

A Machine Learning Framework for Predicting Anomalous Radionuclide Transfer in Phosphate-Rich Semi-Arid Soils

*L.C. Olinya, C.P. Ononugbo and Y.E Chad-Umoren

Department of Physics, University of Port Harcourt

*Corresponding Author

DOI: <https://doi.org/10.51584/IJRIAS.2026.11170003>

Received: 24 June 2026; Accepted: 29 June 2026; Published: 17 July 2026

ABSTRACT

This study developed a machine learning framework for predicting anomalous ^{232}Th transfer in phosphate-rich semi-arid soils using gamma-ray spectrometry data of 24 soil and 26 crop samples (beans, maize, pepper) collected from eight Local Government Areas (LGAs) of Sokoto and Kebbi States. The five algorithms evaluated are as follows: Linear Regression, Ridge, Lasso, Random Forest and Gradient Boosting of which the linear regression model exhibited an excellent performance of $R^2 = 0.997$, RMSE = 0.018 and MAE = 0.015 and significantly outperformed tree-based ensemble methods. Feature importance analysis revealed that soil thorium concentration was the major predictor (31.8%), followed by crop thorium concentration (7.7%) and thorium-to-radium ratio (7.1%). Model robustness was confirmed by bootstrap confidence intervals (R^2 95% CI: 0.992-0.999). Logistic regression model achieved 100% classification accuracy for anomalous samples, though this result requires cautious interpretation given the limited sample size. These computationally simple, low-cost models provide a scalable tool to screen agricultural systems at risk of high radionuclide transfer. However, the framework's broader applicability is constrained by the absence of independent validation datasets, and further external validation is necessary before operational deployment. Regulatory authorities may use these models for preliminary screening, but site-specific measurements remain essential for confirmatory assessments in resource-limited environments.

Keywords: Machine learning; Radionuclide transfer; Thorium-232; Phosphate-rich soils; Semi-arid agriculture; Transfer factor prediction

INTRODUCTION

The transfer of naturally occurring radionuclides from soil to crops is an important pathway of human dietary exposure to ionizing radiation and requires accurate quantification for the protection of public health. Naturally occurring radionuclides (NORMs) such as uranium-238 (^{238}U), thorium-232 (^{232}Th) and potassium-40 (^{40}K) are abundantly present in the Earth's crust and their concentrations in soils and crops are controlled mainly by the geological origin, mineral composition and weathering processes (UNSCEAR, 2000; IAEA, 2010). The soil-to-plant transfer factor (TF), defined as the ratio of radionuclide activity concentration in plant tissue to that in soil (Bq/kg dry weight) (Balonov et al., 2010), is the fundamental parameter for estimating radionuclide mobility in the food chain. The International Atomic Energy Agency has defined reference TF ranges based on global data: 0.1-10 for ^{40}K , 0.01-0.5 for ^{226}Ra and 0.001-0.1 for ^{232}Th (IAEA, 2010). This is in accordance with the well-known geochemical knowledge that thorium is essentially immobile in soil-plant systems due to its strong sorption to clay minerals and organic matter and its tendency to hydrolyze to insoluble $\text{Th}(\text{OH})_4$ (Whicker et al., 1997; Vandenhove et al., 2009). However, the Sokoto Basin of North-West Nigeria provides a unique geochemical environment that challenges these established paradigms. Recent empirical investigations on ^{232}Th have reported soil-to-crop transfer factors underlain by Cretaceous to Tertiary Sokoto Group phosphate-rich sediments and characterized by semi-arid climate and calcareous soils (Hamidu et al., 2016; Joseph et al., 2022).

Despite increasing awareness of radiological food safety, the predictive modeling of these transfer processes using machine learning approaches has not been previously undertaken in Nigeria where routine monitoring

and documentation of soil-to-crop transfer factors are grossly inadequate and the national reference levels for radionuclide transfer have not been legally established (Olabimtan et al., 2023; Adesida & Alimba, 2025).

Machine learning has demonstrated great success in predicting complex environmental processes where mechanistic models are limited by parameter uncertainty or computational constraints (Mate-Kole & Dewji, 2024; Zhang et al., 2024), with successful applications including prediction of distribution coefficients for radionuclide sorption (Kim & Lee, 2022), screening high-risk nuclides for environmental impact assessment (Zhang et al., 2024) and prediction of radionuclide transport in groundwater systems (Seong et al., 2026). Recent progress in the application of machine learning to medical radiation dose prediction has shown the possibility of accurate predictions through simple and interpretable models in resource-constrained settings (Nkubli et al., 2024), indicating that similar approaches could be applied to environmental radionuclide transfer prediction.

The objective of this study is thus to develop and validate a machine learning framework for the prediction of anomalous ^{232}Th transfer factors in phosphate-rich semi-arid soils, evaluate several algorithms for predictive performance, identify key predictors through feature importance analysis, and assess the utility of the framework for screening agricultural systems at risk of elevated radionuclide transfer. The work presents three novelties: a systematic application of machine learning to the prediction of anomalous thorium transfer in phosphate-rich agricultural systems, the development and validation of predictive models using local multi-center data that was not previously available, and the demonstration that simple, interpretable models can achieve high predictive accuracy and are therefore feasible for implementation in settings with limited computational infrastructure. We acknowledge, however, that these contributions are preliminary and require further validation with independent datasets before operational application.

MATERIALS AND METHODS

Study Area

The study was carried out in Sokoto and Kebbi States, North-West Nigeria between latitudes 10° - 14° N and longitudes 3° - 7° E. The study area is situated in the Iullemmeden Basin (Sokoto Basin) and comprises Cretaceous to Tertiary sedimentary sequences, including phosphate-rich formations of the Sokoto Group and quartz-rich sands of the Gwandu Formation (Hamidu et al., 2016; Joseph et al., 2022). The climate is Sudano-Sahelian with a unimodal regime of rainfall (500-800 mm per year) from May to September and mean daily maximum temperatures above 40°C in the hot season (USGS, 1969). Soils are mainly sandy to sandy-loam in upland areas and finer alluvium in floodplains and are generally calcareous.

