INTERPRETABLE AI DRIVEN DEEP LEARNING FOR SOIL FERTILITY CLASSIFICATION

T S Anushya Devi[1], V Vadivu[2]

Abstract

Predicting the appropriate crops for a particular soil category in a specific region is one of the most important issues in precision farming. To solve this problem, different algorithms in machine learning, such as Support Vector Machines (SVM), Decision Tree models, and Random Forest ensembles - have been developed in recent years, which learn the historical data about various soil properties of multiple crops in a particular region to forecast yield productivity. Traditional machine learning methods often struggle with high-dimensional tabular soil data and lack interpretability. This study introduces an interpretable AI driven deep learning framework tailored for tabular datasets to classify the chemical attributes of soil. These attributes include essential macronutrients—nitrogen (N), phosphorus (P), and potassium (K) as well as soil pH, which indicates its acidity or alkalinity. Unlike conventional black-box models, the proposed framework emphasizes interpretability, enabling researchers and farmers to understand the contribution of individual features to the classification process. By harnessing artificial intelligence, the approach delivers real-time, data-informed insights that not only improve the accuracy of soil classification but also contribute to long-term agricultural sustainability through better nutrient management and decision-making.

The objective is to highlights the potential of integrating advanced AI techniques like TabNet in modern agricultural decision-making systems to emphasize the limitations of existing methods and propose innovative recommendations for better soil classification. This work involves preprocessing real soil datasets, training the TabNet model using PyTorch, and evaluating its accuracy and feature importance.

Keywords : Soil classification; Crop yield prediction; Agricultural productivity; Machine learning techniques; Artificial intelligence applications; Deep neural networks; TabNet; PyTorch framework

Department of BCA[1]
PSGR Krishnammal College for Women, Coimbatore, India[1]
tsanushya82@gmail.com[1]

Department of Artificial Intelligence and Data Science[2]
Karpagam Academy of Higher Education, Coimbatore. India[2]
vbkkaviraksha@gmail.com[2]

* Corresponding Author

I. INTRODUCTION

Agriculture is a critical activity for meeting the nation's food demands. Specifically, the application of predictive analytics in soil-related agricultural practices can enhance decision-making and increase crop productivity more efficiently. Classifying or analyzing soils is crucial for making decisions about a variety of agricultural-related matters.

Agricultural productivity heavily depends on accurate knowledge of soil chemical properties. Traditional soil classification methods are often manual and time-consuming approaches grounded in machine learning have displayed potential when it comes to automating this procedure; however, a large number of them are limited by their subpar capacity for interpretation. TabNet, introduced by Arik and Pfister, addresses this challenge by providing attention-based interpretability for tabular data. This paper explores TabNet is suitability for classifying soil chemical parameters and generating insights for fertilizer recommendations. [12]

II. REVIEW OF LITERATURE

A framework was designed for predicting rice and wheat yields depending on the ecological distance scheme. Initially, climate factors, soil factors, and behavioral factors were collected as crop yield impact factors. Subsequently, these elements were integrated with the comprehensive sensitivity index to establish the crop yield predictor. After, the ecological distance scheme was fused with the crop yield predictors for creating the yield prediction framework. but its computational cost and time were high for larger amounts of factors. [1]

A data mining based crop yield prediction system was presented. Initially, soil conditions and atmospheric factors were collected as a database. Then, the database was pre-processed and the most relevant attributes were chosen by the Recursive Feature Elimination (RFE) scheme. Moreover, those attributes were classified by the bagging classifier to predict the suitable crop. But it was computationally intensive and more complex to interpret while using a large-scale database. [2]

A new parametric modeling scheme depending on a

Non-linear Finite Impulse Response (NFIR) method to predict the crop yield from various soil properties. The NFIR was applied to recognize and quantify the correlation between soil characteristics and crop yield. But it needs more soil physical properties and crop-related attributes to increase the prediction performance. [3]

A Neural Network based on Geographically and Temporally Weighted regression (GTWNN) to forecast winter wheat yields. However, an error arose because it failed to consider numerous factors related to crop yield, such as soil characteristics, among others. [4]

Predictive modeling techniques to predict oil palm harvest. Initially, soil moisture, weather, and fresh fruit bunch yield data were collected and pre-processed to eliminate redundant data. Then Then, tree-based ensemble methods were trained for estimating oil palm yield; however, they struggled to capture complex patterns across multisource meteorological and agricultural datasets. [5]

