [31] discuss a model to identify an optimal algorithm for soil fertility prediction. Their study finds that linear regression performs effectively for classifying soils as per their characteristics, achieving a root mean square of 0.0617. Addit mbjhionally, it also concludes that the random forest is most suitable among the various algorithms used in the model for categorising soils based on nutrient content, with an accuracy of 72%. In comparison, linear-kernel support vector machine (SVM) and Gaussian Naive Bayes classifiers achieved lower accuracy scores of 63% and 50.78%, respectively, on the same dataset.

Gaikwad et al. [32] proposed an enhanced SVM-based model that leverages both soil and crop datasets to predict soil fertility and subsequently recommend perfect crops as per soil type. The study reports that the proposed model outperforms existing available models applied to similar datasets, achieving an accuracy of 94.95% compared to earlier implementations, which had accuracy rates of 91.90% and 92.30%, respectively. Saleem et al. [33] proposed a model that leverages environmental parameters such as temperature, humidity, rainfall, wind speed, and sunshine to predict pest outbreaks. Trained on five years of data (2018-2022), the model achieved an accuracy of 94% in weekly predictions, enabling timely intervention strategies for farmers. The study focused on the impact of real-time environmental surveillance for better crop management.

Terhoeven-Urselman et al. [34] evaluated mid-infrared reflectance spectroscopy using 971 soil samples obtained from the ISRIC (International Soil Reference and Information Centre) database. To reduce errors, the researchers applied specialised preprocessing tools and utilised partial least-squares regression for model development. The study presented good accuracy in predicting organic carbon, pH, and cation exchange capacity. Predictions for magnesium and clay content were moderately accurate, while results for sand and exchangeable calcium were less stable. Despite some limitations, the study suggests that a global soil spectral library could help in classifying soil fertility and improving land management.

A recent study by Suchithra and Pai utilised an extreme learning machine (ELIM) [35] to classify key soil parameters at the village level, including organic carbon (OC), available potassium (K), boron (B), available phosphorus (P), and pH. The research compared ELM performance with different activation functions such as Gaussian radial basis, sine-squared, hyperbolic tangent, triangular basis, and hard limit. The Gaussian radial basis function achieved the highest accuracy (above 80%) in four out of five soil classification tasks, while the hyperbolic tangent function achieved the best results (close to 90%) for pH classification.

The application of such machine learning-based classification methods in soil testing significantly reduces fertilizer waste, enhances profitability, minimizes dependency on chemical analysis experts, and improves both soil health and environmental quality.

Citations	[31]	[32]	[33]	[34]	[35]	Proposed Model
Algorithm used	Different classification and regression algorithms like SVM, Random forest, and Gaussian Naive Bayes	SVM	IOT + DNN	PLSR, PCA, Box-Cox transformation	ELM with Hyperbolic Tangent Function for pH and ELM with Gaussian Radial Basis Function for soil nutrients	LGBM
Methodology	Three phases - Data preprocessing, training, and testing. Used 1 soil dataset and 2 modules in the model	Two phases - training and testing. Used 2 datasets - soil dataset and crop dataset	Four phases - Data Collection, and their transfer to the IOT server, Data Preprocessing, Training, and testing	Four Phases - Soil Characterization, Spectral Measurements, Data Processing, Training, and Testing. Used 1 soil dataset	Five phases - Dataset Selection, Preprocessing, Model Selection, Training, Hyperparameter Optimization, Testing, and Performance Analysis. 2 Datasets - one for fertility and one for pH	Five phases -Data Collection, Data Preprocessing, Model Construction, Model Evaluation, and computation of the ROC-AUC score
Evaluation metrics used	Accuracy %, RMSE	Accuracy %	Accuracy %, Recall, Precision, F-measure, Cohen's Kappa	RMSEP, RPD	Accuracy, Precision, Recall, FI-score, and Kappa Statistics	Accuracy %, Recall, F-measure, Precision
Accuracy	72.74%	94.95%	94%	NA	88.50%	95.56%

TABLE 3: Comparison of different soil fertility models.

DNN, Deep Neural Network; ELM, Extreme Learning Machine; IoT, Internet of Things; LGBM, Light Gradient Boosting Machine; NA, Not Available; PCA, Principal Component Analysis; PLSR, Partial Least Squares Regression; RPD, Residual Predictive Deviation; RMSE, Root Mean Square Error; RMSEP, Root Mean Square Error of Prediction; ROC-AUC, Receiver Operating Characteristic-Area Under Curve; SVM, Support Vector Machine

Table 3 analyzes the different articles based on the machine learning algorithms implemented, their methodology used, the accuracy of the approaches, as well as the transparency level, and further evaluation of the proposed model after initial analysis.

Citations	Dataset	Source	Features of Dataset
[31]	ICRISAT V3 Soil Dataset	The dataset consists of diverse soil samples collected from agricultural fields in 13 districts of Andhra Pradesh, India.	Zn, Mn, S, K, S, OC, B, EC, pH, and soil types.
[32]	SRDI Soil Dataset	The soil dataset contains soil samples from 6 sub-districts of Khulna District, Bangladesh.	K, B, Ca, Mg, Zn, S, Cu, Fe, Mn, Soil Organic Matter, pH, Salinity (%).
[33]	Environmental Data (2018-2022)	Data was collected from crop fields in Pakistan, specifically for cotton pest prediction.	Temperature, humidity, sunshine, rainfall, and windspeed.
[34]	ISRIC Dataset	Data was collected by scanning 971 soil samples from ISRIC using mid-infrared spectroscopy.	Organic carbon content, pH value, CEC (cation exchange capacity), exchangeable magnesium concentration, clay content, exchangeable calcium content, and sand content.
[35]	North Central Laterites region soil dataset and Marathwada region dataset	North Central Laterites region was collected from soil samples taken from different locations in North Kerala, and Marathwada dataset was taken from a previous study conducted by Sirsat et al. [36].	B-F, P-F, K-F, OC-F for soil nutrient classification and pH level (SA, HA, MA, SLA) for pH level classification.
Proposed Model	Soil Health Monitoring System	Primary + Secondary collected from different government websites.	Soil Organic Matter (%), Nitrogen (ppm), Phosphorus (ppm), Potassium (ppm), pH Level, Microbial Activity (CFU), Soil Structure (1-5), Moisture Retention (%), Iron (ppm), Zinc (ppm), Temperature (°C), Soil Salinity (dS/m), Weather Condition, Soil Compaction (g/cm^3).

