THE STATE COUNCIL FOR PROFESSORSHIP
INTERDISCIPLINARY PROFESSOR COUNCIL
OF THE EARTH AND MINING SCIENCES

HUS

VNU UNIVERSITY OF SCIENCE
VIETNAM NATIONAL UNIVERSITY, HANOI

# Advances in the Earth, Mining and Environmental Sciences for Safe and Sustainable Development

## Volume 1

## Advanced Technologies and Artificial Intelligence in the Earth and Environmental Sciences

PUBLISHING HOUSE FOR SCIENCE AND TECHNOLOGY

**PUBLISHING HOUSE FOR SCIENCE AND TECHNOLOGY**
A16, 18 Hoang Quoc Viet Road, Cau Giay, Hanoi
Marketing & Distribution Department: **024.22149040**
Editorial Department: **024.37917148**
Administration Support Department: **024.22149041**
Fax: **024.37910147**; Email: **nxb@vap.ac.vn**;
Website: **vap.ac.vn; ebook.vap.ac.vn**

# ADVANCES IN THE EARTH, MINING AND ENVIRONMENTAL SCIENCES FOR SAFE AND SUSTAINABLE DEVELOPMENT

## Volume 1: Advanced Technologies and Artificial Intelligence in the Earth and Environmental Sciences

**VNU University of Science, Vietnam National University, Hanoi**

**The State Council for Professorship, Interdisciplinary Professor Council of the Earth and Mining Sciences**

***Responsible for Publishing***
**Director, Editor in Chief**
**PHAM THI HIEU**

***Editor***
**Nguyen Van Vinh**

***Computing Technique***
**Do Hong Ngan**

***Cover design***
**Do Hong Ngan**

***Corporate publishing:*** **VNU University of Science**
**Address: 334 Nguyen Trai street, Thanh Xuan District, Ha Noi**

**ISBN: 978-604-357-341-1**

Registered number for Publication: 4571-2024/CXBIPH/01-55/KHTNVCN.
Decision number for Publication: 84/QĐ-KHTNCN, was issued on 04/12/2024.
Copyright deposit was completed in 4th quarter, 2024.

## Part 3: Digital Transformation and GeoAI

Doi: 10.15625/vap.2024.0342

# Machine learning analysis Geophysical data to study soil's properties

**Huong Phan[1*], Tung Nguyen[2], Phong Nguyen[2]**

[1]Hanoi University of Mining and Geology, Hanoi, Vietnam (phanthienhuong@humg.edu.vn)

[2]Vietnam Petroleum Institute, Hanoi, Vietnam

**Abstract:** Over the past decades, there has been an increasing demand for land resource information worldwide while soils are recognized as having important influence on global issues such as water supply, food security and environmental degradation. Understanding the soil properties and their spatial variations is fundamental to the sustainable use of soil resources as well as environmental monitoring. Traditionally, soil properties are measured by sample analysis methods but recently soil data are increasingly being collected by geophysical methods with proximal sensors, that are more rapid and cost effective. A main difficulty of this approach is that the physical parameters obtained from geophysical methods are dependent on soil parameters but not explicitly correlated to any of them. The application of AI/ML to infer soil properties from combination of physical parameters is a reasonable and practical solution in this case. In this paper, the authors carried out three types of geophysical surveys including electro-magnetic conductivities, gamma ray spectrum and electrical resistivities at three locations in Nghe An province. Key soil properties including moistures, salinity, sand ratio, porosity was also measured by sample analysis at drilled shallow wells. These datasets are used as input to train predictive models using random forest machine learning techniques. The results show that random forest can give the good prediction with less than 10% errors for all properties. The selected final model is ready for soil parameters prediction from geophysical field data. This approach allows to save up a lot of the field work and the cost compared to conventional methods.