A new spatiotemporal hybrid model called (DRS-RF) for predicting tea yield using remotely sensed hydro-meteorological data. A Dragonfly optimization algorithm and Support vector regression (DRS) was applied to choose the significant characteristics from every variable from the corresponding stations, whereas the RF was used to exploit the predictors' data for predicting the tea yield. But to increase prediction accuracy, supplementary characteristics and advanced deep learning approaches are necessary. [6]

The Aqua Crop framework to produce the Fractional Vegetation Cover (FVC) maize harvest. A reliable FVC optimization database was created depending on an ensemble Kalman filtering adjustment scheme. The RF algorithm identified a regression link between FVC and harvest from long-term time-series records; nevertheless, the use of advanced models is essential to enhance the representation of FVC attributes in harvest prediction. [7]

The use of classifiers for soil parameter analysis applied ML models like Decision Trees, SVMs, and Random Forests for soil classification. Nevertheless, the majority of these models lack interpretability, a crucial aspect for practical agricultural applications. [8]

Sophisticated data-driven methodologies are advancing swiftly, presenting numerous potential applications within the field of soil science. Nevertheless, the role of human expertise and perception is crucial in identifying soil characteristics, particularly the qualitative elements that may be missed by sensing technologies or computer-based models. The solution to this challenge lies in the combination of computer-assisted predictive modeling with human understanding and expertise. [9]

A study aims to predict soil fertility using machine learning classifiers, with a comparative analysis of algorithms such as J48, Random Forest, Decision Table, PART, and Naïve Bayes Soil fertility data from 34 villages in Anand District, Gujarat, was analyzed using the WEKA tool. The dataset included attributes like pH, EC, OC, N, P2O5, K2O, and Fertility Index. Min-Max normalization was applied to preprocess the data. However, this relatively small dataset may limit the generalizability of the findings to other regions or larger datasets. [10]

A hybrid model that applies fuzzy logic with neural approaches to better handle ambiguity and noise in agricultural datasets. This improves the reliability of predictive tasks, including crop classification, yield estimation, fertilizer planning, and pest control. Despite these advantages, the authors noted that the framework remains complex, with overlapping and ambiguous features still presenting challenges. [11]

III. EXISTING WORK

The study 'Soil Quality Prediction for Determining Soil Fertility in Bhimtal Block of Uttarakhand', which aimed to improve regional soil fertility prediction through the use of advanced and interpretable deep learning techniques, his work aims to refine current practices by resolving the gaps in traditional methods.

The goal was to forecast different categories of soil fertility using data from soil measurements, helping both farmers and government officials make smart choices.

Random Forest, a supervised ensemble method, was applied to soil samples obtained from the Bhimtal block, Nainital district, Uttarakhand, India. Soil samples were examined based on various physicochemical parameters: pH, electrical conductivity (EC), and organic carbon (OC), in addition to both macronutrients (N, P, K) and trace elements (Zn, Fe, Mn, Cu). [12]

Future work proposed extending the study to larger regions, The outcomes are contrasted against those produced by deep learning techniques, specifically TabNet and Convolutional Neural Networks (CNN). integrating real-time soil sensor data, and developing mobile-based advisory tools for farmers.

TabNet, a cutting-edge architectural framework for deep learning on tabular datasets, notable for its superior performance and inherent interpretability. TabNet integrates deep learning model developed for organized datasets, providing both robust prediction performance and insights into feature contributions, ensuring transparency in decision-making. Consequently, TabNet holds significant value in

sectors where comprehending model decisions is essential, including finance and healthcare. Moreover, its capacity to manage missing data and its resilience against overfitting further increase its attractiveness. By merging the advantages of deep learning with the requirements of tabular data analysis, TabNet signifies a notable progression in the domain, connecting traditional machine learning approaches with contemporary deep learning methodologies. [13]

IV. PROBLEM DEFINITION

The problem considered for the research work are listed as follows:

- The reliable soil classification was not effective for various numbers of predictions. [13]
- When number of instances was very high, the parameter selection for each classifier was complex that leads the maintainability issue.
- If the size of training data set was increased, the memory cost and resource consumption were also increased.
- The regularization performance was poor and the class similarity was not ensured to increase the performance since it discards important factors such as heterogeneous difficulty.
- Conventional soil fertility classification methods are often limited by low accuracy, lack of interpretability, and inefficiency in handling high-dimensional structured agricultural data. [12]

V. RESEARCH GOALS

The following lists the goals of this research project:

- To make more reliable and feasible soil classification for various numbers of predictions.
- To improve the classification ability and select the parameters for such classifiers with various number of instances.
- To reduce the memory cost and resource consumption with increasing the size of training data.

