

Prediction of Soil Health and Nutrient Deficiency in the Soil using Advanced Machine Learning Techniques

Varsha Ganesh Sonawane¹ and Dinesh D. Patil²

¹Department of Computer Science and Engineering, Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal, India

Email: varshasonawane.15@gmail.com

²Associate Professor and Head, Department of Computer Science and Engineering, Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal, India

Email: dineshonly@gmail.com

Abstract— Predicting soil health is essential for advancing precision agriculture and promoting sustainable land management practices. This study investigates the effectiveness of various machine learning approaches—such as Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), XGBoost, LightGBM, and hybrid models incorporating Voting Classifier and Stacking techniques in classifying soil health. Experimental findings reveal that Random Forest, the Hybrid Model, and the Stacking Model demonstrate superior performance compared to other classifiers, achieving higher accuracy in evaluating soil quality. Notably, the Stacking Model efficiently combines multiple algorithms to enhance prediction capabilities. The study highlights the importance of ensemble learning in soil classification, providing data-driven insights that can optimize agricultural productivity and support sustainable soil management strategies.

Index Terms— Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), XGBoost, LightGBM.

I. INTRODUCTION

Soil health is fundamental to agricultural productivity, environmental stability, and global food security. The physical, chemical, and biological properties of soil collectively determine its capacity to support plant growth, mitigate erosion, and sustain ecosystem services. However, contemporary agricultural practices—such as intensive farming, excessive use of synthetic inputs, and climate change—have exacerbated soil degradation, disrupting nutrient cycles and threatening long-term sustainability. Addressing these issues demands a holistic understanding of soil dynamics to enhance crop yields while promoting environmentally responsible land management.

Nutrient depletion occurs when essential elements like nitrogen (N), phosphorus (P), potassium (K), and micronutrients (e.g., iron, zinc) fall below levels necessary for optimal plant development. Monoculture farming, insufficient crop rotation, and inefficient fertilizer application contribute to this decline, resulting in lower agricultural productivity and weakened plant resilience. Additionally, nutrient-deficient soils experience reduced microbial activity, hindering natural biogeochemical cycles and further diminishing fertility.

Traditional soil nutrient evaluation relies on laboratory-based techniques such as soil spectroscopy, chemical extractions, and commercial test kits. While these methods provide valuable data, they suffer from several drawbacks, including prolonged processing times (often days to weeks), high operational costs due to specialized equipment, and the need for technical expertise. Moreover, their limited spatial resolution often fails to capture field-scale nutrient variability, restricting their practicality for precision agriculture.

To address these challenges, researchers are increasingly adopting machine learning (ML) for soil nutrient analysis. Unlike conventional methods, ML algorithms offer rapid, cost-effective predictions with real-time applicability. Their strength lies in detecting complex, non-linear relationships within soil datasets—patterns that may elude traditional analysis. By correlating diverse soil parameters with nutrient levels, ML models enable data-driven fertilization strategies, optimizing resource use and crop performance.

Despite its potential, ML-based nutrient deficiency prediction faces significant hurdles. Key limitations include sparse or inconsistent agricultural datasets, geographical and seasonal variability, and difficulties in modelling intricate soil-plant interactions. Additionally, algorithmic biases may arise from underrepresented extreme deficiency cases, necessitating robust data preprocessing and model validation. This main objective of this study is developing and evaluating two advanced ML frameworks for soil nutrient deficiency prediction: (1) a hybrid ensemble model integrating multiple algorithms to improve generalization, and (2) a stacked generalization model employing meta-learning for enhanced predictive accuracy. The primary goal is to surpass the performance of traditional ML approaches in nutrient diagnostics.