TABLE 4: Comparison of soil datasets used in the study

ICRISAT, International Crops Research Institute for the Semi-Arid Tropics, Patancheru, Telangana, India; SRDI, Soil Resource Development Institute, Dhaka, Bangladesh; ISRIC, International Soil Reference and Information Centre, Wageningen, the Netherlands; B, Boron; Ca, Calcium; Cu, Copper; Fe, Iron; K, Potassium; Mg, Magnesium; Mn, Manganese; S, Sulfur; Zn, Zinc; OC, Organic Carbon; EC, Electrical Conductivity; pH, Potential of Hydrogen (acidity/alkalinity level); B-F, Boron Fertility; P-F, Phosphorus Fertility; K-F, Potassium Fertility; OC-F, Organic Carbon Fertility; SA, Strongly Acidic; HA, Highly Acidic; MA, Moderately Acidic; SLA, Slightly Acidic; CHU, Colony Forming Unit

Soil Types refers to the classification of soil based on texture, composition, and properties (e.g., sandy, loamy, clayey).

Table *4* presents the comparison of the datasets used across the various machine learning models compared in Table *3*. The comparison focuses on features such as dataset name, data source, and type of features provided. This framework is useful for understanding the differences and the quality of the dataset that impacted the effectiveness and robustness of each model.

Evaluating the model outcomes using these parameters helps in detecting bias, data restrictions, and the strength of the results. The evaluation also assists in understanding the model's flexibility when deployed in new agro-climatic zones or regions.

Problem formulation

Soil health is vital to agricultural productivity, as it directly affects crop yields. Tracking soil nutrient levels is crucial for efficient resource management since nutrients such as phosphorus (P), nitrogen (N), and potassium (K), etc., are required for plant development. However, conventional farming practices often lead to improper nutrient applications due to a lack of knowledge, resulting in soil degradation and reduced crop production.

Let S represent the set of soil samples, where each sample $s_i \in S$ is defined by a feature vector:

$s_i = \{x_1, x_2, ..., x_n\}$, where $x_1, x_2, ..., x_n$ are the values of different soil properties, such as phosphorus, nitrogen, and potassium; pH; soil moisture; soil organic matter content; and other macronutrients. The problem can be defined as a classification problem, where the goal is to predict the health status y_i of each soil sample s_i:

$$y_i = f(s_i) = \begin{cases} 1 \text{ if the soil is healthy} \\ 0 \text{ if the soil is unhealthy} \end{cases} \quad (1)$$

where *f* is a classification function learned from the training data.

To solve this problem, various machine learning models are employed, including LightGBM. Let $D = \{(s_1,y_1), (s_2,y_2), \ldots, (s_m,y_m)\}$ represent the training dataset, where m is the number of samples. The objective is to minimize the classification error:

Minimize: $$L(\Theta) = \frac{1}{m}\sum_{i=1}^{m} L(y_i \hat{y}_i))$$

where $\hat{y}_i = f(s_i;\theta)$ is the predicted label for the i-th sample using model parameters θ, and L is the loss function, typically cross-entropy classification problem.The aggregate accuracy score, Acc, of the model is given by:

$$Acc = \frac{\sum_{i=1}^{m} \mathbb{I}(\hat{y}_i = y_i)}{m} * 100\% \quad (2)$$

where $\mathbb{I}$ is the indicator function:

$$\mathbb{I}(\hat{y}_i = y_i) = \begin{cases} 1 \text{ if } \hat{y}_i = y_i \\ 0 \text{ otherwise} \end{cases} \quad (3)$$

In this work, 13 macronutrients were considered as features, and the LightGBM model yielded an accuracy of 95.56%. The forecasting model provides analysis of soil conditions, empowering farmers to use accurate fertilizer amounts, improving nutrient balance, and increasing crop yields. The proposed model analyses soil conditions and helps farmers to apply precise fertilizer amounts, which would enhance nutrient balance, and boost crop yields.

Proposal

In this study, a machine learning-based model is proposed for monitoring soil health and predicting crop yields to augment agricultural productivity. The model is formulated to analyse soil data and provide practical information that can improve farming practices and increase crop yields.

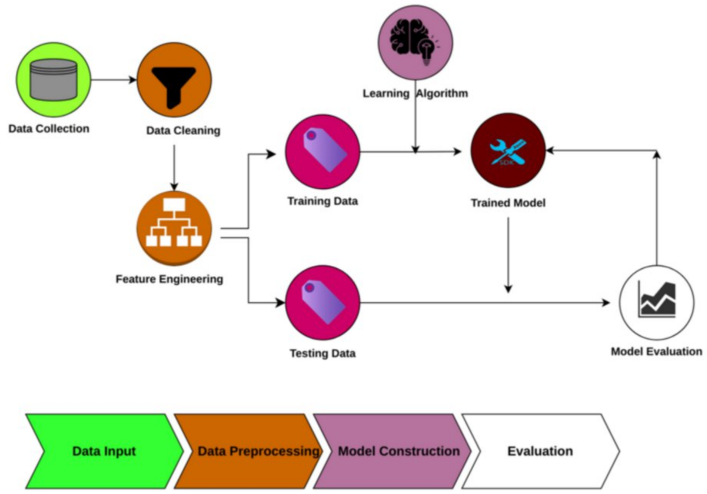

FIGURE 2: Depicts the essential steps in a typical machine learning process used for soil health monitoring

Figure 2 provides a summary of the key stages in a machine learning process designed for soil health monitoring. The workflow begins with the acquisition of soil data, followed by preprocessing to cleanse and prepare the dataset. In the feature engineering and selection stages, critical predictors are identified to enhance model performance. The workflow continues with model selection and training, aiming to find the best algorithm for accurate soil health predictions. Finally, model evaluation and interpretation ensure the model's effectiveness, and insights are used for decision-making, supporting improved soil health management.

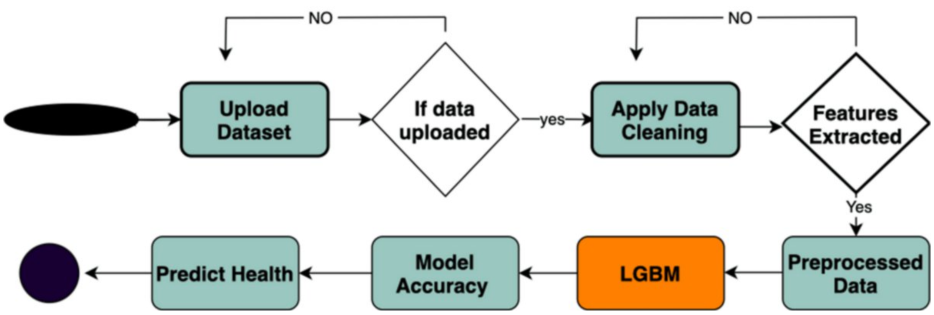

FIGURE 3: Soil health monitoring system workflow

LGBM, Light Gradient Boosting Machine

Figure 3 represents the workflow of the Soil Health Monitoring System. The process starts with dataset uploading, followed by data cleaning and feature extraction. Also, the preprocessed data is then used for training purposes of the LightGBM model, and the model's accuracy is evaluated before making predictions on soil health.

