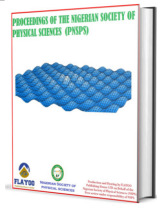
Published by Nigerian Society of Physical Sciences. Hosted by FLAYOO Publishing House LTD

Proceedings of the Nigerian Society of Physical Sciences

Journal Homepage: https://flayoophl.com/journals/index.php/pnspsc



Prediction of heavy metal concentrations in soil using machine learning models

Temitope M. Osobamiro[a,*], Samuel O. Sipeolu[a], Emmanuel F. Ayo[b], Sakeenah O. Abdullah[a]

[a]*Department of Chemical Sciences, Olabisi Onabanjo University, P.M.B. 2002, Ago-Iwoye, Nigeria*
[b]*Department of Computer Sciences, Olabisi Onabanjo University, P.M.B. 2002, Ago-Iwoye, Nigeria*

ABSTRACT

The application of machine learning (ML) models is increasingly used to predict pollutant levels in environmental samples, particularly heavy metals. This study predicted lead (Pb), zinc (Zn), and cadmium (Cd) concentrations in soil samples collected near a plastic recycling facility in Ogun State, Nigeria. Atomic absorption spectrophotometry (AAS) showed low heavy metal concentrations (mg kg^{-1}), with Pb $\leq$ 0.38, Cd $\leq$ 0.40, and Zn $\leq$ 7.55, below the regulatory limits considered in this study. Five regression models—linear regression, Random Forest, Extra Trees, XGBoost, and CatBoost—were evaluated using soil physicochemical properties as predictors. XGBoost produced the highest R^2 values for Pb (0.973) and Cd (0.971), whereas Extra Trees produced the highest R^2 value for Zn (0.957). CatBoost and Random Forest also showed strong predictive ability, with generally low root mean square error (RMSE) and mean absolute error (MAE) values. The feature-importance results indicated that nitrogen, total organic carbon, and organic matter were important predictors of heavy metal concentrations. The findings suggest strongly nonlinear relationships between soil properties and heavy metal concentrations and support ensemble ML models as useful tools for rapid preliminary monitoring. The results should, however, be interpreted cautiously because of the limited number of experimental samples and the use of rule-based synthetic data augmentation.

Keywords: Heavy metals, Machine learning, Soil properties, Ensemble models.

DOI:10.61298/pnspsc.2026.3.336

1. INTRODUCTION

Anthropogenic activities, such as the combustion of fossil fuels, fertilizer application, industrial processes, mining activities, and municipal effluents, increase the deposition of heavy metals in soil and the environment, thereby posing risks to the quality and quantity of agricultural production [1, 2]. These metals may enter the food chain and contribute to diseases such as cardiovascular disorders, cancer, cognitive impairment, renal failure, and neurological disorders [1, 3]. Although heavy metal contamination in soil is often attributed to the extensive use of fertilizers and pesticides, other important sources include landfills, fuel filling stations, mechanic workshops, and dumpsites [4]. The distinct challenges posed by prolonged heavy metal contamination include irreversibility, restricted movement, persistence, non-degradability, and elevated toxicity, which may disrupt ecosystem balance [1].

Indiscriminate refuse disposal is common, especially among middle- and low-income earners in many developing countries,

*Corresponding Author Tel. No.: +234-805-6628-102.
e-mail: osobamiro.temitope@oouagoiwoye.edu.ng (Temitope M. Osobamiro)

including Nigeria. Waste generated in many Nigerian cities is not properly managed, contributing to preventable illness and reduced environmental quality [5]. Because of their versatility, durability, and widespread use, plastics are among the most common solid wastes in dumpsites. Burning, a common practice at many dumpsites, releases toxic pollutants, including heavy metals, which may eventually leach into soil and water bodies [5, 6]. Early detection and prediction are therefore necessary for environmental monitoring, especially in localized soil assessments. Traditional methods, including atomic absorption spectrometry (AAS), inductively coupled plasma mass spectrometry, and colorimetric techniques, can be labor-intensive, time-consuming, and prone to secondary pollution during remediation. These limitations underscore the need for approaches capable of addressing the nonlinear dynamics of heavy metal distribution in soil and water systems [7]. Because detailed soil heavy metal content data are often difficult to obtain, rapid prediction using trained ML models is needed [8].

ML is one of the most common approaches for data prediction and has been applied in several fields because it can learn from multidimensional, large, and complex datasets to form predictive models [3, 8]. Extreme Gradient Boosting (XGBoost) has been used to predict mercury in soils [9]. Linear regression and decision trees have been used to predict cadmium in soil [1], and Random Forest has been used to predict arsenic and iron in groundwater [10]. The rapid development of ML models and their use in environmental science have led to the integration of spectral data and ML, as well as the use of statistical methods such as Pearson correlation coefficient to reduce redundancy and high-dimensional characteristics in spectral datasets [11]. Liu *et al.* [12], for example, analyzed correlations among nine heavy metals.

Despite the increasing global application of ML in soil contamination assessment, studies focusing on its use for contaminant prediction in environmental soils in south-western Nigeria remain limited. This study therefore aimed to evaluate the performance of ensemble ML models, including Random Forest, XGBoost, CatBoost, and Extra Trees, in predicting Pb, Zn, and Cd concentrations in soils around a plastic recycling facility in Obafemi Owode Local Government Area, Ogun State, Nigeria.