II. LITERATURE REVIEW

Adewopo et al. investigated machine learning (ML)-based spatial modelling techniques to assess soil organic carbon (SOC) distribution in Florida's Everglades. Their research underscores SOC's critical role in maintaining soil health and enhancing agricultural output, advocating for precise predictive frameworks to advance sustainable land-use practices [1]. Angelopoulou et al. Conducted a systematic evaluation of remote sensing (RS) methodologies for assessing soil organic carbon (SOC) content. Their comparative analysis examined various techniques including multispectral and hyper spectral imaging, thermal sensors, as well as LiDAR and radar systems. The findings indicate that hyper spectral imaging provides superior accuracy in SOC quantification owing to its enhanced spectral resolution, whereas multispectral approaches present a more economical solution for regional-scale monitoring applications. The study particularly emphasizes the synergistic combination of machine learning algorithms, such as Partial Least Squares Regression (PLSR) and Random Forest (RF), with RS data to significantly enhance SOC estimation precision [2]. Barman and Choudhury conducted a comparative analysis of machine learning and neural network methodologies for soil texture classification. Their study evaluated the performance of Random Forest (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN), with results demonstrating ANN's superior classification accuracy (94.3%). This performance advantage is attributed to ANN's enhanced capacity to capture nonlinear relationships among soil properties. The research incorporated key predictive variables including electrical conductivity measurements, soil moisture content, and spectral reflectance data in their modelling framework [3]. Chakraborty et al. (2021) systematically assessed various machine learning approaches for identifying soil nutrient deficiencies through hyperspectral imaging data. Their comparative study examined Partial Least Squares Regression (PLSR), Random Forest (RF), and Convolutional Neural Network (CNN) algorithms, with results indicating CNN's exceptional predictive capability ($R^2=0.89$) for estimating nitrogen, phosphorus, and potassium concentrations. The researchers attribute CNN's superior performance to its advanced spatial feature extraction from hyper spectral data, enabling more accurate nutrient deficiency detection compared to conventional machine learning methods [4]. Harumarajan et al. (2022) investigated the application of machine learning techniques for digital soil nutrient mapping in semi-arid tropical regions. Their research established electrical conductivity and Normalized Difference Vegetation Index (NDVI) as key indicators of soil fertility. While demonstrating machine learning's effectiveness for large-scale nutrient analysis, the study also revealed challenges related to data availability, particularly in understudied areas. The authors emphasize the need for localized model calibration to improve prediction accuracy in tropical agricultural systems, highlighting this approach as crucial for advancing precision nutrient management practices [5].

Gholami et al. (2021) conducted a comparative evaluation of machine learning algorithms for spatial soil moisture estimation using remote sensing data. Their study examined Random Forest (RF), Support Vector Regression (SVR), and Artificial Neural Network (ANN) approaches, with RF demonstrating superior predictive accuracy (88%). This performance advantage stemmed from RF's effective incorporation of multiple data sources including Landsat-8 spectral indices, topographic features, and climatic variables. The analysis identified

Normalized Difference Vegetation Index (NDVI) and Land Surface Temperature (LST) as the most significant predictors of soil moisture distribution. While confirming machine learning's potential to improve prediction quality, the authors note constraints related to the temporal resolution limitations inherent in satellite-derived datasets [6]. Gupta et al. (2022) performed a comprehensive meta-analysis of 120 international studies examining machine learning applications in soil health evaluation. Their research determined that Random Forest (RF) algorithms demonstrated particularly strong performance in modelling soil organic carbon content and nutrient availability. Interestingly, the study found climatic factors and land-use patterns to be more significant predictors of soil health outcomes than intrinsic soil characteristics. These findings highlight the importance of environmental context in developing accurate soil assessment models [7]. Jafari et al. (2020) found RF most accurate among three ML models (RF, Cubist, k-NN) for predicting SOC in arid zones, using terrain, climate and remote sensing data [8]. Kumar et al. (2021) demonstrated RF's effectiveness for Indian soil micronutrient mapping, with clay content and NDVI emerging as key predictors among various soil and spectral variables [9]. Li et al. (2022) proposed a novel one-dimensional convolutional neural network (1D-CNN) architecture for detecting soil macronutrient deficiencies (nitrogen, phosphorus, potassium) through hyperspectral reflectance data (400-2500 nm wavelength range). The developed model achieved exceptional classification accuracy (92.4%), surpassing conventional machine learning approaches including Support Vector Machines (SVM) and Random Forest (RF) by more effectively capturing subtle spectral patterns. This technological advancement contributes substantially to precision agriculture by facilitating rapid, non-invasive soil analysis. However, the authors suggest integrating multiple sensor modalities could further improve the system's reliability when applied to varied pedological conditions and agricultural environments [10]. Padarian et al. (2020) systematically reviewed machine learning implementations in pedological research through meta-analysis of 346 scientific studies. Their analysis revealed Random Forest as the dominant algorithm, selected primarily for its effectiveness in modelling nonlinear soil property relationships. The review documents successful machine learning deployments across three key domains: digital soil mapping, nutrient content prediction, and soil erosion modelling. However, the authors also identify several persistent challenges, including limited model interpretability, variable data quality concerns, and constraints in spatial model transferability across different geographical regions [11]. Poggio et al. (2021) developed innovative 30-meter resolution global maps of soil nutrient distributions across agricultural areas, filling critical gaps in existing soil databases. By integrating machine learning techniques with comprehensive soil sample analyses, their research generated accurate predictions of essential soil parameters including pH, organic carbon content, and nitrogen levels. The resulting spatial representations of soil fertility variations provide valuable insights for implementing precision agriculture techniques and promoting sustainable land use practices. Their publicly available dataset represents a significant contribution to global food security initiatives and environmental studies. While this pioneering work substantially progresses soil science through its detailed, scalable mapping approach – offering crucial tools for agricultural decision-making at multiple scales – the authors acknowledge inherent prediction uncertainties in data-scarce regions [12]. Seneviratne et al. (2022) developed a machine learning framework combining remote sensing data and environmental variables to detect soil nitrogen deficiency, offering a solution to a key limitation in precision agriculture implementation. Their comparative analysis of predictive algorithms demonstrated Random Forest's superior performance over Support Vector Machines when processing spectral and geospatial datasets. This approach establishes an innovative methodology for large-scale nitrogen monitoring that significantly decreases dependence on conventional laboratory testing while maintaining measurement accuracy. The research particularly emphasizes the economic and logistical advantages of remote sensing integration for comprehensive soil nutrient assessment [13]. Tsakiridis et al. (2020) introduced an innovative deep learning architecture for soil property prediction utilizing hyper spectral imaging technology. Their convolutional neural network (CNN)-based framework processes high-dimensional spectral reflectance data with superior performance compared to conventional machine learning approaches. Experimental results demonstrated significant improvements in predicting three critical soil parameters: organic carbon content, moisture levels, and texture classification. This advancement underscores deep learning's unique capacity to model intricate nonlinear relationships within spectral signatures, establishing a new paradigm for high-precision pedological analysis [14].

