

# ENHANCEMENT OF ACCURACY IN SOIL UREA ESTIMATION USING MACHINE LEARNING TOOLS\*

Sulaxana Vernekar<sup>1, †</sup>, Marlon Sequeira<sup>2, \*, †</sup>, Sameer M. Patil<sup>3, †</sup>, Jivan Parab<sup>2, †</sup> and Prof. Gourish Naik<sup>2, †</sup>

<sup>1</sup> GVM'S GGPR College of Commerce and Economics

<sup>2</sup> Electronics Programme, School of Physical and Applied Sciences, Goa University, Goa, India

<sup>3</sup> Dnyanprassarak Mandals College & Research Centre, Mapusa, Goa India

## Abstract

Soil health is vital for getting a good crop yield. Analysis of available soil nutrients done at the right time not only helps in conservation of soil fertility but can also help in getting a good crop yield by limiting the usage of external inputs to the soil such as fertilizers, water etc. The use of AI in agriculture is being explored in recent times to optimize the crop yield. Machine Learning techniques are used in developing smart soil sensing systems to provide accurate soil nutrients distribution. In this study, a sample of 40 spectral data in the frequency range of 500MHz to 1000MHz was passed to the ParLeS software. The PLSR cross validation in ParLeS gave us an RMSE of 2.87. However, when Ridge regression based on machine learning was applied, we obtained a RMSE of 1.02 with parameter alpha set to 0.005. Thus, we can say conclusively that, machine-learning methods yield better results than traditional methods. In addition, implementation of ParLeS needs LabVIEW type of environment and needs external graphics support, whereas, Ridge regression can be implemented using simple Python environment, which is now a day most often used programming language. The implementation does not require compulsory graphics support.

## Keywords

ParLes, PLSR, Ridge Regression.

## 1. Introduction

Agriculture is the backbone of any thriving economy. Advancements in technology has seen lot of influence on the way agriculture is practiced. Smart farming has paved a way for sustainable agriculture and increasing the crop productivity. Soil fertility plays an important role in crop production [1]. The available nutrients in the soil can highly influence the crop yield. Cultivating crops constantly without proper analysis of the soil can deteriorate its health leading to the soil becoming arid. Smart farming techniques are based on micro management of the farm taking into consideration the spatial and temporal variability exhibited by soil. This enables the proper management of external inputs such as fertilizer and pesticide application etc. to the soil [2]. Proper understanding and knowledge about the soil can enable the farmers to take proper decisions in crop management thus enhancing the crop productivity [3]. There are several issues that need to be


---

SCCTT-2024: International Symposium on Smart Cities, Challenges, Technologies and Trends, 29th Nov 2024, Delhi, India

<sup>1\*</sup> Corresponding author.

<sup>†</sup> These authors contributed equally.

✉ sulaxgoa@gmail.com (S. Vernekar); marlon@unigoa.ac.in (M. Sequeira); sameer@dmscollege.ac.in (S. Patil); jsparab@unigoa.ac.in (J. Parab); gmnaik@unigoa.ac.in (G Naik)

 (S. Vernekar); 0000-0002-7462-3492; 0000-0002-1444-5428 (M. Sequeira); 0000-0002-0848-7349, 0009-0007-3674-1942 (S. Patil); (J. Parab)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

tackled in agriculture such as lack of digitization, food safety issue, ecological problems and inefficient agri-food supply chain. Integration of Industry 4.0 in agriculture can greatly influence productivity, agri-food supply chain efficiency, food safety, and the sustainable use of natural resources[4].

Artificial Intelligence (AI) is the most rapidly growing technology embedded into all aspects of human life. In agriculture AI technologies can be used in precision farming for soil and irrigation management, weather forecasting, plant growth, disease prediction, and animal management [5]. With the exponential growth and development of data processing, information technology, and artificial intelligence, smart farming makes use of cutting-edge innovations to boost productivity and reduce labor stress and automating soil and crop management with AI [6]. Smart soil prediction is a low-cost method of forecasting a soil's performance over a wide range of crops.