**Keywords:** Proximal sensor, geophysics, soil, moisture, salinity, random forest.

## Introduction

In recent years, Vietnam has achieved significant milestones in agriculture. However, despite the fact that 40% of the population is engaged in agriculture, this sector only contributes about 20% to the GDP. According to the World Bank Country Director for Vietnam, to ensure sustainable growth in the future, agriculture needs to shift towards high-tech agriculture that meets the requirements of reducing costs, increasing yields, and achieving better product quality. Moreover, farmers and businesses must be capable of producing reliable, high-quality, safe, and sustainable products. Recognizing these challenges, the Vietnamese government has prioritized investment in agriculture. The Prime Minister has affirmed the importance of applying high technology in agriculture (Article 5 of the High-Tech Law). Vietnam needs to transform its agricultural production methods, incorporating the achievements of the Fourth Industrial Revolution into agriculture, turning Vietnam's agriculture into a 4.0 agriculture, which is characterized by the interconnection and complementarity of science, technology, and engineering, creating a synergistic force to innovate all industries. Applying geophysical methods to build databases for agriculture to serve smart production is also a trend in applying 4.0 technology to agriculture. With the characteristic of continuously collecting information, the large data conversion from geophysical parameters to soil parameters makes the application of AI indispensable.

The application of AI techniques in various scientific fields has increased rapidly, especially in the last 10 years. The adoption of machine learning in Vietnamese agriculture remains nascent, primarily confined to applications like logistics optimization, weather prediction, and plant health monitoring. In contrast, the exploration of soil analysis using machine learning is scarce. Conventional techniques involving physical soil sampling and laboratory testing continue to dominate soil research in the country.

The pedosphere is composed of soils and their connections with the hydrosphere, lithosphere, atmosphere, and biosphere. Soil is the result of a number of processes, factors, and their interactions that create different soil types (classifications) or horizons. The main soil processes are weathering and pedogenesis, while the soil forming factors are parent material, topography, climate, organisms, and time (Jenny, 1994). The factors under the influence of interactive processes create soils with different properties such as mineral texture, porosity, structure, moisture, salinity, etc.

Due to the complex relationship between soil properties and measured physical parameters, for example, electrical conductivity - a physical parameter obtained from the EMI (electromagnetic induction) geophysical method of soil, depends simultaneously on porosity, moisture, and salinity, therefore, the use of ML is a necessary tool for processing and analyzing these data. In the world, there have been some studies applying AI to determine soil properties based geophysical methods (proximal sensors), include the use of convolutional neural networks of Wadoux, 2019 or Padarian et al., 2019 (Westhuizen at al., 2023); ANN (Cockx, 2010; Boadu at al., 2013); K-NN, SVM (Rahman at al, 2018); Random Forest- RF ((Lacoste et al., 2011; Viscarra Rossel et al., 2014; Harris and Grunsky, 2015). In this paper, the authors carried out three types of geophysical surveys including electro-magnetic conductivities, gamma ray spectrum and electrical resistivities at three locations in Nghe An province. Key soil properties including moistures, salinity, sand ratio, porosity were also measured by sample analysis at 100 drilled shallow wells.

## Materials and methods

### Data

Nghe An is a large province with diverse terrain with many potentials and advantages of 3 ecological regions: midland and mountainous areas, plains and coastal areas. Nghe An has two main types of land: alluvial soil in the plains and ferralitic soil in the mountains. The alluvial soil group accounts for 198,139 ha, the red-yellow ferralitic soil group in the mountainous areas is 383,121 ha, the rest are other soil groups. Alluvial soil is mainly deposited by the Ca River and its tributaries, suitable for growing food crops, foodstuffs and short-term industrial crops. Ferralitic soil is mainly concentrated in the midland and mountainous areas, including many types: ferralitic soil on limestone, shale, red basalt soil suitable for growing industrial crops, fruit trees and afforestation. With the characteristics of rapid change in soil properties according to terrain, it is very suitable for assessing the ability to detect signs of change in soil properties and the horizontal resolution of geophysical methods. To capture the soil property variations across different terrains, we selected three diverse sites in Nghe An province: Ductai Farm in the alluvial plain, Diem Nghiep Cooperative in the coastal area, and Duoc Lieu Cooperative in the mountainous region. This selection allows for a comprehensive data collection to support machine learning applications in soil studies (Figure 1). Duc Tai Farm Area: Located at the longitude between 105°36’45”E and 105°36’51”E, latitude 19°7’32”N and 19°7’26”N with relatively flat terrain and altitude above sea level fluctuating around 5m, Duc Tai Farm meets the criteria of being a representative area of the plain. Diem Nghiep Cooperative Area: Located at the meridian between 105°33’11”E and 105°33’17”E, latitude 19°8’10”N and 19°7’59”N with relatively flat terrain and altitude above sea level fluctuating around 11m, Diem Nghiep Cooperative meets the criteria of being a representative area of the coastal plain. Duoc lieu Nghia dan Cooperative: Located at the meridian between 105°33’11”E and 105°33’17”E, latitude 19°18’30”N and 19°18’24”N with relatively flat terrain and altitude ranging from 49 m to 55 m, Duoc lieu Nghia dan Cooperative meets the criteria of being a representative mountainous area with the terrain divided into 3 levels.