2. MATERIALS AND METHODS

2.1. STUDY AREA

The study site was a plastic company located in Obafemi Owode Local Government Area, Ogun State, Nigeria (latitude N 6°56′22″ and longitude E 3°32′15″). Soil samples were collected from different areas within the company premises, including the dumpsite, the active recycling plant, and the parking lot. The soil mostly contained a matrix of plastic fragments mixed with dark, greasy patches, indicating the presence of petroleum product spills. Industrial runoff and effluent from washing and sorting processes could introduce pollutants into the surrounding environment.

2.2. SAMPLE COLLECTION AND LABORATORY ANALYSIS

Samples were collected with a soil auger at a depth of 10–15 cm and placed in labelled polythene bags. The samples were homogenized and sieved through a 2 mm sieve, air-dried away from direct sunlight for five days, and stored for further analysis. The soil physicochemical properties determined included pH, total organic carbon (TOC), organic matter (OM), electrical conductivity (COND), nitrogen, and total dissolved solids (TDS). The pH, COND, and TDS were determined using a calibrated HANNA multimeter (Model 11398). TOC was determined using the modified Walkley–Black wet oxidation method, as described by the Association of Official Analytical Chemists (AOAC) [13]. Following Ojo *et al.* [14], OM was calculated by multiplying TOC by the conventional Van Bemmelen factor of 1.724. Nitrogen content was determined using the Kjeldahl method. Soil digestion was carried out with 10 mL of aqua regia (3:1 HNO_3:HCl), whereas metal concentrations in solution were measured using a Buck Scientific atomic absorption spectrophotometer (Model 210A). Blanks and standards were run after every five determinations to calibrate the instrument [15].

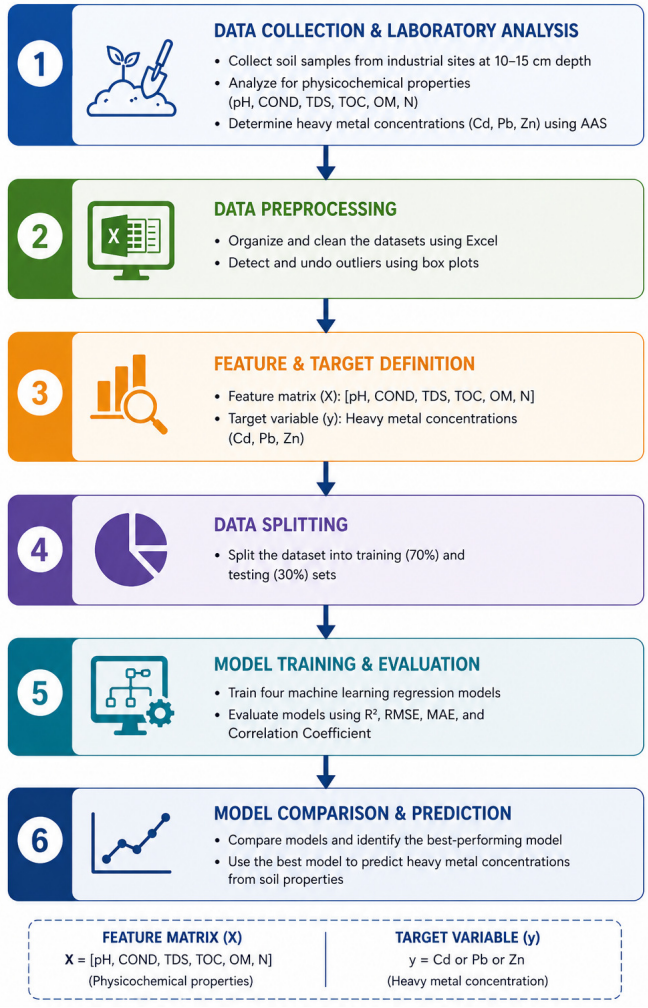


Figure 1. Machine learning workflow for soil prediction.

2.3. MACHINE LEARNING ALGORITHM

The dataset was created using 30 experimental soil samples consisting of the physicochemical parameters pH, COND, TOC, OM, N, and TDS as input variables (X), whereas heavy metal

Table 1. Equations of metrics used to assess model performance [11, 28].

Error metric	Equation
Mean absolute error (MAE)	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Coefficient of determination (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
Root mean square error (RMSE)	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Correlation coefficient (CC)	$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$

concentrations (Pb, Zn, and Cd) were defined as the target variables (y) [1]. Owing to the sample size (n = 30), rule-based synthetic data generation was used to enhance the limited dataset in a controlled manner. This generated 150 synthetic samples under the default parameter settings. The synthetic data were used to train the dataset, whereas the original data were used to evaluate the model following a 70:30 subset split [16]. The generated synthetic samples remained within the observed experimental value ranges while preserving the relationships among OM, TOC, and heavy metals, which were key variables. This approach was used to avoid extrapolation beyond real environmental conditions, improve model-training stability, and maintain the physical and environmental constraints of the dataset [17].

