The University of Tehran Press

Pollution

Print ISSN: 2383-451X
Online ISSN: 2383-4501

https://jpoll.ut.ac.ir/

A Novel Machine Learning Framework for Predicting ^{232}Th Distribution in Radionuclide-Contaminated Soils Using Physicochemical Environmental Factors

Sati Lubis[1] | Haruna Adamu[1,2]✉ | Jamilu Usman[3] | Abdullahi Garba Usman[4,5] | Sani Isah Abba[6]

1. Department of Chemistry, Abubakar Tafawa Balewa University, Gubi Campus, 740102, Bauchi, Nigeria.
2. Department of Environmental Management Technology, Abubakar Tafawa Balewa University, Yelwa Campus, 740272, Bauchi, Nigeria.
3. Interdisciplinary Research Centre for Membranes and Water Security, King Fahd University of Petroleum and Minerals, Dhahran, 31261, Saudi Arabia.
4. Operational Research Center in Healthcare, Near East University, TRNC, Mersin 10, 99138, Nicosia, Turkey.
5. Department of Analytical Chemistry, Faculty of Pharmacy, Near East University, TRNC, Mersin 10, 99138, Nicosia, Turkey.
6. Department of Chemical Engineering, Prince Mohammad Bin Fahd University, Al Khobar, Saudi Arabia.

Article Info

Article type:
Research Article

Article history:
Received: 27 August 2025
Revised: 25 November 2025
Accepted: 06 February 2026

Keywords:
Radionuclides
Soil Chemistry
Machine Learning
Pollution
Land Management

ABSTRACT

This study investigates the role of soil chemistry, specifically pH, organic carbon (OC), organic matter (OM), and cation exchange capacity (CEC), in influencing the mobility and distribution of 232Th radionuclides in abandoned mine soils using advanced machine learning (ML) models. Soil samples were collected from multiple locations across different seasons. Gaussian Process Regression (GPR), Long Short-Term Memory (LSTM) networks, Adaptive Neuro-Fuzzy Inference System (ANFIS), and Random Forest (RF) models were employed to predict 232Th distribution, with feature selection identifying optimal model combinations (C1, C2, and C3). The performance evaluation of machine learning models revealed distinct patterns in predicting 232Th distribution. The results indicate that GPR-C1 exhibited the highest predictive accuracy, with MAPE improving from 8.9909 to 3.0468 and MAE reducing from 3.5236 to 1.6044 during the verification phase. In addition, GPR-C1 emerged as the top-performing model during both training (RMSE = 7.0851, DC = 0.6482) and testing (RMSE = 4.5808, DC = 0.5848), demonstrating its robustness in capturing non-linear relationships between soil properties (pH, OC, OM, CEC) and ^{232}Th mobility. In contrast, RF models (RF-C1, RF-C3) exhibited the poorest performance (training RMSE > 11.5123; testing RMSE > 7.6855), likely due to their inability to resolve complex geochemical interactions, as evidenced by their low DC (<0.2) and PCC (<0.3) values. A notable observation was that several models exhibited lower RMSE in the testing set than in calibration, reflecting the reduced variance within the held-out site–season blocks; however, nested cross-validation and a leave-site-out analysis consistently identified GPR-C1 as the most reliable and accurate model. This aligns with field data showing higher ^{232}Th mobility during wet seasons due to leaching and runoff transport ($p < 0.05$). For instance, testing RMSE (4.5808) of GPR-C1 was significantly lower than its training RMSE (7.0851), reinforcing the role of seasonal dynamics in ^{232}Th redistribution. Therefore, this model demonstrates significant potential for accurately predicting ^{232}Th behaviour and distribution, crucial for environmental risk assessments. Hence, accurate predictions of ^{232}Th distribution can guide targeted remediation efforts and inform land management practices, mitigating risks associated with ^{232}Th exposure.

Cite this article: Lubis, S., Adamu, H., Usman, J., Usman, A.G., & Abba, S.I. (2026). A Novel Machine Learning Framework for Predicting 232Th Distribution in Radionuclide-Contaminated Soils Using Physicochemical Environmental Factors. *Pollution*, 12(1), 327-341.
https://doi.org/10.22059/poll.2025.400497.3073

 Publisher: The University of Tehran Press.
DOI: https://doi.org/10.22059/poll.2025.400497.3073

*Corresponding Author Email: *hadamu2@atbu.edu.ng*

INTRODUCTION

Abandoned tin and columbite mines on Nigeria's Jos Plateau pose a major environmental threat due to persistent radioactive contaminants like ^{232}Th in tailings and degraded soils (Alshahrani et al., 2025; Li et al., 2024). With an extremely long half-life (~14 billion years), ^{232}Th is a critical contaminant that can be redistributed *via* weathering and erosion, leading to soil and water contamination, bioaccumulation in plants, and increased radiological risks to local communities through inhalation or ingestion (Hofmann et al., 2025). Traditional methods for analyzing ^{232}Th distribution (e.g., gamma spectroscopy, ICP-MS) are costly, time-consuming, and require specialized lab equipment, hindering large-scale monitoring (IAEA, 2004; Thakur et al., 2021). The mobility and distribution of ^{232}Th are strongly influenced by key soil properties including pH, organic matter (OM), organic carbon (OC), and cation exchange capacity (CEC) (Sarkar et al., 2021; Ouyang et al., 2023).

Therefore, this study develops a novel ML framework to predict ^{232}Th distribution using these readily measurable soil parameters as predictive features. We employed advanced ML models including Long Short-Term Memory (LSTM), Gaussian Process Regression (GPR), Adaptive Neuro-Fuzzy Inference System (ANFIS), and Random Forest (RF) to analyze complex, non-linear relationships and generate accurate spatial risk maps. Feature selection identified the most significant predictors, resulting in three model combinations (C1, C2, C3). The research provides a cost-effective, data-driven tool for environmental risk assessment, enabling the identification of high-risk zones and informing targeted remediation strategies in resource-limited mining regions.

Recent studies report on practical advances in data-driven modelling for aquatic systems, including Kolmogorov–Arnold Networks, which improved chlorophyll-a prediction in large lakes and capture nonstationary behaviour (Saravani et al., 2025). Cross-basin representation learning enhances generalisation under limited data by pretraining transferable embeddings and fine-tuning at new sites (Zheng et al., 2025). Spatiotemporal graph neural networks use river-network topology to improve multi-station water-quality prediction (Wan et al., 2025). Physics-informed neural models that embed the advection–dispersion equation estimate longitudinal dispersion and velocity with greater data efficiency and interpretability than black-box alternatives (Meng et al., 2024). A recent review consolidates guidance on uncertainty quantification, leakage-safe validation and explainability (Zhi et al., 2024). Taken together, these findings support uncertainty-aware, topology-aware, and physics-guided approaches, and they emphasize explicit uncertainty treatment, cross-site evaluation, and tests for non-stationarity (Saravani et al., 2025; Zheng et al., 2025; Wan et al., 2025; Meng et al., 2024; Zhi et al., 2024).

Notably, the mobility and distribution of ^{232}Th are fundamentally dictated by the complex solution chemistry of Th(IV), which undergoes extensive hydrolysis and is subject to strong, often irreversible sorption onto soil components across variable pH and redox conditions (Aba et al., 2021; Maher et al., 2013). This inherent, non-linear geochemical complexity challenges predictive modeling; however, a critical gap in environmental machine learning studies is the frequent failure to validate complex models against simple, interpretable baselines such as Linear Regression or Generalized Additive Models (GAMs). As a result, to truly justify the advanced Novel Machine Learning Framework proposed here, our study provides essential validation by directly addressing the non-linear physicochemical controls while rigorously comparing the performance gains of our ensemble of models (LSTM, GPR, ANFIS, RF) against these simple baselines. Hence, the primary aim of this study is to develop and validate a novel ML framework capable of accurately predicting the spatial distribution of ^{232}Th contamination in soils using easily measurable physicochemical environmental factors as predictive inputs.