Geophysical data was acquired through three methods: 1) Electromagnetic induction with an EM38-MK2, 2) Gamma-ray spectrometry with a Gamma Surveyor and 3) Resistivity measurement with a GESKA. The EM38-MK2 provides simultaneous measurements of ground conductivity and magnetic susceptibility (In-Phase) with two transmitter receiver coil separation at 1 m and 0.5 m, for 3 effective depth ranges; 1.5 m and 0.75 m in vertical dipole mode and two ranges, 0.75 m and 0.375 m, in horizontal dipole mode. Gamma Surveyor Vario serves for determination of K, U, Th concentrations and natural gamma dose rate in walking. The analysis is based on multichannel spectral measurement with direct K assessment (via K40) and indirect assessment of U and Th (via their daughter products) so the natural balance of isotopes in measured rock is assumed. The spectrometer is detected natural radioisotopes. Resistivity method involves injecting a current between two metal electrodes partially embedded in the ground. The potential difference is measured between another pair of electrodes placed on the surface. The resistivity of the subsurface is then calculated using the measured current, voltage, and electrode geometry. For all 3 sites, a grid system with 10-meter spacing in both the horizontal and vertical directions was established in each study area. For EMI measurements, with grid 10 m × 10 m data was collected at 1-meter intervals along these lines. At each measurement point, both in-phase and quadrature components were recorded in both the vertical and horizontal directions, totaling 2000 data points. Gamma-ray spectrometry measurements were conducted at the grid intersections with a 10-meter spacing in one direction and a 5-meter spacing in the other. Four values (K, U, Th, and total gamma) were obtained at each point, yielding 242 data points.

Resistivity measurements were conducted at the intersections of the 10 m × 10 m grid, resulting in 121 data points. Additionally, 900 soil samples were collected for laboratory analysis (100 sampling points in each area, at each borehole take 3 samples at 3 depths 37.5 cm, 75 cm and 150 cm). The location of geophysical survey lines and samplings points are shown in Figure 1. In this study, we aim to use random forest regression to predict soil properties in the Duoc lieu Nghia Dan Cooperative area. To achieve this, we will train the model using all available data from Nghe An province. Three study sites are illustrated in Figure 1 where geophysical data was acquired. The site 1 is the Ductai Farm, the site 2 is the Diem Nghiep Cooperative, and Duoc lieu Nghia Dan Cooperative is site 3. The site 3 was designated as a validation site to assess the accuracy of the random forest prediction model.

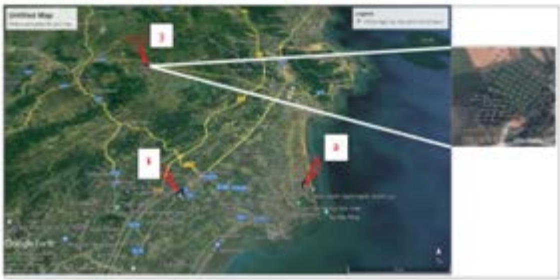


**Figure 1.** Location of study arears in NgheAn province and geophysicl survey with sampling points at Duoclieu Nghiadan Cooperative

## Method

Random Forest is a supervised machine learning algorithm that can be used for regression tasks. It is also one of the most flexible and easy-to-use algorithms. The random forest model (Bergen et al., 2019; Ding et al., 2020) is a collection of "decision trees". In each tree, the dataset is recursively partitioned into increasingly homogeneous subsets with respect to the response variable based on the improvement in the residual sum of squares (RSS). RF creates decision trees on randomly selected data samples, predictions are made from each tree, and the best solution is selected by voting. All predictors are examined at each split to decide which predictor and which value of the predictor best splits the data. Random forest is considered a non-linear, non-parametric method that can handle large p-small n problems, is rarely overfitted, and is relatively robust to outliers and noise (Díaz-Uriarte and Alvarez de Andrés, 2006).