Data Collection

The proposed study utilises the soil health data collected from various government websites in India. The data collected consists of 4,817 instances with 15 attributes that provide comprehensive information on the soil's chemical, physical, and biological properties. Used attributes help evaluate soil health and its impact on agricultural productivity, especially in the context of sustainable farming and crop yield prediction, such as state, district, and 13 soil chemical parameters, like organic matter, moisture retention expressed using percentage (%). Micronutrients like phosphorus, nitrogen, and potassium, iron (Fe), and zinc (Zn) are expressed using parts per million (ppm). The soil salinity is measured in terms of decisemens/cm (dS/cm), temperature in °C, microbial activity in terms of colony forming unit (CFU)/g, soil compaction in g/cm^3, and soil structure on a scale of 1-5, where pH indicates hydrogen ion concentration in the soil.

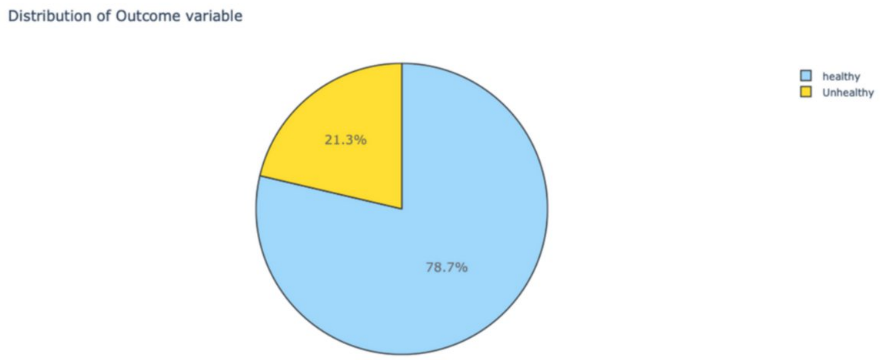

FIGURE 4: Distribution of soil health

Figure 4 shows the proportion of soil samples categorised as "healthy" and "unhealthy" in the dataset. The "healthy" samples are represented using a light blue colour, while the "unhealthy" samples are shown in gold. The chart illustrates the overall balance between the two categories, offering insight into the distribution of the dataset’s target variable for further analysis in soil health monitoring.

The data used in the study was gathered using some recognised soil sampling techniques and collected from multiple districts across various Indian states. Soil samples were also tested in labs, and the final results were uploaded to government databases. The dataset aims to give a regional overview of soil health to help farmers and improve soil management practices. A dataset is important for understanding soil health and fertility, and both factors directly affect crop yield. By analysing soil factors like nutrient levels, pH, and moisture retention, farmers can modify the existing methods by using fertilisers, rotating crops, and managing irrigation techniques to improve crop yield and promote sustainable soil management. The data also helps with precision agriculture by providing region-specific recommendations to optimise resource use and increase productivity.

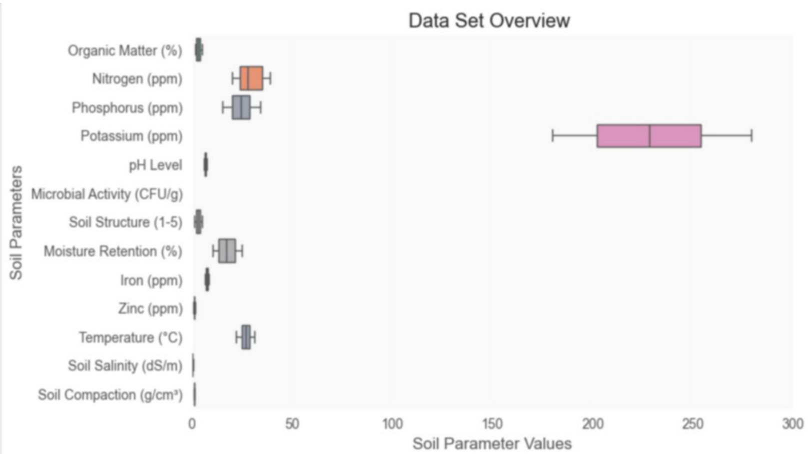

FIGURE 5: Summary of dataset features

Figure 5 shows the variable dispersion. The horizontal layout makes it easy to compare the range, median, and outliers of multiple parameters. Using a Set2 color palette and ggplot2-inspired style, the plot provides clarity of the dataset's structure, helping to identify outliers and differences between parameters based on their values.

Data Preprocessing

The data transformation step plays an important role in organizing the soil health dataset for model training. The procedure outlines the steps followed to clean, transform, and prepare the data for model training [36]. The following preprocessing pipeline ensures that the data is clean, standard, and prepared for use in machine learning algorithms. The used tasks establish a solid base for making an accurate, trustworthy, and reliable model for crop yield prediction.

Setting Up Libraries and Reading the Dataset: The process begins by setting up the essential libraries such as pandas, numpy, scikit-learn, scipy, seaborn, and matplotlib. The dataset, containing all soil health indicators like pH level, phosphorus, nitrogen, and potassium content, moisture, etc., is then read into a pandas frame.

Target Variable Identification: The final response variable for this work is crop yield, and our goal is to predict it based on the soil health indicators. This variable should be isolated and segregated from the feature set.

Handling Missing Values: Missing values in the dataset must be handled appropriately because incomplete data can have a detrimental effect on the model's performance. In this instance, the dataset's missing values were first identified, and then they were imputed using the proper techniques, such as substituting the median for numerical features.

Exploratory Data Analysis: To find better, deeper insights into the dataset, exploratory data analysis (EDA) is used or conducted. This includes charting the distributions of soil health indicators, identifying trends, and understanding relationships between variables. EDA also helps in recognizing some important factors and possible connections between soil health indicators that can impact crop yield. The relationship matrix helps in evaluating the strength of relationships between different indicators [37].