VI. RESEARCH METHODOLOGY

- **Soil Data Collection**: Raw soil chemical data is collected from samples.
- **Data Preprocessing with AI**: AI techniques handle missing data, correct errors, and normalize values.
- **Feature Selection by AI**: AI models automatically pick the most important soil parameters.
- **Deep Learning Model Building**: Construct neural networks suited to the data.
- **Model Training**: AI learns patterns from training data.
- **Model Evaluation**: AI tests model accuracy and fine-tunes it.
- **Prediction/Classification**: AI classifies soil types or fertilizer needs on new samples.
- **Continuous Feedback Loop**: New data and feedback help AI improve over time.

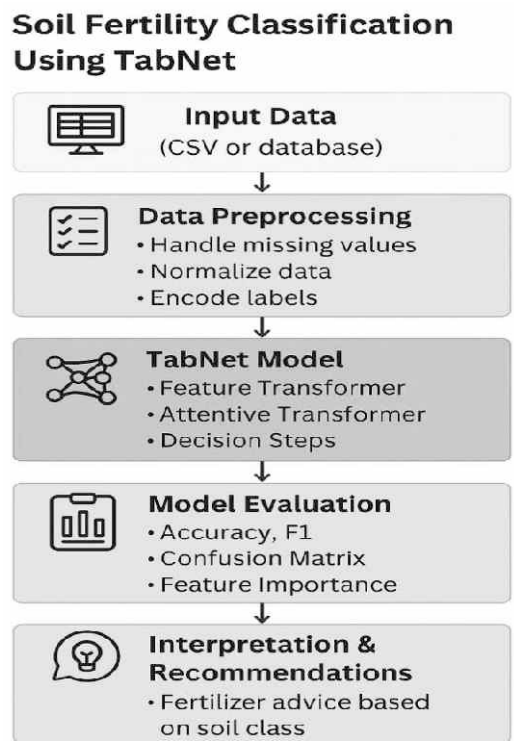

Figure 1: Flow Diagram

A. Architecture of TABNET

The TabNet structure essentially involves several sequential steps, routing the input from one step to the next. The approach to determining the number of steps varies based on the model's capacity. Each step consists of the following components:

- Initially, the entire dataset is fed into the model without any feature modifications. It then undergoes batch normalization prior to being passed through a feature transformer.
- **Feature Transformer:** A defined number (n) of distinct GLU blocks (for instance, 4) are incorporated, and each block includes the following layers

A GLU block consists of the subsequent arrangement of layers:

1. Fully Connected
2. Batch Normalization
3. GLU (Gated Linear Unit)

The GLU activation is defined as:

$GLU(x)=\sigma(x).x$

where σ(x)\sigma(x)σ(x) denotes the sigmoid activation function, which acts as a gate to control the information flow.

In a configuration consisting of four GLU blocks, two blocks are shared while the remaining two operate independently, thereby enhancing both the efficiency and robustness of learning. Additionally, there is a skip connection present between each pair of consecutive blocks. Following each block, we apply normalization using 0.5 and 0.5 to maintain stability and ensure that the variance remains relatively constant. The feature transformer produces two outputs:

1. nd (decision output): This represents the prediction from the specific step, providing continuous values or discrete classes.
2. na (attentive output): It acts as the input for the following attentive transformer, thereby marking the beginning of the next cycle.

- **Attentive Transformer:**

An attentive transformer is organized as a sequence of the following components:

The attentive transformer architecture is structured with an FC layer, then batch normalization, a prior scale layer, and finally a sparsemax layer It takes na as input, which is initially handled by the FC layer, then normalized via the Batch Normalization layer, and subsequently enhanced by the Prior Scales layer before it is processed through the Sparsemax activation.

The layer from the preceding scale consolidates the frequency of utilization for each feature prior to the current decision-making process.

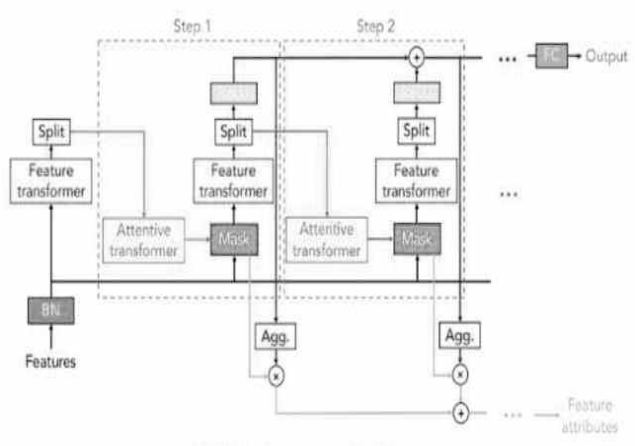

Figure 2: Architecture of TABNET

VII. RESULTS AND DISCUSSIONS

To classify soil types based on physicochemical properties, three conventional machine learning techniques: SVM, RF, and C4.5 Decision Tree were applied. Following preprocessing, the final dataset comprised ten soil samples characterized by forty-five relevant features. The evaluation of soil classification models was carried out using standard performance indicators. These included accuracy for overall correctness, precision for relevance of predictions, recall for sensitivity, and the F1-score as a balanced measure.