MATERIALS AND METHODS

This section provides details of the laboratory procedures employed to investigate the influence of soil chemistry on ^{232}Th mobility and distribution in abandoned mine sites. The focus was on analyzing pH, OC, OM, and CEC, and their impact on ^{232}Th behaviour.

2.1. Sample collection and preparation

Alzubaidi *et al* (2016) obtained a mean soil ^{232}Th concentration ± standard deviation of 133.96 ± 2.92 Bqkg^{-1} with a sampling precision of 1.62 % *rsd*. As target relative standard deviation (RSD_x) was 0.81%, then the target absolute error (E) is 0.0081 x 133.96 Bqkg^{-1} ≈ 1.09 Bqkg^{-1}. Therefore, the formula by using σ_S = 2.92 Bqkg^{-1} and E= 1.09 Bqkg^{-1}, and the t-value for $N \approx 32$ (t ≈ 1.70) is:

$$N = \left(\frac{t\sigma_s}{E}\right)^2 \quad (1)$$

$$N = \left(\frac{1.70\,x\,2.92}{1.09}\right)^2 \approx (4.55)^2 \approx 20.7 \approx 21.$$

Thus, the number of samples required to be taken to obtained accurate representative was approximately (18) per sampling area.

Soil samples were collected from ten villages in Jos South and Barkin Ladi LGAs, Nigeria. At each village, three composite samples were taken: (1) at a mining point (primary source), (2) 200 m downwind, and (3) 400 m downwind, to assess wind-driven dispersion of particulate-bound ^{232}Th (Barnekow et al., 2019). Sampling was conducted monthly both during dry and wet seasons. Each composite sample comprised 21 homogenized subsamples collected from a 0-20 cm depth within a 1 m² area of the study area as shown in Fig. 1, following the guidelines of HJ/166-2004 (China, 2004). Samples were air-dried, homogenized, and sieved to 2 mm for uniformity.

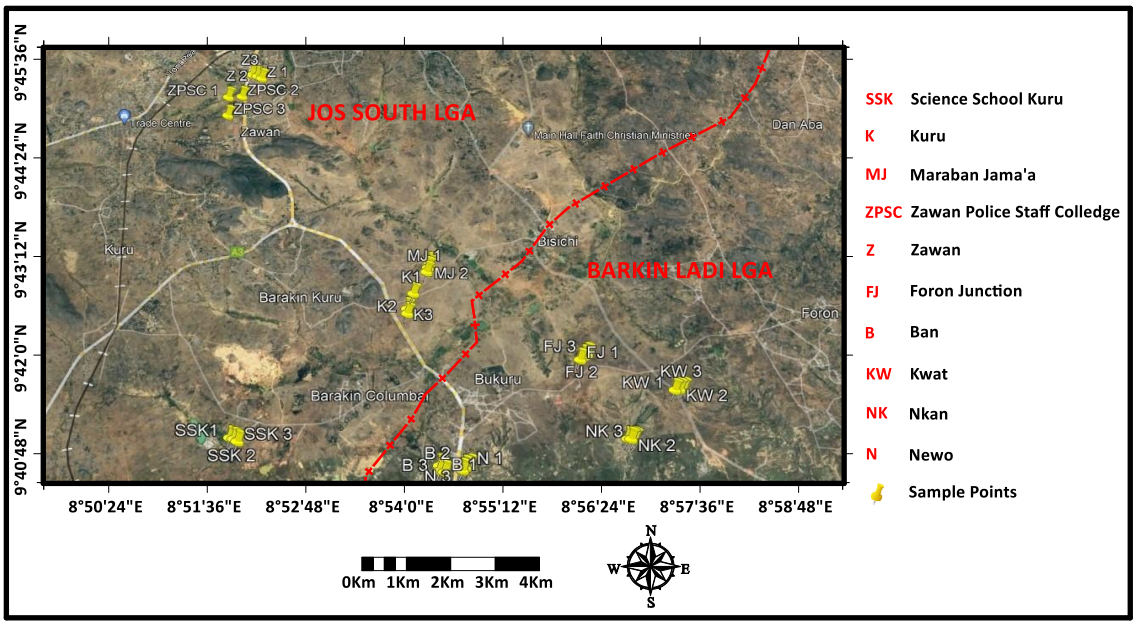


Fig. 1. Sampling points in Barkin Ladi and Jos South Local Government Area of Plateau State, Nigeria. Source: Google Satellite. The map is displayed using the Geographic Coordinate System (GCS) WGS 84.

Analysis of Soil Physicochemical Properties

Soil pH was measured using a Hanna pH meter 209 with a 1:1 (w/v) soil/water mixture, following a 15-minute equilibration period after stirring, as described by (Beretta et al., 2014). The meter was pre-calibrated with pH 4 and 7 buffers. OC content was determined using the Walkley-Black method. In this process, 1g of soil was oxidized with 10 ml of 1N $K_2Cr_2O_7$ and 20 ml of concentrated H_2SO_4. After a 30-minute reaction period, the mixture was diluted, and the excess dichromate was back-titrated with ferrous ammonium sulphate using diphenylamine as an indicator. OM was estimated from the OC results. CEC was measured using the method described by (Gillman and Sumpter, 1986). The method involves equilibrating soil with an $MgCl_2$ solution to fully exchange cations. Excess salt is removed by washing until the endpoint equivalent to 1.5 mM $MgSO_4$ is achieved, which indicates the concentration required for the soil's zero net charge (ZPC effect). This endpoint is determined experimentally by monitoring the electrical conductivity (EC) of the washings until they match the EC of a 1.5 mM $MgSO_4$ solution. Once saturated and washed, the adsorbed Mg^{2+} is displaced using a KCl solution and subsequently quantified (e.g., *via* AAS) to determine the CEC in $cmol_c kg^{-1}$. CEC is a critical property influencing soil structure and nutrient availability (Hazleton and Murphy, 2016). The activity concentration of ^{232}Th was determined by gamma spectrometry using a NaI(Tl) detector. Due to its alpha decay nature, ^{232}Th was quantified indirectly by measuring the gamma emissions from its decay products in secular equilibrium, specifically ^{228}Ac (911 keV) and ^{212}Pb (238 keV). The detector was calibrated for energy and efficiency using ^{137}Cs and ^{60}Co point sources. The ^{232}Th activity was primarily measured *via* ^{228}Ac (911.2 keV) rather than ^{208}Tl (583.2 keV). This is because ^{228}Ac is upstream of the gaseous intermediate ^{220}Rn (Thoron), making it a more reliable proxy for parent ^{232}Th activity and less susceptible to ^{220}Rn loss from the sealed containers. Furthermore, the ^{228}Ac peak at 911.2 keV is in a cleaner spectral region, minimizing interference risks.

