Optimization in Agriculture

Journal Homepage: sciencesforce.com/oia

Optimization Agri. Vol. 2 (2025) 1–9

Paper Type: Original Article

A Comparative Study of Machine Learning Models for Soil Fertility Prediction Based on Soil Properties

Walid Abdullah [1,*]

[1] Faculty of Computers and Informatics, Zagazig University, Zagazig, Sharqia, 44519, Egypt, waleed@zu.edu.eg.

Received: 12 Sep 2024 **Revised:** 31 Dec 2024 **Accepted:** 27 Jan 2025 **Published:** 29 Jan 2025

Abstract

Soil fertility plays an important role in agricultural productivity, accurately predicting soil fertility based on elemental soil properties is essential for optimizing resource management and promoting sustainable farming practices. This study investigates the use of machine learning models to predict soil fertility from a dataset containing various elemental soil properties and several trace elements. The soil fertility is categorized into three classes: "Less Fertile," "Fertile," and "Highly Fertile." We evaluate the performance of each model using key metrics such as accuracy, precision, recall, and F1 score. Among the models tested, the Extra Trees classifier achieved the highest accuracy of 96.17%, followed by Random Forest and LightGBM with similar accuracies of 95.69% and 95.22% respectively. Our results indicate that ensemble models such as Extra Trees and Random Forest perform significantly better than other algorithms like SVM and SGD. These findings demonstrate the potential of machine learning to revolutionize soil fertility assessment, providing a scalable and effective tool for precision agriculture and sustainable soil management.

Keywords: Machine Learning; Soil Fertility; Extra Trees; Random Forest; Soil Analysis; Precision Agriculture.

1 | Introduction

Precision agriculture has emerged as a transformative approach to modern farming, aiming to optimize agricultural inputs and maximize productivity while minimizing environmental impacts [1]. A key component of precision agriculture is the assessment of soil fertility, which determines the soil's capacity to provide essential nutrients for plant growth [2]. Soil fertility is influenced by various factors, including the concentrations of macronutrients like nitrogen (N), phosphorus (P), and potassium (K), as well as micronutrients such as zinc (Zn), iron (Fe), and manganese (Mn) [3]. Comprehensive soil analysis is therefore critical for identifying nutrient deficiencies and tailoring fertilizer applications to specific crop and soil needs. By enabling sustainable resource management, soil fertility assessment directly contributes to enhancing agricultural productivity and mitigating soil degradation [4].

Soil fertility prediction has traditionally relied on laboratory-based chemical analysis and heuristic methods informed by agricultural expertise. While these methods provide valuable insights, they are often time-consuming, labor-intensive, and geographically limited [5, 6]. Moreover, traditional approaches may fail to

Corresponding Author: waleed@zu.edu.eg

https://doi.org/10.61356/j.oia.2025.2477

capture the complex, non-linear interactions between soil properties and fertility levels [7]. As agriculture scales to meet the demands of a growing global population, the need for more efficient, scalable, and accurate methods to predict soil fertility becomes increasingly apparent [8]. These gaps underscore the importance of exploring advanced analytical techniques to complement and enhance traditional soil assessment methods. In recent years, advancements in artificial intelligence (AI) and machine learning (ML) have revolutionized multiple fields, including healthcare [9, 10], finance [11], and environmental science [12]. Machine learning excels at identifying patterns and relationships within large, complex datasets, making it a powerful tool for predictive modeling. In agriculture, ML has been successfully applied to optimize crop yield predictions [13], Crop diseases detection [14], and weather forecasting [15]. When applied to soil analysis, ML models have the potential to analyze vast amounts of elemental and chemical data, accurately classifying soil fertility levels and offering actionable insights for farmers and agronomists.