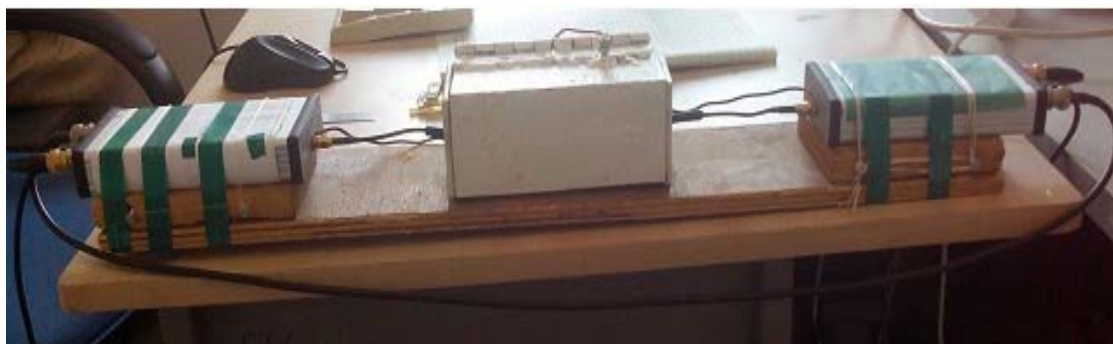
Digital soil mapping (DSM) is used to generate digital maps of the type and quality of soil by combining soil sensing data with environmental factors [7]. Recent years have seen a significant rise in the use of DSM in soil science, which can be attributed to the integration of several ideal factors, including, but not limited to, tremendous interest in quantitative and spatial soil information, the buildup of databases of predicted or interpreted soil properties together with thoroughly known environmental factors, and the development of computational methods combined with computer resources to extract these stores of soil data [8]. Obtaining exact data on soil nutrient composition is a critical step in the implementation of precision agriculture and DSM is providing a potential breakthrough [9]. Artificial Intelligence tools such as fuzzy systems, decision trees, expert knowledge, machine learning algorithms, deep learning methodologies, and other artificial intelligence technologies can be used to provide more accurate forecasts and solutions in DSM [10].

AI models and DSM have been utilized in soil fertility prediction, offering a decision-making tool capable of forecasting the best crop based on soil pH, soil nutrients, soil moisture, environmental variables, and other components [11]. It was observed from a study conducted on prediction of soil nutrients using spectroscopic data that using Machine Learning (ML) techniques greatly improves the accuracy of soil nutrient prediction [12]. ML algorithms were used in a study to find the relationship between independent variables and dependent variables for soil data analysis. The independent variables were moisture, temperature, soil pH, Cation Exchange Capacity(CEC) and the dependent variables were Nitrogen, Phosphorus and Potassium (NPK). This study showed that there exist relationships between Phosphorus, Potassium, soil pH and CEC; Nitrogen and soil moisture and temperature using ML algorithms [13].

In another review study on using machine learning methods for predicting soil properties, agricultural yield, and soil fertility, it was observed that for soil prediction, Random Forest (RF) and deep learning techniques surpass traditional ML algorithms. Depending on the model's inputs, the RF and deep learning techniques can reliably forecast soil conditions and crop to be grown. It was also found from the study that inaccurate data has the ability to reduce forecasting precision. Variations in geographical elements, meteorological circumstances, and farming techniques can hamper the process of generalizing models. Furthermore, selecting relevant characteristics from numerous influencing factors necessitates subject expertise and testing [1].

## 2. Methodology

To obtain the RF spectra of various samples a cell is designed based on the principle of dielectricity. The design details of the cell are discussed in [14].