On the laboratory QA/QC and statistical inference, the γ-spectrometry analysis used rigorous quality control, starting with hermetically sealing and storing samples for a minimum of 30 days to ensure secular equilibrium (> 99%) for ^{226}Ra and ^{232}Th decay series. Counting efficiency was established using a standard multi-nuclide source (e.g., IAEA-375) prepared in the same geometry as the samples. A background spectrum was acquired monthly (counted for 10,000 seconds) and subtracted from all spectra to correct for ambient radiation. Each sample was measured for a minimum counting time of 20,000 seconds to achieve < 10 uncertainty. The combined standard uncertainty for the activity concentration ($Bqkg^{-1}$) was calculated by propagating individual uncertainties (counting statistics, efficiency, calibration) following GUM principles. Finally, the Minimum Detectable Concentration (MDC) for each radionuclide was determined using the Currie (1968) method based on detector efficiency, background, and counting time, reporting specific MDCs for ^{232}Th ($Bqkg^{-1}$), ^{238}U ($Bqkg^{-1}$), and ^{40}K ($Bqkg^{-1}$).

Machine Learning Methods

This study employed four advanced machine learning models to address complex environmental predictions. Gaussian Process Regression (GPR) was used for its non-parametric, Bayesian approach to model non-linear relationships and quantify prediction uncertainty (Abba et al., 2022, 2023), defined by its mean and covariance functions (Gbadamosi et al., 2024; Usman et al., 2023). Random Forest (RF), an ensemble method, builds multiple decision trees via bootstrap sampling to output a mean prediction and mitigate overfitting (Peng et al., 2020; Shang and He, 2018). The Adaptive Neuro-Fuzzy Inference System (ANFIS) integrates neural network learning with fuzzy logic reasoning for function approximation (Abba et al., 2020; Abdullahi, 2020; Zubaidi et al., 2020), utilizing a Sugeno-type system with membership functions (Usman et al., 2021). Long Short-Term Memory (LSTM) networks were selected to capture long-term dependencies in sequential data, overcoming the vanishing gradient problem

common in RNNs (Alamrouni et al., 2022). To ensure robustness, k-fold cross-validation was implemented to manage the challenges of small datasets, including increased uncertainty and overfitting risk (Ferdosi et al., 2023; Xu et al., 2021). Model performance was rigorously evaluated using six statistical metrics: determination coefficient (DC), Pearson correlation coefficient (PCC), mean square error (MSE), mean absolute percentage error (MAPE), mean absolute error (MAE), and root mean square error (RMSE) (Jibril et al., 2024; Uzun-Ozsahin et al., 2023; Usman et al., 2021).

$$DC = 1 - \frac{\sum_{j=1}^{N}\left[(Y)_{obs,j} - (Y)_{com,j}\right]^2}{\sum_{j=1}^{N}\left[(Y)_{obs,j} - \overline{(Y)}_{obs,j}\right]^2} \tag{2}$$

$$PCC = \frac{\sum_{i=1}^{N}\left(Y_{obs} - \overline{Y}_{obs}\right)\left(Y_{com} - \overline{Y}_{com}\right)}{\sqrt{\sum_{i=1}^{N}\left(Y_{obs} - \overline{Y}_{obs}\right)^2 \sum_{i=1}^{N}\left(Y_{com} - \overline{Y}_{com}\right)^2}} \tag{3}$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(Y_{obsi} - Y_{comi})^2 \tag{4}$$

$$MAPE = \frac{1}{N}\left[\sum_{i=1}^{N}\left|\frac{Y_{obsi} - Y_{comi}}{Y_{obsi}}\right|\right] \tag{5}$$

$$MAE = \frac{\sum_{i=1}^{N}\left|Y_{comi} - Y_{obsi}\right|}{N} \tag{6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_{obsi} - Y_{comi}\right)^2}{N}} \tag{7}$$

Where N, Y_{obsi}, $\overline{Y}$ and Y_{comi} *a*re data number, observed data, average value of the observed data and computed values, respectively.

Data splitting, cross-validation and preprocessing

The full dataset was divided into a training set (70%) and a testing set (30%) using a grouped stratified procedure based on site (village) and season (dry or wet). All samples belonging to a given site–season combination were kept together in either the training or the testing set to avoid spatial and seasonal information leaking between the two sets. Model selection and hyperparameter tuning were carried out within the training set using 5-fold grouped cross-validation, where folds were defined on the same site–season groups. In each cross-validation run, four folds were used to train the model and one fold was used for validation, and this process was repeated until each fold had served once as the validation set. All preprocessing and feature-selection steps were embedded inside this cross-validation pipeline. Specifically, any normalization of input variables and the correlation-based feature selection were fitted using only the training portion of each fold, and the resulting transformations were then applied to the corresponding validation or test data. In this way, no information from validation or testing observations was used when selecting features, scaling inputs, or fitting the models, thereby preventing data leakage. After identifying the optimal hyperparameters from the inner 5-fold cross-validation, the final models were retrained on the entire training set and subsequently evaluated once on the held-out testing set, while the calibration and testing results on the full training and testing sets are reported in Tables 1 and 2. To further assess spatial generalizability,

a leave-site-out evaluation was conducted. In this setting, all samples from one village were left out as a test set, and the models were trained on the remaining villages; this procedure was repeated so that each village served once as the test set.

RESULTS AND DISCUSSION

Experimental Results

The results obtained for pH, OC, OM, CEC are presented in Table S1-6. During the wet season (July–September 2021), acidic soil conditions (pH 3.12–6.8) predominated, increasing ^{232}Th mobility due to electrostatic repulsion from positively charged soil particles (Smiciklas and Sljivic-Ivanovic, 2016). Low OC: 0.02–1.62% and OM: 0.01–3.33% contents limited complexation, further enhancing mobility (Habib et al., 2019), while moderate CEC: 1.01–4.91 cmol/kg provided some retention. Concentrations were highest at mining sources (e.g., 102.50 Bq/kg in July) and decreased with distance, though over 50% of samples exceeded the 45 Bq/kg permissible limit (UNSCEAR, 2008). Spatial analysis revealed hotspots in southern Jos South and north-central Barkin Ladi, posing significant exposure risks. A concentration decline from July to August was attributed to leaching from high precipitation (Guagliardi et al., 2016), highlighting seasonal hydrological impacts.

In the dry season (January–March 2022), ^{232}Th mobility remained influenced by acidic pH (3.24–6.91) and generally low OC/OM, though slightly higher OM in some samples (e.g., B2: OC 0.95%) suggested increased complexation and reduced mobility. CEC values (1.30–4.90 cmol/kg) showed a limited direct impact, with mobility more affected by pH and mineralogy (e.g., iron oxides adsorbing ^{232}Th; Amir et al., 2015). Concentrations remained elevated (e.g., 86.74 Bq/kg in February), with artisanal mining activities and harmattan winds identified as key redistribution mechanisms (Pecha et al., 2021; Osman et al., 2022). Most source-point samples exceeded UNSCEAR (2008) limits, underscoring persistent contamination. The minimal concentration changes between February and March indicated reduced erosion, though residual risks persisted due to soil adsorption heterogeneity, particularly in clay-rich areas. The seasonal concentrations and the ^{232}Th distribution maps are presented in the supplementary information.

Preliminary Predictive and feature selection results

The correlation matrix was used as a deterministic feature selection method in this study. Figure 2 demonstrate that all the correlation coefficients are close to zero, ranging from −0.13 to +0.08. This indicates no strong linear relationships between Th and any of the soil parameters (pH, OC, OM, or CEC). Also, regarding predictive modeling, nonlinear machine learning approaches may better capture the subtle dependencies between Th and soil characteristics. But nevertheless, OM (-0.13) showed the highest correlation with the target in the current study, then followed with pH and OC with R-values equal to -0.08 and +0.08 respectively and lastly CEC (-0.05). Hence, the soil characteristics variables used in the current study for modelling ^{232}Th were categorized as follows prior to dwelling into the simulation state. C1 comprised of pH, OC, OM and CEC, while C2 consists of; pH, OC, OM and C3 composed of pH and OM.