This study explores the application of machine learning techniques to predict soil fertility based on elemental soil analysis. A total of twelve machine learning models covering a diverse range of approaches were utilized in this study. The methods include ensemble methods such as Extra Trees, Random Forest, LightGBM, XGBoost, Gradient Boosting, and linear models such as Logistic Regression, Stochastic Gradient Descent (SGD), as well as other algorithms like Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), and AdaBoost. All these models were evaluated and compared using a comprehensive dataset containing key soil properties. The models were assessed based on their accuracy, precision, recall, and F1 score to identify the most effective approaches for soil fertility prediction. By leveraging ML's ability to handle complex data relationships, this study aims to provide a scalable and efficient alternative to traditional soil fertility assessment.

The primary contributions of this study include the development and evaluation of twelve machine learning models tailored for soil fertility prediction, with a focus on comparing their performance in classifying soils into three fertility categories: "Less Fertile," "Fertile," and "Highly Fertile." The study highlights the effectiveness of ensemble models, particularly Extra Trees and Random Forest, in achieving high accuracy and reliability. By applying these advanced methods to real-world soil datasets, the research demonstrates the potential of machine learning to enhance precision agriculture through faster and more accurate soil assessments. This work provides valuable insights and a robust framework for integrating machine learning into soil fertility analysis, contributing to sustainable agricultural practices and improved resource management.

The rest of this paper is structured as follows: Section 2 reviews and discusses related work. Section 3 outlines the methodology including the details of the dataset, preprocessing steps, and presents the machine learning models used. Section 4 presents experimental analysis, experimental setup, and evaluation metrics. Section 5 shows and discusses the results, highlighting the strengths and limitations of the top-performing models and their implications for soil fertility prediction. Finally, Section 6 concludes the paper by summarizing key insights and providing recommendations for future research directions.

2 | Related Work

The prediction of soil fertility is a critical aspect of environmental management, agricultural planning and fields productivity. Over the years, various machine learning techniques have been employed to predict soil fertility with varying degrees of success. In this section, we review the existing literature on soil fertility prediction using machine learning models, highlighting the strengths and limitations, and better understand the effectiveness of these approaches.

In [16], the authors developed a soil fertility index map (SFIm) for Benin using machine learning models to support sustainable land management. The study utilized legacy soil data with eight properties, including nitrogen, pH, and organic matter, and compared Cubist (CB) and Quantile Random Forest (QRF) models. While CB slightly outperformed QRF in accuracy metrics like RMSE and R2, it showed higher uncertainty in predictions. Key predictors included topographic and bioclimatic variables. The analysis revealed widespread low fertility soils, emphasizing the need for targeted agricultural policies and sustainable practices.

Authors in [17] developed predictive fertilization models for potato crops in Eastern Canada using machine learning techniques to optimize nitrogen, phosphorus, and potassium requirements for high tuber yield and quality. Utilizing a dataset of 273 field experiments conducted from 1979 to 2017, they compared models including k-nearest neighbors, random forest, neural networks, and Gaussian processes. Machine learning models achieved R2 values of 0.49-0.59 for yield prediction, outperforming the traditional Mitscherlich model (R2 =0.37). Gaussian processes were identified as the most promising due to their ability to incorporate risk assessment in decision-making. Similarly, machine learning was applied to predict soil fertility elements in Morocco as an alternative to traditional methods [18]. A dataset of 400 soil samples was used test multiple linear regression (MLR), support vector machines (SVM), and random forest (RF) to predict organic matter, potassium (K2O), and phosphorus (P2O5). The models showed satisfactory predictions for soil fertility elements. Cation exchange capacity, carbonates, and texture were identified as key contributors to prediction accuracy, demonstrating the potential of ML in cost-effective soil fertility assessment.

