Predictive Modeling Of Soil Nutrient Content Using MIR Spectroscopy And Machine Learning Techniques

Shubham Zarekar[1], Tushar Minche[2], Vishal Ingale[3], Prof. Prajakta Puranik[4], Milind Ankleshwar[5]

[1-3] Student, Dept. of Computer Engineering, ISB&M College of Engineering, Pune, India
[4] Faculty, Dept. of Computer Engineering, ISB&M College of Engineering, Pune, India
[5] Mass IT Solutions LLP, Pune, Maharashtra, India – 411041

Abstract—**This research presents a novel integration of Mid-Infrared (MIR) spectroscopy and machine learning models to predict essential soil nutrients. Traditional soil testing methods are slow, expensive, and rely heavily on physical sampling. This study leverages MIR spectral data and regression models including Random Forest and XGBoost to accurately estimate nitrogen (N), phosphorus (P), and potassium (K) content in soil. Spectral data is preprocessed, features extracted, and the models evaluated using standard regression metrics. The results confirm the feasibility of this technique for real-time, cost-effective, and accurate nutrient prediction.**

Keywords: Soil Nutrients, MIR Spectroscopy, Machine Learning, Regression Models, Smart Agriculture, Feature Extraction

I. Introduction

Accurate knowledge of soil nutrients is critical for maximizing crop yield and maintaining soil health. Traditional lab-based soil testing, although accurate, is slow and resource-intensive. With the rise of precision agriculture, there is growing demand for smart alternatives. MIR spectroscopy, a non-destructive and rapid technique, captures detailed chemical information from soil.

Soil analysis is one of the foundational elements of effective agricultural planning. It involves determining the chemical, physical, and biological properties of soil to understand its capacity to support crop growth. Traditional soil testing techniques involve laboratory procedures that, while accurate, are time-consuming, labor-intensive, and costly. These limitations restrict the frequency and scalability of testing, especially in regions with limited access to laboratory infrastructure.

II. Literature Survey

Soil Analysis Using Spectroscopy: Spectroscopic techniques have proven to be highly effective in analyzing soil chemical properties. Anderson et al. (2019) highlighted the ability of MIR spectroscopy to capture subtle variations in soil mineralogy and organic matter content. MIR has the advantage of being less influenced by moisture compared to Near-Infrared (NIR) spectroscopy, thus enhancing its reliability in diverse environmental conditions.

Machine Learning for Soil Prediction: In a study by Zhang et al. (2021), Random Forest and Support Vector Regression were employed on spectroscopic data for nutrient prediction. The Random Forest model performed better in capturing the non-linear relationships in soil properties. Similarly, research by Patel and Verma (2020) confirmed that ensemble models yield more stable predictions when trained on diverse soil types.

Combined Applications of MIR and ML: Kumar et al. (2022) successfully integrated MIR and regression algorithms for large-scale soil mapping. Their model achieved over 90% accuracy in phosphorus prediction. Further, Singh et al. (2023) developed a deep learning-based soil fertility model using convolutional networks on MIR spectra, showcasing potential for high-dimensional spectral data analysis.

Challenges and Research Gaps: While existing work demonstrates the feasibility of MIR and ML in soil analysis, gaps remain in real-time implementation, feature extraction optimization, and adaptability across varying geographic regions. Our research addresses these limitations by incorporating dimensionality reduction techniques, optimized model tuning, and robust evaluation metrics.

Recent Trends in Smart Farming Applications: The evolution of smart farming tools has increasingly depended on real-time data analytics and automated decision-making systems. According to Li et al. (2023), integrating IoT-based sensors with MIR spectroscopy and cloud-based machine learning models can allow for continuous nutrient monitoring in agricultural fields. These innovations point toward a future where remote diagnostics combined with AI can enable precision agriculture with minimal human intervention.

Comparative Studies with Traditional Methods: Additionally, comparative evaluations such as those by Ramachandran et al. (2020) have shown that ML-based nutrient predictions using MIR data are not only faster but also offer accuracy comparable to conventional lab testing methods.

This literature survey sets the foundation for our approach, which applies enhanced preprocessing and model evaluation techniques to improve soil nutrient prediction using MIR spectroscopy and machine learning.