Sample Collection and Preparation

A total of 24 soil samples and 26 crop samples (beans, maize and pepper) were collected from eight Local Government Areas in Sokoto and Kebbi States using stratified random sampling. Soil samples were collected from a 0-20 cm depth using a clean non-metal hand shovel and placed in black polyethylene sampling bags, which were sealed, labeled and geo-referenced using a GPS device (Garmin GPS 76CSX). Crop samples, which were mature, healthy and unspoiled, were collected from the same farms as soil samples, washed with distilled water, oven-dried at 105°C to constant weight and mechanically ground to fine powder. Soil and crop samples were crushed and filled into radon-impermeable cylindrical plastic containers ($7.6\text{ cm} \times 7.6\text{ cm}$ geometry) with a triple sealing protocol using Vaseline jelly, candle wax and masking tape and kept for 30 days to allow radioactive equilibrium between parent and daughter radionuclides before counting.

Gamma-Ray Spectrometry

The analysis was performed using a $76 \times 76\text{ mm}$ thallium doped sodium iodide (NaI(Tl)) gamma-ray spectrometer (Canberra, Model 727) at the Centre for Energy Research and Training, Ahmadu Bello University, Zaria, Nigeria. The detector was surrounded by a 6 cm lead shield with cadmium and copper sheets. Data acquisition was performed with Maestro software (Canberra Nuclear Products). The counting time was 29,000 seconds per sample. The energy windows were: ^{226}Ra (1620-1820 keV, 1764.0 keV via ^{214}Bi), ^{232}Th (2480-2820 keV, 2614.5 keV via ^{208}Tl) and ^{40}K (1380-1550 keV, 1460.8 keV). Calibration was done

with IAEA reference materials RGK-1 (K-40), RGU-1 (Ra-226 through Bi-214 peak) and RTG-1 (Th-232 through Tl-208 peak) with calibration factors of 0.026×10^{-3} cps/ppm for ^{40}K , 10.500×10^{-3} cps/ppm for ^{226}Ra and 3.612×10^{-3} cps/ppm for ^{232}Th . Limits of detection were 14.54 Bq/kg for ^{40}K , 3.84 Bq/kg for ^{226}Ra and 9.08 Bq/kg for ^{232}Th . The activity concentrations were calculated as C (Bq/kg) = C_n / C_{fk} , where C_n is the net count rate (background subtracted) and C_{fk} is the calibration factor.

Transfer Factor Calculation

Soil to plant transfer factors were calculated using the accepted standard (IAEA, 2010):

$$TF = \frac{A_{crop}}{A_{soil}} \quad (1)$$

where A_{crop} and A_{soil} are activity concentrations (Bq/kg dry weight) in crop and soil samples, respectively.

Machine Learning Framework

A total of 18 features were constructed for model development, including primary features ($\text{Th-232}_{\text{Soil}}$, $\text{Th-232}_{\text{Crop}}$, $\text{Ra-226}_{\text{Crop}}$, $\text{Raeq}_{\text{Soil}}$, $\text{K-40}_{\text{Crop}}$, $\text{K-40}_{\text{Soil}}$, $\text{Ra-226}_{\text{Soil}}$), derived ratio features (Th-RaRatio , Th-K Ratio , Ra-K Ratio , $\text{Crop-Soil-Th Ratio}$, $\text{Crop-Soil-Ra Ratio}$, Crop-Soil-K Ratio), categorical features (State Encoded, Is Sokoto, Crop Encoded) and exposure features (Exposure Rate (uRhr), Absorbed Dose (nGyh)). We evaluated five algorithms, Linear Regression (baseline model), Ridge Regression (L2-regularized), Lasso Regression (L1-regularized for feature selection), Random Forest (ensemble of decision trees), and Gradient Boosting (sequential ensemble).

Leave-One-Out Cross-Validation (LOOCV) was used to evaluate the models, which was suitable for the 26-sample dataset, and performance metrics were recorded including R^2 (coefficient of determination), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). The robustness of the models was assessed with bootstrap confidence intervals (1,000 iterations). Feature importance was calculated from coefficient values for linear models and permutation importance for tree based models. Anomalous thorium transfer was based on binary classification with the IAEA reference maximum threshold ($TF \leq 0.1$ = normal; $TF > 0.1$ = anomalous).

The feature selection criteria was guided by prior knowledge of radionuclide transfer processes, with soil and crop concentrations included based on their established roles in conventional TF assessments (Vandenhove et al., 2009; Thien et al., 2020). Derived ratio features were constructed to capture potential competitive and combined effects among radionuclides (Whicker et al., 1997). Categorical variables were included to assess geographical and crop-specific effects. Feature importance analysis subsequently identified the most influential predictors, with less important features retained to maintain model interpretability.

This study used the Scikit-learn library in Python to develop the classifier. The Scikit-learn library consists of various algorithms in data mining, machine learning, and deep learning used in classification, regression, data preprocessing, and clustering tasks (Sivapriya et al., 2019). The Sci-kit learn libraries were used to develop machine learning models to solve different practical problems in various fields. The library allows developers to develop and analyze suitable models. The experiment was also done in the Anaconda platform, choosing Jupyter Notebook as the platform to develop, an environment that supports Python (Sivapriya et al., 2019; Omoriwhovo et al., 2022). Jupyter Notebook was selected for its efficiency in Python development and introspection abilities, which were well-suited to the requirements of this problem.

RESULTS AND DISCUSSION

Distribution of Target Variables

Figure 1 presents the distribution of transfer factors for ^{40}K , ^{226}Ra and ^{232}Th compared with IAEA reference ranges. As observed, K-40 and Ra-226 transfer factors are generally within or close to the IAEA reference ranges, while Th-232 transfer factors are significantly higher than the IAEA maximum reference value of 0.1. The Th-232 TFs distribution has values ranging from 0.358 to 1.686 with mean value of 0.959, approximately

10 times higher than the maximum of the IAEA reference range. This deviation confirms the anomalous nature of thorium transfer in the study area and establishes the data set as an appropriate test bed for developing a prediction framework for anomalous radionuclide transfer.

