ÇOMÜ Zir. Fak. Derg. (COMU J. Agric. Fac.)
2025: 13 (1): 12-31
ISSN: 2147–8384 / e-ISSN: 2564–6826
doi: 10.33202/comuagri.1586063

Research Article

AI-Based Prediction of Microelement and Heavy Metal Contents in Central-Southern Anatolian Soils: A Pilot Study

Noyan Eken[1] Enes Efe[2] Kamer Yazar[3] Mehmet Hamurcu[3]
Fatma Gökmen Yilmaz[3] Sait Gezgin[3] Erdoğan Eşref Hakki[3*]

[1]Ege University, Department of Organic Agriculture, Ödemis Vocational Training High School, 35750 Izmir, Türkiye
[2]Hitit University, Department of Electrical and Electronics Engineering, 19169 Çorum, Türkiye
[3]Selcuk University, Department of Soil Science and Plant Nutrition, Faculty of Agriculture, 42280, Konya, Türkiye
*Corresponding author: eehakki@selcuk.edu.tr

Received Date: 15.10.2024 Accepted Date: 13.03.2025

Abstract

The estimation of total microelement and heavy metal concentrations in soil samples taken from the Central-Southern Anatolian Region of Türkiye was conducted using artificial intelligence models. The accurate prediction of microelement contents and heavy metal contents of soils is of importance for agricultural productivity and environmental health. A total of 62 soil samples were analyzed for Boron (B), Iron (Fe), Zinc (Zn), Manganese (Mn), Copper (Cu), Cadmium (Cd), Chromium (Cr), Nickel (Ni) and Lead (Pb). The artificial intelligence models used in this study were Random Forest (RF), Gradient Boosting (GB), and Support Vector Regressor (SVR). Model performance was evaluated based on Mean Absolute Error (MAE), Mean Squared Error (MSE) and R^2 scores. The best performance was achieved for the B and Cu contents, according to the results. In the case of B contents, the GB model provided the best results (MAE: 4.89, MSE: 28.01 and R^2: 0.55), while the RF model showed the highest performance for Cu predictions (MAE: 3.20, MSE: 16.80, and R^2: 0.75). In addition, the results indicate that the artificial intelligence models used in the present study hold promising potential for predicting microelement and heavy metal concentrations in soil samples.
Keywords: Prediction. Total micro element content, Total heavy metal content, Central Southern Anatolia Region soils, Artificial Intelligence

Yapay Zekâ ile Orta-Güney Anadolu Topraklarında Mikroelement ve Ağır Metal Tahmini: Pilot Bir Çalışma

Öz

Türkiye'nin Orta-Güney Anadolu Bölgesi'nden alınan toprak örnekleri üzerinde toplam mikro element ve toplam ağır metal konsantrasyonlarının tahmini yapay zekâ modelleri kullanılarak yapılmıştır. Toprakların mikro element ve ağır metal içeriklerinin doğru şekilde tahmin edilmesi tarımsal verimlilik ve çevre sağlığı açısından büyük önem taşımaktadır. Toplam 62 toprak örneğinde Bor (B), Demir (Fe), Çinko (Zn), Mangan (Mn), Bakır (Cu), Kadmiyum (Cd), Krom (Cr), Nikel (Ni) ve Kurşun (Pb) elementleri analiz edilmiştir. Bu çalışmada yapay zekâ modelleri olarak Random Forest (RF), Gradient Boosting (GB) ve Support Vector Regressor (SVR) kullanılmıştır. Modellerin performansı, Ortalama Mutlak Hata (MAE), Ortalama Kare Hatası (MSE) ve R^2 skorları ile değerlendirilmiştir. En iyi performans B ve Cu içeriklerinde elde edilmiştir. B içeriği tahmininde en iyi sonuç GB modeliyle sağlanmıştır (MAE: 4.89, MSE: 28.01, R^2: 0.55). Cu içeriği tahmininde ise RF modeli üstünlük göstermiştir (MAE: 3.21, MSE: 16.81, R^2: 0.75). Ayrıca sonuçlar, bu çalışmada kullanılan yapay zeka modellerinin toprak örneklerindeki mikroelement ve ağır metal konsantrasyonlarının tahmini için umut verici bir potansiyele sahip olduğunu göstermektedir.
Anahtar Kelimeler: Tahmin, Toplam mikro element içeriği, Toplam ağır metal içeriği, Orta Güney Anadolu Bölgesi toprakları, Yapay Zeka

Introduction

Soil is considered a vital resource for global food production and ecosystem balance. The sustainability of agricultural activities depends on the preservation and optimization of soil quality. The abundance of microelements and heavy metals in the soil directly affects plant growth and agricultural productivity. Although microelements are required by plants in trace amounts, their deficiency or excess can negatively impact crop quality and yield. Additionally, the use of agricultural chemicals and industrialization can lead to the accumulation of heavy metals in the soil, which poses a threat to both plant and human health (Khosravi et al., 2018; Nie et al., 2024). Accurate analysis of soil components is essential for improving agricultural productivity and maintaining ecosystem health. Predicting the chemical properties of soil offers valuable insights into nutrient availability and potential toxicities, forming a foundation for sustainable land management practices.

Gezgin et al. (2002) conducted a comprehensive analysis of soils from Central and Southern Anatolia, focusing on B content and its influencing factors. Their findings revealed significant variability in soil properties, including pH (8.59–9.4), electrical conductivity (0.047–3.34 mS cm^{-1}), lime content (0.17–69.1%), and organic matter content (0.15–8.8%). They also identified diverse soil texture classes, with 51% classified as fine, 44.5% as loamy, and 4.5% as coarse. Extractable microelement concentrations, such as Fe, Zn, Mn, and Cu, varied widely, while extractable B ranged from 0.01–63.9 mg kg^{-1}. Notably, 26.6% of the soils were boron-deficient (<0.5 mg kg^{-1}), whereas 9.9% exhibited toxic boron levels (>0.5 mg kg^{-1}), highlighting the need for targeted soil management strategies.

Similarly, Ozyazici et al. (2017) investigated heavy metal contamination in agricultural soils from the Central and Eastern Black Sea Region. Using geostatistics combined with GIS, they analyzed 3400 surface soil samples and quantified levels of Cd, Co, Cu, Ni, Pb, and Zn. Their results indicated that while Ni and Co concentrations exceeded threshold levels due to natural sources, other heavy metals remained below critical values. The average concentrations of heavy metals followed the order: Ni > Zn > Cu > Pb > Co > Cd. This study emphasized the importance of distinguishing between natural and anthropogenic sources of contamination to develop effective mitigation strategies.

Traditional soil analysis methods, while accurate, are often labor-intensive, time-consuming, and costly. These limitations highlight the potential of advanced data science techniques, such as artificial intelligence (AI) and machine learning (ML). By utilizing large datasets, these technologies can streamline soil analysis, predict microelement and heavy metal concentrations, and enable efficient large-scale assessments, offering a valuable tool for modern agricultural practices. AI models can learn from large datasets and make predictions regarding microelement and heavy metal contents based on soil characteristics. Recent studies have demonstrated that AI has proven to be an effective tool in the fields of agriculture and environmental sciences. For instance, (Shi et al., 2022) and (Khosravi et al., 2018) have shown that AI models integrated with spectral analysis yielded successful results in predicting heavy metal contamination. Random Forest (RF) and Gradient Boosting (GB) algorithms, in particular, are among the widely used models in soil analysis (Nie et al., 2024). However, as far as we know, studies using limited datasets remain scarce, and comprehensive research focusing on the prediction of microelements and heavy metals based on fundamental soil components is still lacking.

The existing literature highlights the use of various AI methods for predicting elements in agricultural soils. (Nie et al., 2024) successfully predicted the distribution of heavy metals such as Cr, Cd, Pb, and As in agricultural soils using the RF model. Similarly, (Shi et al., 2022) developed hybrid AI models to estimate heavy metal content in soils by leveraging remote sensing images. Additionally, (Khosravi et al., 2018) monitored heavy metal pollution using spectral data combined with machine learning techniques. (Luce et al., 2017) used visible near-infrared spectroscopy to predict heavy metal concentrations in soils contaminated with paper mill waste. These studies also underscore the challenges of working with limited datasets (Khosravi et al., 2018; Munnaf and Mouazen, 2021).

This study integrates laboratory analyses and advanced AI modeling to comprehensively evaluate the microelement and heavy metal contents of soils. By leveraging key soil physicochemical properties such as pH, electrical conductivity (EC), lime content, texture, and organic matter from the Central and Southern Anatolian regions, robust AI models were developed and validated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2. The primary objective of this research is to establish a reliable framework for predicting soil microelement and heavy metal concentrations through AI-based approaches. The findings not only contribute to precision agriculture

by enabling efficient and cost-effective soil management strategies tailored to the unique characteristics of these regions but also provide a foundation for future studies exploring the integration of AI in soil science and sustainable agriculture.

Materials and Method

Central and Southern Anatolia Geological Features

The study area, located in the intermediate zone of Central Anatolia, lies between the Northern Anatolian (Anatolides) and Southern Anatolian (Taurides) fold systems. This region features a diverse geomorphology, including basins, tablelands, isolated folds, volcanic zones, and ancient crystalline masses (Lahn, 1949). The central portion of this zone is defined by three major geological units: (1) the Anatolide folds, comprising the Ankara fan, Çankırı-İskilip ranges, and Çorum fan, delineating the western, northern, and northeastern boundaries; (2) the volcanic zone of Hasan Dağı in the southeast; and (3) the inner Tauride folds in the south. Highlands such as Kırşehir-Keskin and Boz Dağlar further divide the region into three depressions: the Middle Kızılırmak-Delice Irmak basin, the Tuz Gölü basin, and the Konya-Ereğli basin. Volcanic activity in the Karaman district, located in the southern portion of the Central Anatolian Plateau, occurred over four distinct phases during the Late Pliocene to Late Pleistocene. The oldest volcanic deposits, dated at 3.2 million years, are found in the Mercik area, followed by volcanic formations in Sızak, Kartallık, and Kızıldağ, dated at 1.95–2.05 million years. A significant caldera-forming event approximately 1.1 million years ago produced andesites, tuffs, pumice, and volcanic breccias, with base surge deposits observed on the caldera's northern flank. The youngest volcanic activity, producing late-stage andesites, is seen in Bozdağ and Değle Dağı. Additionally, the region includes pre-Neogene crystalline limestones and Neogene formations comprising conglomerates, sandstones, and limestones (Koç, 1987). Climatic conditions in the region reflect a semi-arid climate, with annual average temperatures of 11.7°C and 12.1°C for Konya and Karaman provinces, respectively, and long-term annual precipitation averages of 329.7 mm for Konya and 336.7 mm for Karaman (MGM, 2025).