Another machine learning-based approach for soil classification and crop yield prediction was proposed in [19] using various algorithms, including Support Vector Machine (SVM), Random Forest, Naive Bayes, Linear Regression, Multilayer Perceptron (MLP), and Artificial Neural Networks (ANN). Their study aimed to determine soil fertility, recommend crops suitable for specific soil types, and predict crop yield based on soil features. The ANN model, leveraging a deep learning architecture with multiple layers, outperformed other methods, achieving higher accuracy in soil classification and crop yield prediction. This work highlights the potential of machine learning in precision agriculture. Similarly, in [20], a machine learning model was proposed for soil fertility prediction in the Bhimtal block of Uttarakhand, India. The utilized dataset contained soil test reports to classify soil features such as Organic Carbon (OC), Phosphorus (P), Potassium (K), Magnesium (Mn), and Boron (B). The study employed an Artificial Neural Network (ANN) with ReLU and Tanh activation functions, finding that ReLU outperformed Tanh in predicting four out of five soil nutrient parameters. This approach aimed to reduce fertilizer costs and improve efficiency for stakeholders in agriculture.

In [21], an explainable AI (XAI) model was developed for soil fertility prediction using a Random Forest classifier. The model predicts soil fertility based on various physiochemical properties, such as Nitrogen and Organic Carbon concentrations, achieving an accuracy of 97.02%. The model also provides transparent explanations of its predictions through user-friendly graphs. This approach offers insights for improving soil fertility in both the short and long term, demonstrating the effectiveness of XAI in agricultural applications.

These studies highlight the rapid advancements in applying machine learning models in soil fertility prediction across diverse regions and contexts. They demonstrate the versatility of machine learning techniques, including Random Forest, Neural Networks, and support vector machines, in analyzing key soil parameters such as organic carbon, nitrogen, and phosphorus to predict soil fertility and optimize agricultural practices. which help the stakeholders to make data-driven decisions for improved soil management. These efforts showcase the growing potential of machine learning to address soil fertility challenges, reduce costs, and support sustainable land management practices.

3 | Methodology

This section outlines the methodology employed to predict soil fertility based on elemental soil properties using machine learning techniques. The process begins with a description of the dataset, including its attributes and target variable, followed by the data preparation and preprocessing steps, and finally a description of the machine learning models used.

3.1 | Dataset Description

The dataset used in this study is a publicly available dataset. Each data record represents a soil sample described by 12 attributes that capture various elemental and chemical properties of the soil. These attributes include macronutrients such as Nitrogen (N), Phosphorous (P), and Potassium (K); chemical properties like soil pH and electrical conductivity (EC); and micronutrients, including organic carbon (OC), Sulfur (S), Zinc (Zn), Iron (Fe), Copper (Cu), Manganese (Mn), and Boron (B). The target variable, fertility, categorizes soil into three levels: "Less Fertile" (0), "Fertile" (1), and "Highly Fertile" (2). The dataset serves as a valuable resource for investigating the feasibility of predicting soil fertility based on soil properties or soil elements analysis.

3.2 | Dataset Preparation

To prepare the dataset for machine learning, a series of preprocessing steps were conducted. First, to handle the class imbalance, oversampling was applied [22], which increased the representation of the minority class and ensured balanced learning across all classes as illustrated in Figure 1. Next, due to the wide range of values across different attributes, min-max scaling was employed to normalize all features to a range of 0 to 1. This step prevented attributes with larger magnitudes from dominating model training and ensured a consistent scale across all features [23]. Following normalization, the dataset was split into training and testing sets using an 80:20 ratio, allowing the models to train on a substantial portion of the data while retaining a separate subset for evaluating generalization performance.

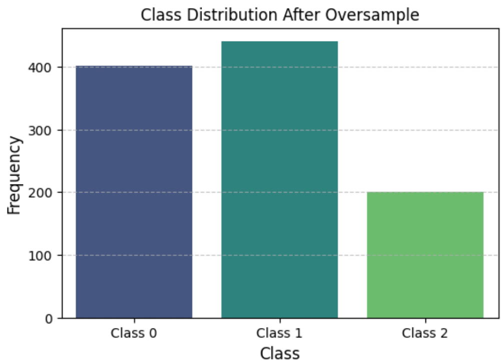


Figure 1. Dataset class distribution.