III. DATASET USED

This study utilizes a dataset containing 1,248 soil and plant nutrient samples, each characterized by 12 critical parameters: Nitrogen (N), Phosphorus (P), Potassium (K), pH, Electrical Conductivity (EC), Organic Carbon (OC), Sulphur (S), Zinc (Zn), Iron (Fe), Copper (Cu), Manganese (Mn), and Boron (B). The target variable,

labelled Output classifies nutrient deficiency into three categories: 0 (less fertile), 1 (fertile), and 2 (Highly fertile). The dataset captures a broad spectrum of agricultural conditions, with significant variations in nutrient concentrations—for example, N (6–138 ppm), P (4.4–122.8 ppm), and K (11–887 ppm). This dataset is highly suitable for training machine learning models to predict nutrient deficiencies, facilitating precision agriculture. By analysing soil health, nutrient interactions, and deficiency thresholds, it supports data-driven decisions for sustainable farming and optimized fertilizer application, ultimately enhancing agricultural productivity.

IV. PROPOSED METHODOLOGY

This study develops a hybrid machine learning model to predict soil nutrient deficiencies. Soil samples undergo pre-processing for missing values, outliers, and feature scaling. Five base models are employed: Logistic Regression (captures linear trends), Random Forest (handles non-linear relationships), SVM (manages high-dimensional data), XGBoost, and LightGBM (optimize gradient boosting). These models are fine-tuned via cross-validation. A Voting Classifier with soft voting combines these models, leveraging their probabilistic outputs to enhance prediction robustness. The ensemble method mitigates individual model weaknesses, improving generalization.

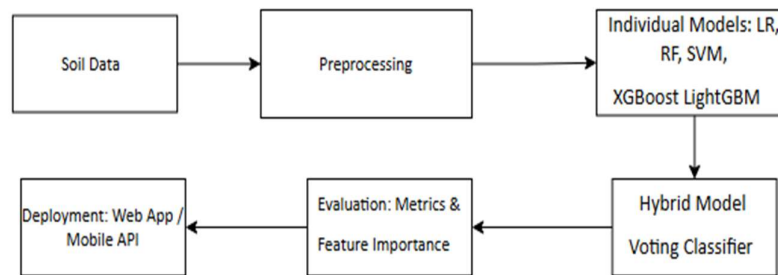


Fig. 1. Proposed Methodology

The above flowchart shows the proposed methodology for soil health and nutrient deficiency. Performance of model is assessed using accuracy, precision, recall, F1-score, and AUC-ROC, with comparisons to standalone models. Feature importance analysis identifies critical soil parameters, ensuring model interpretability.