Soil Sampling

A total of 62 soil samples, each weighing at least 1 kg, were collected through random sampling from various districts in the Central and Southern Anatolian regions, specifically representing the soils of Konya Province. The samples were taken from Seydişehir (10 samples), Karapınar (14 samples), Selçuklu and Meram (11 samples combined), Karatay (10 samples), Cihanbeyli (3 samples), Kulu (7 samples), Hüyük (2 samples), and Sarayönü (1 sample) districts. Additionally, 4 soil samples were collected from Karaman Province. These samples were taken homogeneously from a depth of 0–30 cm to represent the region and were subsequently analyzed for essential soil properties, including pH, electrical conductivity (EC), lime percentage, organic matter percentage, and texture. These variables were used to estimate the concentrations of microelements and heavy metals. This research focused on predicting microelement and heavy metal contents in soils from agricultural lands in the Central-Southern Anatolia region of Türkiye (Figure 1), covering approximately 324.61 ha in the Central Anatolia Region (Konya Province) and 47.22 ha in the Southern Anatolia Region (Karaman Province). To assess model performance, standard evaluation metrics such as MAE, MSE, and R^2 scores were applied. The results highlight the potential of AI techniques in predicting soil microelements and heavy metals, even with limited data. This demonstrates how AI can significantly improve soil quality monitoring and management, ultimately supporting more sustainable agricultural practices in the region. Furthermore, the integration of AI methods with soil analysis provides an innovative approach for better understanding and managing soil health. It offers valuable insights into sustainable farming practices in areas with diverse environmental conditions.

Figure 1. Location map of the study area

Soil Physico-Chemical Analyses

Upon arrival at the laboratory, the soil samples were air-dried at room temperature and sieved through a 2 mm mesh. The following key soil properties were analyzed:

- **pH:** The pH of each soil sample was measured using a potentiometric method by preparing a soil-water mixture (1:2.5 ratio) (Jackson, 1958).
- **Electrical Conductivity (dS cm^{-1}):** Electrical conductivity was determined using an EC meter, which measures the soil solution's ability to conduct electricity (Jackson, 1958).
- **Lime Content ($CaCO_3$ %):** The lime content was determined by adding hydrochloric acid (HCl) to the soil samples, and the percentage of calcium carbonate ($CaCO_3$) was calculated based on the amount of carbon dioxide (CO_2) (Hızalan and Ünal, 1966).
- **Organic Matter (%):** Organic matter content was quantified using the Smith-Weldon Method, which involves the oxidation of organic material by dichromate (Smith and Weldon, 1941).
- **Texture Class (%):** The texture analysis was determined according to the Bouyocous method (Gee, 1986).
- **Total Micro Elements and Heavy Metals Contents (mg kg^{-1}):** The total concentrations of trace elements and heavy metals were determined using the aqua regia extraction method and quantified with a Varian Vista ICP-OES instrument (Kick et al., 1980).

Artificial Intelligence Statical Analyses

The statistical analysis involved developing machine learning models to predict the microelement and heavy metal contents of the soil based on its physicochemical properties. The independent variables (X) included pH, EC, lime percentage, organic matter percentage, and soil texture, while the dependent variables (y) were the concentrations of microelements (B, Fe, Zn, Mn, Cu) and heavy metals (Cd, Cr, Ni, Pb).

Data Normalization: Data normalization is an essential preprocessing step in machine learning to ensure that all features contribute equally to the model. In this study, we employed the StandardScaler method, which transforms the features by subtracting the mean and dividing by the standard deviation, resulting in data with a mean of 0 and a standard deviation of 1. This scaling process is crucial, especially for machine learning algorithms like Support Vector Regressor (SVR), which are sensitive to the magnitude of feature values. Without normalization, features with larger magnitudes could dominate the learning process, leading to biased predictions. By applying this technique, we ensured that the model treated each soil property (pH, EC, lime, and organic matter) equally, improving the overall model accuracy and stability.

Training and Testing Data: In this study, an 80:20 data split was applied, dividing the dataset into training (80%) and testing (20%) sets (Gholamy et al., 2018). The training data was used to train the machine learning models, while the testing data was held out to evaluate the model's generalization performance on unseen data. This method ensures that the model learns from the training data and

generalizes well to new data, thereby preventing overfitting. According to (Hastie et al., 2009), partitioning the dataset in this way is a standard approach in statistical learning and is critical to achieving a robust model evaluation. (Kohavi, 1995) also emphasized that separating training and testing data is vital for ensuring unbiased performance estimates in machine learning models. By utilizing a separate testing set, the model's ability to make predictions on data it has not encountered before is assessed, providing a clearer picture of its generalization capabilities.

Machine Learning Models: To model the relationships between soil properties and the concentrations of microelements (B, Fe, Zn, Mn, and Cu) and heavy metals (Cd, Cr, Ni, and Pb), we employed three different machine learning algorithms:

Random Forest: Random Forest (RF) is a robust ensemble learning algorithm introduced by (Breiman, 2001). It addresses the limitations of decision trees, such as overfitting and high variance, by constructing multiple decision trees and aggregating their outputs. Each decision tree in the forest is trained on a random subset of the data through a process known as bootstrap sampling, where samples are drawn with replacement to ensure diversity among the trees. At each node within the trees, a random subset of features is considered during the split—a technique referred to as random feature selection. This approach reduces correlation between the trees, enhances generalization, and minimizes overfitting.

For regression tasks, the final prediction $\hat{\boldsymbol{y}}$ is obtained by averaging the predictions from all decision trees:

$$\hat{y} = \frac{1}{m}\sum_{i=1}^{m} T_i(X) \quad (1)$$

Where $\boldsymbol{T_i(X)}$ is the prediction from the $\boldsymbol{i}$-th tree, and $\boldsymbol{m}$ is the total number of trees in the forest. For classification tasks, the final class prediction $\hat{\boldsymbol{y}}$ is determined by the majority vote across the trees:

$$\hat{y} = mode(T_1(X), T_2(X), \ldots . T_m(X)) \quad (2)$$

To evaluate the best splits at each node, the Gini impurity for classification is commonly minimized, defined as:

$$I_G(t) = 1 - \sum_{i-1}^{k} p_i^2 \quad (3)$$

Where $\boldsymbol{p_i}$ is the proportion of class $\boldsymbol{i}$ at node $\boldsymbol{t}$, and $\boldsymbol{k}$ is the number of classes. In the case of regression, splits are chosen to minimize the MSE. This ensemble method improves both accuracy and generalization of the model by combining predictions from multiple trees, making it robust against overfitting and highly effective for a wide range of tasks.

Gradient Boosting: Gradient Boosting (GB) is an iterative machine learning algorithm used for both regression and classification tasks. Its primary goal is to build a model by sequentially adding weak learners, typically decision trees, to correct the errors made by previous trees. GB aims to minimize a loss function using gradient descent, which measures the difference between the actual and predicted values (Sigrist, 2021).

The algorithm starts with an initial model, usually a simple prediction (like the mean for regression), and then adds trees iteratively to improve predictions. At each step, the residuals (errors from the previous model) are used to fit a new decision tree, with the goal of reducing these errors. The new tree is added to the existing model with a weight controlled by a learning rate, which prevents overfitting by limiting how much each tree contributes.

The core equation for GB is:

$$\mathrm{F}_{m+1}(x) = \mathrm{F}_m(x) + \eta . h_m(x) \quad (4)$$

Where:

- $\mathbf{F_m(x)}$ is the prediction from the model at the $\boldsymbol{m}$-th iteration.
- $\boldsymbol{h_m(x)}$ is the new decision tree fitted to the residuals.
- $\boldsymbol{\eta}$ is the learning rate, a hyperparameter that controls how much of the new tree's prediction is added to the model.

The algorithm minimizes a specified loss function $\boldsymbol{L(y, \hat{y})}$, such as MSE for regression or log loss for classification. The residuals used to fit the new tree are the negative gradients of this loss function with respect to the model's predictions. In mathematical terms, for each data point $\boldsymbol{i}$, the residuals are:

$$r_i(x) = -\frac{\partial \mathrm{L}(y_i, \mathrm{F}_m(x_i))}{\partial \mathrm{F}_m(x_i)} \quad (5)$$

Thus, the new decision tree $\boldsymbol{h_m(x)}$ is fitted to these residuals to minimize the error at each step (Ke et al., 2017).

Support Vector Regressor (SVR): One of the machine learning models used in this study is the SVR, a powerful algorithm designed to address regression problems. SVR is built on the foundational principles of the Support Vector Machine (SVM) algorithm, which is commonly used for classification tasks, but extended here to predict continuous values (Awad et al., 2015). SVR is particularly useful for modeling both linear and non-linear relationships between input features and the target variable, utilizing support vectors to form the regression line.

The main objective of SVR is to learn the underlying patterns in the data while minimizing errors within a certain margin of tolerance, known as the epsilon-insensitive loss function. Unlike traditional regression approaches that minimize the squared error, SVR ignores small deviations within a defined epsilon margin and focuses only on larger deviations outside this margin. This approach allows SVR to balance accuracy and model simplicity.