The correlation analysis serves as the foundational justification for employing a non-linear machine learning approach, which is the core element of the study's Novel Machine Learning Framework. The results clearly demonstrated that the relationship between ^{232}Th distribution and the selected physicochemical soil factors (pH, OC, OM, CEC) is not linear, as evidenced by the correlation coefficients ranging narrowly from -0.13 to +0.08. This lack of strong linear dependency confirms that conventional statistical methods would be inadequate for accurate prediction, thereby validating the necessity for the proposed non-linear ML models (LSTM, GPR, ANFIS, RF) to capture the subtle, complex interactions governing Th(IV) behaviour. While all correlations were weak, organic matter (OM) showed the strongest influence (R =

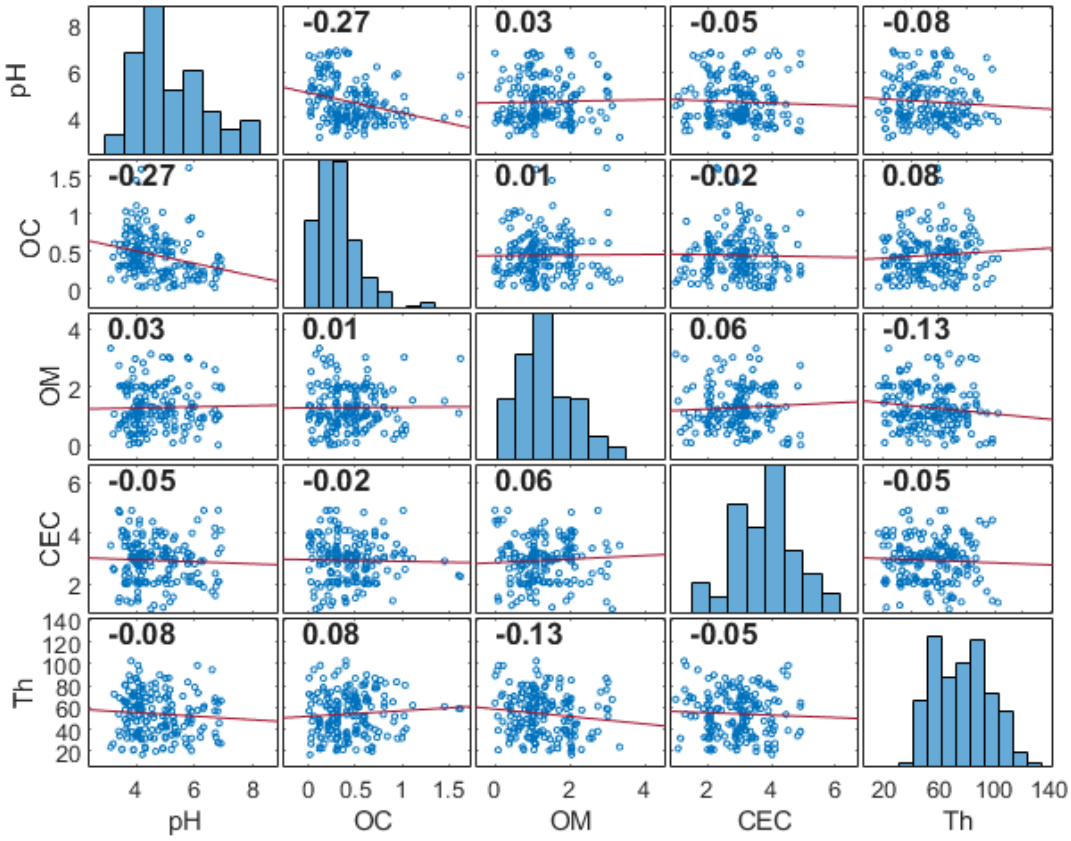


Fig. 2: Correlation matrix for deterministic feature selection

-0.13), followed by pH and OC ($R \approx \pm$ 0.08). Based on this deterministic feature selection, three distinct modeling scenarios (C1: all features, C2: pH, OC, OM, and C3: pH, OM) were established. This strategic partitioning of feature sets ensures a systematic evaluation of feature parsimony within the ML framework, allowing the study to determine the minimum number of input variables required to achieve optimal predictive performance.

Predictive ML Results

Table 1 reveals the performance of various machine learning models in calibrating predictions for ^{232}Th mobility and distribution. Among them, the GPR model with feature combination C1 (GPR-C1) demonstrated the strongest predictive accuracy, achieving the highest DC=0.6482 and PCC=0.8051, alongside the lowest error metrics (MSE: 50.1991, MAPE: 8.9909, MAE: 3.5236, RMSE: 7.0851). The RF-C2 also demonstrated strong performance DC=0.6162, PCC=0.7850, making it comparable to the top GPR model. In contrast, models such as GPR-C3 and certain LSTM and ANFIS configurations performed poorly. This analysis underscores that model selection and optimal feature combinations are critical for enhancing predictive accuracy, a finding supported by studies on the importance of feature selection for identifying the most relevant predictors (Guyon & Elisseeff, 2003).

Based on the performance evaluation, the GPR model with feature combination C1 (GPR-C1) is confirmed as the most promising model for predicting ^{232}Th distribution (Fig. 3). This result underscores that model selection and optimal feature combinations are crucial for achieving the highest predictive accuracy. The integration of such machine learning models offers a faster and more cost-effective alternative to traditional methods for predicting radionuclide behaviour (Nourani et al., 2020). This research contributes to the field by providing a robust predictive

Table 1. Predictive Estimation of ^{232}Th using ML models for Training Phase

	DC	PCC	MSE	MAPE	MAE	RMSE
GPR-C1	0.6482	0.8051	50.1991	8.9909	3.5236	7.0851
GPR-C2	0.2037	0.4514	113.6338	13.3072	5.2443	10.6599
GPR-C3	0.0844	0.2905	130.6623	14.4237	5.6721	11.4308
LSTM-C1	0.2109	0.4592	112.6114	13.1636	5.1625	10.6118
LSTM-C2	0.1035	0.3216	127.9421	13.6010	5.5032	11.3111
LSTM-C3	0.3578	0.5982	91.6446	10.2284	4.4727	9.5731
ANFIS-C1	0.2261	0.4755	110.4396	12.7216	5.1974	10.5090
ANFIS-C2	0.1855	0.4307	116.2291	13.0157	5.2635	10.7810
ANFIS-C3	0.1154	0.3396	126.2441	13.8246	5.4976	11.2358
RF-C1	0.0713	0.2670	132.5324	14.5454	5.7696	11.5123
RF-C2	0.6162	0.7850	54.7675	8.2650	3.5841	7.4005
RF-C3	0.0713	0.2670	132.5324	14.5454	5.7696	11.5123