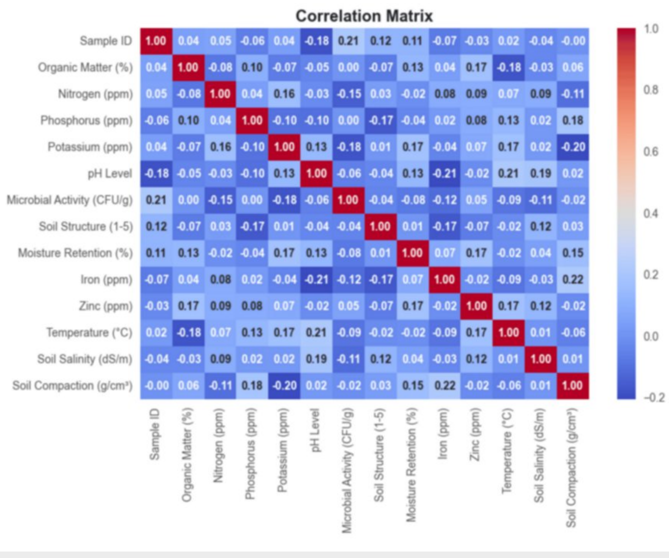

FIGURE 6: Correlation matrix shows how features in the dataset are related

Figure 6 visually represents the relationships between different attributes of the soil health dataset. Each cell represents a correlation coefficient; values close to 1 reflect a strong positive correlation, and values close to -1 show a strong negative correlation. Lighter shades represent weaker correlations, and the darker shades indicate stronger associations, further helping in identifying the attribute dependencies and patterns that are important for understanding soil health and guiding further analysis. Darker blue and red hues represent stronger negative and positive correlations, respectively. Of particular interest are a positive relationship between soil compaction and moisture retention (r = 0.22) and a negative relationship between potassium and soil compaction (r = -0.20). This correlation matrix aids in the determination of interdependencies between features that affect soil health classification.

Dataset Preparation: Before training the machine learning model, numerical attributes were normalised, and categorical variables were encoded. The StandardScaler is applied to standardise numerical attributes such as pH level, moisture retention, nutrient concentrations, etc. The standardisation process makes sure that all attributes have equal contributions to the model and are scaled correctly. In case there are any categorical variables in the dataset, like soil structure, they are encoded using LabelEncoder. Additionally, outliers present were also addressed by detecting and managing extreme values to avoid distorting predictions of the model. Feature selection was also performed in such a manner that less relevant features were removed based on EDA, and correlation analysis was also performed to enhance model efficiency and accuracy.

Model Construction

In this stage, the objective is to develop a strong predictive model using the LightGBM algorithm. The framework was made using a systematic approach that included data preparation, hyperparameter tuning, and model evaluation [36-40]. This section describes the steps involved in data splitting, model training, and the role of each phase in the process.

Preparing Features (X) and Target (y): After preprocessing, the given dataset is split into input features (X) and the target variable (y). The features are then scaled and encoded, and then stored in X_Scaled, while the target variable (y) remains the same. This step enables the model to understand how soil health attributes affect crop production.

Data Splitting for Model Training and Evaluation: Furthermore, the dataset is divided into two parts: 70% for training the model and 30% for testing the model. The training part is used to train the model, while the testing part is used to check how well the model performs on new or unfamiliar data.

Model Training

After Data splitting, the model was trained using LightGBM classifier [26]. During the training process, the model learned different patterns and relationships between different attributes that correlate with the target variable. To improve the performance of the model and to avoid overfitting, hyperparameters were also fine-tuned using the cross-validation method. Once the training process was complete, the model was tested on new data to evaluate its accuracy and overall performance.

Model Evaluation

Evaluating the model is also an important step in machine learning because it finds out how well the trained model works and whether it can correctly predict new, unseen data. Different methods and metrics were used to examine the effectiveness of the LightGBM model on the test dataset. The following evaluation metrics were calculated to evaluate the performance of the model:

Accuracy: This metric indicates the ratio of correctly predicted values to the total values. It provides a general idea of the model's effectiveness. The accuracy of the model can be calculated using the formula:

$$Accuracy = \frac{X+Y}{X+Y+W+Z} \qquad (4)$$

where X = True Positives, Y = True Negatives, W = False Positives, and Z = False Negatives.

Precision: Precision measures the fraction of true positive predictions concerning all positive predictions made by the model. The precision of the model can be calculated using the formula:

$$Precision = \frac{X}{X+Z} \qquad (5)$$

where X = True Positives and Z = False Positives

Recall: Recall is used to measure the ratio between true positive and the sum of true positive and false negative, and can be calculated using the formula :

$$Recall = \frac{X}{X+Y} \qquad (6)$$

where X = True Positives and Y = False Negatives

F1-Score: The F1-score is defined as the ratio of twice precision and recall to the sum of precision and recall. It is especially useful when there is an unequal distribution of classes, and can be calculated using this formula:

$$F1 = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{7}$$

where:

P = $\frac{TP}{TP + FP}$

R = $\frac{TP}{TP + FN}$

ROC-AUC Score: The ROC-AUC score defines the area under the ROC curve. The ROC-AUC score ranges between 0 and 1, where 0.5 indicates random guessing, and 1 indicates perfect performance. It reflects how well a model can produce relative scores for differentiating between positive and negative instances at various classification thresholds.

Confusion Matrix: The confusion matrix summarizes the model's predictions by comparing several correct and incorrect classifications. It reflects the counts of true positives, true negatives, false positives, and false negatives, which allows an evaluation of the model's performance for each class.

Algorithm: LightGBM classifier

LightGBM is an ensemble learning method based on decision trees, designed to be efficient and scalable. It uses the principles of gradient boosting decision trees and is optimized for speed and memory efficiency.

Input: Training dataset $D = \{(s_1,y_1), (s_2,y_2), ..., (s_m,y_m)\}$, where $s_i \in \mathrm{R}^n$ represents a feature vector, y_i represents the class label, learning rate η, number of boosting iterations T, and maximum depth d.

Output : Trained LightGBM model $F_T(s)$.

Initialize: $$F_0(s) = \arg min_\gamma \sum_{i=1}^{m} L(y_i, \gamma) \tag{8}$$

where L is the function that determines the difference between predicted and actual values, typically the logarithmic difference for classification.

$$L(y_i, \hat{y}) = -\sum_{j=1}^{c} \mathbb{I}(y = j) \log(\hat{y}_j) \tag{9}$$

for i = $1 \mapsto$ T, do

1. Compute the residuals for each sample:

$$r_i^{(t)} = -\left.\frac{\partial L(y_i, F(s_i))}{\partial F(s_i)}\right|_{F(s_i)=F_{t-1}(s_i)}, \quad i = 1, 2, \ldots, m \tag{10}$$

2. Fit a decision tree h_t to the residuals:

$$h_t = \arg\min_h \sum_{i=1}^{m} (r_i^{(t)} - h(s_i))^2 \tag{11}$$

3. Update the model:

$$F_t = F_{t-1}(s) + \eta h_t(s) \tag{12}$$

end

Output: Final model $F_T(s)$, where the predicted probability of class j for a given sample s is:

$$\hat{y}_j = \frac{\exp(F_T^j(s))}{\sum_{k=1}^{c} \exp(F_T^k)(s)}, j = 1, 2, ..., c \quad (13)$$

Detailed Explanation:

1. Initialization: The initial model $F_0(s)$ is set as a constant that minimizes the loss function L, where y_i represents the actual outcome and $\hat{y_i}$ represents the predicted probability.