SVR is capable of handling non-linear relationships through the use of kernel functions. In this study, we employed popular kernels such as linear, polynomial, and the Radial Basis Function (RBF). These kernels transform the input data into a higher-dimensional space, where non-linear relationships can be more easily modeled with a linear decision boundary.

The regularization parameter (C) controls the trade-off between the model's complexity and its tolerance to errors. A smaller value of (C) allows for a simpler model with greater tolerance for errors, while a larger (C) places more emphasis on minimizing errors, potentially leading to overfitting.

The epsilon parameter defines a margin of tolerance around the predicted function, within which deviations from the true target values are ignored. By tuning epsilon, the model can be adjusted to disregard small variations in the data, resulting in a more generalized model.

SVR performs exceptionally well with high-dimensional datasets, as it focuses on the most informative data points (support vectors) to construct the model. However, the performance of SVR is highly sensitive to the choice of kernel function and the tuning of hyperparameters such as (C) and the epsilon parameter. Additionally, SVR can be computationally expensive when applied to large datasets due to the complexity of solving quadratic optimization problems.

In this study, SVR was employed to model the relationships between the chemical properties of soil samples (pH, EC, organic matter, and lime percentage) and the concentrations of microelements and heavy metals. The results indicate that SVR effectively captures both linear and non-linear relationships, significantly improving prediction accuracy (Awad et al., 2015).

Hyperparameter Tuning: To enhance the predictive accuracy of each machine learning model, GridSearchCV was employed for hyperparameter optimization. GridSearchCV systematically tests multiple combinations of hyperparameters to identify the set that maximizes model performance (Bergstra and Bengio, 2012). This technique is crucial in machine learning, as the choice of hyperparameters significantly influences model accuracy, robustness, and generalization capabilities.

For instance, in the RF model, critical hyperparameters such as the number of trees (n_estimators) and the maximum depth of each tree (max_depth) were adjusted to improve performance. Similarly, for the GB model, hyperparameters such as the learning rate and the number of boosting stages were fine-tuned to enhance predictive capabilities. In each model, a wide range of values for each parameter was tested.

The GridSearchCV method utilizes cross-validation to evaluate model performance across different subsets of the data. This ensures that the model's predictive accuracy is consistent and not biased toward any particular subset of the training data (Hastie et al., 2009). Cross-validation also helps prevent overfitting, where a model performs well on the training data but fails to generalize to unseen data. Through this method, we ensured that the models developed were robust and generalized well across varying data samples.

Data Augmentation Techniques: Given the relatively small size of the dataset used in this study, data augmentation techniques were applied to artificially expand the training data and improve the model's generalization ability. These techniques create additional training samples by introducing slight variations to the existing data, enabling the model to learn more robust patterns and reduce the risk of overfitting. The following augmentation methods were employed:

- **Random Sampling:** This technique generates new samples by adding small amounts of noise to the original data. By slightly perturbing the feature values, this method increases the variety of data points, helping the model better capture variability and learn from diverse scenarios (Hastie et al., 2009).
- **Jittering:** Jittering involves adding random Gaussian noise to the input features, creating slightly altered versions of the original samples. This prevents the model from memorizing specific patterns in the data and enhances its ability to generalize to unseen data (Geman et al., 1992). Jittering is particularly useful for models prone to overfitting when dealing with a limited number of data points.
- **Bootstrap Sampling:** This method involves creating new training sets by sampling with replacement from the original dataset. Bootstrap sampling enables the construction of multiple variations of the training data, allowing the model to learn from diverse subsets (Tibshirani and Efron, 1993). It has been widely used in ensemble methods to enhance robustness by training the model on different subsets of data, thereby reducing overfitting and improving generalization.

These augmentation techniques were crucial in improving the performance of the machine learning models in this study, allowing them to learn from limited data and make more accurate and robust predictions.

Evaluation Metrics and Performance Criteria: In this study, various evaluation metrics were used to assess the performance of the machine learning models employed for predicting the concentrations of microelements and heavy metals in soil samples. The chosen metrics were critical for understanding the model's accuracy, error rate, and overall predictive capability. These metrics are widely used in regression problems and offer a comprehensive view of the model's performance across various aspects (Wang et al., 2022).

The following performance metrics were used:

The Mean Absolute Error (MAE) measures the average magnitude of the errors between predicted values $\hat{\boldsymbol{y}}$ and actual values $\boldsymbol{y}$, without considering the direction of the errors. It provides a straightforward interpretation of the model's accuracy by averaging the absolute differences between predicted and observed values. A lower MAE indicates better model performance.

The formula for MAE is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

where:

- n is the number of observations,
- $\boldsymbol{y_i}$ represents the actual value, and
- $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ represents the predicted value.

MAE gives a clear insight into the model's accuracy by showing the average error magnitude, without amplifying larger errors as is the case with squared errors.

The Mean Squared Error (MSE) is another common metric for evaluating regression models. Unlike MAE, MSE squares the error terms, which places greater emphasis on larger errors. This makes it a more sensitive metric to large deviations from the true values.

The formula for MSE is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (7)$$

MSE provides an overall view of the error distribution by penalizing larger deviations, which can highlight where the model struggles with extreme values. A lower MSE indicates better model performance, with zero indicating a perfect fit.

The R-squared (R^2) metric, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables. R^2 provides an indication of the goodness-of-fit of the model. A value closer to 1 indicates that the model explains a large proportion of the variance, while values closer to 0 indicate that the model fails to explain much of the variability in the data.

The formula for R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2} \quad (8)$$

Where $\hat{\boldsymbol{y}}$ is the mean of the actual values.

An R^2 value of 1 represents a perfect fit, meaning that the model captures all the variance in the data. However, R^2 can sometimes give misleading results in the presence of overfitting or for non-linear models, which is why it is used alongside other metrics like MAE and MSE.

Results and Discussion

Phsico-Chemical Properties of Soil

In addition to the physico-chemical properties of the soil samples, the maximum, minimum, and mean values of the total microelement and heavy metal contents are presented in Table 1.

Table1. The summary statistics of the soil physico-chemical properties analyzed

Regions from Which Soil Samples Were Collected	Soil Sample Code	pH	EC (dS cm^{-1})	$CaCO_3$ (%)	Organic Matter (%)	Texture Class
	T1	7.60	0.26	1.39	1.43	CL
	T2	7.40	0.19	1.39	1.20	CL
	T3	7.80	0.11	1.39	1.22	CL
	T4	7.10	0.20	1.39	1.48	CL
	T5	7.10	0.22	1.39	1.26	CL
	T6	7.30	0.28	1.39	0.62	CL
	T7	7.90	0.25	1.40	1.07	
Central	**T8**	7.90	0.14	3.49	1.24	CL
Anatolia	**T9**	8.00	0.11	2.10	0.93	L
Region	**T10**	6.50	0.15	1.39	1.37	CL
	T11	7.90	0.10	35.77	1.27	CL
	T12	7.60	0.15	34.37	1.59	CL
	T13	7.90	0.18	21.04	1.05	CL
	T14	8.00	0.23	27.36	1.20	L
	T15	8.00	0.17	35.07	1.20	CL
	T16	8.00	0.10	30.86	1.21	CL
	T17	8.00	0.21	23.85	1.71	CL
	T18	8.00	0.17	49.18	1.28	CL
	T19	8.00	0.18	28.93	1.39	CL
	T20	7.90	0.24	32.27	1.76	L
	T21	8.10	0.17	37.88	1.44	L
	T22	7.80	0.39	23.15	1.03	L
	T23	7.80	0.23	13.33	1.43	CL
	T24	8.00	0.10	14.73	0.94	CL
	T25	8.00	0.29	5.01	1.37	CL
	T26	7.90	0.19	5.72	0.88	C
	T27	8.00	0.24	22.2	0.64	CL
	T28	7.80	0.20	13.5	1.18	CL
	T29	7.80	0.16	16.19	0.70	CL
	T30	8.00	0.12	12.2	0.86	CL
Central	**T31**	8.10	0.16	8.58	1.20	CL
Anatolia	**T32**	8.00	0.15	18.6	1.27	CL
Region	**T33**	8.00	0.22	15.7	1.41	CL
	T34	7.90	0.21	7.87	1.07	CL
	T35	8.20	0.11	36.50	0.84	CL
	T36	8.02	0.34	41.50	1.66	CL
	T37	8.10	0.19	42.20	1.24	CL
	T38	7.90	0.36	44.30	1.14	CL

Table 1. cont.	**T39**	8.10	0.12	44.30	1.29	CL
	T40	8.00	0.16	41.50	1.30	CL
	T41	8.10	0.08	49.30	1.01	CL
	T42	7.90	0.13	35.00	1.46	CL
	T43	8.00	0.12	16.50	1.07	CL
	T44	8.00	0.16	12.20	1.20	CL
	T45	8.00	0.18	8.58	1.31	CL
	T46	7.90	0.21	35.00	1.08	CL
	T47	7.80	0.33	4.29	0.79	L
	T48	7.80	0.17	7.15	1.47	CL
	T49	7.90	0.10	4.29	1.29	L
	T50	7.80	0.19	7.15	1.19	CL
Central Anatolia Region	**T51**	7.90	0.16	8.57	1.01	CL
	T52	7.90	0.19	8.57	1.07	CL
	T53	7.80	0.17	10.70	0.90	L
	T54	7.90	0.15	9.29	0.93	CL
	T55	7.80	0.18	7.86	1.01	CL
	T56	7.60	0.10	2.08	1.26	CL
	T57	7.40	0.09	1.39	1.33	CL
	T58	8.10	1.90	23.00	2.20	SL
	T59	7.80	0.22	7.63	1.31	CL
Southern Anatolia Region	**T60**	7.80	0.19	20.8	1.26	CL
	T61	7.80	0.14	9.71	1.15	CL
	T62	8.10	0.16	3.47	1.37	CL
Minimum Value		6.50	0.08	1.39	0.62	
MaximumValue		8.20	1.90	49.3	2.20	
Average		**7.84**	**0.21**	**17.63**	**1.21**	