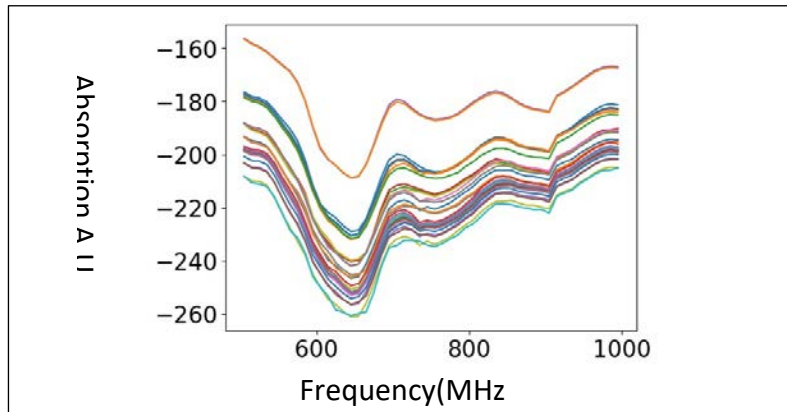
**Figure1:** Experimental Setup

The experimental setup consists of a cell which is placed inside the iron box at the centre as shown in figure 1. A Signal Hound tracking generator USBTG44A and a Signal Hound spectrum analyzer USB-SA124B were used for obtaining the RF response. The sample is placed in the cell and RF signal from the tracking generator is injected into the cell through the central copper wire. The strength of the signal reduces due to dielectric loss offered by the sample solution as the signal propagates towards the receiver end.

The RF spectrum analyzer connected at the receiver end of the cell captures signal proportional to the radiation loss due to the sample solution. The cell has a capacity of holding 15ml of liquid. Soil samples were prepared in the laboratory by mixing 5 different components urea, potash, sodium chloride, calcium carbonate and phosphate in distilled water. Molar solutions for each of the component was prepared and for 15ml of water the amount of each component required to be added was calculated. It was found that amount of urea required was 225mg/15ml. Similarly, for the remaining components the amount required to be added for 15ml of water was calculated. The amount of each component to be added is shown in table 1.

**Table 1:** Concentrations denotation table

Concentrations denotation	Concentration(mg/15ml)				
	Urea	Potash	Phosphate	Lime	Salt
0.5	112.5	139.7	1890	187.5	109.87
1	225	279.4	3780	375	219.74
1.5	337.5	419.1	5670	562.5	329.61
2	450	558.8	7560	750	439.48
3	675	838.2	11340	1125	659.25



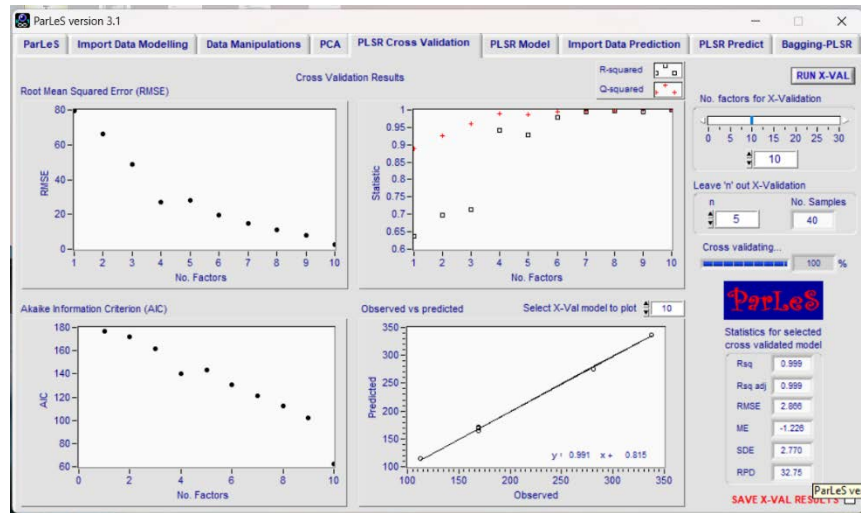
**Figure 2:** RF Spectra of 40 samples in the frequency range 500MHz-1000MHz

Figure 2 shows the RF spectra of 40 samples which were used in building the model for soil urea estimation. These samples were prepared by adding different concentrations of the components taken as per the table 1.

The soil urea estimation using this spectral data of 40 samples was done using two methods. The first method was using the ParLeS software based on Partial Least Squares Regression (PLSR) model. The second method was using Machine Learning Algorithm i.e. Ridge Regression.