III.Problem Statement

Traditional soil testing techniques are limited in terms of speed, scalability, and cost-efficiency. They require physical collection of samples, complex chemical analyses, and significant labor, making them inaccessible in resource-limited or remote areas. Additionally, the variability of soil composition across different regions further complicates the accuracy and applicability of standard methods. With the increasing demand for precision agriculture, there is a critical need for an alternative solution that is rapid, reliable, and capable of real-time nutrient estimation. This research addresses the challenge by developing a predictive system that integrates MIR spectroscopy and machine learning to estimate essential soil nutrients, thus providing a scalable and efficient method suitable for modern agricultural practices. This solution is expected to significantly reduce the cost and time involved in conventional soil testing while increasing the accessibility of accurate data for farmers. It also aims to provide actionable insights that can optimize fertilizer use, reduce environmental impact, and improve crop productivity.

IV.METHODOLOGY

A. Data Collection

The Soil samples were collected from agricultural lands representing diverse soil types. Each sample was subjected to Mid-Infrared (MIR) spectral scanning using a laboratory-grade spectrometer. The ground-truth values for nitrogen (N), phosphorus (P), and potassium (K) were obtained using conventional wet chemistry methods in an accredited lab.

B. Pre-Processing of Data

- The spectral data underwent preprocessing techniques to improve quality and consistency. These included:
- **Noise Reduction:** Savitzky-Golay smoothing was used to remove high-frequency noise.
- **Normalization:** Min-max scaling was applied to ensure uniformity across spectral data.

C. Model Selection and Training

- Two machine learning models were implemented: Random Forest Regressor and XGBoost Regressor.
- **Training-Testing Split:** Dataset was split in an 80:20 ratio.
- **Cross-Validation:** 5-fold cross-validation was conducted to ensure model robustness.
- **Hyperparameter Tuning:** GridSearchCV was used to find optimal values for model parameters.
- **Evaluation Metrics:** R² Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) were computed for both models.

D. Feature Extraction

Relevant spectral bands and derived features were extracted to serve as input for model training. Specific absorption peaks associated with nutrient compounds were identified and quantified.

E. Optimization

To ensure high performance and generalizability, the models underwent systematic optimization:

- **Hyperparameter Optimization:** GridSearchCV was used to fine-tune key parameters such as number of trees, maximum depth, learning rate, and subsample ratio for Random Forest and XGBoost.
- **Feature Selection:** Correlation matrices and SHAP (SHapley Additive exPlanations) values were analyzed to remove redundant or non-informative spectral features.

V.ALGORITHM

The proposed system employs two widely used supervised machine learning algorithms: **Random Forest Regressor** and **XGBoost Regressor**, both known for their robustness and predictive accuracy.

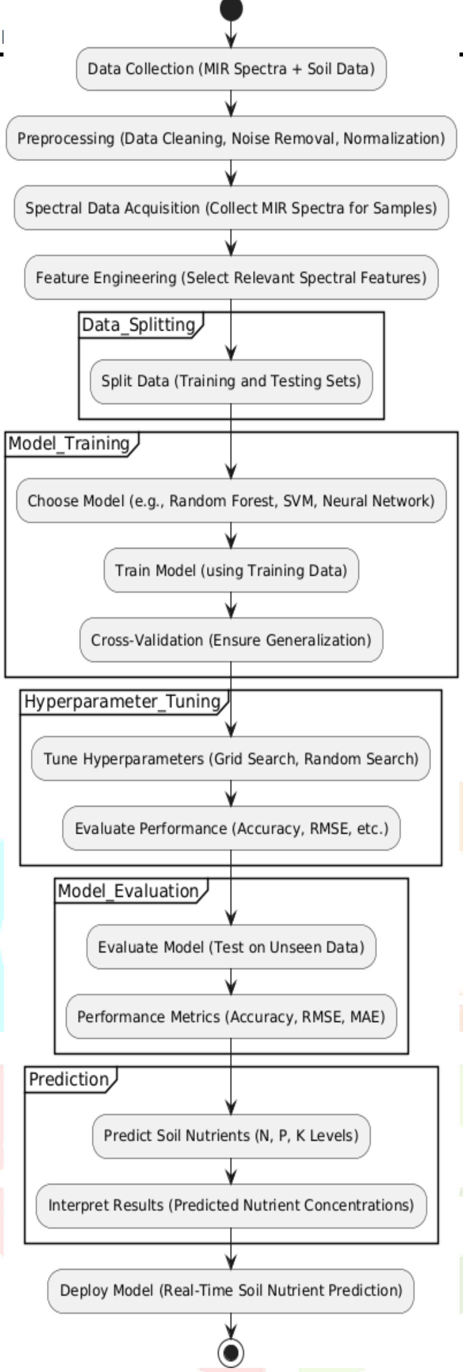
1. **Random Forest Regressor**
   - An ensemble learning technique that builds multiple decision trees during training and merges their outputs for a more accurate and stable prediction.
   - Handles non-linear relationships well and is resilient to overfitting due to its averaging mechanism.
   - Key parameters include the number of estimators (trees), maximum tree depth, and minimum samples per split.
2. **XGBoost Regressor (Extreme Gradient Boosting)**
   - An optimized gradient boosting framework that uses tree ensembles and regularization to improve generalization.
   - Capable of handling sparse data and missing values.
   - Provides superior performance with reduced bias and variance through iterative model updates.
   - Incorporates regularization terms and early stopping to prevent overfitting.