3.3 | Machine Learning Models

For the prediction of soil fertility, 12 machine learning models were employed, representing a diverse array of algorithms. These included ensemble methods such as Extra Trees, Random Forest, Gradient Boosting, LightGBM, and XGBoost, which can capture complex relationships through the aggregation of multiple decision trees [24]. Linear models, including Logistic Regression and Stochastic Gradient Descent (SGD), were also tested for their efficiency in handling linear relationships [25]. Additionally, Support Vector Machine (SVM), a kernel-based model known for its performance in high-dimensional spaces, and K-Nearest Neighbors (KNN), a distance-based algorithm, were evaluated [26, 27]. The study also incorporated boosting models like AdaBoost, which iteratively combines weak classifiers, and neural network-based approaches such as the Multi-Layer Perceptron (MLP), which excels at capturing non-linear patterns [28]. All models were implemented using the scikit-learn library, with hyperparameter tuning conducted to optimize their performance. The models were evaluated using asset of significant evaluation metrics, providing a comprehensive assessment of their predictive capabilities. These methods laid the groundwork for identifying the most effective model for soil fertility prediction and demonstrating the potential of machine learning in advancing precision agriculture.

4 |Experimental Analysis

In this section, we outline the methodology used to assess the performance of the machine learning models in predicting soil fertility based on soil properties analysis. The following subsections detail the experimental setup, including model training, and evaluation, as well as the evaluation metrics employed to measure the models' effectiveness. These components are critical for understanding how the models were developed, optimized, and assessed in the context of this research.

4.1 |Experimental Setup

The soil properties values in the dataset were normalized using standard scaling to ensure that all features had a similar scale, improving the performance of distance-based algorithms like KNN model [29]. Normalization helps avoid bias in the model's performance, as it prevents features with larger numerical ranges from dominating the distance calculations, thus improving the overall effectiveness of the algorithm. Following the normalization, the dataset was split into two distinct subsets: one for training the machine learning models and another for testing their performance. The data was partitioned in an 80/20 ratio, with 80% of the samples used for training the models and the remaining 20% reserved for testing the models' generalization ability. This division is standard practice in machine learning to evaluate the models on unseen data, ensuring that the models do not overfit to the training data. The experiments were implemented using Python programming language (version 3.10), along with the widely used Scikit-learn library (version 1.5) which provides a robust set of tools for building and evaluating machine learning models [30]. Initially, each machine learning model was trained and evaluated using the default hyperparameters provided by Scikit-learn. This allowed for a baseline assessment of each model's performance. Afterward, hyperparameter optimization was performed to fine-tune the models and enhance their performance by adjusting the settings for each algorithm, such as the number of trees in a random forest or the learning rate in gradient boosting.

4.2 |Evaluation Metrics

To evaluate the performance of the proposed hybrid models, we employed four essential metrics: accuracy, precision, recall, and F1-score. These metrics offer a multifaceted view of the model's overall performance, considering both its effectiveness and efficiency. The mathematical expressions for each of the evaluation metrics are as follows:

$$Accuercy = (True\ Positives + True\ Negatives) / Total\_predictions \tag{1}$$

$$Precision = True\ Positives / (True\ Positives + False\ Positives) \tag{2}$$

$$Recall = True\ Positives / (True\ Positives + False\ Negatives) \tag{3}$$

$$F1 - score = 2 * (Precision * Recall) / (Precision + Recall) \tag{4}$$

These metrics offer distinct ways to evaluate model performance. Accuracy provides an overall measure of the model's performance across all classes, while also addressing how well the model deals with class imbalances. Precision and recall are particularly useful for assessing the model's ability to accurately identify positive cases. In scenarios where a trade-off might occur between precision and recall, the F1-score serves as a harmonizing measure, reflecting the balance between these two metrics and helping to highlight the model's ability to achieve both high precision and recall simultaneously.