ParLeS is a chemometrics software for multivariate modelling and prediction. It provides users with various algorithm options to transform, preprocess and pretreat spectra. It may be used to implement principal components analysis (PCA); partial least squares regression (PLSR) with leave-n-out cross validation; and bootstrap aggregation-PLSR (bagging-PLSR). ParLeS facilitates the implementation of a large number of preprocessing techniques as well as bagging-PLSR, which can improve the robustness and accuracy of PLSR models. Other unique features of ParLeS include the provision of a number of assessment statistics and graphical output as well as a user-friendly interface and functionality [15].

In this study, a sample of 40 spectral data in the frequency range of 500MHz to 1000MHz was passed into the ParLeS software. The PLSR cross validation technique was used for soil urea estimation where the  $n=5$  was chosen for the cross validation. Using this an RMSE of 2.87 was obtained. A screenshot of the ParLeS software is as shown in Figure 3.



**Figure 3:** Screenshot of ParLeS software

Ridge regression, also known as L2 regularization, is one of the many regularization techniques applied to linear regression models. Regularization is a statistical method used to prevent errors due to the overfitting of training data. Ridge regression is specifically tailored to address multicollinearity in regression analysis, which is crucial when developing machine learning models with many parameters, particularly when these parameters are significantly weighted.

A standard, multiple-variable linear regression equation is:

$$Y = X_0 + B_1X_1 + B_2X_2 + \dots + B_nX_n \dots\dots(1)$$

In the above equation, Y represents the expected value or the dependent variable, X is the predictor or independent variable, B denotes the regression coefficient linked to that independent variable, and X0 is the value of the dependent variable when the independent variable is zero, also referred to as the y-intercept. It's important to observe how the coefficients illustrate the relationship between the dependent variable and a specific independent variable. The best-fitting line for a given dataset is obtained by calculating coefficients for each independent variable that result in the smallest residual sum of squares (also called the sum of squared errors).

The Residual sum of squares (RSS) represents how well a linear regression model matches the training data and is represented by the formula:

$$RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \dots\dots(2)$$

This formula is used to calculate the accuracy of model predictions against the expected values in the training data. If the Residual Sum of Squares (RSS) is zero, it indicates that the model perfectly predicts the dependent variables. If two or more variables have a strong linear correlation, high-value coefficients are generated, causing the model's output to be sensitive to minor changes in the input data. This indicates that the model overfitted on a single training dataset and is unable to correctly generalise to new test datasets. This causes the model to be unstable.

Multicollinearity exists when two or more predictors have a near-linear relationship or are highly correlated, which results into unreliable and unstable estimates of regression coefficients. Ridge

regression is a procedure for eliminating the bias of coefficients and reducing the mean square error by shrinking the coefficients of a model towards zero in order to solve problems of overfitting or multicollinearity that are normally associated with ordinary least squares regression.

Ridge regression corrects for high-value coefficients by introducing a regularization term (often called the penalty term) into the RSS function. This penalty term denoted as L2, is the sum of the squares of the model's coefficients. It is represented in the formulation:

$$RSS_{L2} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^p B_j^2 \dots\dots(3)$$