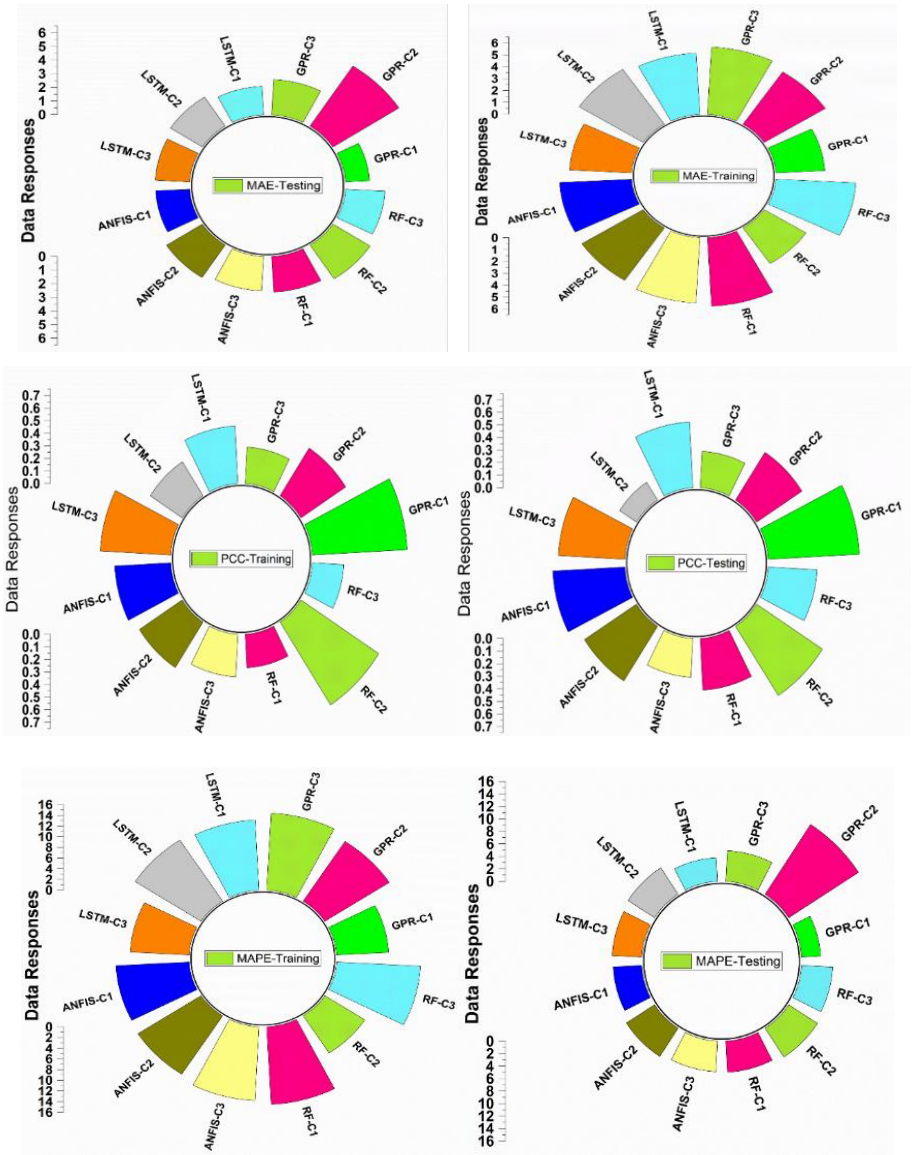


Fig. 3. Radial-based estimation of ^{232}Th for several performance predictive skills.

Table 2. Predictive estimation of ^{232}Th using ML models for the testing (verification) phase.

	DC	PCC	MSE	MAPE	MAE	RMSE
GPR-C1	0.5848	0.7647	20.9833	3.0468	1.6044	4.5808
GPR-C2	0.1987	0.4464	113.6838	13.3572	5.2943	10.7099
GPR-C3	0.0849	0.2914	54.9021	4.9178	2.5849	7.4096
LSTM-C1	0.2753	0.5247	36.6230	3.7982	2.0929	6.0517
LSTM-C2	0.0279	0.1669	51.9410	5.0570	2.5355	7.2070
LSTM-C3	0.3106	0.5573	45.2622	4.7257	2.3116	6.7277
ANFIS-C1	0.3658	0.6048	47.2104	4.5866	2.2832	6.8710
ANFIS-C2	0.2676	0.5173	69.1083	5.3597	2.8365	8.3131
ANFIS-C3	0.0955	0.3091	55.4399	4.9060	2.5431	7.4458
RF-C1	0.1689	0.4109	59.0666	4.9893	2.6191	7.6855
RF-C2	0.4292	0.6551	72.2224	5.4900	2.9528	8.4984
RF-C3	0.1675	0.4093	59.0833	5.0048	2.6377	7.6866

framework for ^{232}Th mobility, which is vital for identifying high-risk areas, informing land management practices, and developing targeted remediation strategies to mitigate health risks in abandoned mine sites.

Based on the testing phase results (Table 2), the GPR model with feature combination C1 (GPR-C1) demonstrated the highest predictive accuracy for ^{232}Th concentrations, achieving the best DC=0.5848, PCC=0.7647, and the lowest error metrics (MSE: 20.9833, MAPE: 3.0468, MAE: 1.6044, RMSE: 4.5808). In contrast, models like GPR-C3 and LSTM-C2 performed poorly. These findings, supported by the scatter plots and histograms in Fig. 4, confirm that GPR-C1 is the most promising model and underscore the critical importance of model selection and optimal feature combinations for achieving superior predictive performance. The accuracy of GPR-C1 highlights its potential for practical application in environmental risk assessment, enabling the identification of high-risk areas and guiding targeted remediation strategies to mitigate health risks from radionuclide exposure in abandoned mine sites.

A comparison of error metrics (MAPE, MAE) between training and testing phases reveals significant differences in model generalization. GPR-C1 demonstrated superior performance and robustness, with its error metrics substantially decreasing from training (MAPE: 8.9909, MAE: 3.5236) to testing (MAPE: 3.0468, MAE: 1.6044). LSTM-C1 and ANFIS-C1 also showed significant improvement on unseen data. In contrast, models like GPR-C2, RF-C1, and RF-C3 maintained high errors in both phases, indicating poor performance and an inability to generalize. These quantitative findings are visually confirmed by the residual histograms in Fig. 5. GPR-C1 exhibited a narrow, symmetric distribution of errors, indicative of an accurate and well-fitting model. Conversely, poorer-performing models such as GPR-C2, GPR-C3, and the RF variants displayed wider, skewed error distributions, revealing model biases and fit issues. This analysis stresses the critical importance of model selection and feature combination, with GPR-C1 emerging as the most accurate and reliable model for predicting ^{232}Th distribution.

The evaluation of the machine learning models revealed that GPR-C1 demonstrated superior predictive accuracy for ^{232}Th behaviour, achieving the highest skill during the testing phase (DC = 0.5848; PCC = 0.7647) and showing a marked improvement in error metrics from training to testing (MAE: 3.5236 to 1.6044; MAPE: 8.9909% to 3.0468%). LSTM-C1 and ANFIS-C1 also exhibited reasonably good performance, with intermediate DC and PCC values, whereas RF-

based models showed comparatively low skill (DC < 0.20; PCC < 0.42) across both phases. The superior performance of models using the optimal feature combination (C1) underscores the importance of appropriate feature selection for enhancing model skill, consistent with previous findings on variable selection and model performance (Inakollu et al., 2009).