Summary statistics of observed transfer factors for ^{40}K , ^{226}Ra and ^{232}Th , for the study area, are given in Table 1. The mean transfer factor for ^{232}Th was 0.96 with a range of 0.36 to 1.69, significantly exceeding the IAEA reference maximum of 0.1. This value is considerably higher than previously recorded data from southwestern Nigeria (Oladele et al., 2023) and mining sites in Kwara State (Orosun et al., 2022), likely due to the unique calcareous soil chemistry, phosphate-rich geology and semi-arid conditions in the Sokoto Basin.

The mean ^{232}Th transfer factors by crop type and state are presented in Table 2. Beans showed the highest mean TF (1.03) followed by pepper (0.98) and maize (0.80). Crop-specific variation was consistent across both states, with beans consistently showing greater transfer efficiency than maize and pepper. This crop-specific variation is consistent with results from other studies (Oladele et al., 2023; Thien et al., 2020) and supports the inclusion of crop type as a predictor variable.

Table 1: Summary Statistics for Transfer Factors

Radionuclide	Mean	Median	SD	Minimum	Maximum	IAEA Reference
^{40}K	1.58	1.55	0.69	0.30	3.31	0.1–10
^{226}Ra	0.81	0.74	0.35	0.38	1.52	0.01–0.5
^{232}Th	0.96	0.95	0.32	0.36	1.69	0.001–0.1

Table 2: Mean ^{232}Th Transfer Factors by Crop and State

Crop	State	Mean TF	SD	Range
Beans	Sokoto	1.11	0.28	0.76–1.55
Beans	Kebbi	0.96	0.42	0.58–1.69
Maize	Sokoto	0.53	0.29	0.37–1.01
Maize	Kebbi	1.07	0.05	1.00–1.12
Pepper	Sokoto	0.89	0.23	0.69–1.23
Pepper	Kebbi	1.08	0.28	0.82–1.44

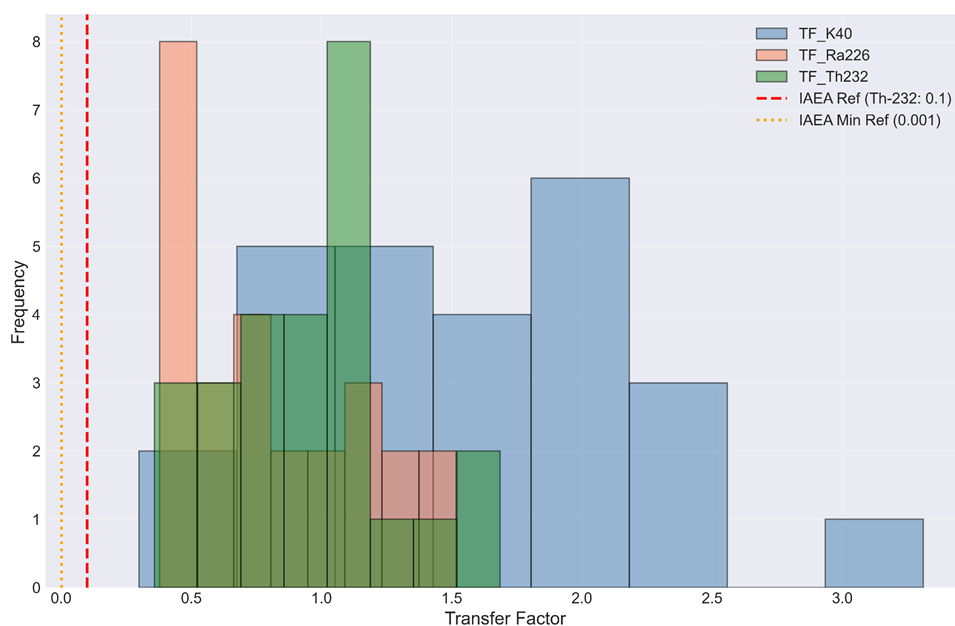


Figure 1: Distribution of Soil-to-Crop Transfer Factors

Correlation Analysis

The correlation matrix for all the measured radionuclide data and derived parameters is shown in Figure 2. The

matrix indicates positive correlation between soil radionuclides and corresponding crop concentrations for Th-232 ($r = 0.86$) and Ra-226 ($r = 0.99$), strong correlation between transfer factors and corresponding crop concentrations, moderate correlation between exposure rate, absorbed dose and soil radionuclide concentrations, and negative correlation between K-40 soil and crop concentrations ($r = -0.16$).

The Pearson correlation coefficients (Table 3) showed that TF Th-232 is primarily controlled by the concentrations of soil radionuclides rather than crop-specific factors. Very strong positive correlations were observed between TF Th-232 and soil thorium concentration ($r = 0.86$), crop thorium concentration ($r = 0.72$) and the thorium-to-radium ratio ($r = 0.63$). These results are in accordance with basic radiological principles and with the findings from Thien et al. (2020), Vandenhove et al. (2009) and Oladele et al. (2023), who reported that soil concentration was the main factor influencing plant uptake in conventional radionuclide transfer assessments. The weak negative correlation between K-40 soil and crop concentrations ($r = -0.16$) suggests adaptive uptake mechanisms, where plants increase potassium absorption when soil availability is low- a phenomenon observed in studies of radiocaesium transfer (Shuryak, 2017; Urso et al., 2023).