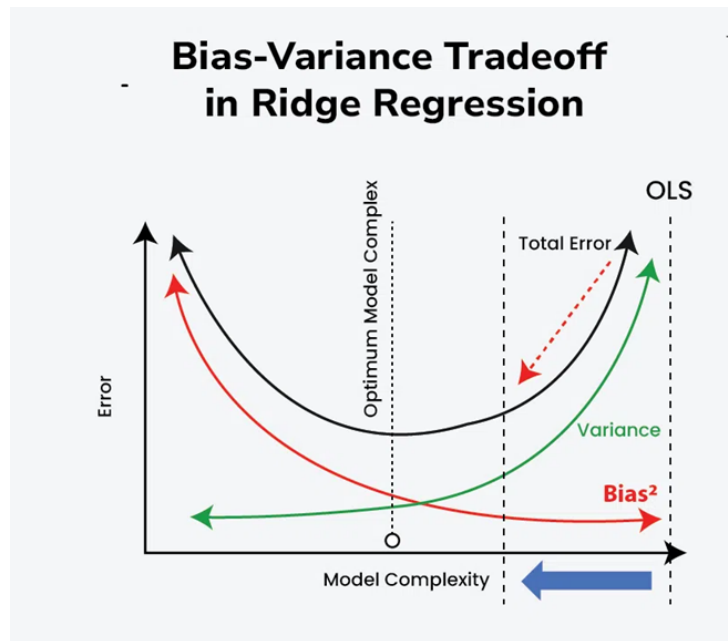
The L2 penalty term reduces all coefficients to balance the high ones. This process is utilized in ridge regression, to calculate new coefficients that minimize the residual sum of squares (RSS) for a model, thereby reducing overfitting.

Ridge regression doesn't reduce all coefficients equally and is proportional to their original magnitude. As the lambda ( $\lambda$ ) parameter increases, coefficients with higher values diminish more rapidly than those with lower values, resulting in a greater penalty for the former [16].

In machine learning, ridge regression is used to reduce overfitting that results from model complexity. Model complexity can be due to a model possessing too many features and features possessing too much weight. Feature weight refers to a given predictor's effect on the model output.

In machine learning terms, ridge regression amounts to adding bias into a model for the sake of decreasing that model's variance. Bias measures the average difference between predicted values and true values and variance measures the difference between predictions across various realizations of a given model. As bias increases, a model predicts less accurately on a training dataset. As variance increases, a model predicts less accurately on other datasets. Bias and variance thus measure model accuracy on training and test sets respectively. To reduce the model bias and variance, ridge regression technique can be used [16].

Using Ridge regression technique allows control over the bias-variance trade-off. Increasing the value of  $\lambda$  increases the bias but reduces the variance, while decreasing  $\lambda$  does the opposite. The goal is to find an optimal  $\lambda$  that balances bias and variance, leading to a model that generalizes well to new data.



**Figure 4:** Bias Variance Tradeoff

Selection of an appropriate value for the ridge parameter  $k$  is crucial in ridge regression, as it directly influences the bias-variance tradeoff and the overall performance of the model. There are several methods for the selection of ridge parameter:

### 1. Cross-Validation

Cross-validation is one of the most popular method used in the selection of the ridge parameter. In this method, the dataset is divided into multiple subsets, and the model is trained on some subsets while being validated on the remaining ones. The process is repeated over multiple iterations, and the average performance across all iterations is used to determine the optimal value of  $\lambda$ .

- **K-Fold Cross-Validation:** The dataset is divided into  $K$  subsets (folds). The model is trained on  $K$ - folds and validated on the remaining fold. This process is repeated  $K$  times, with each fold being used as the validation set once. The average performance across all folds is used to select  $\lambda$ .
- **Leave-One-Out Cross-Validation (LOOCV):** A special case of  $K$ -fold cross-validation where  $K$  equals the number of observations. Each observation is used as a validation set once, and the model is trained on the remaining observations. This method is computationally intensive but provides an unbiased estimate of the model's performance.

**2. Grid Search:** This method defines a grid of possible values for  $\lambda$  and the ridge regression model is trained for each value of  $\lambda$ . The performance of the model is evaluated for each value of  $\lambda$  from the grid and the one with the best performance is then selected as the ridge parameter.

### 3. Bayesian Optimization:

Bayesian optimization is used to efficiently explore the space of possible  $\lambda$  values and find the optimal value. This method can be more efficient than grid search for large search spaces.