RF differs from a decision tree model in that it grows many trees instead of a single decision tree and averages the results. The number of trees is called ntree. The power of the model is achieved by two main reasons: (1) by varying the input data, i.e., taking a bootstrapped sample of size two-thirds of the full dataset to train each tree, and (2) by varying the tree model structure, i.e., randomly selecting a subset of predictors from all predictors at each tree node. The size of the predictor subset - here called mtry - remains the same for the entire forest.

Two important variable importance measures are returned by random forest: (1) the increase in node purity, which indicates the improvement in the RSS splitting criterion as a measure of importance for the splitting variable, and (2) permutation importance (also called decrease in MSE). Variable importance is often used to aid in the interpretation of the dataset by detecting interactions between predictor variables, identifying important predictor variables, and as a filter to eliminate predictor variables.

The RF algorithm operates in four steps (Figure 2):

- Randomly select samples from the given dataset.
- Build a decision tree for each sample and obtain prediction results from each decision tree.
- Vote for each prediction result.
- Select the most predicted result as the final prediction.

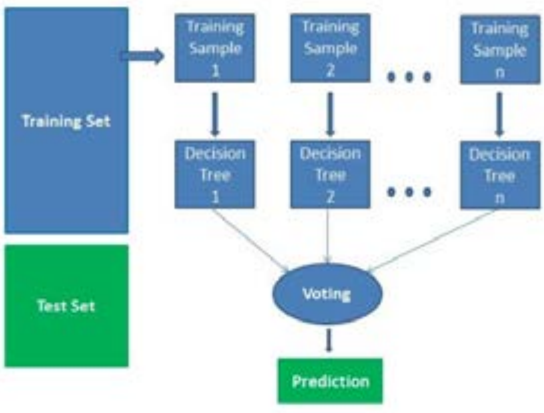


**Figure 2.** algorithm diagram random forest

In the beginning, data analysis revealed that the distribution of geophysical data followed a Gaussian pattern. This finding indicates that the collected data is well-suited for training machine learning models due to its adherence to the typical characteristics of physical parameters.

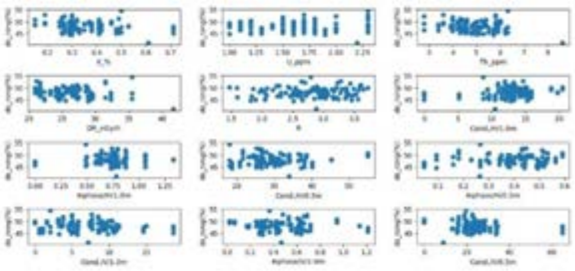


**Figure 3.** Cross plot between soil porosity and physical parameters obtained from geophysical methods

The relationship between soil properties and physical parameters obtained from geophysical methods is complex. Figure 3 shows a result indicating the relationship between porosity and physical parameters such as K, Th, U, total gamma, and electrical conductivity. These relationships demonstrate that there is no linear relationship between physical parameters and soil properties. In other words, a single physical parameter cannot directly determine soil properties. This confirms the role of machine learning in soil property research.

For the purpose of random forest regression algorithm, the dataset is divided into. a training set (90%) and a testing set (10%). The number of fold K = 10.