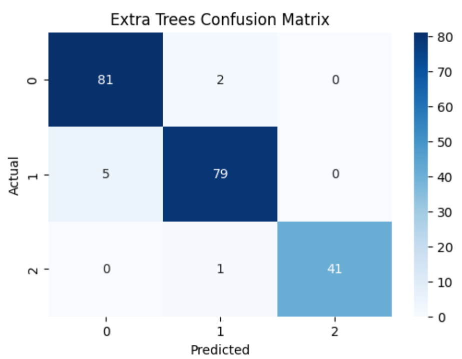
5 |Results and Discussion

Machine learning models offer significant potential for accurately predicting soil fertility levels by leveraging complex data relationships. This section presents the evaluation of twelve machine learning models based on their performance metrics, including accuracy, precision, recall, and F1-score. The results are discussed in terms of overall performance, with particular attention to the strengths of ensemble methods, and their implications for precision agriculture. The performance metrics for all twelve models are summarized in Table 1. Among the evaluated models, ensemble techniques demonstrated superior performance, with Extra Trees achieving the highest accuracy of 96.17%, followed closely by Random Forest (95.69%) and LightGBM (95.22%). These models not only exhibited excellent accuracy but also high precision, recall, and F1-scores, making them the most reliable choices for soil fertility prediction. On the other hand, linear models such as Logistic Regression and SGD recorded lower accuracies, indicating their limitations in handling the non-linear relationships present in soil datasets.

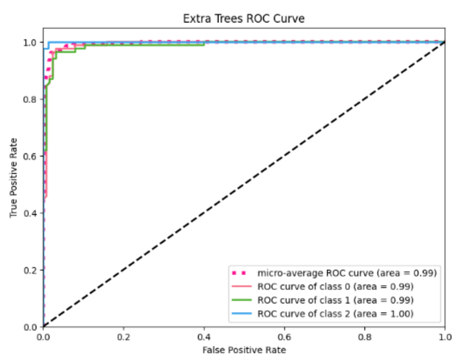
Table 1. Performance metrics of machine learning models for soil fertility prediction.

Model	Accuracy	Precision	Recall	F1-Score
Decision tree (DT)	0.9234	0.9194	0.9363	0.9264
Random Forest (RF)	0.9569	0.9570	0.9642	0.9600
Extra Trees (ET)	**0.9617**	**0.9684**	**0.9642**	**0.9661**
Support Vector Machine (SVM)	0.7560	0.7835	0.7560	0.7700
Logistic Regression (LR)	0.7608	0.7280	0.6545	0.6280
K-Nearest Neighbors (KNN)	0.8278	0.8166	0.8406	0.8243
AdaBoost	0.8373	0.8410	0.8212	0.8289
XGBoost	0.9426	0.9413	0.9522	0.9462
Gradient Boosting	0.9378	0.9348	0.9483	0.9402
LightGBM	0.9522	0.9468	0.9602	0.9524
Stochastic Gradient Descent SGD	0.6938	0.6346	0.6349	0.6267
Multiple layer Perceptron (MLP)	0.8804	0.8881	0.8686	0.8770

The ensemble models consistently outperformed traditional methods and simpler models like Logistic Regression and SGD. The top-performing model, Extra Trees, achieved the highest accuracy of 96.17%, and its confusion matrix and ROC curve are presented in Figure 2. These visualizations further confirm the model's robustness in accurately classifying soil samples into fertility categories. The confusion matrix for Extra Trees, shown in Figure 2 (a), demonstrates a high level of agreement between predicted and actual soil fertility categories. The model correctly classified most of the samples, with minimal misclassifications. The ROC curve in Figure 1 (b) further underscores the model's ability to distinguish between different fertility categories, with an area under the curve (AUC) nearing 1.



(A)



(B)

Figure 2. Performance Analysis of the Extra Trees Model: (a) Confusion Matrix, (b) ROC Curve.

Furthermore, the comparative diagram in Figure 3 provides a comparative visualization of the accuracy of all twelve models and highlights the overall ranking of model performances based on accuracy. Ensemble methods such as Extra Trees, Random Forest, LightGBM, and XGBoost occupy the top ranks, clearly outperforming linear models and simpler classifiers like SVM and KNN. These results emphasize the strength of ensemble learning in capturing complex data patterns for soil fertility prediction.