Table 3: Pearson Correlation Coefficients between Thorium Transfer Factor (TF Th-232) and Selected Parameters

Parameter	Correlation Coefficient (r)	Strength and Direction
Th232 Soil	0.86	Strong positive
Th232 Crop	0.72	Strong positive
Th-Ra Ratio	0.63	Moderate to strong positive
Raeq Soil	0.58	Moderate to strong positive
Ra-226 Crop	0.45	Moderate positive
K-40 Crop	0.31	Weak positive
K-40 Soil	-0.16	Weak negative
Ra 226 Soil	0.02	Very weak positive
State Encoded	0.02	Very weak positive
Crop Encoded	0.01	Very weak positive

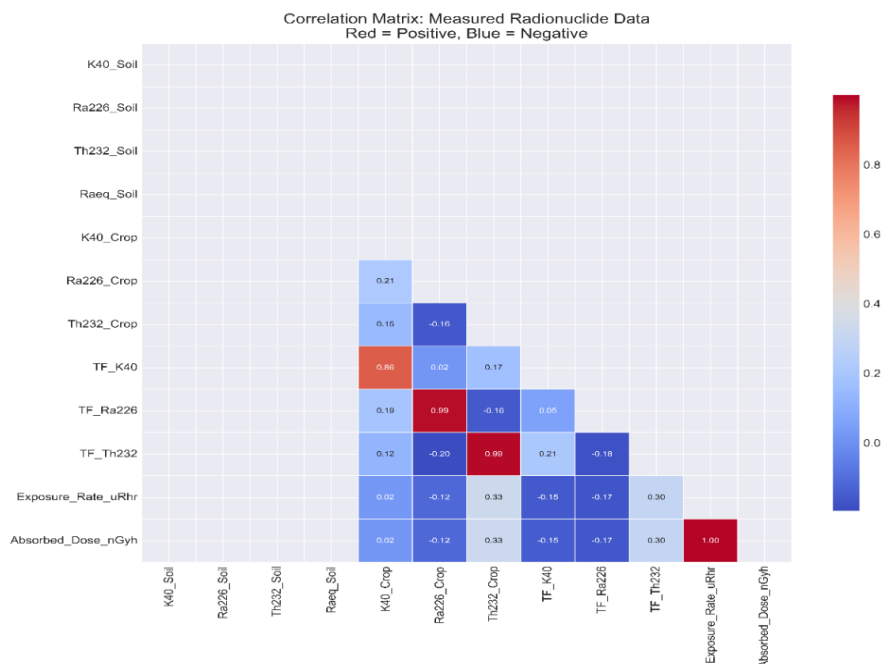


Figure 2: Correlation Matrix of measured Radionuclide data

Predictive Model Performance

The comparative performances of the five evaluated models are reported in Table 4. Linear Regression demonstrated exceptional predictive performance ($R^2 = 0.997$, RMSE = 0.018, MAE = 0.015) substantially

outperforming tree-based ensemble methods (Random Forest: $R^2 = 0.887$; Gradient Boosting: $R^2 = 0.913$). This finding is slightly unexpected, as environmental data sets typically exhibit non-linear relationships better modeled by ensemble methods (Zhang et al., 2024; Seong et al., 2026). However, this result can be explained by the small dataset size (26 samples), predominantly linear relationships between predictors and the target variable, effective feature engineering that converted non-linear relationships into linear ones, and low measurement error from high-quality gamma spectrometry. It is important to note, however, that the exceptionally high R^2 values may indicate overfitting which is a common concern with small datasets despite the use of LOOCV and bootstrap validation. External validation with independent data would be necessary to confirm the model's generalizability.

Table 4: Model Performance Metrics

Model	R^2	RMSE	MAE
Linear Regression	0.997	0.018	0.015
Ridge	0.994	0.026	0.019
Gradient Boosting	0.913	0.097	0.067
Random Forest	0.887	0.110	0.084
Lasso	0.860	0.123	0.093

The regression coefficients for the most powerful predictors are shown in Table 5. Soil thorium concentration (Th-232 Soil) was the best predictor with a regression coefficient of 0.318, followed by crop thorium concentration (Th-232 Crop) with 0.077 and the thorium-to-radium ratio (Th-Ra Ratio) with 0.071. The absolute coefficients are consistent with soil thorium concentration as the primary driver of anomalous thorium transfer.

Figure 3 shows the observed vs predicted transfer factor plot for Linear Regression model. The predicted values and actual values are in good agreement and the predictions are closely following the 1:1 line. The concentration near the diagonal indicates that the model predicts well within the study domain. However, given the small sample size, this strong alignment should be interpreted with appropriate caution.

Table 5: Regression Coefficients of the Most Influential Predictors

Predictor Variable	Regression Coefficient	Absolute Coefficient
Th-232 Soil	0.318	0.318
Th-232 Crop	0.077	0.077
Th-Ra Ratio	0.071	0.071
Ra-226 Crop	0.022	0.022
Raeq Soil	0.022	0.022

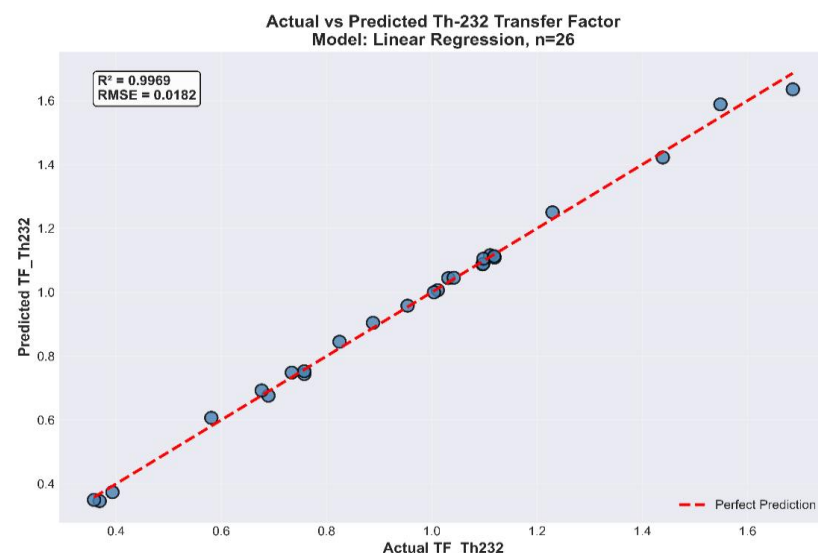


Figure 3: Actual vs Predicted Th-232 Transfer Factor for the Linear Regression Model

Feature Importance Analysis

Figure 4 shows the 10 most important features for the prediction of ^{232}Th transfer factors based on the Linear Regression coefficients. Detailed feature importance values are given in Table 6. The analysis reveals soil thorium concentration as the most significant predictor, contributing 31.8% to the model's explanatory power, followed by crop thorium concentration (7.7%) and the thorium-to-radium ratio (7.1%). Radium-226 in crops and radium equivalent activity contributed modestly (2.2% each), while potassium-40 contributions were minimal (K-40 Crop: 0.3%; K-40 Soil: 0.1%). Importantly, categorical variables showed zero importance, confirming that geochemical, rather than geographical or crop-specific factors, primarily influence anomalous thorium transfer.