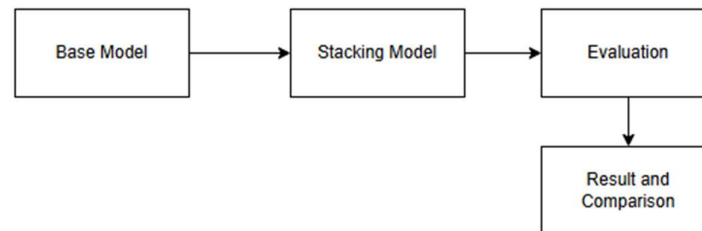


Fig. 2. Stacking Model

The above figure 02 outlines a comparative analysis employing a stacking ensemble model, an advanced machine learning technique designed to enhance predictive performance by integrating outputs from multiple base models. In this study, LR, RF, SVM, XGBoost, and LightGBM serve as the base models (Level-0). These models generate initial predictions, which are subsequently utilized as input features for the meta-model (Level-1). The meta-model, typically a higher-level algorithm such as logistic regression or gradient boosting, is trained to optimally combine the base predictions, thereby refining the final output. The primary objective of this stacked generalization approach is to achieve superior accuracy compared to individual standalone models by leveraging the collective strengths of diverse algorithms while mitigating their individual limitations.

V. RESULT AND DISCUSSION

The performance of five machine learning models—Logistic Regression, Random Forest, SVM, XGBoost, and LightGBM—was evaluated on a dataset with three classes 0, 1, and 2.

TABLE I. PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	Accuracy	Precision (Macro Avg)	Recall (Macro Avg)	F1-Score (Macro Avg)	Precision (Weighted Avg)	Recall (Weighted Avg)	F1-Score (Weighted Avg)
Logistic Regression	0.83	0.56	0.59	0.57	0.79	0.84	0.81
Random Forest	0.88	0.59	0.63	0.61	0.84	0.89	0.86
SVM	0.84	0.56	0.60	0.58	0.79	0.84	0.82
XGBoost	0.88	0.92	0.69	0.72	0.89	0.89	0.87
LightGBM	0.88	0.92	0.72	0.76	0.89	0.89	0.88

The above table demonstrates the comparative analysis for performance of machine learning models. Five machine learning model such as LR, SVM, RF, XGBoost, and LightGBM are tested on a multiclass imbalanced dataset. In this work, and the results show notable variations in their classification abilities. In comparison to SVM and Logistic Regression, the ensemble approaches of Random Forest, XGBoost, and LightGBM are performed better, with the greatest accuracy of 0.8864.

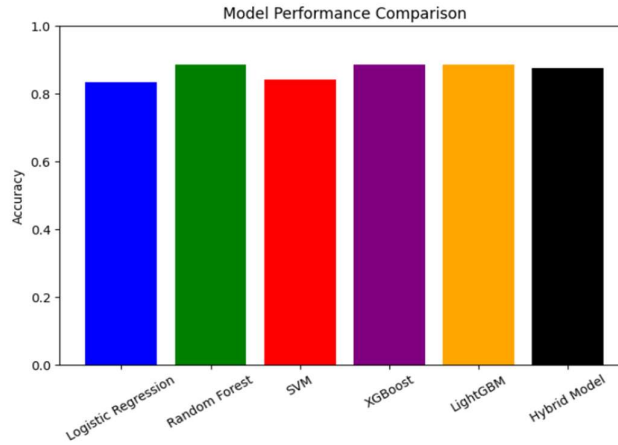


Fig. 3. Model Performance Comparison

The above bar graph illustrates a comparison of multiple machine learning models in terms of their predictive accuracy. The evaluated models consist of LR, RF, SVM, XGBoost, LightGBM, and a Hybrid Model. The results indicate that XGBoost and LightGBM achieve the highest accuracy scores, with the Hybrid Model performing slightly lower but still competitively. Random Forest also demonstrates robust performance, exceeding both SVM and LR in predictive capability.