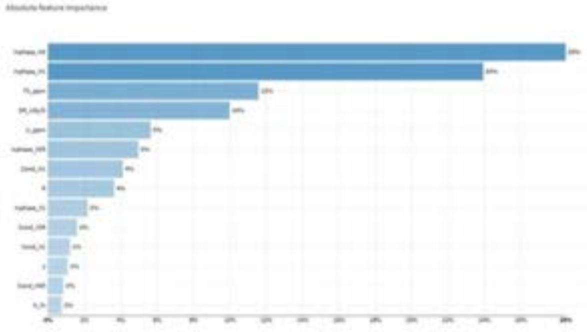
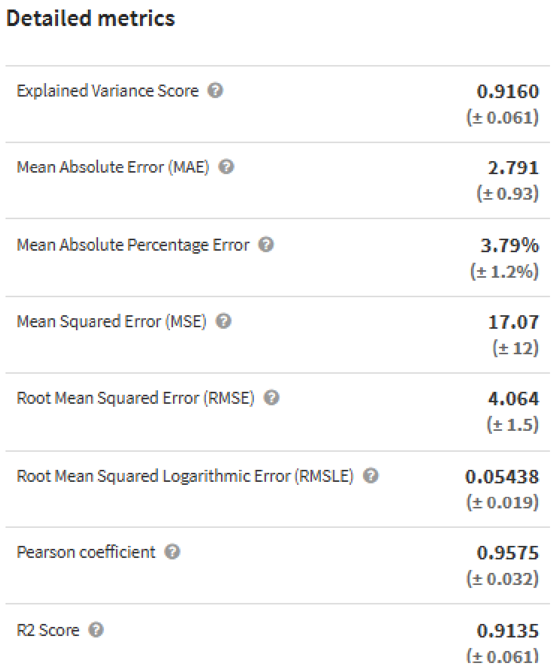


**Figure 4.** Important feature

Important features were tested to select those that have a significant impact on soil properties. Physical parameters with a correlation coefficient below 5% were eliminated. Figure 4 shows the results after testing, indicating that only a few physical parameters were used.

After identifying the input data, including the target variable and input variables, the machine learning model training process was initiated. The next step was the independent evaluation (blind test). If the blind test results met the requirements, the training process was completed; otherwise, the machine learning model was retrained. The result of the Random Forest regression is the average value from the models. Table 1 is the example of one of these model evaluations, the model to predict the sand ratio in the Duoclieu Cooperative -Nghia dan area at a depth of 37.5 cm.

**Table 1.** Evaluation test



**Detailed metrics**

| | |
|---|---|
| Explained Variance Score | 0.9160<br>(± 0.061) |
| Mean Absolute Error (MAE) | 2.791<br>(± 0.93) |
| Mean Absolute Percentage Error | 3.79%<br>(± 1.2%) |
| Mean Squared Error (MSE) | 17.07<br>(± 12) |
| Root Mean Squared Error (RMSE) | 4.064<br>(± 1.5) |
| Root Mean Squared Logarithmic Error (RMSLE) | 0.05438<br>(± 0.019) |
| Pearson coefficient | 0.9575<br>(± 0.032) |
| R2 Score | 0.9135<br>(± 0.061) |

## Results and Discussion

In this paper, we present some results of soil properties predictions including salinity (Figure 5), moisture (Figure 6) porosity (Figure 7) and sand content (figure 8) in the area of Duoc lieu Cooperative Nghia Dan at a depth of 0.375 m. While Random Forest (RF) seems relatively simple with uncomplicated mathematical formulas, it is quite sensitive to training data. Hyperparameters, such as the number of trees, need to be tuned. In our training, the number of trees varied from 100 to 200. Bootstrapping was naturally selected. The maximum number of features also needs consideration as it determines branching. Figure 4 suggests that limiting the features to EMI data, specifically the in-phase component of H and the gamma spectra of Th and total gamma, is optimal. The RF model, trained on data from 3 regions (300 data points), can accurately predict soil properties. The model evaluation results are acceptable. For example, the model achieved an $R^2$ score of 0.91 for sand prediction, indicating a strong fit to the data. The $R^2$ for porosity was slightly lower at 0.67 was acceptable, however, suggesting that the model captured 67% of the variability in the porosity data.