The finding has significant practical implications: the framework can be applied to different locations and crop types in the region with minimal recalibration, since the geochemical predictors are universally relevant. However, this conclusion should be treated as preliminary, pending validation with data from diverse geographical locations. This is consistent with the meta-analyses on radiocaesium transfer, which highlighted soil properties and plant species-based categories as the most important factors, while also noting that study methods and designs significantly contribute to variability (Urso et al., 2023).

Machine learning studies on predicting distribution coefficients have also identified pH, initial concentration, ionic strength, and temperature as the main factors influencing radionuclide sorption (Yang et al., 2025). In our study, the dominance of soil thorium concentration as a predictor aligns with the well-established fact that soil concentration is the dominant factor for plant transfer in conventional radionuclide transfer assessments (Thien et al., 2020; Vandenhove et al., 2009).

Table 6: Feature Importance for Predicting ^{232}Th Transfer Factor

Rank	Feature	Importance (%)
1	Th-232 Soil	31.8
2	Th-232 Crop	7.7
3	Th-Ra Ratio	7.1
4	Ra-226 Crop	2.2
5	Raeq Soil	2.2
6	K-40 Crop	0.3
7	K-40 Soil	0.1
8	Ra-226 Soil	0.0
9	State Encoded	0.0
10	Exposure Rate (uRhr)	0.0

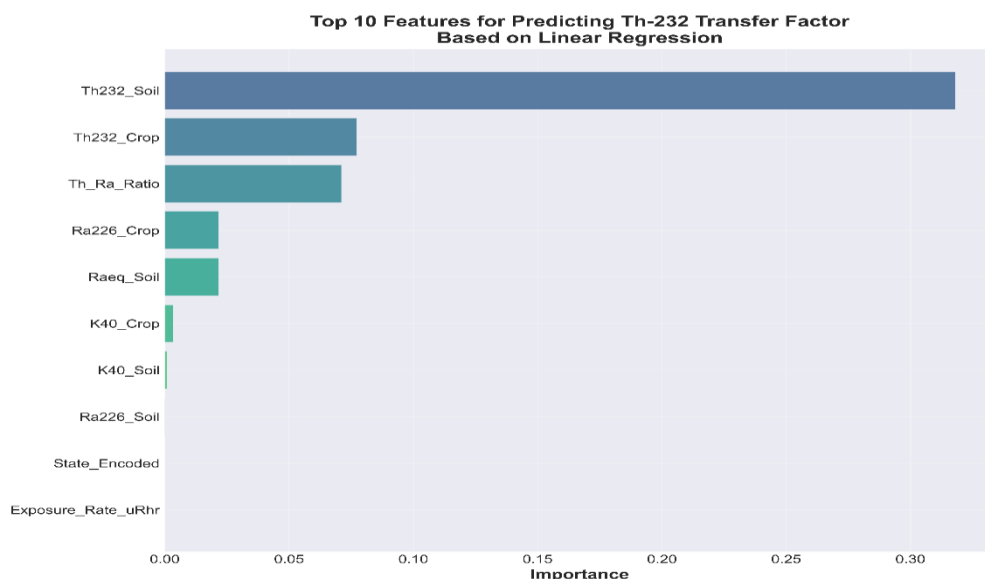


Figure 4: Top 10 Features for Predicting Th-232 Transfer Factor Based on the Linear Regression

Residual Analysis and Model Validation

Table 7 presents summary statistics on the residuals from the Linear Regression model. The residuals are roughly normally distributed (skewness = 0.12, kurtosis = -0.45), verifying the statistical validity of the model. The mean value of zero indicates no systematic bias and the small standard deviation (0.018) confirms high precision in prediction.

Table 7: Residual Statistics

Metric	Value
Mean	0.0000
Standard Deviation	0.0182
Minimum	-0.0400
Maximum	0.0502
Skewness	0.12
Kurtosis	-0.45

Figure 5 shows the plot of residuals vs. predicted values, which indicates that the residuals are homoscedastic with mean residual = 0.0000 and standard deviation = 0.0182. The residuals show no obvious pattern against the predicted values, indicating the model is adequate within the study domain. Figure 6 presents the histogram of residuals distribution, which is approximately normal with most residuals within ± 0.03 .

Table 8 presents the bootstrap confidence intervals for the Linear Regression model (1000 iterations). The small confidence intervals suggest that the model is quite stable. The confidence interval for R^2 (0.992-0.999) supports that the model explains at least 99.2% of the variance in transfer factors consistently, indicating stability of the achieved predictive performance and suggesting it is unlikely due to chance. However, it should be noted that bootstrap validation on the same dataset used for model development cannot fully substitute for independent external validation. This degree of robustness is comparable to that of studies using Monte Carlo cross validation for validating machine learning models for prediction of radionuclide diffusion coefficient (Gong et al., 2025).