A total of five regression models were used to predict heavy metal concentrations: linear regression, Extra Trees, XGBoost Regressor, Random Forest, and CatBoost. This selection included linear, tree-based, and boosting algorithms. Linear regression was included as a baseline model for performance comparison. Hyperparameter optimization was performed with GridSearchCV using a predefined parameter grid to identify the optimal configuration. The parameters included the number of estimators, maximum tree depth, and learning rate for the tree-based models and boosting algorithms. After hyperparameter optimization, k-fold cross-validation (k = 5) was used to assess model stability [3, 18]. Feature importance was used to determine the most relevant features for predicting the target variables Cd, Pb, and Zn [10]. The optimized models were applied to the independent test dataset, and the predicted values were compared with observed values.

Linear regression is a statistical method used to examine the relationship between one dependent variable and one or more independent variables [19]. It is useful because it provides both interpretation and prediction. Multiple linear regression extends the basic model by allowing several predictors to be analyzed simultaneously, thereby making it possible to study more complex relationships [20]. Linear regression also produces statistical measures, such as the regression coefficient and coefficient of determination, which show the direction, strength, and explanatory power of the model [21]. Despite its usefulness, linear regression works best when assumptions such as linearity, independence, and constant error variance are satisfied.

Random Forest, as proposed by Breiman [22], is an ensemble model based on regression-tree algorithms that forms a forest when constructed from many regression trees [11]. Random Forest can predict soil properties and classes, handle a large number of predictor variables, identify nonlinear relationships, and account for interactions among variables [9]. It works by selecting a random subset of features for each tree, after which the best feature is selected for each node split [23].

Extra Trees, also known as Extremely Randomized Trees, is similar to Random Forest but uses a more randomized approach. It trains each base estimator with a random subset of features and selects random threshold values for splitting nodes, making it more random than Random Forest and potentially less prone to overfitting [23]. XGBoost is an algorithm that fits new base learners using second-order derivatives of the loss function and is often considered more precise and effective than conventional gradient boosting [24]. CatBoost, or categorical boosting, is a boosting ensemble model that handles categorical features and produces oblivious decision trees as base learners, thereby reducing overfitting and training time [24].

2.4. MODEL EVALUATION METRICS

Three metrics were used to evaluate prediction accuracy: R^2, RMSE, and MAE. The R^2 value indicates model stability; the closer the value is to 1, the better the model. A value of 0.7 is generally considered to indicate good model performance [25]. RMSE has the same order of magnitude as the sampled data, and smaller RMSE values indicate higher estimation accuracy. MAE is the average absolute difference between the estimated and measured values; smaller MAE values indicate higher model accuracy [11]. Correlation analysis was applied to examine the linear relationships between heavy metal concentrations and soil physicochemical parameters, providing preliminary insight into variable relevance. Correlation coefficient (CC) values closer to +1 indicate strong positive correlation, whereas values closer to -1 indicate strong negative correlation [26, 27]. The metric equations are presented in Table 1 [11, 28].

In Table 1, y_i is the measured value of heavy metal content, $\hat{y}_i$ is the estimated value of heavy metal content, n is the number of samples, $\bar{y}$ is the mean of observed values, and i is the sample point.

2.5. SOFTWARE

All analyses were performed using the Python programming language. ML models and evaluation procedures were implemented using libraries including Scikit-learn, XGBoost, LightGBM, and CatBoost.

2.6. LIMITATIONS

Because only 30 experimentally measured samples were available, rule-based synthetic data generation was used to enhance the dataset. This method may introduce bias and cannot fully replace real observations. In addition, nested cross-validation was not implemented, which may result in slightly optimistic performance estimates. Therefore, the results should be interpreted as preliminary.

3. RESULTS AND DISCUSSION

3.1. SOIL PHYSICOCHEMICAL PROPERTIES

Soil properties, including pH, TDS, TOC, COND, OM, and N, which are known to influence heavy metal availability in soil,

Table 2. Physicochemical parameters of the sampled soils.