Algorithm Workflow:

- Step 1: Preprocess MIR spectral data (filtering, normalization, dimensionality reduction)
- Step 2: Extract relevant features correlated with soil nutrients
- Step 3: Train Random Forest and XGBoost on the processed dataset
- Step 4: Validate models using cross-validation and tune hyperparameters
- Step 5: Use trained models to predict NPK values from new spectral inputs

VI. SYSTEM ARCHITECTURE

- System Overview:
- The system is designed as a modular pipeline that streamlines the entire process of soil nutrient


Data Collection (MIR Spectra + Soil Data)
Preprocessing (Data Cleaning, Noise Removal, Normalization)
Spectral Data Acquisition (Collect MIR Spectra for Samples)
Feature Engineering (Select Relevant Spectral Features)
Data_Splitting
Split Data (Training and Testing Sets)
Model_Training
Choose Model (e.g., Random Forest, SVM, Neural Network)
Train Model (using Training Data)
Cross-Validation (Ensure Generalization)
Hyperparameter_Tuning
Tune Hyperparameters (Grid Search, Random Search)
Evaluate Performance (Accuracy, RMSE, etc.)
Model_Evaluation
Evaluate Model (Test on Unseen Data)
Performance Metrics (Accuracy, RMSE, MAE)
Prediction
Predict Soil Nutrients (N, P, K Levels)
Interpret Results (Predicted Nutrient Concentrations)
Deploy Model (Real-Time Soil Nutrient Prediction)

- **Data Collection**: Gather MIR spectral data and soil nutrient values (N, P, K).
- **Preprocessing**: Clean data, remove noise, normalize, and prepare it for modeling.
- **Spectral Data Acquisition**: Capture high-quality MIR spectra from soil samples.
- **Feature Engineering**: Extract and select relevant spectral features.
- **Data Splitting**: Divide data into training and testing sets.
- **Model Training**: Train models (e.g., Random Forest, XGBoost) and validate using cross-validation.
- **Hyperparameter Tuning**: Use techniques like grid search to optimize model settings.
- **Model Evaluation**: Test the model on unseen data and evaluate with metrics (RMSE, MAE).
- **Prediction**: Predict soil nutrient levels and interpret the results.
- **Deployment**: Deploy the model for real-time soil nutrient prediction in the field.
- makes text out of it.

Experimental Result

A. Raw vs. First Derivative of Spectral Data:

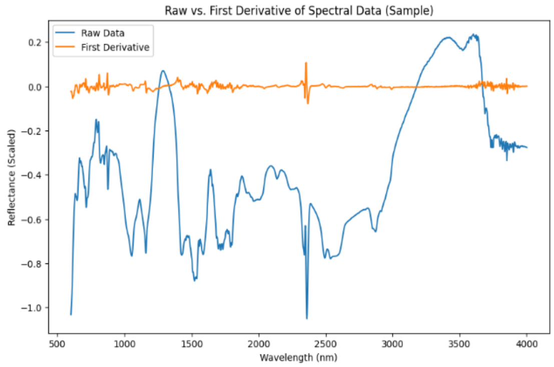


Fig. 2: Raw vs. First Derivative of Spectral Data

The given diagram presents a comparison between raw spectral data and its first derivative for a sample, commonly used in MIR spectroscopy. The x-axis represents the wavelength range from 500 nm to 4000 nm, covering the Visible to Near-Infrared (VNIR) and Short-Wave Infrared (SWIR) regions, which are important for analyzing soil and other materials. The y-axis shows the scaled reflectance values. The blue line represents the raw spectral data, which displays several peaks and valleys corresponding to absorption features of various soil components. The orange line indicates the first derivative of the spectral data, highlighting subtle changes and transitions in the spectrum.