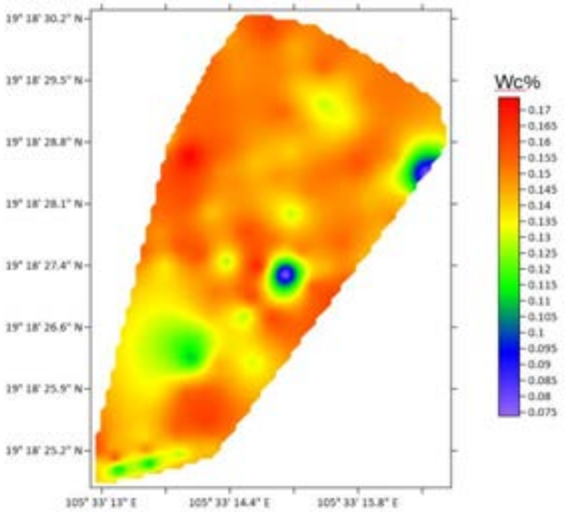


**Figure 5.** Predicted salinity content at Duoc lieu Cooperative Nghia Dan at a depth of 37.5 cm

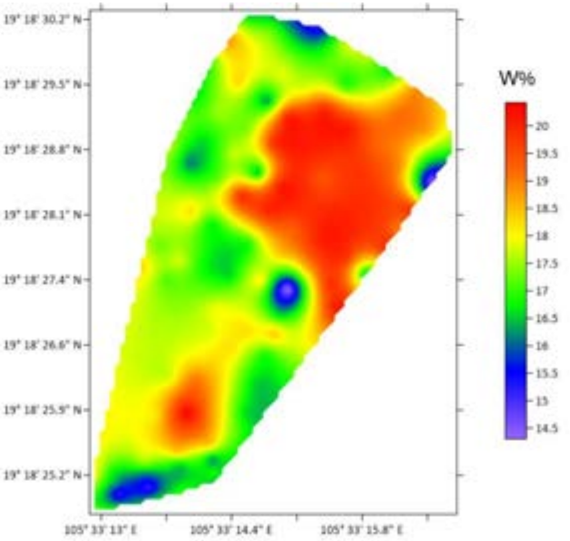


**Figure 6.** Predicted moisture content at Duoc lieu Cooperative Nghia Dan at a depth of 37.5 cm

The higher R² for sand can be attributed to the relatively simpler relationship between sand content and the geophysical properties measured. In contrast, porosity is influenced by a more complex interplay of geological, hydrological, and biological factors. Additionally, the accuracy of input data, particularly for porosity-related variables, might have influenced the model's performance. Finally, simplifying assumptions made about the relationships between variables could have introduced errors in porosity prediction.

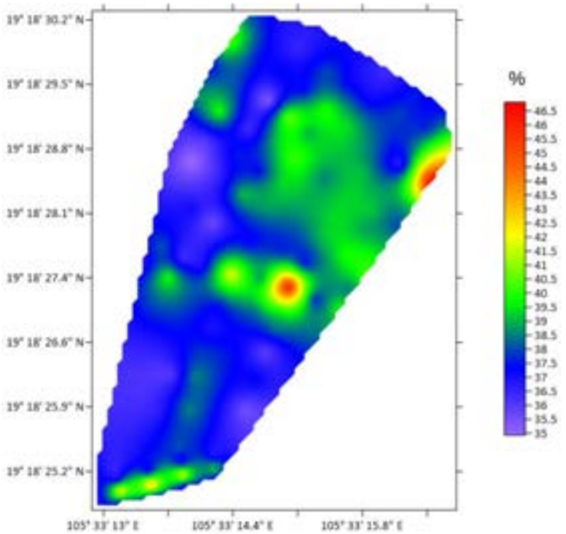


**Figure 7.** Predicted porosity at Duoc lieu Cooperative Nghia Dan at a depth of 37.5 cm

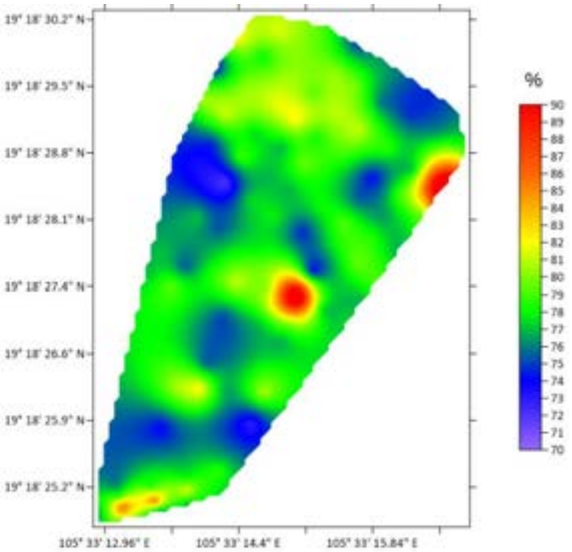


**Figure 8.** Predicted sand content at Duoc lieu Cooperative Nghia Dan at a depth of 37.5 cm

To improve the model's performance, several strategies can be considered:

- **Collect additional data:** Increasing the number of sampling points, especially in areas with high porosity variability, this may not be feasible in practice, particularly when the number of existing sampling points is already large.