2. Gradient Calculation: At each boosting iteration t, compute the negative gradient of the loss function concerning the current model $F_{t-1}(s)$ to determine the residuals $r_i^{(t)}$. This is given by:

$$r_i^{(t)} = \frac{\partial L\left(y_i, F\left(s_i\right)\right)}{\partial F(s_i))} \quad (14)$$

3. Tree Fitting: Fit a new decision tree $h_t(s)$ to the residuals. Its objective is to mitigate the squared difference between the residuals and the predictions made by the tree:

$$h_t = \arg\min_h \sum_{i=1}^{m} ({r_i}^{(t)} - h(s_i))^2 \quad (15)$$

4. Model Update: The model is updated by adding the new tree weighted by the learning rate η:

$$F_t(s) = F_{t-1}(s) + \eta h_t(s) \quad (16)$$

Here, η controls how much impact each tree has on the final model.

5. Prediction: The final model $F_T(s)$ generates a score for each class j, and the softmax function is used to convert these scores into probabilities:

$$\hat{y}_j = \frac{exp(F_T^j(s))}{\sum_{k=1}^{c} exp(F_T^K(s))} \quad (17)$$

where c is the total number of classes.

Parameters	Description
$D = \{(s_1, y_1), \ldots, (s_m, y_m)\}$	Training dataset containing m samples.
$s_i \in \mathbb{R}^n$	Feature vector for the i-th sample with n features.
$y_i \in \{1, 2, \ldots, c\}$	True class label of the i-th sample, where c is the total number of classes.
η	Learning rate – controls the contribution of each new tree to the model.
T	Number of boosting iterations (trees to be added).
d	Maximum depth of each decision tree (controls tree complexity).
$F_t(s)$	The model after t iterations, mapping a sample s to a real-valued prediction vector.
$F_0(s)$	Initial model prediction, typically a constant minimizing the loss function over all samples.
$h_t(s)$	The decision tree trained at iteration t, which predicts residuals.
r_i^(t)	The residual or pseudo-gradient for sample i at iteration t.
$L(y_i, \hat{y}_i)$	Loss function measuring the difference between true label y_i and predicted probability vector $\hat{y}_i$.
$L(y_i, \hat{y}) = -\Sigma\, I(y_i = j) \log(\hat{y}_j)$	Multiclass log loss , with I(·) as the indicator function.
$\partial L(y_i, F(s_i))/\partial F(s_i)$	Gradient of the loss with respect to the model's output, used to compute residuals.
r_i^(t)	Pseudo-residual for sample i, computed using the current model's output.
$\hat{y}_j$	Softmax output: predicted probability that sample s belongs to class j.

TABLE 5: Summary of parameters and mathematical components in the LightGBM classification algorithm

LightGBM, Light Gradient Boosting Machine

The main parameters, variables, and functions utilised in the development and application of the LightGBM classification algorithm, which operates iteratively, are briefly summarised in Table 5. The algorithm works iteratively. At each step, it fits a new tree to approximate the gradient of the loss and updates the model accordingly. The softmax function is used in the end to convert raw scores $F_T^j(s)$ into normalised class probabilities $\hat{y}_j$, where $F_T^j(s)$ is the final model score for class *j* for input sample *s*.

Results

In this experiment, a dataset focusing on detailed information on various soil health parameters is used, including organic matter moisture retention, micronutrients like nitrogen, phosphorus, potassium, iron, and zinc, soil salinity, temperature, microbial activity, soil compaction, soil structure, etc., to get the health status of soil. The dataset samples from multiple geographical locations, making it suitable for identifying trends and predicting soil health classifications. For preprocessing, modeling, evaluation, and visualization, Python 3.10 was utilized with some powerful libraries such as Pandas and NumPy for data handling and manipulation. LabelEncoder was used for encoding categorical variables and StandardScaler for numerical features normalization. Seaborn and Matplotlib were used for generating statistical plots and heatmaps, while Plotly and Squarify supported interactive visual analysis.

The dataset went through preprocessing, which includes handling missing values, feature standardization, and outlier removal. Relevant parameters were selected based on their importance, and additional derived features, such as the soil fertility index and nutrient imbalance score, were calculated to augment the analysis. Model development and evaluation were performed using scikit-learn and LightGBM. LightGBM was trained and evaluated. The dataset was split into training and testing subsets using an 80:20 ratio. Hyperparameter optimization was conducted using both GridSearchCV and RandomizedSearchCV, while cross-validation and ensemble learning were also implemented using K-Fold and VotingClassifier. Model performance was assessed using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, with tools like classification report, confusion matrix, and ROC curve.

The LightGBM classifier for the soil health monitoring system displayed outstanding performance with an accuracy of 95.56%. A detailed review of the confusion matrix of the work presented highly accurate classification, with only a few mistakes made between "healthy" and "unhealthy" soil categories. The model achieved a weighted precision of 95.79% and a weighted recall of 95.56%, which together minimised false negatives and guaranteed accurate identification of true positives. The weighted F1-score was 95.35%, indicating a balanced performance in terms of both precision and recall. Furthermore, the AUC score of 0.90 shows the model's strong ability to differentiate between different soil health classes. Important attributes of soil health, like organic matter, nitrogen, pH levels, potassium, microbial activity, etc., confirm the model's relations with known agricultural practices.

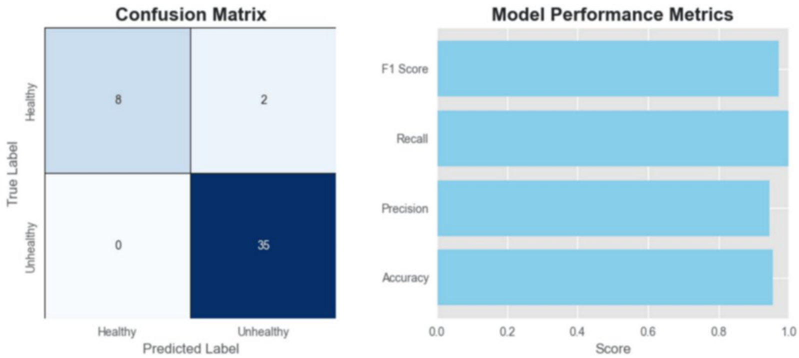

FIGURE 7: Confusion matrix and model performance metrics of the proposed model, reflecting its performance in terms of correct and incorrect predictions, and the evaluation metrics like accuracy, precision, recall, and F1-score