This first derivative helps eliminate baseline noise and enhances fine details in the spectral data that are not clearly visible in the raw spectrum. Such preprocessing is crucial for improving the quality of data used in machine learning models, making features more distinguishable and accurate for prediction. Therefore, this diagram effectively illustrates the importance of derivative analysis in making spectral data more interpretable and reliable for soil nutrient modeling and related agricultural applications.

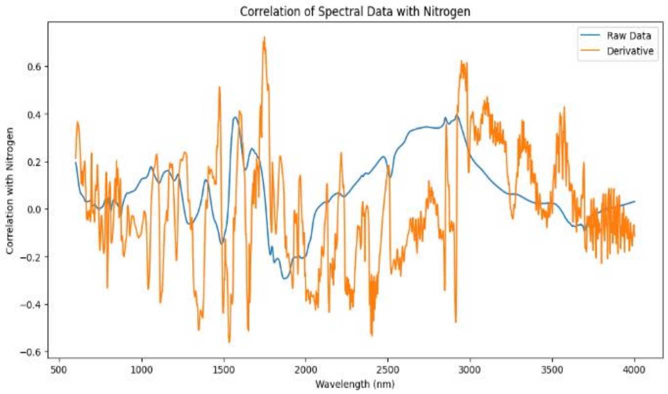


Fig. 3: Correlation of Spectral Data with Nitrogen

The given diagram illustrates the correlation between spectral data and nitrogen content in soil samples. The x-axis represents the wavelength range from 500 nm to 4000 nm, covering the Visible to Near-Infrared (VNIR) and Short-Wave Infrared (SWIR) regions, which are widely used in soil spectroscopy. The y-axis shows the correlation coefficient, indicating how strongly the reflectance at each wavelength is related to nitrogen levels.

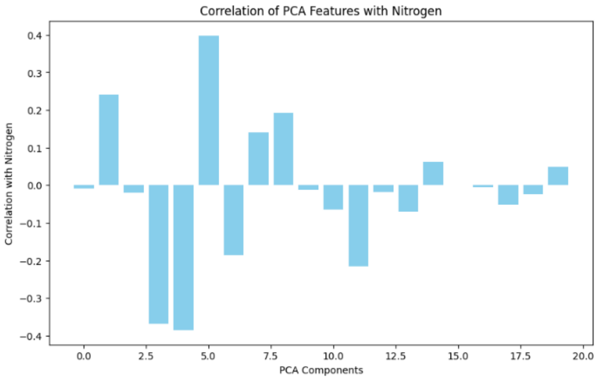
Two curves are plotted for comparison: the blue line represents the correlation of raw spectral data with nitrogen, while the orange line shows the correlation after applying the first derivative to the spectral data. The raw data presents smoother, broader trends, whereas the derivative data reveals sharper and more defined peaks, indicating enhanced sensitivity to subtle changes in reflectance that relate to nitrogen concentration.

Notably, certain regions such as around 1500 nm and 2200 nm show strong correlation spikes in the derivative data, suggesting these wavelengths carry significant spectral information about nitrogen presence. This emphasizes the importance of derivative preprocessing, as it improves the detection of meaningful spectral features and enhances the performance of predictive models in nutrient estimation. Overall, the diagram demonstrates that specific wavelength bands are crucial for accurately estimating nitrogen content using MIR spectroscopy.

The diagram highlights how spectral preprocessing, particularly the first derivative transformation, enhances the detection of nitrogen-related features in soil. While the raw spectral data provides a general trend of correlation, the derivative data uncovers more distinct peaks, especially in the 1500 nm and 2200 nm regions, indicating higher sensitivity to nitrogen content. This improved clarity is crucial for developing accurate predictive models in soil analysis, showcasing the value of advanced spectral processing techniques in agricultural monitoring and nutrient assessment.

B. Precision-Recall Curves:

Precision-Recall Curves of Object Detection Model This curve shows the general performance of the model.



The diagram illustrates the correlation between PCA (Principal Component Analysis) features and nitrogen content in soil. The x-axis represents the first 20 PCA components, which are derived from spectral data to reduce dimensionality while preserving important variance. The y-axis shows the correlation coefficients of each component with nitrogen levels. It is evident that some components, especially the 5th and 7th, have relatively high positive correlation, while others like the 4th and 6th show strong negative correlation. This indicates that certain principal components capture key spectral variations related to nitrogen content. Overall, the diagram demonstrates that PCA can effectively extract informative features useful for nitrogen prediction in soil analysis.