Sample	pH	COND (mS cm^{-1})	TDS (ppm)	TOC (%)	OM (%)	Nitrogen (%)
1	6.16 ± 0.19	7.06 ± 0.19	44.38 ± 2.17	2.58 ± 0.01	4.44 ± 0.01	0.98 ± 0.01
2	7.06 ± 0.19	6.46 ± 0.19	18.38 ± 2.17	2.49 ± 0.01	4.28 ± 0.01	0.96 ± 0.01
3	6.46 ± 0.19	6.56 ± 0.19	88.38 ± 2.17	2.64 ± 0.01	4.54 ± 0.01	1.22 ± 0.01
4	6.56 ± 0.19	6.76 ± 0.19	108.38 ± 2.17	2.64 ± 0.01	4.54 ± 0.01	1.23 ± 0.01
5	6.76 ± 0.19	6.86 ± 0.19	58.38 ± 2.17	2.64 ± 0.01	4.54 ± 0.01	1.07 ± 0.01
6	6.86 ± 0.19	6.76 ± 0.19	50.38 ± 2.17	2.64 ± 0.01	4.54 ± 0.01	1.11 ± 0.01
7	6.76 ± 0.19	6.86 ± 0.19	74.38 ± 2.17	2.94 ± 0.01	5.06 ± 0.01	1.06 ± 0.08
8	6.86 ± 0.19	7.06 ± 0.19	84.38 ± 2.17	4.14 ± 0.01	7.12 ± 0.01	1.00 ± 0.04
9	7.06 ± 0.19	6.86 ± 0.19	44.38 ± 2.17	4.44 ± 0.01	7.64 ± 0.01	1.24 ± 0.01
10	6.86 ± 0.19	7.06 ± 0.19	103.38 ± 2.17	3.54 ± 0.01	6.10 ± 0.01	1.00 ± 0.01
11	7.06 ± 0.19	6.96 ± 0.19	64.38 ± 2.17	1.74 ± 0.01	3.00 ± 0.01	1.21 ± 0.01
12	6.96 ± 0.19	7.06 ± 0.19	80.38 ± 2.17	2.34 ± 0.01	4.03 ± 0.01	1.24 ± 0.01
13	7.06 ± 0.19	7.06 ± 0.19	68.38 ± 2.17	3.24 ± 0.01	5.57 ± 0.01	1.23 ± 0.01
14	7.06 ± 0.19	7.16 ± 0.19	77.38 ± 2.17	2.64 ± 0.01	4.54 ± 0.01	1.23 ± 0.01
15	7.16 ± 0.19	6.46 ± 0.19	50.38 ± 2.17	2.04 ± 0.01	3.52 ± 0.01	0.97 ± 0.01
16	6.46 ± 0.19	6.76 ± 0.19	40.38 ± 2.17	2.19 ± 0.01	3.78 ± 0.01	1.26 ± 0.01
17	6.76 ± 0.19	6.56 ± 0.19	33.38 ± 2.17	2.94 ± 0.01	5.06 ± 0.01	1.27 ± 0.01
18	6.56 ± 0.19	6.56 ± 0.19	44.38 ± 2.17	2.79 ± 0.01	4.81 ± 0.01	1.26 ± 0.01
19	6.56 ± 0.19	6.76 ± 0.19	31.38 ± 2.17	2.87 ± 0.01	4.95 ± 0.01	0.99 ± 0.01
20	6.76 ± 0.19	6.36 ± 0.19	55.38 ± 2.17	2.93 ± 0.01	5.05 ± 0.01	0.99 ± 0.01
21	6.36 ± 0.19	5.46 ± 0.19	125.38 ± 2.17	1.74 ± 0.01	3.00 ± 0.01	1.05 ± 0.01
22	5.46 ± 0.19	5.36 ± 0.19	430.38 ± 2.17	1.59 ± 0.01	2.74 ± 0.01	1.22 ± 0.01
23	5.36 ± 0.19	5.56 ± 0.19	490.38 ± 2.17	1.55 ± 0.01	2.67 ± 0.01	1.23 ± 0.01
24	5.56 ± 0.19	5.46 ± 0.19	287.38 ± 2.17	1.76 ± 0.01	3.03 ± 0.01	1.11 ± 0.01
25	5.46 ± 0.19	5.46 ± 0.19	433.38 ± 2.17	1.57 ± 0.01	2.71 ± 0.01	1.23 ± 0.01
26	5.46 ± 0.19	5.56 ± 0.19	332.38 ± 2.17	1.71 ± 0.01	2.94 ± 0.01	1.18 ± 0.01
27	5.56 ± 0.19	5.66 ± 0.19	427.38 ± 2.17	1.56 ± 0.01	2.69 ± 0.01	1.22 ± 0.01
28	5.66 ± 0.19	5.66 ± 0.19	173.38 ± 2.17	1.75 ± 0.01	3.02 ± 0.01	1.24 ± 0.01
29	5.66 ± 0.19	5.66 ± 0.19	159.38 ± 2.17	1.74 ± 0.01	3.00 ± 0.01	1.23 ± 0.01
30	5.66 ± 0.19	6.26 ± 0.19	164.38 ± 2.17	1.86 ± 0.01	3.20 ± 0.01	1.23 ± 0.01
Standard	6.00–8.50[a]	≤ 4.00[b]	≤ 500.00[b]	3.00–6.00[b]	≥ 2.00[b]	0.10–0.50

Note: [a]USEPA [33]; [b]USDA [34].
USEPA, United States Environmental Protection Agency; USDA, United States Department of Agriculture.

were determined in this study. The results are presented in Table 2. The soil pH values ranged from slightly acidic to neutral, with an average pH of 6.4, which is within the acceptable USEPA range [29]. This result is similar to those reported by Olayinka *et al.* [4] and Abbas *et al.* [30], but lower than that obtained by Joseph *et al.* [2]. The slightly acidic result indicates that landfill activities may affect soil pH, which could influence nutrient availability for plant development [6].

Electrical conductivity was consistently elevated across all samples, with minimal spatial fluctuation and an average of 6.4 mS cm^{-1}, exceeding the recommended limit. This indicates the presence of more soluble salts in the soil [2]. TDS showed a wide range of values, with lower concentrations observed in samples 1–20 and markedly higher concentrations in samples 21–30. Despite this variability, all values remained below the recommended threshold, indicating that dissolved solids did not exceed regulatory limits, although localized enrichment was evident in the latter samples.

The average TOC and OM contents were 2.5% and 4.2%, respectively, falling within or below the recommended ranges. These values were similar to those reported by Yerima *et al.* [31]. The TOC values of most samples were below the recommended USDA range, with a few samples, particularly samples 8–10, having values within the recommended range. In contrast, a noticeable decline in OM was observed in samples 21–30, where values approached the lower recommended boundary. OM is important because it enhances the cation exchange capacity of soil, improves soil structure, retains nutrients, and minimizes erosion [31]. The relatively high OM values may be due to the deposition of plant residues, organic waste, or sewage waste that accompanied the plastic waste at the company premises [32]. Total nitrogen concentrations were uniformly high across all samples, with an average of 1.15%, exceeding the recommended range. Nitrogen is an essential macronutrient for soil health, and high N content indicates the presence of nitrogenous compounds that may support crop growth and yield [6].