These findings have significant environmental implications. The high accuracy of top-performing models like GPR-C1 provides a robust, cost-effective alternative to traditional methods for predicting ^{232}Th distribution, directly enhancing environmental risk assessment and the ability to guide targeted remediation efforts in abandoned mine sites. This aligns with broader research highlighting the efficacy of ML models in handling complex, non-linear environmental data (Nourani et al., 2020). The study confirms that advanced, kernel-based models like GPR are particularly practical tools for risk prediction and remediation planning in complex soil systems. Based on the evidence of the high accuracy of the GPR-C1 model and

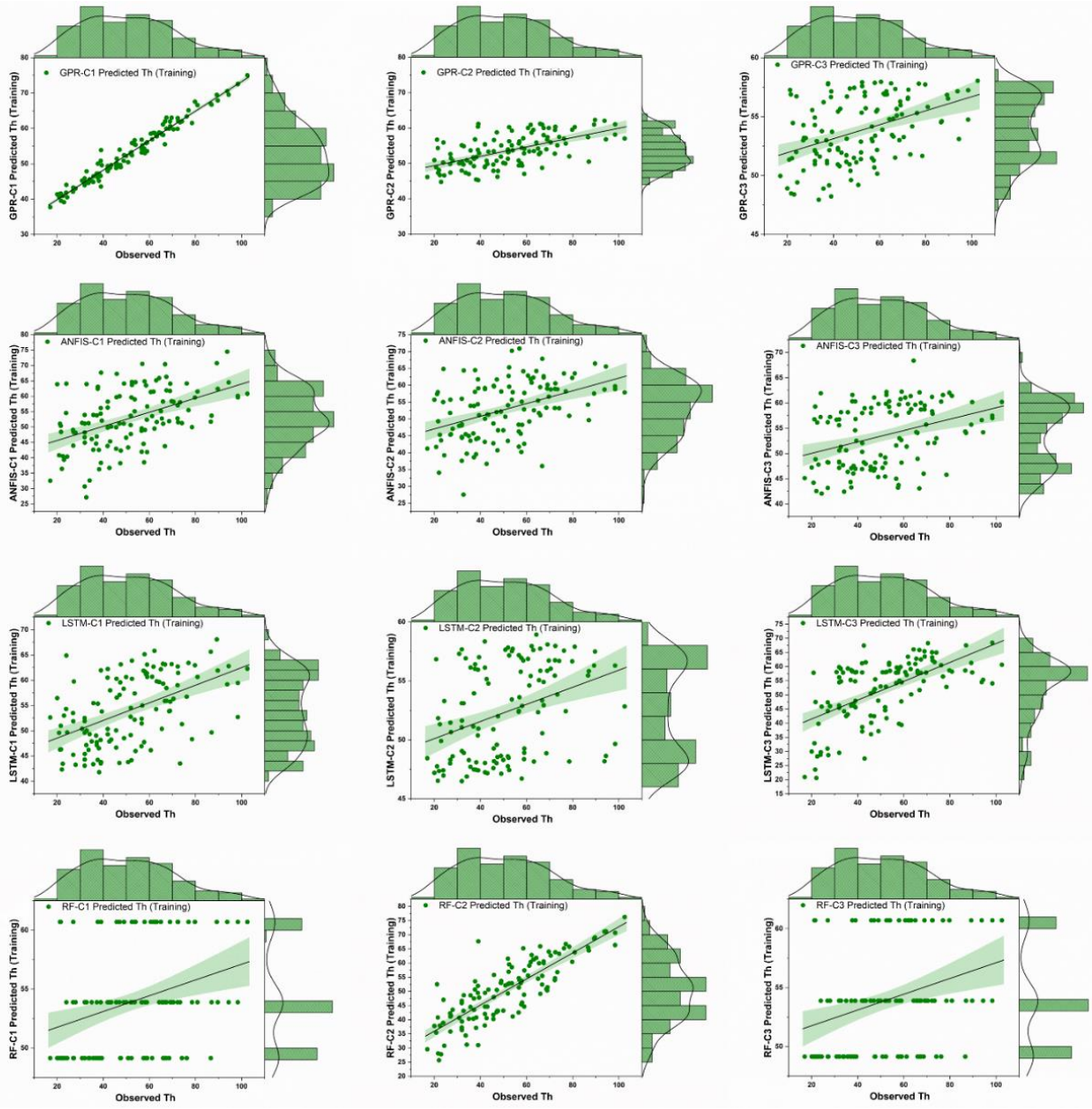
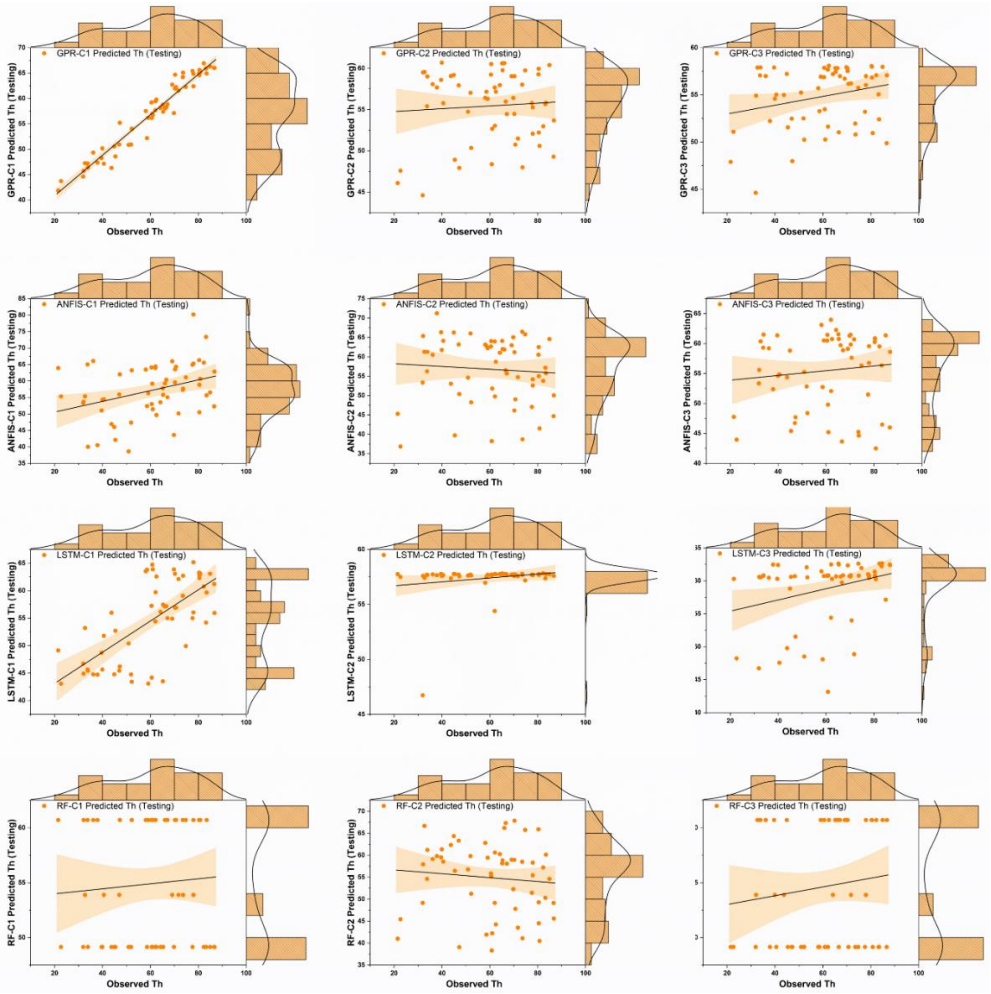


Fig. 4. The predictive comparison in terms of goodness-of-fit between the observed and experimental ^{232}Th concentrations.



Continued Fig. 4. The predictive comparison in terms of goodness-of-fit between the observed and experimental ^{232}Th concentrations.