Abbrevations.:C:Clay, L:Loam, CL:Clay Loam, SL:Sandy Loam

The pH values of the soils sampled from the Central and Southern Anatolia regions ranged from 6.50 to 8.20. It was determined that 90.32% of the soils were mildly alkaline, while 9.68% were neutral. The EC values of the soils ranged from 0.08 to 1.90 dS/cm, with 88.71% classified as highly saline and 11.29% as very saline. The lime content of the soils ranged from a minimum of 1.39% to a maximum of 49.3%, with 25.81% classified as calcareous, 32.26% as moderately calcareous, 14.52% as highly calcareous, and 27.42% as very highly calcareous. The organic matter content ranged from 0.62% to 2.20%, with 17.74% classified as very low, 80.65% as low, and 1.62% as moderate. Texture analysis revealed that 82.26% of the soils were clay loam, 14.52% were loam, and 1.61% were sandy loam and clay. In a study by (Gezgin et al. (2002), the pH values of soils in the Central Anatolia region were similarly reported as alkaline. The EC values indicated that 13.5% of the soils were classified as highly saline. Lime content in their study was distributed as 13.6% lightly calcareous, 53.3% calcareous, and 33.1% highly calcareous. Regarding organic matter content, 87.5% of the soils had low levels, while 12.5% had adequate organic matter. Soil texture was reported as 51% fine-textured, 44.5% loamy, and 4.5% coarse-textured. A comparison of the results reveals notable differences between the present study and those of (Gezgin et al. (2002). While both studies agree on the alkaline nature and lime content of the soils, this study shows significantly higher salinity levels, likely due to recent changes in climate and agricultural practices. The increased salinity may be attributed to climate change, improper fertilization, and unsustainable farming practices over the years. Additionally, the low organic matter content observed in both studies underscores the negative impact of these practices on soil health. These findings highlight the urgent need to adopt sustainable farming methods that address soil salinity and organic matter depletion to maintain soil quality and productivity in the long term.

Total Microelement and Total Heavy Metal Contents of the Soils

In addition to the total macroelement, microelement, and heavy metal contents of the soil samples, the maximum, minimum, and mean values are also presented in Table 2.

Table2. The summary statistics of the total microelements and heavy metals in the soil are presented based on the analysis.

Regions fromWhich Soil Samples Were Collected	Soil Sample Code	Micro Elements mg kg^{-1}						Heavy Metals mg kg^{-1}			
		B	Fe	Zn	Mn	Cu	Cd	Cr	Mo	Ni	Pb
	T1	39.09	28702.06	69.99	681.45	32.59	0.57	41.05	0.75	27.86	173.88
	T2	42.47	29491.09	72.38	932.36	42.92	0.56	67.34	1.56	51.68	270.01
	T3	46.63	34286.72	76.63	750.98	22.10	0.60	28.50	0.39	18.24	79.80
	T4	35.50	25553.44	56.74	529.98	22.38	0.52	32.13	0.77	20.07	124.24
	T5	38.79	26318.76	63.52	534.49	24.56	0.20	34.06	0.70	23.83	146.89
	T6	48.64	26793.83	69.90	473.65	28.49	0.54	55.78	1.45	34.40	166.36
	T7	43.01	26919.50	80.72	485.04	36.15	0.59	40.38	1.50	25.72	117.69
Central Anatolia	T8	26.89	19926.86	43.76	303.27	11.05	0.27	20.24	0.52	13.18	38.05
Region	T9	44.92	31848.74	77.04	639.77	18.93	0.69	23.99	0.10	13.89	84.78
	T10	42.49	32314.10	61.40	786.01	25.39	0.71	39.43	1.17	22.08	138.27
	T11	30.89	14209.91	40.04	306.61	15.48	0.34	58.12	0.66	44.31	68.30
	T12	36.98	17179.08	48.22	376.53	24.96	0.32	69.46	0.55	55.71	82.39
	T13	48.17	20169.91	68.28	484.64	31.70	0.49	72.97	0.38	67.44	94.95
	T14	38.91	17060.75	53.13	378.09	23.29	0.60	55.19	0.15	52.42	85.39
	T15	32.24	12072.04	29.38	250.56	14.72	0.26	39.84	0.43	37.66	49.02
	T16	35.77	17213.84	48.47	376.15	21.71	0.31	52.01	0.41	51.07	80.77
	T17	42.00	19300.58	50.47	277.25	21.14	0.30	57.48	0.45	51.99	88.50
	T18	29.55	11334.03	36.53	326.33	15.24	0.09	35.23	0.94	32.00	52.98
	T19	45.92	18961.87	60.06	482.84	26.10	0.49	63.56	0.19	57.60	105.20
	T20	42.87	16509.16	53.66	393.85	21.58	0.36	49.32	1.59	48.42	82.54
	T21	41.87	16970.93	51.93	403.77	23.49	0.50	49.80	1.62	47.83	82.62
	T22	46.61	18356.38	56.47	469.87	29.15	0.41	56.59	0.87	57.10	94.17
	T23	48.95	20652.37	63.54	520.41	27.35	0.57	62.10	1.27	65.77	97.99
	T24	55.64	24569.53	76.74	682.21	37.06	0.51	75.85	0.58	87.86	143.75
	T25	58.06	29005.27	168.25	961.38	46.03	0.64	204.05	0.33	308.79	133.24
Central Anatolia	T26	46.85	27279.78	83.37	779.31	31.01	0.85	204.83	1.20	270.60	131.40
Region	T27	37.71	19982.59	79.00	479.52	23.21	0.48	91.92	0.49	129.41	93.38
	T28	43.26	25194.24	75.60	774.71	32.54	0.65	87.55	0.94	82.74	123.39
	T29	34.57	19385.96	54.04	464.71	20.67	0.27	42.53	0.76	37.86	104.67
	T30	40.24	23186.58	59.85	496.52	22.76	0.52	34.28	0.37	29.09	65.10
	T31	71.03	26740.35	102.13	752.06	41.94	0.72	87.59	1.57	54.68	1210.81
	T32	41.59	18750.10	65.13	680.55	31.38	0.48	53.59	0.63	47.35	143.36
	T33	61.94	23505.50	94.06	747.09	32.40	0.55	73.83	1.02	59.27	185.55
	T34	52.51	25385.70	72.52	708.26	32.11	0.60	87.03	0.33	63.80	188.08
	T35	37.99	15262.54	64.35	360.55	19.85	0.28	44.60	0.58	39.57	79.64
	T36	48.31	12307.52	45.20	336.89	21.63	0.20	33.51	0.54	34.42	76.88
	T37	40.76	15045.98	46.48	370.52	18.13	0.36	40.11	1.25	43.15	79.40
	T38	29.77	10405.22	32.40	266.32	14.72	0.33	28.23	0.84	29.27	45.28
	T39	31.05	12350.35	33.56	315.38	17.43	0.24	35.50	0.16	32.66	59.78
	T40	35.57	13853.99	37.10	322.01	14.79	0.28	39.64	0.00	37.55	48.39
	T41	29.07	9814.94	28.17	254.79	15.19	0.28	26.99	0.81	24.87	66.34
	T42	40.89	15790.39	49.65	386.28	23.29	0.43	46.60	0.85	41.96	86.59
	T43	50.35	20580.58	64.82	481.78	26.45	0.33	68.18	0.72	60.85	114.76
	T44	53.72	23020.15	71.64	582.87	29.41	0.45	81.13	0.82	71.02	103.05
	T45	55.60	23611.77	79.98	629.53	33.71	0.60	84.80	1.01	82.03	124.19
Central Anatolia	T46	42.78	12079.60	58.12	304.38	18.29	0.15	37.72	0.51	36.28	69.50
Region	T47	57.25	20036.12	57.34	511.45	28.91	0.31	70.13	0.36	70.38	143.21
	T48	52.56	24675.46	59.17	686.44	34.90	0.45	123.74	0.45	114.81	105.72
	T49	50.32	23088.50	61.39	612.34	33.08	0.56	95.06	0.61	89.61	104.19
	T50	52.79	28578.70	72.19	734.90	47.69	0.62	182.29	0.28	166.41	72.85
	T51	55.05	27914.73	75.54	738.95	43.69	0.68	167.24	0.83	164.56	80.87
	T52	50.83	25995.88	67.21	605.05	38.42	0.53	137.10	0.44	129.31	140.23
	T53	45.77	23448.25	58.85	619.74	35.89	0.38	101.08	0.36	94.19	96.68
	T54	48.22	25806.45	60.93	650.41	42.89	0.33	154.74	0.78	127.16	58.03
	T55	43.88	26199.80	63.76	613.84	46.63	0.49	238.28	0.44	192.36	54.08
	T56	45.33	28344.04	86.59	676.49	27.85	0.58	61.57	1.01	43.34	161.70
	T57	42.89	27785.50	115.08	752.11	29.09	0.85	62.01	2.49	42.18	298.44
	T58	72.22	14852.05	126.95	428.22	55.79	4.54	70.83	2.21	69.56	2591.52
Southern Anatolia	T59	60.28	26818.84	81.61	812.41	35.52	0.67	74.09	0.60	67.24	165.76
Region	T60	36.67	19364.32	55.51	527.88	26.44	0.57	54.14	0.28	52.40	94.05

Table 2. cont.