3.2. HEAVY METAL CONCENTRATIONS

The average measured Pb, Cd, and Zn concentrations were 0.18, 0.19, and 5.94 mg kg^{-1}, respectively, in the collected soil samples. These values were below the international regulatory thresholds considered in this study. The low concentrations indicate limited contamination from the plastic recycling operation at the time of sampling. However, continuous monitoring remains necessary because heavy metals are cumulative in soils. The analytical results are shown in Table 3.

The concentrations followed a clear and consistent order across all samples, with Zn showing the highest values and Pb and Cd occurring at much lower levels. Zinc concentrations were relatively low and stable, generally ranging between about 5.0 and 7.5 mg kg^{-1}, with slightly higher values observed in samples S11–S20. Lead concentrations were very low throughout

Table 3. Concentrations of heavy metals at the study site.

S/N	Zn (mg kg^{-1})	Pb (mg kg^{-1})	Cd (mg kg^{-1})
S1	5.40 ± 0.00	0.15 ± 0.07	0.05 ± 0.00
S2	5.25 ± 0.07	0.15 ± 0.00	0.15 ± 0.07
S3	5.23 ± 0.11	0.10 ± 0.00	0.05 ± 0.00
S4	5.08 ± 0.04	0.25 ± 0.07	0.08 ± 0.04
S5	5.18 ± 0.04	0.13 ± 0.04	0.35 ± 0.07
S6	5.05 ± 0.07	0.15 ± 0.00	0.15 ± 0.00
S7	5.18 ± 0.04	0.15 ± 0.07	0.35 ± 0.07
S8	5.25 ± 0.00	0.15 ± 0.00	0.15 ± 0.07
S9	5.30 ± 0.00	0.10 ± 0.00	0.15 ± 0.14
S10	5.08 ± 0.04	0.25 ± 0.07	0.03 ± 0.04
S11	6.18 ± 0.04	0.20 ± 0.07	BDL
S12	7.55 ± 0.07	0.20 ± 0.07	BDL
S13	6.35 ± 0.07	0.20 ± 0.00	0.15 ± 0.00
S14	6.30 ± 0.00	0.15 ± 0.00	BDL
S15	6.10 ± 0.00	0.38 ± 0.04	0.40 ± 0.00
S16	6.50 ± 0.14	0.25 ± 0.07	BDL
S17	6.40 ± 0.00	0.23 ± 0.18	0.05 ± 0.00
S18	6.55 ± 0.34	0.33 ± 0.04	0.15 ± 0.00
S19	6.40 ± 0.00	0.23 ± 0.11	0.35 ± 0.07
S20	6.25 ± 0.07	0.15 ± 0.00	0.15 ± 0.07
S21	6.23 ± 0.11	0.05 ± 0.00	0.33 ± 0.11
S22	5.89 ± 0.04	0.25 ± 0.07	0.08 ± 0.04
S23	5.68 ± 0.04	BDL	0.16 ± 0.07
S24	5.55 ± 0.07	BDL	0.20 ± 0.07
S25	6.18 ± 0.04	BDL	0.35 ± 0.07
S26	6.75 ± 0.00	0.15 ± 0.00	0.15 ± 0.07
S27	6.30 ± 0.00	0.10 ± 0.00	0.40 ± 0.00
S28	6.08 ± 0.04	0.25 ± 0.07	BDL
S29	6.43 ± 0.39	0.13 ± 0.04	BDL
S30	6.55 ± 0.07	0.15 ± 0.00	0.20 ± 0.07
Standard	300[a]/140[b]	400[a]/85[b]	0.43[a]/0.80[b]

Note: [a]USEPA [33]; [b]DPR [35]
USEPA, United States Environmental Protection Agency;
DPR, Department of Petroleum Resources
EGASPIN, Environmental Guidelines & Standards for Petroleum Industry in Nigeria;
BDL, below detection limit.

Table 4. Model performance assessment for the evaluated metals.

Model	Index	Pb	Zn	Cd	Feature importance
Random Forest	R^2 (Real 30)	0.797	0.343	0.844	N (Zn), N, OM (Cd), TDS (Pb)
	R^2 (Augmented)	0.448	0.008	0.568	
	RMSE	0.039	0.504	0.052	
	MAE	0.031	0.409	0.042	
	CC	0.944	0.585	0.947	
XGBoost	R^2 (Real 30)	0.973	0.763	0.971	OM (Zn, Pb), TOC (Cd)
	R^2 (Augmented)	0.599	0.406	0.531	
	RMSE	0.014	0.303	0.022	
	MAE	0.012	0.214	0.015	
	CC	0.993	0.900	0.990	
CatBoost	R^2 (Real 30)	0.906	0.911	0.912	OM (Zn, Pb), N (Cd)
	R^2 (Augmented)	0.755	0.640	0.640	
	RMSE	0.027	0.185	0.039	
	MAE	0.020	0.148	0.030	
	CC	0.970	0.980	0.977	
Extra Trees	R^2 (Real 30)	0.929	0.957	0.928	N (Zn), OM (Cd, Pb)
	R^2 (Augmented)	0.769	0.811	0.709	
	RMSE	0.023	0.129	0.035	
	MAE	0.016	0.098	0.026	
	CC	0.975	0.988	0.985	
Linear regression	R^2 (Real 30)	-0.058	0.343	0.217	TOC, OM (Cd, Zn, Pb)
	R^2 (Augmented)	0.026	0.008	0.039	
	RMSE	0.090	0.504	0.116	
	MAE	0.069	0.409	0.102	
	CC	0.319	0.585	0.532	