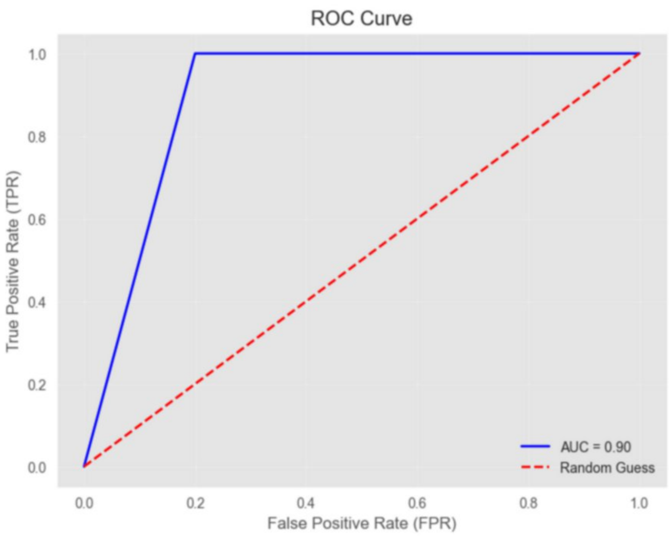

FIGURE 8: ROC-AUC curve for the soil health monitoring system model, showing the balance between true positive rate (TPR) and false positive rate (FPR) to find out the model's classific

ROC-AUC, Receiver Operating Characteristic - Area Under the Curve

Figures 7 and 8 show an evaluation of model performance through various visualizations such as a confusion matrix, model performance matrix, and ROC curve. The confusion matrix offers an understanding of the algorithm's accuracy by showing correct and incorrect classifications. Performance metrics show precision, recall, and F1-score and evaluate the model's effectiveness in predicting soil health outcomes. The ROC curve provides the model's competence in distinguishing between classes, with AUC indicating overall performance. Finally, the feature importance scores highlight key predictors in the dataset, identifying which features most significantly impact soil health predictions.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed Model	95.56	95.79	95.56	95.35
Random Forest	91.35	90.2	91.1	90.65
SVM	89.47	88.1	89	88.54
ELM	86.84	85.3	87.1	86.19
KNN	77.78	60.49	78	68.06

TABLE 6: Comparison of classification models for soil health prediction

SVM, Support Vector Machine; ELM, Extreme Learning Machine; KNN, K-Nearest Neighbors

Table 6 shows performance measures of different machine learning models tested on the soil health dataset. The proposed model shows the best accuracy, precision, recall, and F1-score compared to classical classifiers like Random Forest, SVM, ELM, and KNN. Measures are computed by using standard definitions in terms of confusion matrices and weighted averages for adjusting class imbalance. The improved performance can be traced to the fact that LightGBM can efficiently manage feature interactions, is highly resistant to overfitting via gradient-based one-side sampling , and performs well on handling imbalanced data. LightGBM benefiting because of accelerated training, improved regularization, and histogram-based splitting optimization, which are especially favorable in structured tabular data like measurement of soil parameters. Metrics were calculated utilizing standard test methods on confusion matrices and weighted means to account for class imbalance inherent within the dataset.

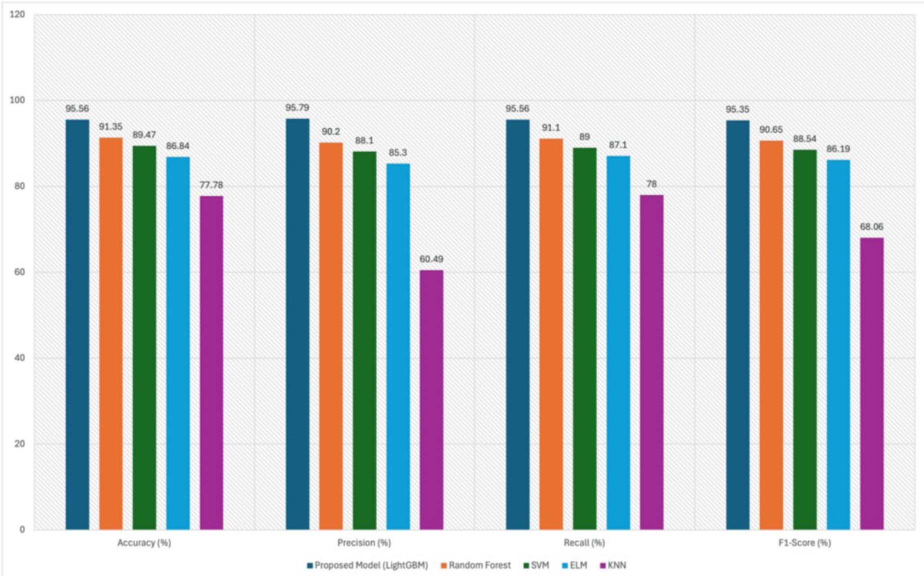

FIGURE 9: Performance comparison of machine learning models for soil health classification

SVM, Support Vector Machine; ELM, Extreme Learning Machine; KNN, K-Nearest Neighbors

Figure 9 presents the comparative study of five machine learning models, i.e., LightGBM, Random Forest, SVM, ELM, and KNN, on the soil health dataset. The bar chart shows performance in terms of evaluation metrics - accuracy, precision, recall, and F1-score. The proposed model performs better than other models on all measures with better classification ability and generalizability for binary soil health prediction.

Discussion

The LightGBM model's great performance in classifying soil health status can be ascribed to its natural capacity to control class imbalance and capture complicated non-linear relationships properly. Particularly helpful for soil datasets where chemical and physical properties often affect each other in hidden ways, LightGBM's gradient boosting framework precisely handles data with order and smooth values. Its leaf-wise tree growth and built-in regularization mechanisms likely contributed to the model's robustness and generalization. The proposed model's 95.6% accuracy exceeds related studies on soil classification, which usually report accuracy in the range of 72% to 94%. Despite the dataset's class imbalance, with a healthy-to-unhealthy ratio of approximately 78:22, the model managed to generalize well. This is especially important since the LightGBM classifier managed the imbalance to guarantee correct forecasts for both classes.

Where traditional models like Gaussian Naive Bayes and SVM often struggle with skewed data, LightGBM maintained strong precision and perfect recall for the minority class, correctly identifying all unhealthy samples with no false negatives, a crucial advantage for agricultural decision-making. The weaker results of basic models, particularly Gaussian Naive Bayes, support the idea that it does not work well with the complex relationships in agricultural data. By contrast, LightGBM's capacity for capturing feature dependencies enabled it to surpass less complex models in terms of accuracy and reliability.