T61	39.27	17438.66	56.71	503.76	27.12	0.40	63.95	0.82	68.83	81.75
T62	49.19	25310.12	82.97	638.91	41.49	0.61	58.40	0.62	64.24	88.63
Minimum Value	26.89	9814.94	28.17	250.56	11.05	0.09	20.24	0.00	13.18	38.05
Maximum Value	72.22	34286.72	168.25	961.38	55.79	4.54	238.28	2.49	308.79	2591.52
Average	**44.69**	**21530.84**	**65.42**	**538.94**	**28.39**	**0.53**	**71.44**	**0.76**	**67.45**	**164.82**

The total microelement contents of soil samples from the Central and Southern Anatolia regions exhibited a wide range of variability. B content ranged from 26.89 mg kg^{-1} to 72.22 mg kg^{-1}, while Fe levels spanned from 9,814.94 mg kg^{-1} to 34,286.72 mg kg^{-1}. Zn content varied between 28.17 mg kg^{-1} and 168.25 mg kg^{-1}, Mn ranged from 250.56 mg kg^{-1} to 961.38 mg kg^{-1}, and Cu levels ranged from 11.05 mg kg^{-1} to 55.79 mg kg^{-1}. When classified according to (Taylor, 1964) and (Taylor and McLennan, 1985), all soil samples (100%) were found to have sufficient B levels, exceeding the threshold of 10 mg kg^{-1}. However, Fe content in all samples was below the sufficient level of 44,100 mg kg^{-1}, indicating a deficiency. Zn deficiency was observed in 82% of the soils (levels below 78.89 mg kg^{-1}), while 18% had adequate Zn levels. Mn levels were insufficient in all soil samples, falling below the threshold of 1,900 mg kg^{-1}. Regarding Cu, 37% of the soils were deficient, while 63% had sufficient levels. The total heavy metal contents of the soils also showed significant variability. Cd content ranged from 0.09 mg kg^{-1} to 4.54 mg kg^{-1}, Cr varied from 20.24 mg kg^{-1} to 238.28 mg kg^{-1}, and Mo was detected between 0.00 mg kg^{-1} and 2.49 mg kg^{-1}. Ni ranged from 13.18 mg kg^{-1} to 308.79 mg kg^{-1}, while Pb ranged from 38.05 mg kg^{-1} to 2,591.52 mg kg^{-1}. According to the Soil Pollution Parameters Regulation published by (T.O.B., 2010), 98% of the soils had cadmium concentrations below the threshold of 3 mg kg^{-1}, with only 2% exceeding the toxicity threshold. Cr concentrations were below the 100 mg kg^{-1} threshold in 86% of the samples, while 14% exceeded it. Mo concentrations were below the 10 mg kg^{-1} threshold in all samples, indicating no toxicity. Ni concentrations exceeded the 75 mg kg^{-1} threshold in 23% of the samples, while 77% were below this level. Pb content was below the 300 mg kg^{-1} threshold in 97% of the soils, with only 3% exceeding the toxicity level. These findings provide a comprehensive assessment of microelement and heavy metal contents in soils from the Central Anatolia region, with significant implications for agriculture and environmental management. In comparison, the findings of (Gezgin et al., 2002) highlighted micronutrient deficiencies, particularly in B (26.6% of soils), Fe (86.3%), and Zi (61.0%), suggesting potential limitations in soil fertility that could negatively impact nutrient-sensitive crops. Similarly, (Günal et al., 2012) reported substantial variability in heavy metal concentrations, including Ni, Pb, Cd, Co, and Cr, and noted correlations with soil properties influenced by natural (geogenic) factors and human activities (anthropogenic). (Günal et al., 2022) emphasized the role of soil structure in land suitability for wheat cultivation, demonstrating the utility of geostatistical methods for generating suitability maps. Key factors such as soil pH, electrical conductivity, and organic matter significantly influenced soil fertility, crop yield, and heavy metal uptake. (Ozyazici et al., 2017) contributed to understanding heavy metal contamination by identifying that some soils exceeded permissible limits for Ni and cobalt. They observed that the generally acidic soils (pH 4.5–5.5) enhanced metal solubility and bioavailability, with agricultural practices, such as excessive fertilization, further contributing to increased levels of cadmium, copper, and zinc. Our study confirms the wide variability in microelement and heavy metal concentrations in soils from the Central and Southern Anatolia regions, reflecting both natural and anthropogenic influences. High levels of Cr, Ni, and Pb, alongside deficiencies in Fe, Zn, and Mn, highlight the need for targeted soil management strategies. Metal solubility, driven by factors like soil pH and organic content, plays a critical role in bioavailability and toxicity. The findings align with previous research, such as (Ozyazici et al., 2017), emphasizing the need for integrated soil management approaches that consider metal solubility dynamics and interactions between geogenic and anthropogenic factors. Regular monitoring, sustainable agricultural practices, and careful fertilizer management are essential for mitigating contamination risks and promoting sustainable land management.

Machine Learining Applications in Central and Southern Anatolia Soils

The performance of the machine learning models was evaluated using MAE, MSE, and R^2 score. The results are presented in Table 3 and visualized in Figures 2–11 for each element (B, Fe, Zn, Mn, Cu, Cd, Cr, Mo, Ni, Pb).

Table 3. Performance of Machine Learning Models with Different Data Augmentation Techniques for Predicting Microelements and Heavy Metals in Soil Samples

Elements	Augmentation Technique	Models	MAE	MSE	R^2 Score
B	Random Sampling	RF	6,23	45,84	0,26
		GB	6,15	44,96	0,28
		SVR	6,76	72,93	0,01
	Jittering	RF	5,17	34,39	0,45
		GB	6,15	45,29	0,27
		SVR	6,76	72,88	0,01
	Bootstrap Sampling	RF	5,10	33,97	0,45
		GB	**4,89**	**28,01**	**0,55**
		SVR	6,94	80,37	0,01
Fe	Random Sampling	RF	**15,91**	**42,90**	**0,87**
		GB	16,81	70,59	0,78
		SVR	50,06	325,22	0,00
	Jittering	RF	17,38	62,25	0,81
		GB	17,61	58,14	0,82
		SVR	50,09	325,41	0,00
	Bootstrap Sampling	RF	20,75	80,23	0,75
		GB	20,87	82,48	0,75
		SVR	50,12	331,65	0,01
Zn	Random Sampling	RF	19,33	702,94	0,01
		GB	26,09	1793,14	0,01
		SVR	9,97	145,20	0,09
	Jittering	RF	17,82	681,78	0,01
		GB	19,08	1026,01	0,01
		SVR	**9,96**	**144,49**	**0,10**
	Bootstrap Sampling	RF	24,75	1504,91	0,01
		GB	25,76	1657,38	0,01
		SVR	10,02	156,70	0,02
Mn	Random Sampling	RF	**6,55**	**95,86**	**0,53**
		GB	8,97	152,93	0,25
		SVR	10,13	147,03	0,27
	Jittering	RF	8,33	141,39	0,30
		GB	8,44	180,08	0,11
		SVR	10,14	147,05	0,27
	Bootstrap Sampling	RF	10,28	337,09	0,01
		GB	13,71	481,35	0,01
		SVR	8,83	116,61	0,42
Cu	Random Sampling	RF	**3,21**	**16,81**	**0,75**
		GB	3,64	25,88	0,62
		SVR	4,46	33,19	0,51
	Jittering	RF	3,23	18,07	0,73
		GB	5,00	48,08	0,29
		SVR	4,41	32,47	0,52
	Bootstrap Sampling	RF	4,31	33,27	0,51
		GB	4,58	43,40	0,36
		SVR	4,62	35,31	0,48
Cd	Random Sampling	RF	0,10	0,02	0,01
		GB	0,13	0,02	0,01
		SVR	**0,07**	**0,01**	**0,44**
	Jittering	RF	0,10	0,02	0,01
		GB	0,12	0,02	0,01
		SVR	0,07	0,01	0,44
	Bootstrap Sampling	RF	0,11	0,02	0,01
		GB	0,12	0,02	0,01
		SVR	0,09	0,01	0,01

Table 3. cont.

Cr	Random Sampling	RF	37,02	235,18	0,01
		GB	27,71	172,33	0,12
		SVR	25,53	158,12	0,19
	Jittering	RF	34,16	197,84	0,01
		GB	34,35	237,16	0,01
		SVR	25,50	158,16	0,19
	Bootstrap Sampling	RF	44,60	413,52	0,01
		GB	40,77	441,56	0,01
		SVR	**25,34**	**150,79**	**0,23**
Mo	Random Sampling	RF	0,40	0,21	0,01
		GB	0,50	0,40	0,01
		SVR	0,33	0,15	0,01
	Jittering	RF	0,36	0,16	0,01
		GB	0,39	0,18	0,01
		SVR	0,33	0,15	0,01
	Bootstrap Sampling	RF	**0,34**	**0,14**	**0,01**
		GB	0,44	0,26	0,01
		SVR	0,50	0,38	0,01
Ni	Random Sampling	RF	47,72	405,85	0,01
		GB	36,76	300,01	0,01
		SVR	27,04	154,05	0,15
	Jittering	RF	44,85	382,45	0,01
		GB	41,07	473,20	0,01
		SVR	27,05	154,87	0,15
	Bootstrap Sampling	RF	60,00	820,83	0,01
		GB	46,91	721,99	0,01
		SVR	**27,39**	**148,29**	**0,18**
Pb	Random Sampling	RF	52,89	1154,50	0,01
		GB	54,11	1432,36	0,01
		SVR	29,91	144,91	0,11
	Jittering	RF	47,84	863,68	0,01
		GB	59,63	1195,52	0,01
		SVR	29,95	144,26	0,11
	Bootstrap Sampling	RF	60,66	1640,73	0,01
		GB	102,85	5403,52	0,01
		SVR	**28,16**	**118,17**	**0,27**

The best performance for B was achieved by GB using Bootstrap Sampling, with an MAE of 4.89 and an R^2 score of 0.55 (Figure 2). These results indicate good predictive capability, as the model successfully reduced the error while capturing the variance in B concentrations.