#### 4. Information Criteria:

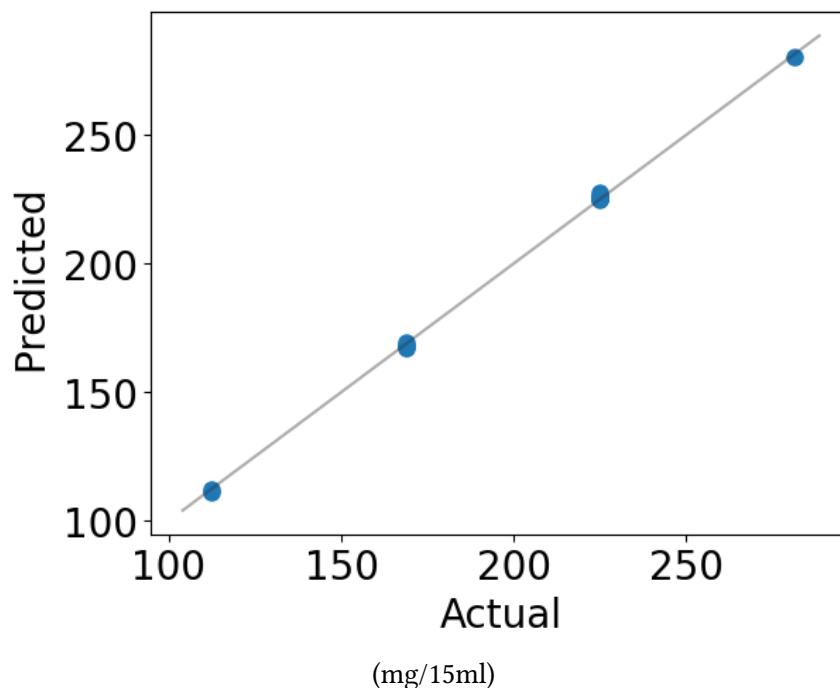
Use information criteria like Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) to select the optimal value of  $\lambda$ . These criteria balance model fit and complexity.

#### 5. Domain Knowledge:

- Incorporate domain knowledge about the problem to guide the choice of  $\lambda$ . For example, if you know that overfitting is a significant concern, you might choose a larger value of  $\lambda$  [17].

Ridge regression was implemented using the sklearn python package in the python programming language. The sklearn package includes the Application Programming Interface (API) interface to implement the same. The `linear_model.Ridge()` API is used to implement the ridge regression. The only parameter supplied to the API is `alpha` with a value of 0.005. Here it may be noted that the `alpha` is equivalent to  $\lambda$  specified above. The other parameters have default values. The parameters with the default values are `copy_X=True`, `fit_intercept=True`, `tol=0.0001`, `max_iter=None`, `positive=False`, `solver='auto'`, and `random_state=None`.

The dataset consisting of 40 samples was used for training and testing the ML model using Ridge regression. With the parameter `alpha` set to 0.005, the RMSE obtained using this technique was found to be 1.02.



**Figure 5:** Typical graph showing the actual and predicted urea values



### 3. Result and Discussion

It may be seen that the analysis using Ridge technique (Which is a machine learning based tool for regression analysis) gives excellent performance with error as low as 1.02. Whereas, the error in traditional technique of ParLeS regression is 2.87, which is nearly three times more than that of Ridge technique. As mentioned earlier that, in addition to the advantage of less error, the implementation of the algorithm can be done in a simpler computational platform, not necessarily requiring complicated LabVIEW back end. This is reflected in the table 2. The regression graph shown in figure 5 show good agreement between the actual and predicted values using the ridge regression technique.

Table 2: Result obtained using various methods

Model Name	PLSR	Ridge Regression
RMSE	2.87	1.02

### 4. Conclusion

In this article we studied the application of Ridge Regression Technique for analysis of urea in the soil for better productivity of the crops. In past we had done such analysis using ParLeS (which is propriety and not an open source software). The results obtained were encouraging with errors as low as 1.02mg/15ml. Therefore, we conclude here that Ridge Technique is far superior to the traditional technique of regression analysis.