The analysis of feature importance identifies that microbial activity, organic matter, pH level, nitrogen level, potassium, soil compaction, retention of moisture, phosphorus, and iron were the most important predictors of soil health. These findings are consistent with agronomic knowledge and show that the model is not only accurate but also transparent. This explainability can assist agronomists in better customizing soil management and fertilization practices. For example, low nitrogen can direct the application of nitrogenous fertilizers like urea or ammonium sulfate. Organic matter deficiency can indicate the addition of compost, green manure, or other organic amendments to increase soil structure and biological processes. Where pH levels range away from the ideal, soil amendments such as lime (to increase pH) or elemental sulphur (to decrease pH) can be utilized. Potassium and phosphorus content can also guide the application of site-specific NPK fertilizers. These insights guide farmers in making informed decisions and result in enhanced nutrient equilibrium, optimization of resources, and increased crop yields. The research provides early soil health diagnoses to help prompt an immediate response and enhance crop yield, showing promising potential for applying machine learning to precision agriculture.

Also, in order to test the robustness of the proposed model based on LightGBM, we compared the model with baseline classification models like Random Forest, SVM, ELM, and KNN. The comparative results evidently indicate that the proposed model performs better than these models with respect to accuracy, precision, recall, and F1-score. The reasons behind such performance lie in its capacity for dealing with intricate feature interactions, regularization, and speed during training on large-scale structured data. Although with these encouraging results, one of the limitations of this work is that there was no geospatial and real-time field testing. The model was trained and tested over a tabular dataset, ignoring spatial variation and dynamic soil states. Future refinements could involve the incorporation of georeferenced soil data and IoT sensor inputs to facilitate real-time monitoring and localized decision-making.

Although the confusion matrix shows a 100% recall for unhealthy soil samples, and this implies that the model has learned well enough to recognize the minority class. When we looked at the distribution of the features for these samples, we could see that the unhealthy class had a wide range of values in important parameters like pH, nitrogen, phosphorus, and potassium. Yet, this performance rests on test data with a defined structure, and under real-world conditions, edge cases like borderline nutrient concentrations or mixed deficiency profiles can still cause difficulties.

Conclusions

This study aimed to assess soil health based on various soil parameters using the LightGBM Classifier. In-depth analysis of key parameters on which soil health depends, such as organic matter, nitrogen, phosphorus, potassium, pH levels, microbial activity, etc., was conducted to evaluate the performance of the LightGBM model in the context of soil health monitoring. The findings reveal that the LightGBM classifier achieved an exceptional accuracy of 95.56% by showing a higher degree of generalisability and robustness when applied to a soil health monitoring system. This high performance was achieved using a reduced set of features, demonstrating the model’s efficiency, highlighting the viability of leveraging existing models for soil health assessment without significantly increasing the complexity of our solutions.

Experiments were conducted on well-established datasets, which ensure that results contribute more value to the soil health monitoring methodologies. The robust performance of the model proposed reveals encouraging outcomes in developing better and more precise tools for soil health assessment. This proposed work not only shows the viability of applying the LightGBM model to soil health monitoring but also highlights the importance of continuous research in this critical area of research. A combination of diverse techniques will most probably bring richer understandings and advancements to agricultural sustainability. The proposed LightGBM model shows improved accuracy in predicting soil health. In future work, we plan to benchmark the proposed model on more soil datasets covering varied geographic locations and crop systems in order to evaluate the model's generalizability and robustness under different agro-climatic settings. Furthermore, incorporating real-time sensor inputs and geospatial heterogeneity in future work would considerably enhance the model's flexibility and usability in precision agriculture.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Anmol Chaudhary

Acquisition, analysis, or interpretation of data: Anmol Chaudhary, Prabh Deep Singh, Rohan Verma, Manmohan Sharma

Drafting of the manuscript: Anmol Chaudhary

Critical review of the manuscript for important intellectual content: Prabh Deep Singh, Rohan Verma, Manmohan Sharma

Supervision: Prabh Deep Singh, Rohan Verma, Manmohan Sharma

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

References

1. Gowdy J: Our hunter-gatherer future: Climate change, agriculture and uncivilization. Futures. 2020, 115:102488. 10.1016/j.futures.2019.102488
2. Timsina J, Wolf J, Guilpart N, et al.: Can Bangladesh produce enough cereals to meet future demand?. Agricultural Systems. 2018, 163:36-44. 10.1016/j.agsy.2016.11.003
3. Hartemink AE: Assessing soil fertility decline in the tropics using soil chemical data. Advances in Agronomy. 2006, 89:179-225. 10.1016/s0065-2113(05)89004-2
4. Sujatha M, Jaidhar CD: Machine learning-based approaches to enhance the soil fertility—A review. Expert Systems with Applications. 2024, 240:122557. 10.1016/j.eswa.2023.122557
5. Pramanik SK, Hossain MS, Islam SMM: Noncontact soil moisture estimation using continuous wave radar and deep learning. IEEE Sensors Journal. 2024, 24:28419-28426. 10.1109/jsen.2024.3430531
6. Yüzügüllü O, Fajraoui N, Liebisch F: Soil texture and pH mapping using remote sensing and support sampling. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2024, 17:12685-12705. 10.1109/jstars.2024.3422494
7. Muthusamy S, Ramu SP: IncepV3Dense: Deep ensemble based average learning strategy for identification of micro-nutrient deficiency in banana crop. IEEE Access. 2024, 12:73779-73792. 10.1109/access.2024.3405027
8. Lee W, Lee J: Tree-based modeling for large-scale management in agriculture: Explaining organic matter content in soil. Applied Sciences. 2024, 14:1811. 10.3390/app14051811
9. Umpierrez-Failache M, Rahim AA, Betancor L, Ghoshal S: Oryza sativa as a tool for assessing arsenic efficacy of arsenic remediation of agricultural soils by sulfidated zerovalent iron nanoparticles. IEEE Transactions on NanoBioscience. 2022, 21:157-165. 10.1109/tnb.2021.3105281
10. Fikry I, Eltawil A, Gheith M: A robust crop rotation optimization model with water scarcity and net return uncertainty considerations. IEEE Access. 2021, 9:128938-128950. 10.1109/access.2021.3113125
11. Sakin E, Yanardağ Hİ, Fırat Z, Çelik A, Beyyavaş V, Cun S: Some indicators for the assessment of soil health: a mini review. MAS Journal of Applied Sciences. 2024, 9:297-310. 10.5281/zenodo.11665000
12. Tian A, Fu C, Su X, Yau HT, Xiong H: Estimation of low organic matter content in desert soil of arid area based on fractional order Sprott chaotic circuit and gray theory. IEEE Access. 2020, 8:25001-25013. 10.1109/access.2020.2970513
13. Tan Y-J, Jiang Q, Yu L, Liu H, Zhang B: Reducing the moisture effect and improving the prediction of soil organic matter with VIS-NIR spectroscopy in black soil area. IEEE Access. 2021, 9:5895-5905. 10.1109/access.2020.3048794
14. Zhao W, Wu Z, Yin Z, Li D: Reducing moisture effects on soil organic carbon content estimation in Vis-NIR

spectra with a deep learning algorithm. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2023, 16:7733-7748. 10.1109/jstars.2023.3287583