the site, with several samples recording values at or below the detection limit. Cadmium showed low but detectable concentrations in many samples, with small fluctuations across the site and slightly higher values in a few locations, such as S15 and S27. These values were below guideline limits established by the USEPA for residential soils [33]. When evaluated against Nigerian DPR/EGASPIN and USEPA standards, all heavy metal concentrations fell within acceptable limits for agricultural and industrial soils [33, 35]. This finding suggests that, despite the presence of industrial activities, soils within the plastic company premises had not accumulated heavy metals to levels that pose immediate regulatory concern.

Table 5. XGBoost predictions from the test dataset for the first 10 samples.

Cd		Pb		Zn	
Actual	Predicted	Actual	Predicted	Actual	Predicted
0.05	0.110	0.15	0.147	5.4	5.524
0.15	0.169	0.15	0.146	5.25	5.276
0.05	0.088	0.1	0.098	5.225	5.269
0.075	0.087	0.25	0.226	5.075	5.167
0.35	0.246	0.125	0.120	5.175	5.194
0.15	0.194	0.15	0.139	5.05	5.255
0.35	0.263	0.15	0.141	5.175	5.409
0.15	0.148	0.15	0.142	5.25	5.377
0.15	0.141	0.1	0.115	5.3	5.431
0.025	0.071	0.25	0.235	5.075	5.226

3.3. MACHINE LEARNING MODEL PREDICTIONS

The performance of the ML models was comparable, although XGBoost and Extra Trees showed the strongest predictive capability among the evaluated models. As shown in Table 4, XGBoost produced the highest R^2 values for Pb (0.973) and Cd (0.971), whereas Extra Trees produced the highest R^2 value for Zn (0.957). The closer the R^2 and CC values are to 1, the better the model, and the smaller the RMSE and MAE values, the more accurate the model [11, 25, 27]. Table 5 shows strong similarity between the first 10 actual and predicted values in the test dataset, with Pb showing the closest agreement and Cd showing the least agreement among the three metals.

All the nonlinear ensemble models outperformed linear regression, indicating that soil physicochemical properties and heavy metal concentrations were related through nonlinear patterns. The results suggest nonlinear interactions among soil properties, OM, TOC binding effects, and metal concentrations. Across the models, OM was an important feature, probably because it is one of the primary factors controlling heavy metal mobility, distribution, and remediation behavior [36, 37]. Similar findings have been reported in related ML studies. Hu *et al.* [38] used Extra Trees for low-cost determination of soil heavy metal concentrations, Keçeci *et al.* [1] used a tree-based model to predict Cd in soil, and Ma *et al.* [39] compared multiple models and observed the strong performance of Random Forest. The strong performance of tree-based and boosting models in the present study is consistent with recent findings showing that ML approaches, particularly tree-based models, effectively capture nonlinear relationships in soil heavy metal datasets [40].

4. CONCLUSION

Soil samples collected within the premises of the plastic recycling facility in Obafemi Owode Local Government Area showed low concentrations of Pb, Cd, and Zn relative to regulatory limits. XGBoost achieved the highest predictive performance for Pb and Cd, whereas Extra Trees performed best for Zn. Model accuracy was generally consistent with the performance reported in related industrial-soil ML studies. Pb demonstrated a strong statistical association with soil properties, whereas Cd showed moderate predictability consistent with its heterogeneous environmental distribution. These findings support the application of ensemble ML models as predictive support tools for rapid heavy metal assessment. Such tools should be used alongside laboratory validation to provide a data-driven foundation for environmental monitoring strategies in south-western Nigeria. Because this study is a preliminary investigation of heavy metal concentrations in soil, larger datasets are needed for external validation.

DATA AVAILABILITY

The data will be available on request from the corresponding author.

References

[1] M. Keçeci, F. Gökmen, M. Usul, C. Koca & V. Uygur, ''Prediction of cadmium content using machine learning methods'', Environmental Earth Sciences **83** (2024) 362. https://doi.org/10.1007/s12665-024-11672-5.

[2] E. Joseph, J. Azorji, O. Nwachukwu, S. Iheagwam, J. Okere, K. Ukeje & D. Anamnah, ''Assessment of physicochemical characteristics and heavy metal concentration in soils and plants in selected refuse dumpsites within Nkwerre LGA, Imo State, Southeast Nigeria'', South Asian Research Journal of Natural Products **3** (2020) 26. Available online: https://www.sarpublication.com/.

[3] K. N. Palansooriya, J. Li, P. D. Dissanayake, M. Suvarna, L. Li, X. Yuan, B. Sarkar, D. C. W. Tsang, J. Rinklebe, X. Wang & Y. S. Ok, ''Prediction of soil heavy metal immobilization by biochar using machine learning'', Environmental Science & Technology **56** (2022) 4187. https://pubs.acs.org/doi/10.1021/acs.est.1c08302.