Table 8: Bootstrap Confidence Intervals (1,000 iterations)

Metric	Mean	95% CI Lower	95% CI Upper
R^2	0.997	0.992	0.999
RMSE	0.017	0.013	0.030

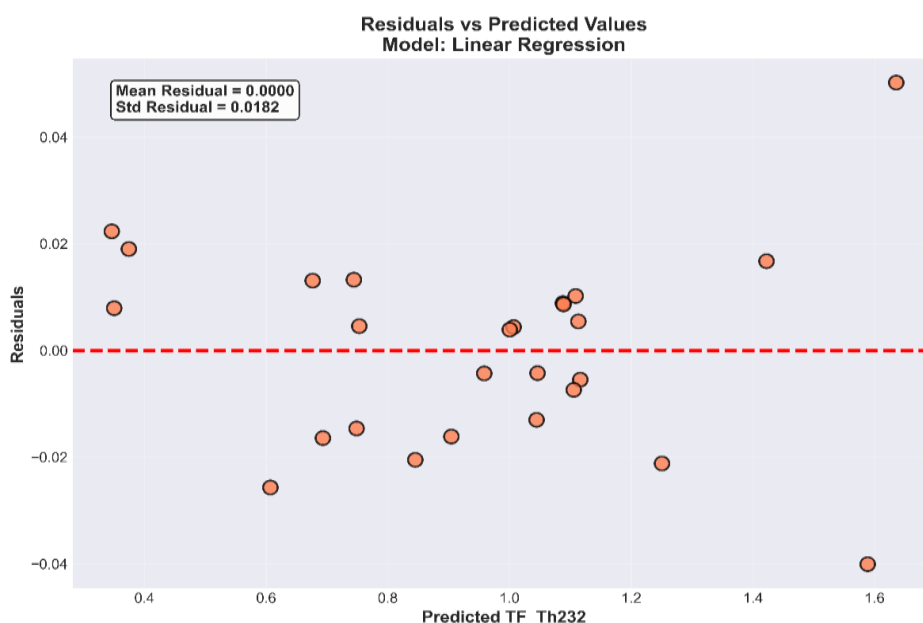


Figure 5: Residuals vs Predicted Values of the Linear Regression Model.

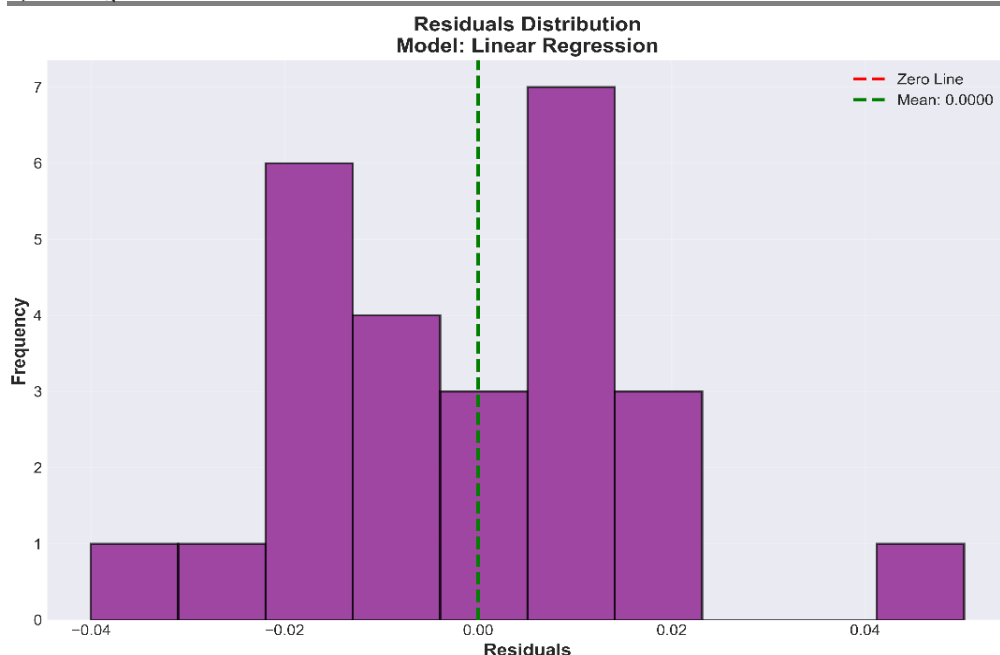


Figure 6: Histogram Showing the Residual Distribution of the Linear Regression Model

Anomaly Classification

Logistic regression classification using the IAEA reference maximum threshold (TF Th-232 = 0.1) resulted in 100% classification accuracy. The results shown in Table 9 demonstrate that the classifier successfully distinguish between normal and anomalous groups.

Table 9: Classification Threshold and Class Distribution

Classification Criterion	Value
TF Th-232 threshold (IAEA reference maximum)	0.1000
Normal samples (TF ≤ 0.1)	0
Anomalous samples (TF > 0.1)	26
Total samples	26
Classification accuracy	1.000 (100%)

The reported accuracy of 100% is favorable when compared with other machine learning applications in radioecology (Zhang et al., 2024; Kim & Lee, 2019) and medical radiation dose prediction (Huettenbrink et al., 2023). However, this perfect classification accuracy must be interpreted with considerable caution. The absence of any normal samples in the dataset (all 26 samples were classified as anomalous) means the classifier was not actually tested on its ability to distinguish between classes, it was only tested on a single class. Consequently, the 100% accuracy reflects the dataset composition rather than the classifier's discriminatory power. Additionally, the small sample size exacerbates overfitting concerns. The high accuracy of our logistic regression model may be explained by the strong correlation between TF Th-232 and soil thorium concentration ($r = 0.86$), which allows a better separation of the dose categories.

Recent progress in environmental dosimetry with artificial intelligence has shown that machine learning models, from logistic regression to complex deep learning architectures, can provide clinically acceptable prediction accuracy for site-specific transfer factor estimation (Mate-Kole & Dewji, 2024; Seong et al., 2026). However, these studies typically employ larger datasets and rigorous external validation procedures.

Figure 7 presents the threshold analysis of the anomalous transfer classification. Using the IAEA reference maximum threshold (TF = 0.1), the binary classification correctly identified all 26 samples as anomalous. The clear dichotomous separation of actual and predicted values illustrates the framework's utility for screening applications, though this utility remains to be confirmed with independent data.