15. Bhatt MK, Labanya R, Joshi HC: Influence of long-term chemical fertilizers and organic manures on soil fertility - A review. Universal Journal of Agricultural Research. 2019, 7:177-188. 10.13189/ujar.2019.070502
16. Jino Ramson SR, León-Salas WD, Brecheisen Z: A self-powered, real-time, LoRaWAN IoT-based soil health monitoring system. IEEE Internet of Things Journal. 2021, 8:9278-9293. 10.1109/jiot.2021.3056586
17. Al-Shahari EA, Aldehim G, Aljebreen M, Saad Alqurni J, Salama AS, Abdelbagi S: Internet of things assisted plant disease detection and crop management using deep learning for sustainable agriculture. IEEE Access. 2025, 13:3512-3520. 10.1109/access.2024.3397619
18. Salam MA, Rahman QM, Ang SP, Wen F: Soil resistivity and ground resistance for dry and wet soil. Journal of Modern Power Systems and Clean Energy. 2015, 5:290-297. 10.1007/s40565-015-0153-8
19. Zhang Z, Fan Y, Zhang A, Jiao Z: Baseline-based soil salinity index (BSSI): a novel remote sensing monitoring method of soil salinization. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2023, 16:202-214. 10.1109/jstars.2022.3223935
20. Ren L, Qiu J: Influence of boron-containing oxides on rhizosphere soil properties. 2012 2nd International Conference on Remote Sensing, Environment and Transportation Engineering, Nanjing, China. 2012, 1-4. 10.1109/rsete.2012.6260597
21. Narayan OP, Kumar P, Yadav B, Dua M, Johri AK: Sulfur nutrition and its role in plant growth and development. Plant Signaling & Behavior. 2022, 18:2030082. 10.1080/15592324.2022.2030082
22. AlZubi AA: Artificial intelligence and internet of things for sustainable farming and smart agriculture. IEEE Access. 2023, 11:78686-78692. 10.1109/access.2023.3298215
23. Meena SK, Kumar A, Meena KR, et al.: Advanced and emerging techniques in soil health management. Advancements in Microbial Biotechnology for Soil Health. Microorganisms for Sustainability. Bhatia RK, Walia A (ed): Springer, Singapore; 2024. 50:343-362. 10.1007/978-981-99-9482-3_15
24. Chandra H, Pawar PM, Elakkiya R, Tamizharasan PS, Muthalagu R, Panthakkan A: Explainable AI for soil fertility prediction. IEEE Access. 2023, 11:97866-97878. 10.1109/access.2023.3311827
25. Wang YJ, Chen Y: An improved farmland fertility algorithm for global function optimization. IEEE Access. 2020, 8:111850-111874. 10.1109/access.2020.3002555
26. Parganiha V, Verma M: An effective soil analysis and crop yield prediction based on optimised light GBM in smart agriculture. Journal of Agronomy and Crop Science. 2024, 210:e12726. 10.1111/jac.12726
27. Kakkar S, Jaya Lakshmi A, Durai CAD, et al.: Enhancing energy efficiency and classification modeling through a combined approach of LightGBM and stratified KFold cross-validation. Electric Power Components and Systems. 2024, 1-19. 10.1080/15325008.2024.2315213
28. Friha O, Ferrag MA, Shu L, Maglaras L, Wang X: Internet of things for the future of smart agriculture: a comprehensive survey of emerging technologies. IEEE/CAA Journal of Automatica Sinica. 2021, 8:718-752. 10.1109/jas.2021.1003925
29. Nguyen-Tan T, Le-Trung Q: A novel 5G PMN-driven approach for AI-powered irrigation and crop health monitoring. IEEE Access. 2024, 12:125211-125222. 10.1109/access.2024.3452719
30. Shaikh TA, Mir WA, Rasool T, Sofi S: Machine learning for smart agriculture and precision farming: towards making the fields talk. Archives of Computational Methods in Engineering. 2022, 29:4557-4597. 10.1007/s11831-022-09761-4
31. Keerthan Kumar TG, Shubha C, Sushma SA: Random forest algorithm for soil fertility prediction and grading using machine learning. International Journal of Innovative Technology and Exploring Engineering. 2019, 9:1301-1304. 10.35940/ijitee.l3609.119119
32. Gaikwad S, Aiwale A, Rekade V, Kalunge V: Soil Classification and Crop Suggestion using Machine Learning Techniques. Soil classification and crop suggestion using machine learning techniques. 2022, 10:4984-4986. 10.22214/ijraset.2022.43524
33. Saleem RM, Bashir RN, Faheem M, Haq MA, Alhussen A, Alzamil ZS, Khan S: Internet of things based weekly crop pest prediction by using deep neural network. IEEE Access. 2023, 11:85900-85913. 10.1109/access.2023.3301504
34. Terhoeven-Urselmans T, Vagen T-G, Spaargaren O, Shepherd KD: Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library. Soil Science Society of America Journal. 2010, 74:1792-1799. 10.2136/sssaj2009.0218
35. Suchithra MS, Pai ML: Improving the prediction accuracy of soil nutrient classification by optimizing extreme learning machine parameters. Information Processing in Agriculture. 2020, 7:72-82. 10.1016/j.inpa.2019.05.003
36. Sirsat MS, Cernadas E, Fernández-Delgado M, Khan R: Classification of agricultural soil parameters in India. Computers and Electronics in Agriculture. 2017, 135:269-279. 10.1016/j.compag.2017.01.019
37. Erlin, Yunianta A, Wulandhari LA, Desnelita Y, Nasution N, Junadhi: Enhancing rice production prediction in Indonesia using advanced machine learning models. IEEE Access. 2024, 12:151161-151177. 10.1109/access.2024.3478738
38. Perera AS, Witharana C, Manos E, Liljedahl AK: Hyperparameter optimization for large-scale remote sensing image analysis tasks: a case study based on permafrost landform detection using deep learning. IEEE Access. 2024, 12:43062-43077. 10.1109/access.2024.3379142
39. Condran S, Bewong M, Islam MZ, Maphosa L, Zheng L: Machine learning in precision agriculture: a survey on trends, applications and evaluations over two decades. IEEE Access. 2022, 10:73786-73803. 10.1109/access.2022.3188649
40. Asif MR, Bording TS, Barfod AS, et al.: Effect of data pre-processing on the performance of neural networks for 1-D transient electromagnetic forward modeling. IEEE Access. 2021, 9:34635-34646. 10.1109/access.2021.3061761