[4] O. O. Olayinka, O. O. Akande, K. Bamgbose & M. T. Adetunji, ''Physicochemical characteristics and heavy metal levels in soil samples obtained from selected anthropogenic sites in Abeokuta, Nigeria'', Journal of Applied Sciences and Environmental Management **21** (2017) 883. https://doi.org/10.4314/jasem.v21i5.14.

[5] O. O. Eseyin, G. J. Udom & I. C. Osu, ''Heavy metal concentration and physicochemical parameters in soil and plants near unengineered dumpsites in Port Harcourt, Nigeria'', Journal of Geography, Environment and Earth Science International **19** (2019) 1. Available online: https://journaljgeesi.com/index.php/JGEESI/article/view/376.

[6] T. A. Gyamfi, B. Koomson & E. Bessah, ''Assessment of burnt polyethylene impact on physicochemical and biological properties of soil at Esereso-Adagya landfill, Ghana'', Research Square (2025). https://doi.org/10.21203/rs.3.rs-7464330/v1.

[7] R. A. Wuana & F. E. Okieimen, ''Heavy metals in contaminated soils: a review of sources, chemistry, risks and best available strategies for remediation'', International Scholarly Research Notices **2011** (2011) 402647. https://doi.org/10.5402/2011/402647.

[8] W. Cao & C. Zhang, ''A collaborative compound neural network model for soil heavy metal content prediction'', IEEE Access **8** (2020) 129497. https://doi.org/10.1109/ACCESS.2020.3009248.

[9] A. Suleymanov, R. Suleymanov, A. Kulagin & M. Yurkevich, ''Mercury prediction in urban soils by remote sensing and relief data using machine learning techniques'', Remote Sensing **15** (2023) 3158. https://doi.org/10.3390/rs15123158.

[10] T. M. T. Huynh, C. F. Ni, Y. S. Su, V. C. N. Nguyen, I. H. Lee, C. P. Lin & H. H. Nguyen, ''Predicting heavy metal concentrations in shallow aquifer systems based on low-cost physicochemical parameters using machine learning techniques'', International Journal of Environmental Research and Public Health **19** (2022) 12180. https://doi.org/10.3390/ijerph191912180.

[11] S. Shi, M. Hou, Z. Gu, C. Jiang, W. Zhang, M. Hou & Z. Xi, ''Estimation of heavy metal content in soil based on machine learning models'', Land **11** (2022) 1037. https://doi.org/10.3390/land11071037.

[12] J. Liu, Y. Zhang, H. Wang & Y. Du, ''Study on the prediction of soil heavy metal elements content based on visible near-infrared spectroscopy'', Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **199** (2018) 43. https://doi.org/10.1016/j.saa.2018.03.040.

[13] Association of Official Analytical Chemists (AOAC), ''Official Methods of Analysis'', 18th ed., AOAC, Arlington (2015) 806.

[14] I. O. Ojo, J. O. Ojo & O. Oladele, ''Analysis of heavy metals and some physicochemical parameters in soil of major industrial dumpsites in Akure Township, Ondo State, Nigeria'', International Journal of Chemistry **7** (2015) 55. https://doi.org/10.5539/ijc.v7n1p55.

[15] J. Haware & H. Pramond, ''Determination of specific heavy metals in fruit juices using atomic absorption spectrophotometer (AAS)'', International Journal of Research in Chemistry and Environment **4** (2014) 163. Available online: https://ijrce.org/index.php/ijrce/article/view/31.

[16] F. S. de Oliveira & R. Stefani, ''Evaluating the use of synthetic data for machine learning prediction of self-healing capacity of concrete'', AI in Civil Engineering **4** (2025) 25. https://doi.org/10.1007/s43503-025-00074-6.

[17] S. Palaniappan, R. Logeswaran, S. Khanam & Y. Zhang, ''Machine learning model for predicting net environmental effects'', Journal of Informatics and Web Engineering **4** (2025) 243. https://doi.org/10.33093/jiwe.2025.4.1.18.

[18] H. Castro-Gutiérrez, C. Robles-Algarín & A. Polo, ''Data augmentation and machine learning for heavy metal detection in mulberry leaves using laser-induced breakdown spectroscopy (LIBS) spectral data'', Processes **13** (2025) 1688. https://doi.org/10.3390/pr13061688.

[19] N. Roustaei, ''Application and interpretation of linear-regression analysis'', Medical Hypothesis, Discovery & Innovation in Ophthalmology **13** (2024) 151. https://doi.org/10.51329/mehdiophthal1506.

[20] G. Heinze, M. Baillie, L. Lusa, W. Sauerbrei, C. O. Schmidt, F. E. Harrell & M. Huebner, ''Regression without regrets: initial data analysis is a prerequisite for multivariable regression'', BMC Medical Research Methodology **24** (2024) 178. https://doi.org/10.1186/s12874-024-02294-3.

[21] T. Alkhalifah, H. Wang & O. Ovcharenko, ''ML real: bridging the gap between training on synthetic data and real data applications in machine learning'', Artificial Intelligence in Geosciences **3** (2022) 101. https://doi.org/10.1016/j.aiig.2022.09.002.