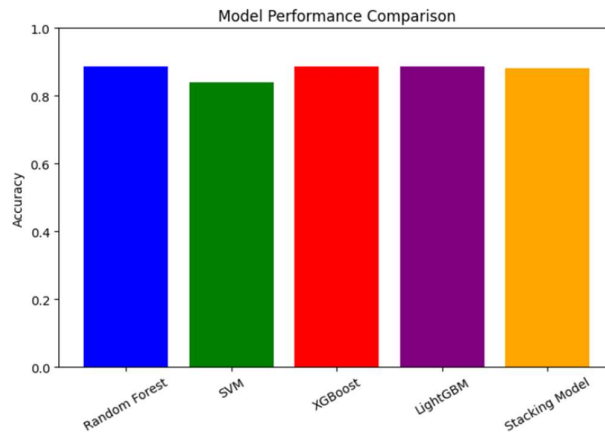


Fig. 4. Model Performance Comparison

A comparison of the prediction accuracy of various machine learning models is shown in the above bar graph. RF, SVM, XGBoost, LightGBM, and a Stacking Model are all used in the analysis. The results show that the

Stacking Model, LightGBM, and XGBoost attain the highest accuracy levels, whereas RF performs marginally worse. SVM, on the other hand, has comparatively less accuracy, suggesting that, for the dataset in question, its capacity for generalization may be less reliable than that of ensemble-based techniques.

VI. CONCLUSION

This research assesses the efficacy of machine learning models such as Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), XGBoost, LightGBM, and hybrid ensemble techniques (Voting Classifier and Stacking Model)—in predicting soil health. A comparative evaluation demonstrated that Random Forest, the Hybrid Model, and the Stacking Model delivered the highest accuracy, showcasing their effectiveness in analysing intricate soil health patterns. The Stacking Model improved predictive performance by intelligently combining multiple base classifiers. The results emphasize the value of ensemble learning for accurate soil health evaluation, offering data-driven solutions to enhance precision agriculture and sustainable land management through advanced soil quality classification.

REFERENCES

- [1] Adewopo, J. B., Van Zomeren, C., Bhomia, R. K., Almaraz, M., Bacon, A. R., Eggleston, E., ... & Osborne, T. Z. (2021). Machine learning approaches for spatial modeling of agricultural soil organic carbon in the Everglades, USA. *Science of The Total Environment*, 787, 147592.
- [2] Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., & Bochtis, D. (2020). Remote sensing techniques for soil organic carbon estimation: A review. *Remote Sensing*, 12(5), 803.
- [3] Barman, U., & Choudhury, R. D. (2020). Soil texture classification using machine learning and neural network approaches. *Measurement*, 152, 107292.
- [4] Chakraborty, S., Elarabawy, M., & Abd-Elrahman, A. (2021). Machine learning approaches for predicting soil nutrient deficiencies using hyperspectral imaging. *Computers and Electronics in Agriculture*, 180, 105900.
- [5] Dharumarajan, S., Lalitha, M., & Hegde, R. (2022) Digital mapping of soil nutrients using machine learning algorithms in semi-arid tropics. *Environmental Monitoring and Assessment*, 194(3), 1-17.
- [6] Gholami, V., Booij, M. J., & Tehrani, E. N. (2021) Spatial soil moisture prediction using machine learning and remote sensing data. *Journal of Hydrology*, 603, 126919.
- [7] Gupta, S., Hengl, T., Lehmann, P., Bonetti, S., & Or, D. (2022), Soil health assessment using machine learning: A global meta-analysis. *Geoderma*, 405, 115385.
- [8] Jafari, A., Khademi, H., Finke, P., & Van de Wauw, J. (2020) Spatial prediction of soil organic carbon using machine learning in arid regions. *Catena*, 185, 104287.
- [9] Kumar, N., Reddy, G. P. O., & Chatterji, S. (2021) Machine learning-based digital mapping of soil micronutrients in Indian agro-ecosystems. *Ecological Informatics*, 61, 101243.
- [10] Li, S., Wu, L., Chen, J., & Zhang, X. (2022) A deep learning approach for soil nutrient deficiency classification using hyperspectral imaging. *Computers and Electronics in Agriculture*, 193, 106697.
- [11] Padarian, J., Minasny, B., & McBratney, A. B. (2020) Machine learning and soil sciences: A review. *Geoderma*, 375, 114500.
- [12] Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021) Soil nutrient maps at 30 m resolution for global agricultural land. *Scientific Data*, 8(1), 1-11.
- [13] Seneviratne, M., Gunaratne, G. P., & Liyanage, L. P. K. (2022) Predicting soil nitrogen deficiency using machine learning and remote sensing data. *Environmental Science and Pollution Research*, 29(12), 17845-17858.
- [14] Tsakiridis, N. L., Keramaris, K. D., Theocharis, J. B., & Zalidis, G. C. (2020) A deep learning framework for predicting soil properties using hyper spectral imaging. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 4434-4449.
- [15] Zhang, Y., Hartemink, A. E., & Huang, J. (2022) Machine learning for digital soil mapping: Applications, challenges, and prospects. *Earth-Science Reviews*, 232, 104107.