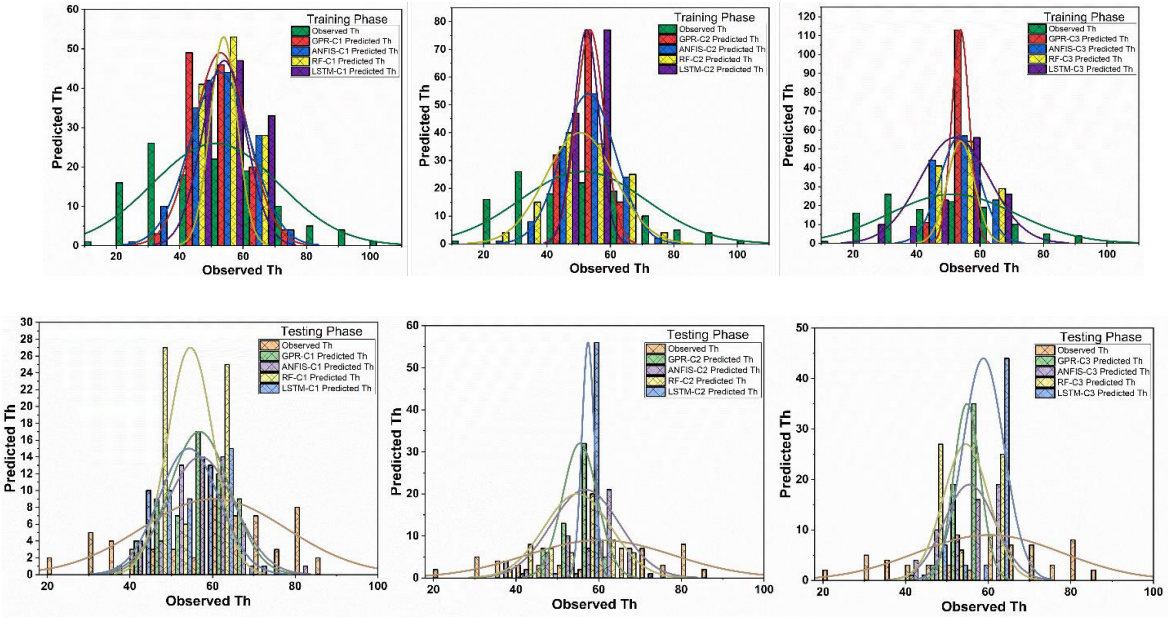


Fig. 5. Cumulative distribution plot based on probability prediction of ^{232}Th.

its ability to robustly predict ^{232}Th distribution in complex, non-linear systems, the predictive method should be viewed as a highly reliable, first-pass screening and prioritization tool, but not a direct replacement for confirmatory direct measurement methods like gamma spectroscopy. However, it can serve as a superior alternative to traditional statistical models. But, due to regulatory requirements and the need for absolute quantification, it should function as a secondary measurement method used to intelligently guide and minimize the required number of confirmatory direct measurements (gamma spectroscopy).

CONCLUSION

This study successfully developed and validated a novel ML framework to predict the spatial distribution of ^{232}Th in soils from abandoned mining areas. The research confirmed that while key soil properties (pH, OC, OM, CEC) exhibit only weak linear correlations with ^{232}Th concentration, advanced ML models are capable of capturing the underlying non-linear and complex geochemical relationships governing its mobility.

Through rigorous validation against simple baselines and the use of grouped cross-validation to prevent data leakage, the Gaussian Process Regression (GPR) model, utilizing the full feature set (pH, OC, OM, CEC), was identified as the most robust and accurate predictor. The GPR-C1 model demonstrated superior performance in the testing phase (DC = 0.5848; MAPE = 3.05%), proving its ability to generalize well to unseen data. This performance features the critical importance of both model selection and optimal feature combination for accurate environmental forecasting.

The study makes a significant methodological contribution by demonstrating that advanced ML models like GPR provide a substantial predictive advantage over simpler linear models for this complex geochemical problem. Consequently, this research provides a cost-effective, data-driven tool for environmental managers. The framework enables the reliable identification of high-risk contamination zones and can effectively guide targeted and effective remediation strategies. Therefore, this offers a practical solution for monitoring and provides cost-effective framework for environmental risk assessment and management of radionuclide-contaminated sites, particularly in resource-limited regions.

FUNDING

No funding was obtained for this study.

DECLARATION ON CONFLICT OF INTEREST

The authors declare that there is not any conflict of interest of whatsoever regarding the submission and this research work.

DATA AVAILABILITY STATEMENT

The data that have been used to support the findings of this study can be made available and shared upon request.

REFERENCES

Aba, A., Al-Boloushi, O., Ismaeel, A., & Al-Tamimi, S. (2021). Migration behaviour of radiostrontium and radiocesium in arid-region soil. *Chemosphere*, *281*, 130953.

Abba, S. I., Benaafi, M., Usman, A. G. and Aljundi, I. H. (2022). Inverse groundwater salinization

modeling in a sandstone's aquifer using stand-alone models with an improved non-linear ensemble machine learning technique. *Journal of King Saud University-Computer and Information Sciences*, *34*(10), 8162-8175.

Abba, S. I., Benaafi, M., Usman, A. G. and Aljundi, I. H. (2023). Sandstone groundwater salinization modelling using physicochemical variables in Southern Saudi Arabia: Application of novel data intelligent algorithms. *Ain Shams Engineering Journal*, *14*(3), 101894.

Abba, S. I., Usman, A. G. and Selin, I. (2020). Simulation for response surface in the HPLC optimization method development using artificial intelligence models: A data-driven approach. *Chemometrics and Intelligent Laboratory Systems*, *201*, 104007.

Abdullahi, H. U., Usman, A. G., Abba, S. I. and Abdullahi, H. U. (2020). Modelling the absorbance of a bioactive compound in HPLC method using artificial neural network and multilinear regression methods. *Dutse Journal of Pure Applied Science, 6*, 362-371.

Alamrouni, A., Aslanova, F., Mati, S., Maccido, H. S., Jibril, A. A., Usman, A. G. and Abba, S. I. (2022). Multi-regional modeling of cumulative COVID-19 cases integrated with environmental forest knowledge estimation: A deep learning ensemble approach. *International Journal of Environmental Research and Public Health*, *19*(2), 738.

Alshahrani, B., Fares, S., Salman, M., & Korna, A. H. (2025). Assessment of natural radioactivity levels in black sand and sand sediments in the Mediterranean coast region, Egypt. *Environmental Challenges*, *18*, 101061.

Alzubaidi, G., Hamid, F. B. and Abdul Rahman, I. (2016). Assessment of natural radioactivity levels and radiation hazards in agricultural and virgin soil in the state of Kedah, North of Malaysia. *The Scientific World Journal*, *2016*(1), 6178103.

Amir, M. N. I., Ismail, N. I., Wood, A. K., Saat, A. and Hamzah, Z. (2015). Effectiveness of mineral soil to adsorb the natural occurring radioactive material (norm), uranium and thorium. In *AIP Conference Proceedings* (Vol. 1659, No. 1). AIP Publishing.

Barnekow, U., Fesenko, S., Kashparov, V., Kis-Benedek, G., Matisoff, G., Onda, Y., ... & Varg, B. (2019). Guidelines on soil and vegetation sampling for radiological monitoring, technical reports series no. 486 of International Atomic Energy Agency (IAEA) Vienna.

Beretta, A., Bassahum, D., & Musselli, R. (2014). ¿ Medir el pH del suelo en la mezcla suelo: agua en reposo o agitando?. *Agrociencia (Uruguay)*, *18*(2), 90-94.

Ferdosi, H., Abbasianjahromi, H., Banihashemi, S. and Ravanshadnia, M. (2023). BIM applications in sustainable construction: scientometric and state-of-the-art review. *International Journal of Construction Management*, *23*(12), 1969-1981.

Gbadamosi, A., Adamu, H., Usman, J., Usman, A. G., Jibril, M. M., Salami, B. A. and Abba, S. I. (2024). New-generation machine learning models as prediction tools for modeling interfacial tension of hydrogen-brine system. *International Journal of Hydrogen Energy*, *50*, 1326-1337.