### 5. References

- [1] Folorunso, O.; Ojo, O.; Busari, M.; Adebayo, M.; Joshua, A.; Folorunso, D.; Ugwunna, C.O.; Olabanjo, O.; Olabanjo, O. 2023. Exploring Machine Learning Models for Soil Nutrient Properties Prediction: A Systematic Review. *Big Data Cogn. Comput.* 7, 113. DOI: 10.3390/bdcc7020113
- [2] Andreas Kamilaris , Francesc X. Prenafeta-Boldú. 2018. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70-90.DOI: 10.1016/j.compag.2018.02.016.
- [3] Jain, S., Sethia, D. 2023. A Review on Applications of Artificial Intelligence for Identifying Soil Nutrients. In: Saini, M.K., Goel, N., Shekhawat, H.S., Mauri, J.L., Singh, D. (eds) *Agriculture-Centric Computation. ICA 2023. Communications in Computer and Information Science*, vol 1866. Springer, Cham. DOI: 10.1007/978-3-031-43605-5\_6

- [4] Liu, Y., Ma, X., Shu, L., Hancke, G. P., & Abu-Mahfouz, A. M. 2020. From Industry 4.0 to Agriculture 4.0: Current Status, Enabling Technologies, and Research Challenges. *IEEE Transactions on Industrial Informatics*, 1–1. DOI:10.1109/tii.2020.3003910
- [5] Shaikh, F.K.; Memon, M.A.; Mahoto, N.A.; Zeadally, S.; Nebhen, J. 2021. Artificial intelligence best practices in smart agriculture. *IEEE Micro* , 42, 17–24.
- [6] Chen, Q.; Li, L.; Chong, C.; Wang, X. AI-enhanced soil management and smart farming. *Soil Use and Management*. 2022, 38, 7–13. DOI: 10.1111/sum.12771
- [7] Dobos, E. 2006. *Digital Soil Mapping: As a Support to Production of Functional Maps*; Office for Official Publication of the European Communities: Luxembourg
- [8] Wadoux, A.M.C.; Minasny, B.; McBratney, A.B. 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Sci. Rev.* , 210, 103359
- [9] Dong, W.; Wu, T.; Sun, Y.; Luo, J. 2018. Digital mapping of soil available phosphorus supported by AI technology for precision agriculture. In *Proceedings of the 2018 7th International Conference on Agro-geoinformatics (Agro-geoinformatics)*, Hangzhou, China, pp. 1–5.
- [10] Khaledian, Y.; Miller, B.A. Selecting appropriate machine learning methods for digital soil mapping. *Appl. Math. Model.* 2020, 81, 401–418.
- [11] Shahare, Y.; Gautam, V. Soil Nutrient Assessment and Crop Estimation with Machine Learning Method: A Survey. In *Cyber Intelligence and Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 253–266.
- [12] Trontelj ml., J.; Chambers, O. Machine Learning Strategy for Soil Nutrients Prediction Using Spectroscopic Method. *Sensors* 2021, 21, 4208. DOI:10.3390/ s21124208
- [13] Umm E Farwa, Ahsan Ur Rehman, Qasim Khan, S. ., & Khurram, M. (2020). Prediction of Soil Macronutrients Using Machine Learning Algorithm. *International Journal of Computer (IJC)*, 38(1), 1–14.
- [14] S. R. Vernekar, I. A. P. Nazareth, J. S. Parab and G. M. Naik, "RF spectroscopy technique for soil nutrient analysis," 2015 International Conference on Technologies for Sustainable Development (ICTSD), Mumbai, 2015, pp. 1-4. doi: 10.1109/ICTSD.2015.7095878
- [15] Raphael A. Viscarra Rossel, "ParLeS: Software for chemometric analysis of spectroscopic data", *Chemometrics and Intelligent Laboratory Systems*, Volume 90, Issue 1, 2008, Pages 72-83, ISSN 0169-7439, <https://doi.org/10.1016/j.chemolab.2007.06.006>.
- [16] Jacob Murel, Eda Kavlakoglu, "What is ridge regression?", URL: [www.ibm.com/topics/ridge-regression](http://www.ibm.com/topics/ridge-regression).
- [17] "What is Ridge Regression?" URL: [www.ibm.com/topics/ridge-regression](http://www.ibm.com/topics/ridge-regression).