A. Model Performance

For robust evaluation on the small dataset, the models were assessed using both Leave-One-Out and 5-fold cross-validation approaches. Their performance was quantified with widely used metrics, including accuracy, precision, recall, and the F1-score.

B. Performance Metrics And Comparision With Confusion Matrix:

Evaluation results indicated that Random Forest yielded the most accurate predictions, with an accuracy of 40%. In contrast, C4.5 exhibited the weakest performance, reflecting its sensitivity to overfitting when trained on a small number of soil samples.

Overall, the models struggled with low accuracy, mainly due to limited sample size and high dimensionality.

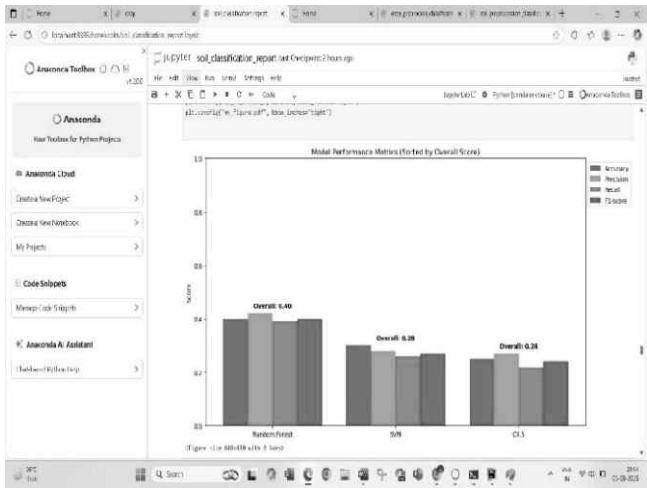

Figure 3: Performance Metrics

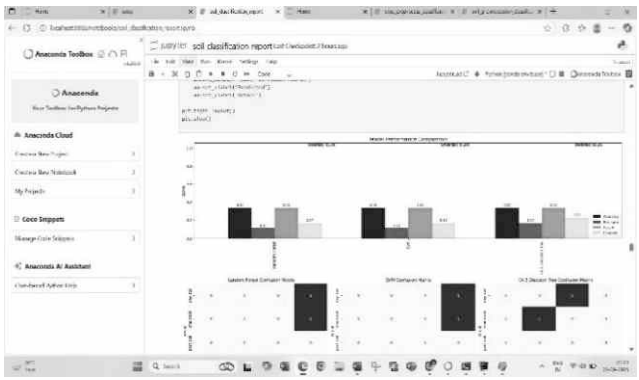

Figure 4: Confusion Matrix

C. Prediction Analysis

The predicted soil types were compared with actual labels to identify misclassification trends.

- Random Forest correctly predicted most silt and peat soils.
- SVM confused clay soil with chalk soil, indicating feature overlap.
- C4.5 produced frequent misclassifications due to the low sample-to-feature ratio.

While traditional machine learning approaches provided initial results, this study proposes leveraging TabNet to achieve improved results in future investigations.

TabNet represents a deep learning architecture tailored for structured data, utilizing sequential attention mechanisms to pinpoint the most relevant characteristics during each step of the decision-making process. The future of this research will involve implementing PyTorch TabNet on an expanded soil dataset, with the objective of enhancing accuracy, interpretability, and the reliability of recommendations for fertilizer and crop selection systems.