Gillman, G. P. and Sumpter, E. A. (1986). Modification to the compulsive exchange method for measuring exchange characteristics of soils. *Soil Research*, *24*(1), 61-66.

Guagliardi, I., Rovella, N., Apollaro, C., Bloise, A., Rosa, R. D., Scarciglia, F. and Buttafuoco, G. (2016). Modelling seasonal variations of natural radioactivity in soils: A case study in southern Italy. *Journal of Earth System Science*, *125*, 1569-1578.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*, 1157-1182.

Habib, M. A., Basuki, T., Miyashita, S., Bekelesi, W., Nakashima, S., Phoungthong, K. and Techato, K. (2019). Distribution of naturally occurring radionuclides in soil around a coal-based power plant and their potential radiological risk assessment. *Radiochimica Acta*, *107*(3), 243-259.

Hazelton, P. A and Murphy, B.W (2007) Interpreting Soil Test Results: What Do All The Numbers Mean?. CSIRO Publishing: Melbourne. Pp 51.

Hofmann, P., Achatz, M., Fohlmeister, J., Schmidt, K., Berg, T., & Sarvan, I. (2025). Levels of naturally occurring radionuclides in foods from the first German total diet study. *Science of The Total Environment*, *965*, 178653.

IAEA-TECDOC-1415. (2004). Soil sampling for environmental contaminants.

Inakollu, P., Philip, T., Rai, A. K., Yueh, F. Y. and Singh, J. P. (2009). A comparative study of laser induced breakdown spectroscopy analysis for element concentrations in aluminum alloy using artificial neural networks and calibration methods. *Spectrochimica Acta Part B: Atomic Spectroscopy*, *64*(1), 99-104.

Jibril, M. M., Malami, S. I., Jibrin, H. B., Muhammad, U. J., Duhu, M. A., Usman, A. G. and Abba, S.

I. (2024). New random intelligent chemometric techniques for sustainable geopolymer concrete: Low-energy and carbon-footprint initiatives. *Asian Journal of Civil Engineering*, *25*(2), 2287-2305.
Li, H., Wang, Q., Zhang, C., Su, W., Ma, Y., Zhong, Q., ... & Xiao, T. (2024). Geochemical distribution and environmental risks of radionuclides in soils and sediments runoff of a uranium mining area in South China. *Toxics*, *12*(1), 95.
Maher, K., Bargar, J. R., & Brown Jr, G. E. (2013). Environmental speciation of actinides. *Inorganic Chemistry*, *52*(7), 3510-3532.
Meng, Y., Jiang, J. and Wu, J. (2024). A physics-enhanced neural network for estimating longitudinal dispersion coefficient and average solute transport velocity in porous media. Geophysical Research Letters, 51, e2024GL110683.
Nourani, V., Gökçekuş, H., Umar, I. K., & Najafi, H. (2020). An emotional artificial neural network for prediction of vehicular traffic noise. *Science of the Total Environment*, *707*, 136134.
Osman, R., Dawood, Y. H., Melegy, A., El-Bady, M. S., Saleh, A. and Gad, A. (2022). Distributions and risk assessment of the natural radionuclides in the soil of Shoubra El Kheima, South Nile Delta, Egypt. *Atmosphere*, *13*(1), 98.
Ouyang, N., Zhang, P., Zhang, Y., Sheng, H., Zhou, Q., Huang, Y., & Yu, Z. (2023). Cation exchange properties of subsurface soil in mid-subtropical China: Variations, correlation with soil-forming factors, and prediction. *Agronomy*, *13*(3), 741.
Pecha, P., Tichý, O. and Pechová, E. (2021). Determination of radiological background fields designated for inverse modelling during atypical low wind speed meteorological episode. *Atmospheric Environment*, *246*, 118105.
Peng, F., Wen, J., Zhang, Y. and Jin, J. (2020, September). Monthly streamflow prediction based on random forest algorithm and phase space reconstruction theory. In *Journal of Physics: Conference Series* (Vol. 1637, No. 1, p. 012091). IOP Publishing.
Saravani, M. J., Noori, R., Jun, C., Kim, D., Bateni, S. M., Kianmehr, P. and Woolway, R. I. (2025). Predicting chlorophyll-a concentrations in the world's largest lakes using Kolmogorov–Arnold Networks. Environmental Science & Technology, 59(3), 1801–1810.
Sarkar, B., Mukhopadhyay, R., Ramanayaka, S., Bolan, N., & Ok, Y. S. (2021). The role of soils in the disposition, sequestration and decontamination of environmental contaminants. *Philosophical Transactions of the Royal Society B*, *376*(1834), 20200177.
Shang, Z. and He, J. (2018). Predicting Hourly $\mathbf {PM} _ {2.5} $ Concentrations Based on Random Forest and Ensemble Neural Network. In *2018 Chinese Automation Congress (CAC)* (pp. 2341-2345). IEEE.
Smičiklas, I., & Šljivić-Ivanović, M. (2016). Pollutants Mobility with Implication to Remediation Strategies. *Soil Contamination: Current Consequences and Further Solutions*, 253.
Thakur, P., Ward, A. L., & González-Delgado, A. M. (2021). Optimal methods for preparation, separation, and determination of radium isotopes in environmental and biological samples. *Journal of Environmental Radioactivity*, *228*, 106522.
UNSCEAR. (2008): Sources and Effects of Ionizing Radiation: Report to the General Assembly, With Scientific Annexes, 2, 1–219. United Nations, New York
Usman, A. G., Tanimu, A., Abba, S. I., Isik, S., Aitani, A. and Alasiri, H. (2023). Feasibility of the Optimal Design of AI-Based Models Integrated with Ensemble Machine Learning Paradigms for Modeling the Yields of Light Olefins in Crude-to-Chemical Conversions. *ACS omega*, *8*(43), 40517-40531.
Usman, A. G., Işik, S., & Abba, S. I. (2021). Hybrid data-intelligence algorithms for the simulation of thymoquinone in HPLC method development. *Journal of the Iranian Chemical Society*, *18*(7), 1537-1549.
Uzun-Ozsahin, D., Precious Onakpojeruo, E., Bartholomew Duwa, B., Usman, A. G., Isah Abba, S. and Uzun, B. (2023). COVID-19 Prediction Using Black-Box Based Pearson Correlation Approach. *Diagnostics*, *13*(7), 1264.
Wan, H., Xiang, L., Cai, Y., Xie, Y. and Xu, R. (2025). Temporal and spatial feature extraction using graph neural networks for multi-point water quality prediction in river network areas. Water Research, 281, 123561.
Xu, X., Mumford, T. and Zou, P. X. (2021). Life-cycle building information modelling (BIM) engaged framework for improving building energy performance. *Energy and Buildings*, *231*, 110496.

Zheng, Y., Zhang, X., Zhou, Y. and Zhang, Y.-P. (2025). Deep representation learning enables cross-basin water quality prediction under data-scarce conditions. npj Clean Water, 8(1), 466.

Zhi, W., Appling, A. P., Golden, H. E., Podgorski, J. and Li, L. (2024). Deep learning for water quality. Nature Water, 2, 228–241.

Zubaidi, S. L., Al-Bugharbee, H., Ortega-Martorell, S., Gharghan, S. K., Olier, I., Hashim, K. S. and Kot, P. (2020). A novel methodology for prediction urban water demand by wavelet denoising and adaptive neuro-fuzzy inference system approach. *Water*, *12*(6), 1628.