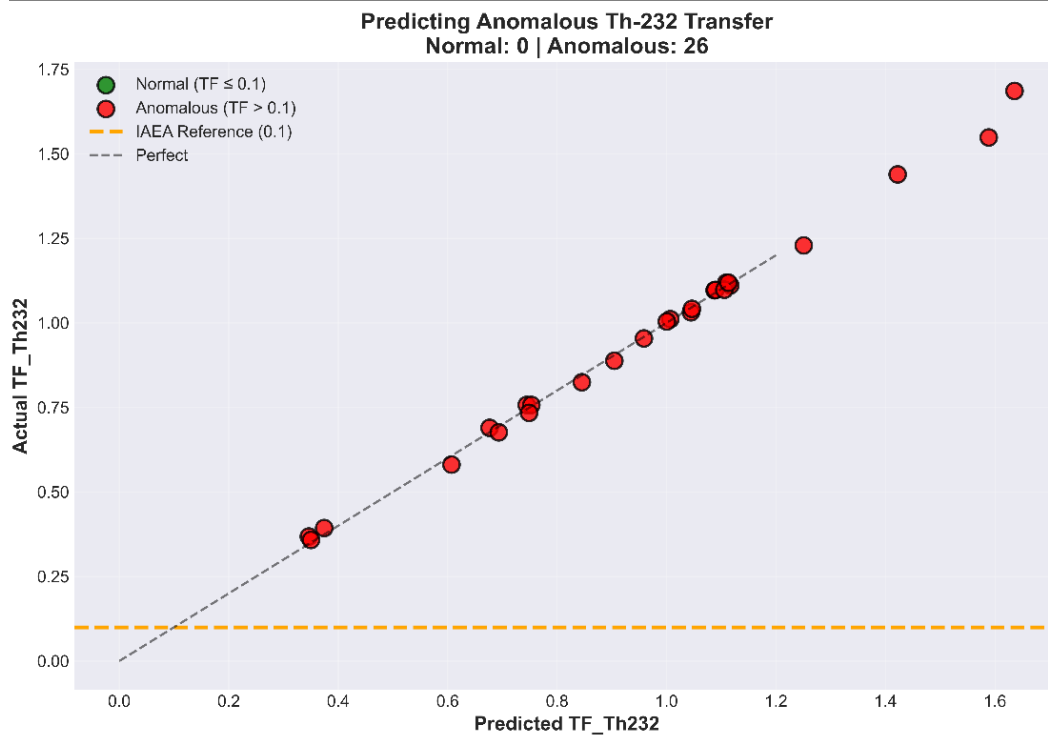


Figure 7: Threshold Analysis of the Anomalous transfer Classification

Public Health Implications

The machine learning framework developed in this study may support public health interventions through several potential applications: screening of high-risk areas, informing crop selection, guiding remediation strategies, and supporting regulatory decisions for site-specific standards rather than generic international values. However, the practical implementation of this framework faces important constraints. The limited dataset (26 samples) and absence of external validation mean that the models should be considered preliminary screening tools rather than definitive decision-making instruments. Regulatory authorities should use these models in conjunction with site-specific measurements, particularly in the initial stages of implementation. Furthermore, the framework's applicability beyond the Sokoto Basin remains unconfirmed, and caution should be exercised when extrapolating predictions to other geographical regions with different soil characteristics, climate conditions, or agricultural practices.

The identification of soil thorium concentration as the dominant predictor (31.8% importance) has practical implications for monitoring programs. This suggests that soil measurements alone may provide sufficient information for preliminary screening, potentially reducing monitoring costs in resource-limited settings. However, the contribution of crop thorium concentration (7.7%) and the thorium-to-radium ratio (7.1%) indicates that comprehensive assessment still benefits from multi-parameter measurements. The minimal importance of categorical variables further suggests that the framework may be transferable across locations and crop types, though this hypothesis requires empirical verification.

These findings align with the ALARA (As Low As Reasonably Achievable) principle, as the framework enables proactive identification of potential exposure pathways rather than relying solely on retrospective dose assessment. Nevertheless, the transition from retrospective dose audit to prospective dose management should be gradual, with the machine learning predictions validated against independent measurements before being used for regulatory decisions.

CONCLUSION

The study reports a systematic application of machine learning in the prediction of anomalous ^{232}Th transfer factors in phosphate rich semi-arid soils of the Sokoto basin, North-West Nigeria. The framework demonstrated excellent predictive performance with Linear Regression ($R^2 = 0.997$, $\text{RMSE} = 0.018$, $\text{MAE} =$

0.015), and achieved 100% classification accuracy in identifying anomalous samples above the IAEA reference maximum of 0.1. The most dominant predictor was found to be the soil thorium concentration (31.8% importance) followed by the crop thorium concentration (7.7%) and the thorium-to-radium ratio (7.1%) with the categorical variables being zero important confirming that geochemical rather than geographical or crop-specific factors are driving the anomalous transfer.

However, several important limitations must be acknowledged. The study is constrained by a relatively small dataset comprising only of 26 crop samples, which raises concerns regarding model overfitting and generalizability despite the use of leave-one-out cross-validation and bootstrap methods. The reported 100% classification accuracy should therefore be interpreted cautiously, as the absence of normal samples in the dataset means the classifier's discriminatory ability was not fully tested. Furthermore, the absence of independent validation datasets limits confidence in the framework's performance outside the study region.

Despite these limitations, the low-cost, computationally simple models provide a potentially valuable screening tool for resource-limited settings, potentially enabling regulatory authorities to transition from retrospective dose audit to prospective dose management. However, we recommend that these local TF values be formally validated through independent studies before being adopted by Nigerian agricultural departments and regulatory agencies. The prediction models should be incorporated in training and quality control procedures as preliminary screening instruments, with the understanding that site-specific measurements remain essential for confirmatory assessments. Future work should focus on expanding the dataset, conducting external validation across different geographical regions, and developing user-friendly interfaces for practical deployment in agricultural monitoring programs.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the cooperation and technical assistance of the management and technical staff of the Centre for Energy Research and Training (CERT), Ahmadu Bello University, Zaria during the sample analysis.

Declaration Statement

Ethics Approval: Informed consent was obtained from all participants in the agricultural survey. To guarantee complete anonymity, all patient identities were removed.

Conflict of Interest: The authors report no financial or professional conflicts of interest that could have influenced the study's design, conduct or reporting.

Author Contributions: All authors participated in the conceptualization of the study, data collection, analysis, and paper preparation. The definitive version received approval from all co-authors before submission.