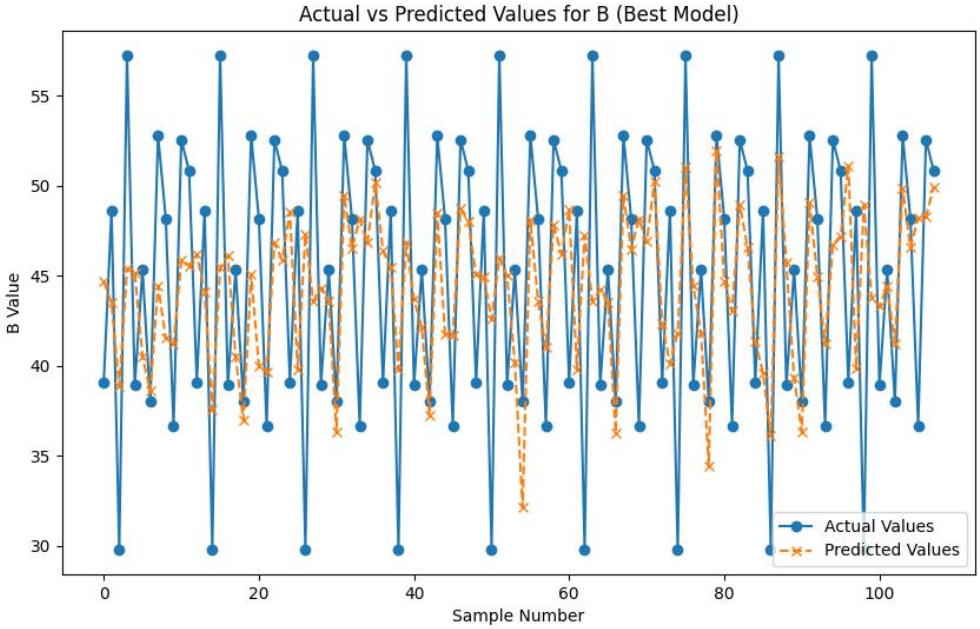

Figure 2. Actual vs Predicted Values for B using the Best Model

For Fe, RF with Random Sampling provided the best performance, achieving an MAE of 15.91 and an R^2 score of 0.87 (Figure 3). The high R^2 score indicates that the model was highly accurate in capturing the variability in iron concentrations, although the MAE remains relatively high due to the large magnitude of the values.

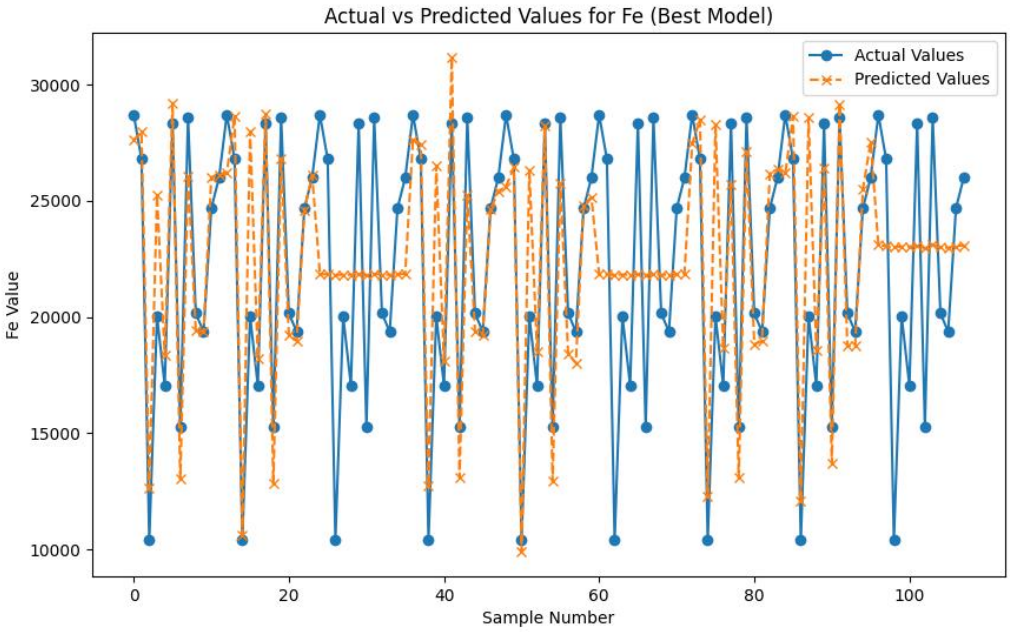

Figure 3. Actual vs Predicted Values for Fe using the Best Model

The best model for Zn was SVR combined with Jittering, yielding an MAE of 9.96 and an R^2 score of 0.10 (Figure 4). Despite the relatively low R^2 score, the model effectively minimized prediction errors, indicating its ability to capture the general trends in Zinc concentrations.

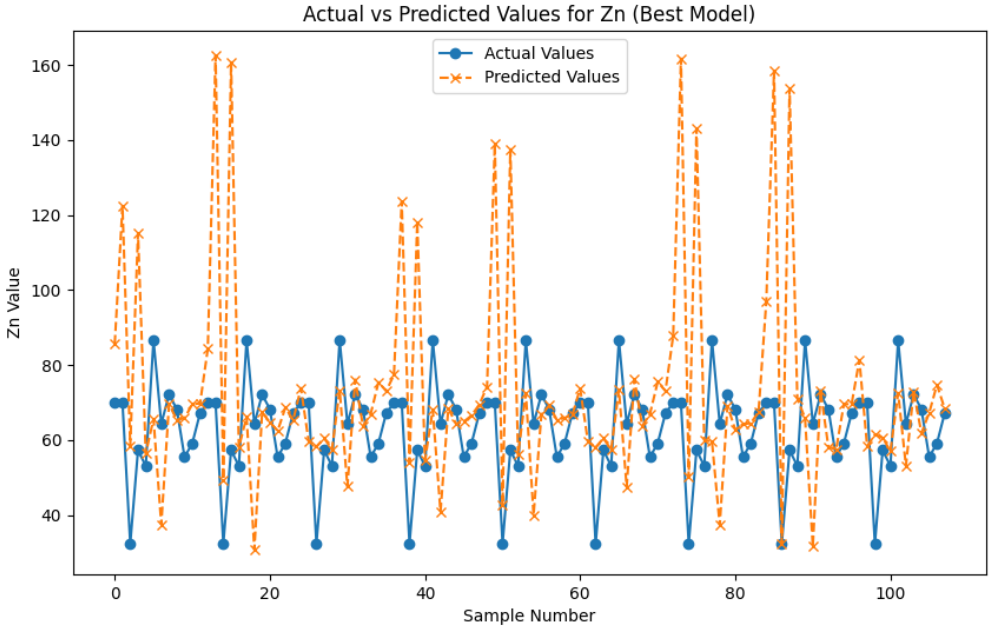

Figure 4. Actual vs Predicted Values for Zn using the Best Model

For Mn, RF using Random Sampling provided the best results, with an MAE of 6.55 and an R^2 score of 0.53 (Figure 5). This model effectively captured the variability in the data and made accurate predictions for manganese concentrations.

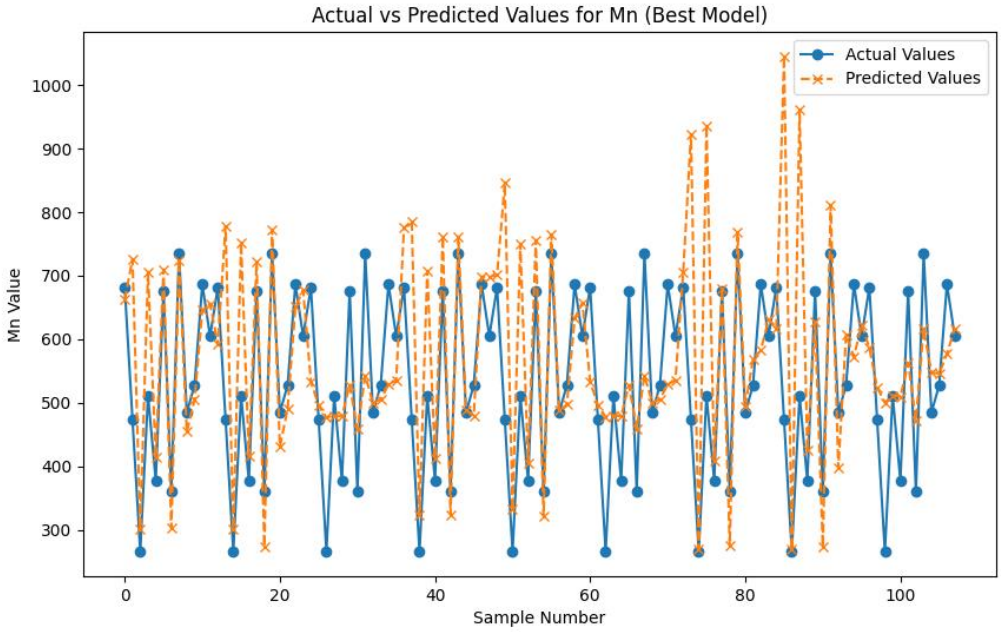

Figure 5. Actual vs Predicted Values for Mn using the Best Model

The most accurate model for Cu was RF with Random Sampling, achieving an MAE of 3.21 and an R^2 score of 0.75 (Figure 6). This model demonstrated the highest predictive accuracy, as evidenced by the strong alignment between actual and predicted values.

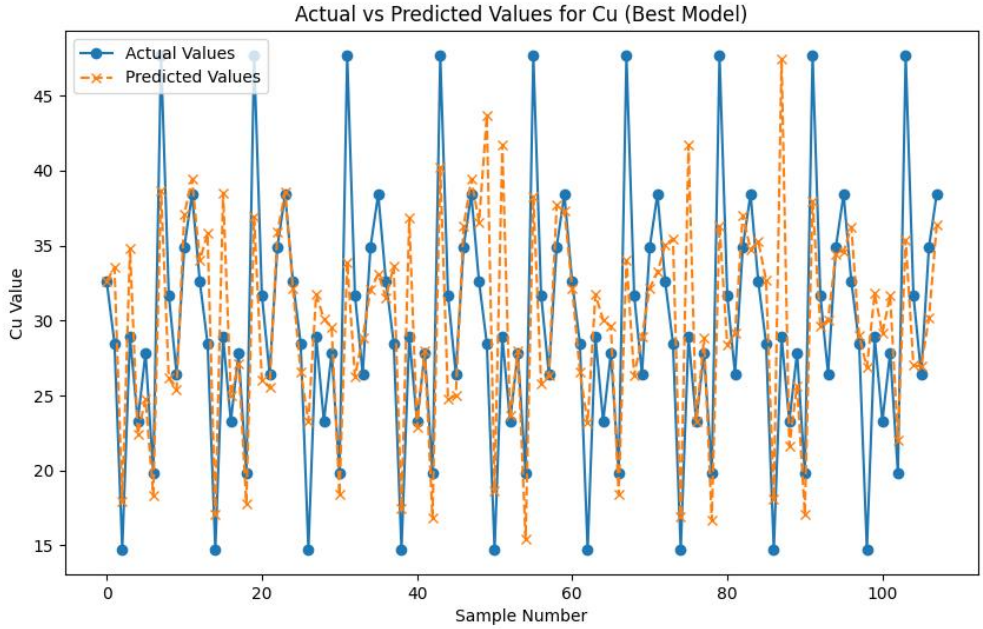

Figure 6. Actual vs Predicted Values for Cu using the Best Model

For Cd, SVR with Random Sampling yielded the best performance, with an MAE of 0.07 and an R^2 score of 0.44 (Figure 7). The model effectively minimized the error, making it well-suited for predicting cadmium concentrations in the soil.