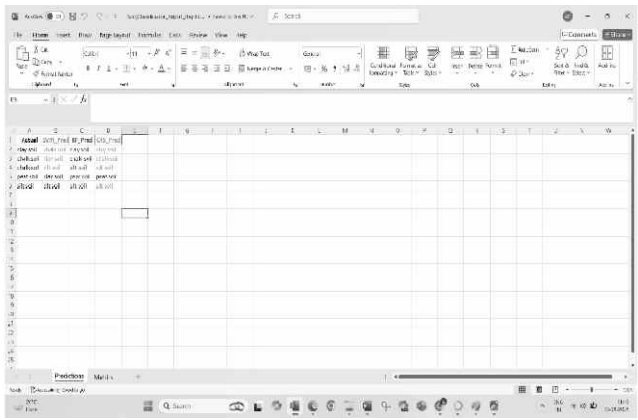

Figure 5: Algorithm Analysis

VIII. CONCLUSION

As a critical sector of the Indian economy, agriculture is highly susceptible to the impacts of changing climatic conditions. Key challenges include insufficient nutrient levels, suboptimal farming practices, and limited water supply. Proper fertilizer application and maintaining suitable soil properties can help mitigate these issues. The present study aims to support decision-making to enhance agricultural productivity and soil quality by leveraging TabNet, an interpretable deep learning framework implemented using PyTorch. TabNet is designed to efficiently apply deep learning techniques to structured datasets, which remain prevalent in many agricultural and data-driven applications. The outcomes of this research could inform management strategies for soil fertility, fertilizer usage, and crop yield optimization. Furthermore, this approach may be extended in future studies to improve nitrogen levels and overall soil fertility at the village level.

REFERENCES

[1]. Tian, L., Wang, C., Li, H., & Sun, H. (2020). Yield prediction model of rice and wheat crops based on ecological distance algorithm. Environmental Technology & Innovation, 20, 1-12.

[2]. Suruliandi, A., Mariammal, G., & Raja, S. P. (2021). Crop prediction based on soil and environmental characteristics using feature selection techniques. Mathematical and Computer Modelling of Dynamical Systems, 27(1), 117-140.

[3]. Whetton, R. L., Zhao, Y., Nawar, S., & Mouazen, A. M. (2021). Modelling the influence of soil properties on crop yields using a non-linear NFIR model and laboratory data. Soil Systems, 5(1), 1-15.

[4]. Feng, L., Wang, Y., Zhang, Z., & Du, Q. (2021). Geographically and temporally weighted neural network for winter wheat yield prediction. Remote Sensing of Environment, 262, 1-53.

[5]. Khan, N., Kamaruddin, M. A., Ullah Sheikh, U., Zawawi, M. H., Yusup, Y., Bakht, M. P., & Mohamed Noor, N. (2022). Prediction of oil palm yield using machine learning in the perspective of fluctuating weather and soil moisture conditions: evaluation of a generic workflow. Plants, 11(13), 1-19.

[6]. Jui, S. J. J., Ahmed, A. M., Bose, A., Raj, N., Sharma, E., Soar, J., & Chowdhury, M. W. I. (2022). Spatiotemporal hybrid random forest odel for tea yield prediction using satellite-derived variables. Remote Sensing, 14(3), 1-18.

[7]. Cui, Y., Liu, S., Li, X., Geng, H., Xie, Y., & He, Y. (2022). Estimating maize yield in the black soil region of northeast China using land surface data assimilation: integrating a crop model and remote sensing. Frontiers in Plant Science, 13, 1-18.

[8]. Sirsat, M. S., Cernadas, E., Fernández-Delgado, M., & Khan, R. (2017). Classification of agricultural soil parameters in India. Computers and Electronics in Agriculture, 135, 269-279.

[9]. David C. Weindorf.,&Somsubhra Chakraborty.(2024), Balancing machine learning and artificial intelligence in soil science with human perspective and experience. Volume 34, Issue 1,(pp.9-12). Elsevier.

[10]. R.S. Parmar1, V. Mehra and G. J. Kaman (2022), Analyzing Soil Fertility Using Data Mining Techniques Gujarat Journal of Extension Education Vol. 34 : Issue 2.

[11]. Vasanthanageswari S and Prabhu P (2025), Deep Neuro

Fuzzy Model for Crop Yield Prediction, Journal of Machine and Computing. Vol. 5 : Issue 1.

[12]. S. Ö. Arik and T. Pfister (2021), TabNet: Attentive Interpretable Tabular Learning. Proceedings of the AAAI Conference on Artificial Intelligence. Vol . 35 : Issue 8,(P.no: 6679-6687).

[13]. Janmejay Pant1,*, Pushpa Pant2 , R. P. Pant1 , Ashutosh Bhatt3 , Durgesh Pant4 , Amit Juyal5 (2021),Soil Quality Prediction For Determining Soil Fertility In Bhimtal Block Of Uttarakhand (India) Using Machine Learning. International Journal of Analysis and Applications Volume 19, Number 1 (2021), 91-109 URL: https://doi.org/10.28924/2291-8639 DOI: 10.28924/2291-8639-19-2021-91