- **Incorporate additional variables:** Gather data on factors such as organic matter content and mineral composition to enhance the model's explanatory power. This not only improves the model but also addresses the needs of agronomists who are interested in carbon exchange and mineral composition.

- **Explore more complex models:** Experiment with neural networks or gradient boosting algorithms.

- **Conduct in-depth data analysis:** Identify potential outliers and handle them appropriately; explore other correlations between variables.

Overall, the model demonstrates promising results, particularly for predicting sand content. Future research could focus on improving porosity prediction by addressing the limitations outlined above and exploring potential applications in fields such as soil science and agriculture.

## Conclusion

Our research shows that the integration of geophysical techniques (electromagnetic induction, gamma-ray spectrometry, and electrical resistivity) with 4.0 technologies random forest machine learning can create detailed maps of soil properties. This approach offers a significant advantage over traditional sampling methods by providing rapid, non-destructive, and spatially continuous data. The results highlight the potential for using geophysical methods to predict various soil properties and pave the way for future research exploring other machine learning algorithms and big data analysis in geophysics.

## Acknowledgement

This work is based on the research supported by the National Research Foundation of Ministry of

Science and Technology (Grant number: KC4.0-32/19-25).

## References

Bergen K.J., Johnson P.A., De Hoop M.V., Beroza G.C. (2019). Machine learning for data-driven discovery in solid earth geoscience. Science, 363(6433): 1–10.

Boadu F.K., Owusu-Nimo F., Achampong F., Ampadu S.I. (2013). Artificial neural network and statistical models for predicting the basic geotechnical properties of soils from electrical measurements. Near Surface Geophysics, 11(6): 599-612.

Cockx L., Meirvenne M.V., Vitharana U.W.A., Vancoillie F.M.B., Verbeke L.P.C., Simpson D., Saey T. (2010). Proximal Soil Sensing. Rossel R.A.V., McBratney A.B., Minasny B. (eds). Springer and New York. (ISBN_978-90-481-8858-1_), 446 p.

Díaz-Uriarte R., Alvarez de Andrés S. (2006). Gene selection and classification of microarray data using random forest. BMC bioinformatics, 7(3): 1-13.

Ding J., Yang S., Shi Q., Wei Y., Wang F. (2020). Using apparent electrical conductivity as indicator for investigating potential spatial variation of soil salinity across seven oases along Tarim River in Southern Xinjiang, China. Remote Sensing, 12(16); 2601; doi:10.3390/rs12162601.

Harris J.R., Grunsky E.C. (2015). Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. Computers & Geosciences, 80: 9-25.

Jenny H. (1994). Factors of soil formation: a system of quantitative pedology. Illustrated, reprint (eds). Dover and New York. (ISBN _ 0-486-68128-9_), 281 p.

Lacoste M., Lemercier B., Walter C. (2011). Regional mapping of soil parent material by machine learning based on point data. Geomorphology, 133(1–2): 90–99.

Rahman S., Mitra K.C., Nohidul Islam S.M. (2018). Soil classification using machine learning methods and crop suggestion based on soil series, 21st International Conference of Computer and Information Technology (ICCIT): 21-23.

Viscarra Rossel R.A., Webster R., Kidd D. (2014). Mapping gamma radiation and its uncertainty from weathering products in a Tasmanian landscape with a proximal sensor and random forest kriging. Earth Surf. Process. Landforms, 39(6): 735–748.

Westhuizen S., Heuvelink G.B., Hofmeyr D.P. (2023). Multivariate random forest for digital soil mapping. Geoderma. 431: 116365. https://doi.org/10.1016/j.geoderma.2023.116365.