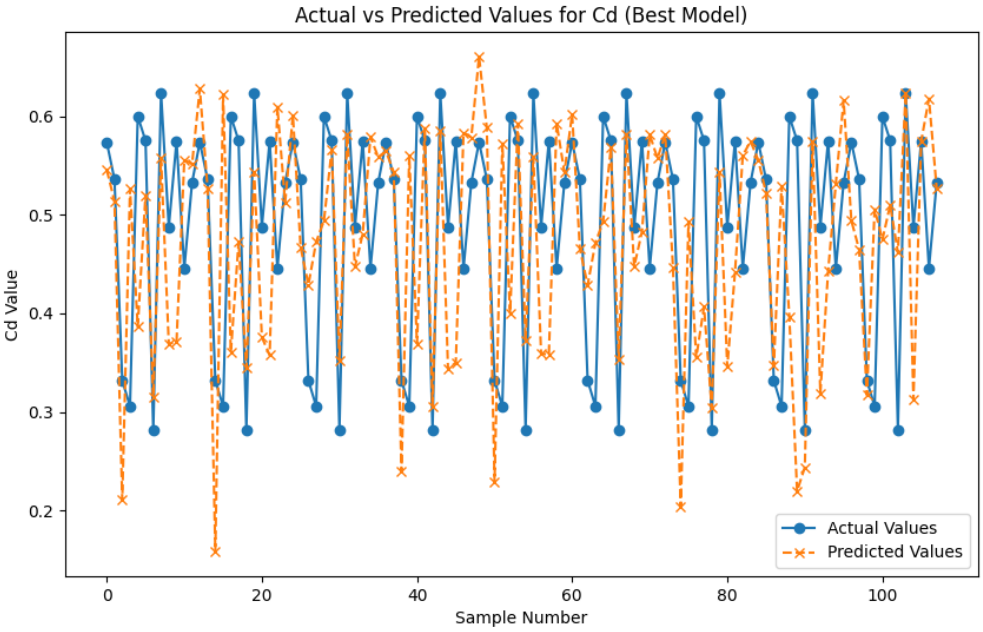

Figure 7. Actual vs Predicted Values for Cd using the Best Model

The best-performing model for Cr was SVR with Bootstrap Sampling, achieving an MAE of 25.34 and an R^2 score of 0.23 (Figure 8). While the error was minimized, the relatively low R^2 score suggests that the model struggled to capture the full variability in Cr concentrations.

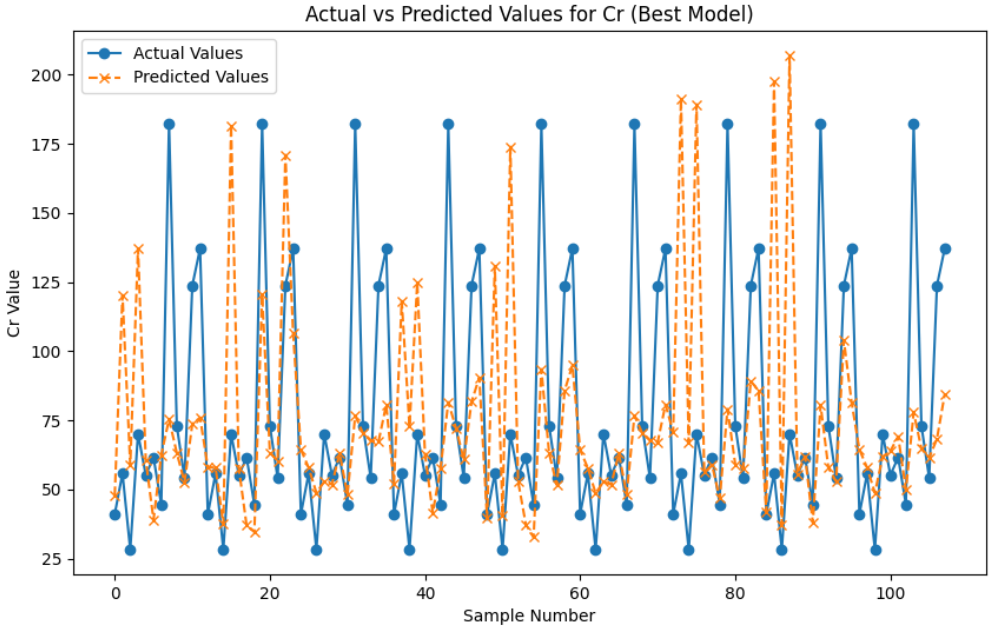

Figure 8. Actual vs Predicted Values for Cr using the Best Model

For Mo, RF with Bootstrap Sampling performed best, with an MAE of 0.34 and an R^2 score of 0.01 (Figure 9). The negative R^2 score indicates that the model struggled to generalize well for molybdenum, highlighting the need for further refinement.

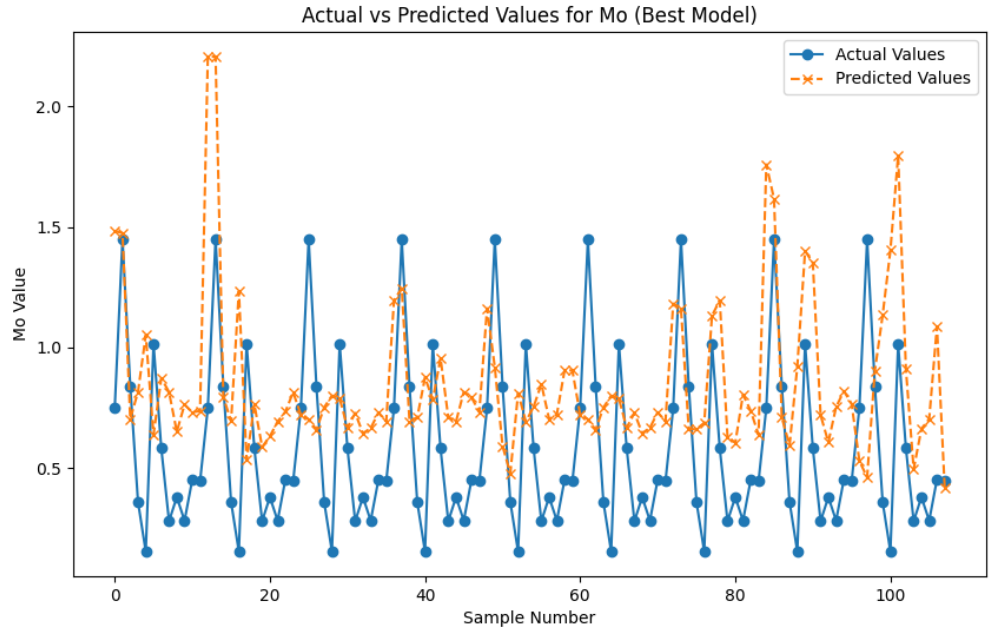

Figure 9. Actual vs Predicted Values for Mo using the Best Model

The best model for Ni was SVR combined with Bootstrap Sampling, with an MAE of 27.39 and an R^2 score of 0.18 (Figure 10). Although the model captured some variability, the low R^2 score suggests that the predictions could be further improved.

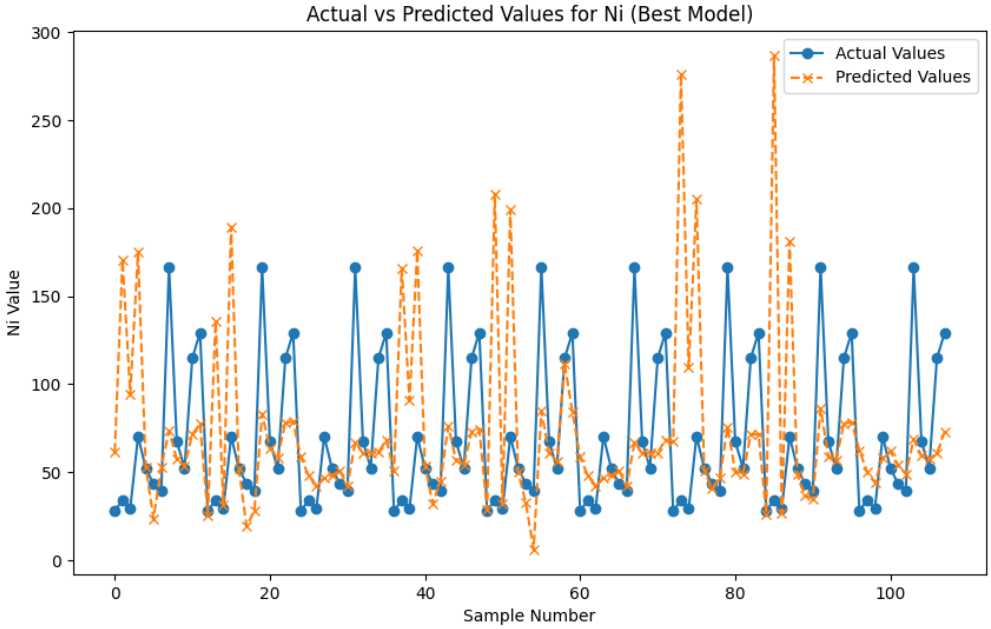

Figure 10. Actual vs Predicted Values for Ni using the Best Model

For Pb, SVR with Bootstrap Sampling provided the best results, achieving an MAE of 28.16 and an R^2 score of 0.27 (Figure 11). The relatively low R^2 score indicates that while the model minimized errors, there is still room for improvement in capturing the full variability of lead concentrations.