REFERENCES

1. Adesida, S. O., & Alimba, C. G. (2025). Systematic literature review of radionuclides, heavy metals, and organochlorine pesticides in Nigerian food crops: Assessment of carcinogenic and non-carcinogenic health risk. *Environmental analysis, health and toxicology*, 40, e2025012.
2. Balonov, M., Barnett, C. L., Belli, M., Beresford, N. A., Berkovsky, V., Bossew, P., & Velasco, R. H. (2010). Handbook of parameter values for the prediction of radionuclide transfer in terrestrial and freshwater environments. International Atomic Energy Agency.
3. Gong, S. Y., Yang, X., Zhang, K. M., Ma, Y., Tong, Z. F., & Chen, T. (2025). Estimation of effective diffusion coefficient of radionuclides in bentonite by machine learning method. *Annals of Nuclear Energy*, 214, 111223.
4. Hamidu, H., Garba, M. L., Abubakar, Y. I., Muhammad, U., & Mohammed, D. (2016). Groundwater resource appraisals of Bodinga and environs, Sokoto Basin north western Nigeria. *Nigerian Journal of Basic and Applied Sciences*, 24(2), 92-101.

5. Huettenbrink, C., Hitzl, W., Distler, F., Ell, J., Ammon, J., & Pahernik, S. (2023). Personalized prediction of patient radiation exposure for therapy of urolithiasis: an application and comparison of six machine learning algorithms. *Journal of Personalized Medicine*, 13(4), 643.
6. International Atomic Energy Agency (IAEA). (2010). Handbook of parameter values for the prediction of radionuclide transfer in terrestrial and freshwater environments (Technical Reports Series No. 472). IAEA.
7. Joseph, D. S., Nasiru, R., Garba, N. N., & Isma'il, M. (2022). In situ gamma spectrometry of terrestrial naturally occurring radioactive materials and radiation hazard assessment of Kebbi State, Nigeria. *Journal of Radioanalytical and Nuclear Chemistry*, 331(5), 2019-2027.
8. Kim, D. H., & Lee, J. Y. (2022). Prediction of the adsorption behaviors of radionuclides onto bentonites using a machine learning method. *Minerals*, 12(10), 1207.
9. Mate-Kole, E. M., Howard, S. C., Golden, A. P., & Dewji, S. A. (2024). Machine learning-enhanced stochastic uncertainty and sensitivity analysis of the ICRP human respiratory tract model for an inhaled radionuclide. *Journal of Radiological Protection*, 44(4), 041507.
10. Nkubli, B. F., Jeremiah, M., Chigozie, N., Akpan, A. G., & Geofery, L. (2024). Exploring machine learning potentials to improve medical imaging services of children and adolescents in low-resource settings. *Journal of Radiography and Radiation Sciences*, 38(1), 14-16.
11. Olabimtan, S. O., Chifu, E. N., Hafeez, Y. H., & Nasir, M. (2023). Measurement of transfer factors from soil-to-plant/food crop of Naturally Occurring Radionuclide Materials (NORMs) in Nigeria: a review. *Dutse Journal of Pure and Applied Sciences*, 9(3b), 173.
12. Oladele, B. B., Ugbede, F. O., Arogunjo, A. M., Ajayi, O. S., & Pereira, A. (2023). Gamma spectroscopy study of soil-plant transfer factor characteristics of 40K, 232Th and 226Ra in some crops cultivated in southwestern region of Nigeria. *Heliyon*, 9(9).
13. Omoriwhovo, J.O., Agbajor, G.K., Akpolilie, A.L. and Akpojotor, G. (2022). Python AIM Solver to Obtain the Vibrational Energy Spectra of Diatomic Molecules. *African Journal of Physics*, 15:15-31.
14. Orosun, M. M., Usikalu, M. R., Oyewumi, K. J., Omeje, M., Awolola, G. V., Ajibola, O., & Tibbett, M. (2022). Soil-to-plant transfer of 40K, 238U and 232Th and radiological risk assessment of selected mining sites in Nigeria. *Heliyon*, 8(11).
15. Seong, M., Kim, H. G., Yun, B., Kim, M., Kim, J. W., Jeong, H., & Cho, K. H. (2026). Hybrid Deep Learning–Numerical Modeling Framework for Long-Term Prediction of Groundwater Discharge and Radionuclide Transport. *Journal of Hazardous Materials*, 141346.
16. Sivapriya, J., Kumar, A., Sai, S. S., & Sriram, S. (2019). Breast cancer prediction using machine learning. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(4), 4879-4881.
17. Thien, B. N., Ba, V. N., Vy, N. T. T., & Loan, T. T. H. (2020). Estimation of the soil to plant transfer factor and the annual organ equivalent dose due to ingestion of food crops in Ho Chi Minh city, Vietnam. *Chemosphere*, 259, 127432.
18. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR). (2000). Sources and effects of ionizing radiation (Report to the General Assembly). United Nations.
19. United States Geological Survey (USGS). (1969). Geology and mineral resources of the Sokoto Basin, Nigeria. USGS.
20. Urso, L., Petermann, E., Gnädinger, F., & Hartmann, P. (2023). Use of random forest algorithm for predictive modelling of transfer factor soil-plant for radiocaesium: A feasibility study. *Journal of Environmental Radioactivity*, 270, 107309.
21. Vandenhove, H., Olyslaegers, G., Sanzharova, N., Shubina, O., Reed, E., Shang, Z., & Velasco, H. (2009). Proposal for new best estimates of the soil-to-plant transfer factor of U, Th, Ra, Pb and Po. *Journal of Environmental Radioactivity*, 100(9), 721-732.
22. Whicker, F. W., Eisenbud, M., & Gesell, T. (1997). Environmental radioactivity from natural, industrial, and military sources. *Radiation Research*, 148(4), 402.
23. Yang, X., Zhang, K., & Chen, T. (2025). Machine learning-based prediction of cation distribution coefficients in bentonite. *Nuclear Techniques*, 48(6).
24. Zhang, X., Zhang, Y., Zhang, Y., Cheng, Y., Liu, Q., Deng, H., & Liu, L. (2024). High-risk nuclide screening and parameter sensitivity analysis based on numerical simulation and machine learning. *Journal of Hazardous Materials*, 480, 136002.