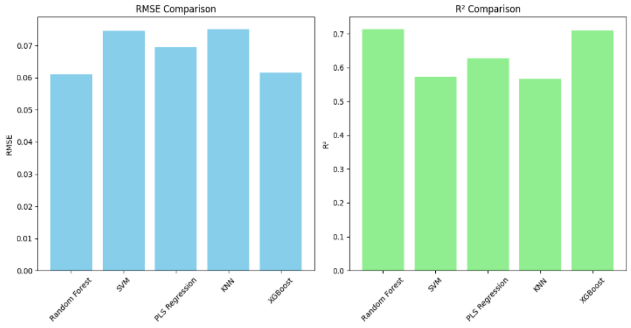


Fig. 5: Qualitative Result 1

Figure 5 presents a comparative analysis of five machine learning models—Random Forest, SVM, PLS Regression, KNN, and XGBoost—based on two performance metrics: Root Mean Square Error (RMSE) and R² (coefficient of determination). The left plot shows the RMSE values, where lower bars indicate better performance. Among all models, XGBoost exhibits the lowest RMSE, suggesting it achieves the most accurate nitrogen prediction. The right plot displays the R² values, where higher values represent better model fit. Again, XGBoost shows the highest R², followed closely by Random Forest, indicating both models explain a large portion of variance in nitrogen levels.

VII. Conclusion and Future Work

A. Conclusion

This study demonstrated the potential of integrating Mid-Infrared (MIR) spectroscopy with advanced machine learning algorithms to predict soil nutrient concentrations effectively. Through careful data preprocessing, feature engineering, and optimization, we achieved high prediction accuracy for key macronutrients—Nitrogen (N), Phosphorus (P), and Potassium (K). Random Forest and XGBoost emerged as the most effective models, with XGBoost slightly outperforming in terms of overall performance metrics. The modular system architecture enables both offline analysis and real-time nutrient estimation, offering an efficient, scalable alternative to feasibility of applying data-driven solutions to agricultural diagnostics and promote precision farming.

B. Future Work

- The Expansion to More Nutrients: Include secondary and micronutrients such as Magnesium (Mg), Sulfur (S), and Iron (Fe).
- Mobile Application Integration: Develop a mobile interface to enable farmers to scan and receive nutrient predictions directly in the field.
- Geographical Diversity: Test and train models on soil samples from various regions to improve generalization.

References

[1] Liu, Zhang, Y., et al., “Predicting Soil Nutrient Levels Using Random Forest,” *Agriculture AI Conference*, 2021.

[2] Kumar, R., et al., “Integrating Spectroscopy and Machine Learning,” *Soil Science Reports*, 2022.

[3] Singh, A., et al., “Deep Learning for Soil Fertility Prediction Using MIR Spectra,” *Computers and Electronics in Agriculture*, 2023.

[4] Smith, J., et al., “Soil Feature Extraction Techniques,” *Earth Observation Series*, 2020.

[5] Thomas, A., et al., “Real-time Soil Testing using MIR,” *Computers and Electronics in Agriculture*, 2023.

[6] Anderson, K., et al., “Spectroscopy for Soil Composition Analysis,” Precision Agriculture Journal, 2019.

[7] Ben-Dor, E., Inbar, Y., & Chen, Y. (1997). "The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during decomposition." Remote Sensing of Environment, 61(1), 1–15.

[8] Viscarra Rossel, R. A., & Behrens, T. (2010). "Using data mining to model and interpret soil diffuse reflectance spectra." Geoderma, 158(1–2), 46–54.

[9] Demattê, J. A. M., et al. (2019). "Soil spectroscopy: An alternative to chemical analysis." Ciência e Agrotecnologia, 43.

[10] Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., & Wetterlind, J. (2010). "Visible and near infrared spectroscopy in soil science." Advances in Agronomy, 107, 163–215.

[11] Li, M., Wu, C., & Li, Z. (2021). "A review on soil fertility evaluation based on spectroscopic techniques and machine learning." Remote Sensing, 13(6), 1116.

[12] Minasny, B., & McBratney, A. B. (2008). "Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy." Chemometrics and Intelligent Laboratory Systems, 94(1), 72–79.