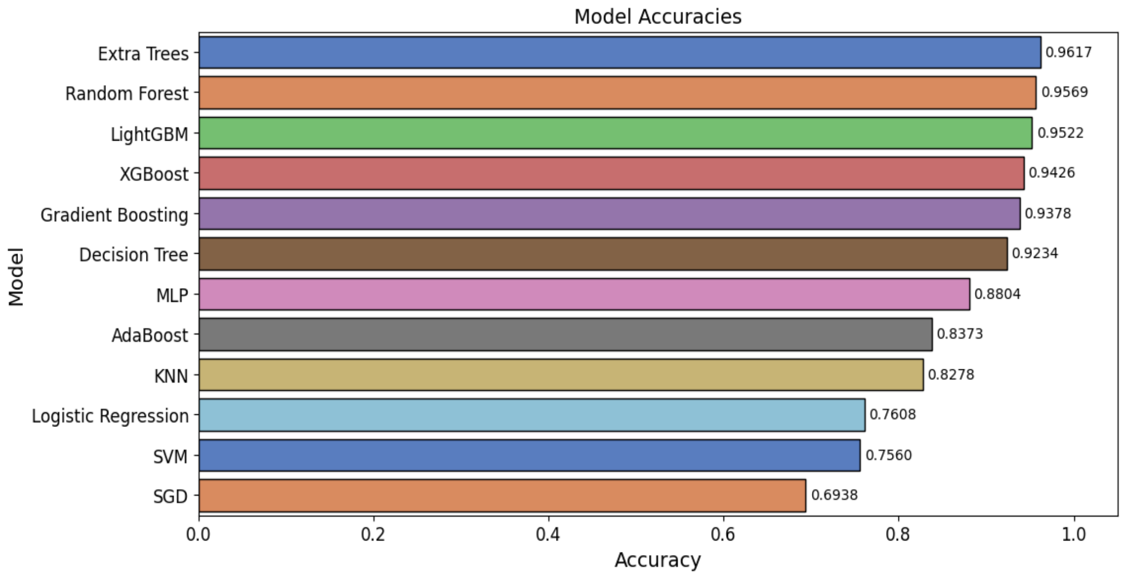


Figure 3. Comparative accuracy of machine learning models for soil fertility prediction.

6 | Conclusion and Future Work

This study demonstrates the effectiveness of machine learning models in predicting soil fertility based on elemental soil properties, providing an efficient and scalable alternative to traditional soil assessment methods. By evaluating twelve machine learning models across key metrics such as accuracy, precision, recall, and F1 score, the research highlights the superior performance of ensemble methods like Extra Trees, Random Forest, and LightGBM. Among these, the Extra Trees classifier achieved the highest accuracy of 96.17%, showcasing its ability to handle complex interactions and relationships within the dataset. The findings underscore the potential of leveraging advanced data-driven techniques to support precision agriculture, enabling more informed decision-making in soil management and resource allocation. Despite these promising results, there are several avenues for future research. Expanding the dataset to include additional soil properties and diverse geographical regions would improve the generalizability of the models. Incorporating temporal data to capture seasonal variations in soil fertility could further enhance prediction accuracy. Moreover, exploring hybrid machine learning approaches, integrating deep learning, or leveraging explainable AI techniques could provide deeper insights into the key factors influencing soil fertility. These advancements will not only strengthen the utility of machine learning in agriculture but also contribute to sustainable farming practices and global food security.

Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Duhan, J.S., et al. (2017). Nanotechnology: The new perspective in precision agriculture. Biotechnology reports, 15, 11-23.
[2] Cambardella, C. and D. Karlen. (1999). Spatial analysis of soil fertility parameters. Precision agriculture, 1(1), 5-14.
[3] Javed, A., et al. (2022). Soil fertility: Factors affecting soil fertility, and biodiversity responsible for soil fertility. International Journal of Plant, Animal and Environmental Sciences, 12(1), 21-33.
[4] Bhardwaj, D., et al. (2014). Biofertilizers function as key player in sustainable agriculture by improving soil fertility, plant tolerance and crop productivity. Microbial cell factories, 13, 1-10.
[5] Nanni, M.R. and J.A.M. Demattê. (2006). Spectral reflectance methodology in comparison to traditional soil analysis. Soil science society of America journal, 70(2), 393-407.
[6] Demattê, J.A.M., et al. (2019). Soil analytical quality control by traditional and spectroscopy techniques: Constructing the future of a hybrid laboratory for low environmental impact. Geoderma, 337, 111-121.
[7] Poppiel, R.R., A.F. da Silveira Paiva, and J.A.M. Demattê. (2022). Bridging the gap between soil spectroscopy and traditional laboratory: Insights for routine implementation. Geoderma, 425, 116029.
[8] Bhat, S.A. and N.-F. Huang. (2021). Big data and ai revolution in precision agriculture: Survey and challenges. Ieee Access, 9, 110209-110222.
[9] Chakraborty, C., et al. (2024). From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare. Current Research in Biotechnology, 7, 100164.
[10] Ramírez, J.G.C., M.M. Islam, and A.I.H. Even. (2024). Machine Learning Applications in Healthcare: Current Trends and Future Prospects. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 1(1).
[11] Goodell, J.W., et al. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. Journal of Behavioral and Experimental Finance, 32, 100577.
[12] Zhong, S., et al. (2021). Machine learning: new ideas and tools in environmental science and engineering. Environmental science & technology, 55(19), 12741-12754.
[13] Rashid, M., et al. (2021). A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. IEEE access, 9, 63406-63439.
[14] Abdullah, W., C. Kim, and Y. Nam. (2024). A Fresh Look at Tomato Leaf Disease Recognition using Vision Transformers. Optimization in Agriculture, 1, 105-114.
[15] Bochenek, B. and Z. Ustrnul. (2022). Machine learning in weather prediction and climate analyses—applications and perspectives. Atmosphere, 13(2), 180.
[16] Hounkpatin, K.O., et al. (2022). Assessment of the soil fertility status in Benin (West Africa)–Digital soil mapping using machine learning. Geoderma Regional, 28, e00444.
[17] Coulibali, Z., A.N. Cambouris, and S.-É. Parent. (2020). Site-specific machine learning predictive fertilization models for potato crops in Eastern Canada. PloS one, 15(8), e0230888.
[18] Al Masmoudi, Y., et al. (2022). Use of machine learning in Moroccan soil fertility prediction as an alternative to laborious analyses. Modeling Earth Systems and Environment, 8(3), 3707-3717.
[19] Yadav, J., S. Chopra, and M. Vijayalakshmi. (2021). Soil analysis and crop fertility prediction using machine learning. Machine Learning, 8(03).
[20] Pant, J., et al. (2021). Soil quality prediction for determining soil fertility in Bhimtal Block of Uttarakhand (India) using machine learning. International Journal of Analysis and Applications, 19(1), 91-109.
[21] Chandra, H., et al. (2023). Explainable AI for soil fertility prediction. IEEE Access, 11, 97866-97878.
[22] Islam, A., et al. (2022). KNNOR: An oversampling technique for imbalanced datasets. Applied soft computing, 115, 108288.
[23] Singh, D. and B. Singh. (2022). Feature wise normalization: An effective way of normalizing data. Pattern Recognition, 122, 108307.

[24] Suthaharan, S. and S. Suthaharan. (2016). Decision tree learning. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, 237-269.
[25] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. Paper presented at the Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers.
[26] Cover, T. and P. Hart. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.
[27] Cortes, C. (1995). Support-Vector Networks. Machine Learning.
[28] Rumelhart, D.E., G.E. Hinton, and R.J. Williams. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.
[29] Raju, V.G., et al. (2020). Study the influence of normalization/transformation process on the accuracy of supervised classification. Paper presented at the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT).
[30] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.