[22] L. Breiman, ''Random forests'', Machine Learning **45** (2001) 5. https://doi.org/10.1023/A:1010933404324.

[23] M. Ghazwani & M. Y. Begum, ''Computational intelligence modeling of hyoscine drug solubility and solvent density in supercritical processing: gradient boosting, extra trees, and random forest models'', Scientific Reports **13** (2023) 10046. https://doi.org/10.1038/s41598-023-37232-8.

[24] A. Alazba & H. Aljamaan, ''Software defect prediction using stacking generalization of optimized tree-based ensembles'', Applied Sciences **12** (2022) 4577. https://doi.org/10.3390/app12094577.

[25] F. Xia, T. Fan, Y. Chen, D. Ding, J. Wei, D. Jiang & S. Deng, ''Prediction of heavy metal concentrations in contaminated sites from portable X-ray fluorescence spectrometer data using machine learning'', Processes **10** (2022) 536. https://doi.org/10.3390/pr10030536.

[26] M. F. Ioni , S. M. Radu & E. C. Dunca, ''Correlation analysis of heavy metal concentrations in the tailing dumps Branch 1 and 2 Lupeni using Pearson coefficient matrix'', Mining Revue **30** (2024) 22. https://doi.org/10.2478/minrv-2024-0023.

[27] O. A. Al-Khashman, A. O. Al-Khashman, N. R. J. Hynes, H. M. Alnawafleh & P. S. Velu, ''Assessment of heavy metals contamination of topsoil and street dust around cement factory in southern Jordan'', Journal of Environmental Protection **15** (2024) 672. https://doi.org/10.4236/jep.2024.156038.

[28] M. M. Mukaka, ''A guide to appropriate use of correlation coefficient in medical research'', Malawi Medical Journal **24** (2012) 69. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC3576830/.

[29] US EPA, ''Regional Screening Levels (RSLs)–Generic Tables (June 2017)'', United States Environmental Protection Agency, Washington, DC (2017). Available online: https://www.epa.gov/risk/regional-screening-levels-rsls-generic-tables-june-2017.

[30] O. R. Abbas, K. B. Al-Paruany & M. H. Mashjel, ''Distribution of heavy metals, polycyclic aromatic hydrocarbons, micro-plastics particles, and their potential contamination of soil in a selected area in Baghdad City, Iraq'', Iraqi Journal of Science **66** (2025) 3818. https://doi.org/10.24996/ijs.2025.66.9.25.

[31] E. A. Yerima, B. N. Hikon, C. V. Ogbodo, H. Ataitiya & J. D. Ani, ''Chemi-

cal speciation and mobility of heavy metals in soils around Nasara Sack and Packaging Company, Akwanga, Nigeria'', Journal of Advances in Chemistry **16** (2019) 5379. Available online: https://rajpub.com/index.php/jac/article/view/8434.

[32] M. Tefera, F. Gebreyohannes & M. Saraswathi, ''Heavy metal analysis in the soils in and around Robe town, Bale Zone, Southeast Ethiopia'', Eurasian Journal of Soil Science **7** (2018) 251. https://doi.org/10.18393/ejss.403004.

[33] United States Environmental Protection Agency (USEPA), ''Supplemental guidance for developing soil screening levels for Superfund sites'', OSWER 9355.4-24, United States Environmental Protection Agency, Washington, DC (2002). Available online: https://www.epa.gov/superfund/superfund-soil-screening-guidance.

[34] United States Department of Agriculture (USDA), ''Soil survey manual'', USDA Handbook No. 18 (2003). Available online: https://www.nrcs.usda.gov/resources/guides-and-instructions/soil-survey-manual.

[35] Department of Petroleum Resources, ''Environmental guidelines and standards for the petroleum industry in Nigeria (EGASPIN)'', Department of Petroleum Resources (2002). Available online: https://www.aziza.com.ng/wp-content/uploads/2020/06/environmental-guidelines-and-standards-for-the-petroleum-industry-in-nigeria-egaspin-2002.pdf.

[36] S. S. Ramos-Romero, H. R. Benavides-Rosales & J. J. Peña-Chamorro, ''Advances in modelling the transport of heavy metals in agricultural soils and their leaching into groundwater: an integrative critical review'', Frontiers in Environmental Science **14** (2026) 1764394. https://doi.org/10.3389/fenvs.2026.1764394.

[37] Y. Wan, J. Liu, Z. Zhuang, Q. Wang & H. Li, ''Heavy metals in agricultural soils: sources, influencing factors, and remediation strategies'', Toxics **12** (2024) 63. https://doi.org/10.3390/toxics12010063.

[38] T. Hu, Q. Chen, Z. Lin, C. Qi & L. Chai, ''Machine learning enables low-cost determination of soil heavy metal concentrations'', ACS ES&T Engineering **5** (2025) 3085. https://doi.org/10.1021/acsestengg.5c00463.

[39] W. Ma, K. Tan & P. Du, ''Predicting soil heavy metal based on random forest model'', in: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE (2016) 4331. https://doi.org/10.1109/IGARSS.2016.7730129.

[40] T. Hu, M. Wu, Q. Chen, L. Chai & C. Qi, ''Machine learning uncovers dominant fractions of heavy metal(loid)s in global soils'', Communications Earth & Environment **7** (2026) 214. https://doi.org/10.1038/s43247-026-03221-8.