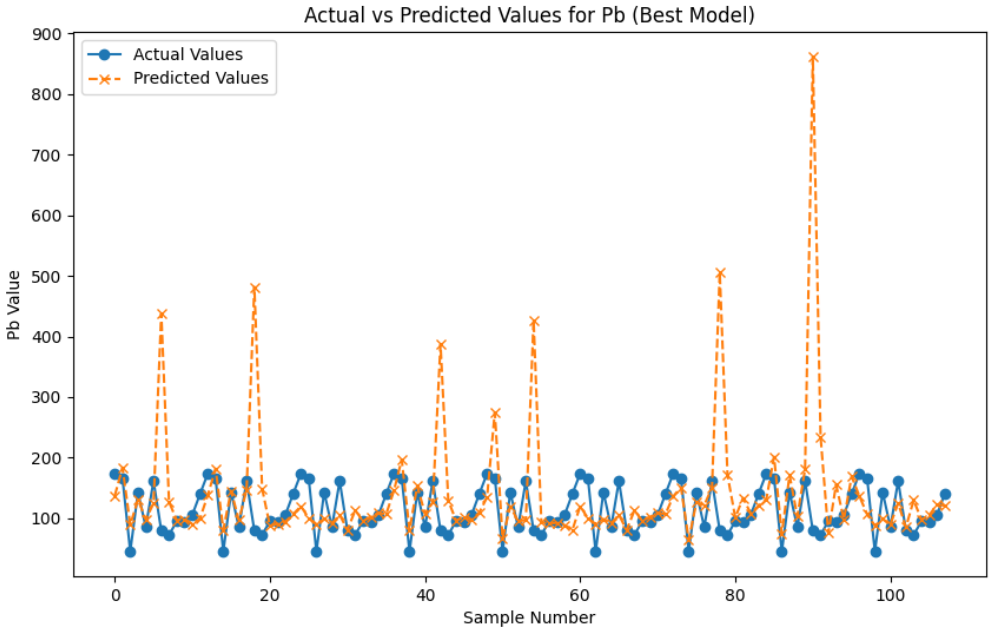

Figure 11. Actual vs Predicted Values for Pb using the Best Model

The results of this study demonstrate that machine learning models, particularly RF, GB, and SVR, can effectively predict microelement and heavy metal concentrations in soil samples. However, the performance of these models varied depending on the element being predicted and the data augmentation technique used. For elements such as B, Fe, and Cu (copper), the models demonstrated strong predictive capabilities, as indicated by relatively high R^2 scores and low MAE values. This

suggests that the concentrations of these elements are more easily predicted based on the available soil features. Additionally, the use of Random Sampling and Bootstrap Sampling improved the models' ability to generalize from the limited dataset. On the other hand, elements such as Mo, Cr, and Ni presented challenges for the models, as evidenced by lower R^2 scores and higher error values. These results may be attributed to the limited size of the dataset and the complexity of predicting the concentrations of these elements based solely on the soil features provided. The negative R^2 score for molybdenum suggests that the model struggled to generalize, indicating the need for either more complex modeling approaches or additional features to improve prediction accuracy. Additionally, the results highlight the importance of selecting appropriate data augmentation techniques. Jittering and Bootstrap Sampling proved effective in improving model performance for certain elements, such as Zn and Pb, respectively. These techniques expanded the dataset and allowed the models to train on a more varied sample set, ultimately enhancing their generalization capability.

Despite some challenges, the overall findings suggest that machine learning models, when combined with appropriate data augmentation techniques, hold significant potential for predicting soil properties, even with limited data. Future studies could benefit from larger datasets and the inclusion of additional soil features to further refine these models and enhance their predictive accuracy.

Conclusion

This study underscores the potential of machine learning models, specifically Random Forest, Gradient Boosting, and Support Vector Regressor, for predicting microelement and heavy metal concentrations in soils from Türkiye's Central-Southern Anatolian region. The integration of data augmentation techniques, such as Random Sampling, Jittering, and Bootstrap Sampling, significantly enhanced the models' performance, particularly under the constraints of limited datasets.

Despite these advancements, the study faced several limitations, including the small dataset size and the limited diversity of soil properties analyzed. These factors restricted the models' ability to generalize, especially for elements like Mo and Ni. To address these challenges, future research should focus on collecting larger and more diverse datasets, integrating additional soil properties, and exploring advanced modeling approaches, such as deep learning and hybrid techniques.

Moreover, incorporating spatial and temporal data, as well as considering the impacts of geogenic and anthropogenic factors, could provide a more comprehensive understanding of soil characteristics. Such advancements would facilitate the development of robust predictive models, contributing to sustainable agricultural practices and improved environmental management. This study highlights the promise of machine learning as a cost-effective and efficient tool for soil analysis, offering valuable insights for agricultural and environmental applications.

Acknowledgements: Our research article was supported by project-based initiatives under the Mevlana Student Exchange Program (Project No: MEV-2017-36), the Scientific and Technological Research Council of TurkeyTUBİTAK (Project No: 116O786) and the Selçuk University Scientific Research Projects Coordination Unit (Project No: 18401058).

Researchers' Contribution Rate Declaration Summary

The authors declare that they have contributed equally to the article.

Conflict to Interest Declaration

The authors declare that there is no conflict of interest between them.

References

Awad, M., Khanna, R., 2015. Support vector regression. In: Efficient learning machines: Theories, concepts, and applications for engineers and system designers, pp. 67–80.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13 (2).

Breiman, L., 2001. Random forests. Mach. Learn. 45: 5–32.

Gee, G.W., 1986. Particle size analysis. In: Methods of soil analysis/ASA and SSSA.

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. Neural Comput. 4 (1): 1–58.

Gezgin, S., Dursun, N., Hamurcu, M., Harmankaya, M., Önder, M., Sade, B., 2002 Boron content of cultivated soils in Central-Southern Anatolia and its relationship with soil properties and irrigation water quality. Boron in plant and animal nutrition. 391-400.

Gholamy, A., Kreinovich, V., Kosheleva, O., 2018. Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. Int. J. Intell. Technol. Appl. Stat. 11 (2): 105–111.

Günal, H., Acir, N., Budak, M., 2012. Heavy metal variability of a native saline pasture in arid regions of Central Anatolia. Carpathian Journal of Earth and Environmental Sciences. 7: 183–193.

Günal, H., Kılıç, O.M., Ersayın, K., Acir, N., 2022. Land suitability assessment for wheat production using analytical hierarchy process in a semi-arid region of Central Anatolia. Geocarto International. 37: 16418–16436.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The elements of statistical learning: Data mining, inference, and prediction. Springer.

Hızalan, E., Ünal, H., 1966. Topraklarda önemli kimyasal analizler. AÜ Ziraat Fakültesi Yayınları. 278: 5–7.

Jackson, M., 1958. Soil chemical analysis. Prentice Hall, Englewood Cliffs, NJ, pp. 183–204.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30.

Khosravi, V., Ardejani, F.D., Yousefi, S., Aryafar, A., 2018. Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods. Geoderma. 318: 29–41.

Koç, Ş., 1987. Karadağ (Karaman) Volkanitlerinin Jeolojisi Ve "Base Surge" Oluşukları. Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi. 2.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of IJCAI, Montreal, Canada, pp. 1137–1145.

Lahn, E., 1949. On the geology of Central Anatolia. Türkiye Jeoloji Bülteni. 2: 90–107.

Luce, M.S., Ziadi, N., Gagnon, B., Karam, A., 2017. Visible near-infrared reflectance spectroscopy prediction of soil heavy metal concentrations in paper mill biosolid- and liming by-product-amended agricultural soils. Geoderma. 288: 23–36.

MGM, 2025. Meteoroloji Genel İklim Verileri. Meteoroloji Genel Müdürlüğü, Ankara, Turkey.

Munnaf, M.A., Mouazen, A.M., 2021. Development of a soil fertility index using on-line Vis-NIR spectroscopy. Comput. Electron. Agric. 188: 106341.

Nie, S., Chen, H., Sun, X., An, Y., 2024. Spatial distribution prediction of soil heavy metals based on random forest model. Sustainability. 16 (11): 4358.

Ozyazici, M.A., Dengiz, O., Ozyazici, G., 2017 Spatial distribution of heavy metals density in cultivated soils of Central and East Parts of Black Sea Region in Turkey. Eurasian Journal of Soil Science. 6: 197-205.

Shi, S., Hou, M., Gu, Z., Jiang, C., Zhang, W., Hou, M., Li, C., Xi, Z., 2022. Estimation of heavy metal content in soil based on machine learning models. Land. 11 (7): 1037.

Sigrist, F., 2021. Gradient and Newton boosting for classification and regression. Expert Systems With Applications. 167: 114080.

Smith, H.W., Weldon, M.D., 1941. A comparison of some methods for the determination of soil organic matter.

T.O.B., 2010. Toprak Kirlilik Parametreleri Yönetmeliği.

Taylor, S., 1964. Abundance of chemical elements in the continental crust: A new table. Geochimica et Cosmochimica Acta. 28: 1273–1285.

Taylor, S., McLennan, S., 1985. The continental crust: Its composition and evolution. Geoscience Texts. 312.

Wang, Y., Zhao, Y., Xu, S., 2022. Application of VNIR and machine learning technologies to predict heavy metals in soil and pollution indices in mining areas. Journal of Soils and Sediments. 22 (10): 2777–2791.