Journal of Hazardous Materials Advances 16 (2024) 100480

Contents lists available at ScienceDirect

Journal of Hazardous Materials Advances

journal homepage: www.elsevier.com/locate/hazadv

Exploring soil pollution patterns in Ghana's northeastern mining zone using machine learning models

Daniel Kwayisi [a,b], Raymond Webrah Kazapoe [c], Seidu Alidu [d], Samuel Dzidefo Sagoe [e], Aliyu Ohiani Umaru [f], Ebenezer Ebo Yahans Amuah [g,h,*], Prosper Kpiebaya [i]

[a] Department of Geology, University of Johannesburg, Auckland Park Kingsway Campus, South Africa
[b] Department of Earth Science, University of Ghana, Legon-Accra, Ghana
[c] Department of Geological Engineering, University for Development Studies, Nyankpala, Ghana
[d] Ghana Geological Survey Authority, P.O. Box M80, Accra, Ghana
[e] Department of Environment and Sustainability Sciences, University for Development Studies, Ghana
[f] Department of Geology, University of Maiduguri, Maiduguri, Borno State, Nigeria
[g] Department of Environmental Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[h] Department of Civil Engineering, Takoradi Technical University, P. O. Box 256, Takoradi, Ghana
[i] Department of Soil Science, University for Development Studies, P. O. Box TL 1882, Nyanpkala, Ghana

ARTICLE INFO

Keywords:
Machine learning
Galamsey
Gold mining
Environmental degradation
Pollution indices

ABSTRACT

This study assessed the pollution status and effectiveness of machine learning models in predicting pollution indices in soils from a mining area in Northeastern Ghana. 552 soil samples were analysed with an Energy Dispersive X-ray Fluorescence (ED-XRF) spectrometer for their elemental concentrations. Four pollution indices; Nemerow Integrated Pollution Index (NIPI), degree of contamination (Cdeg), modified degree of contamination (mCdeg) and Pollution Load Index (PLI). Additionally, the Multivariate Adaptive Regression Splines (MARS) machine learning approach were used. The high CV%, skewness, and kurtosis values show a high degree of variability and uneven distribution patterns which denotes dispersed hotspots that can be interpreted as an influence of gold anomalies and illegal mining activities in the area. V (120.86 mg/L), Cr (242.42 mg/L), Co (30.92 mg/L) Ba (337.62 mg/L), and Zn (35.42 mg/L) recorded values higher than the global and regional contaminant thresholds. The NIPI shows that 46.74% and 26.81% of samples are slightly and moderately polluted respectively. The Cdeg analysis supports these findings, with 36.96% and 41.49% of samples classified as having "moderate" to "considerable" contamination, respectively. The PLI indicates progressive soil quality deterioration (43.84%) of samples reflecting substantial environmental disturbance. The pollution indices show the effect of illegal mining on Shaega, Buin and other areas in the eastern boundary of the study. The MARS models developed for the study demonstrated high predictive capabilities with an R^2 value of 0.9665 for model 1 (NIPI), and RMSE and MAE values of 0.8227 and 0.4287 respectively. For model 2 (Cdeg), R^2 value of 0.9863, RMSE and MAE of 1.0416 and 0.6181, respectively. Model 3 (mCdeg) produced an R^2 value of 0.9844, RMSE and MAE of 0.1225 and 0.0670. These findings suggest MARS models can be an integral tool for soil quality analysis in cooperation with pollution indices. The study suggests that remedial and legislative measures be implemented to address the issue of illegal mining in the area.

1. Introduction

Ghana's vibrant gold mining industry is known to be associated with significant environmental and ecological risk factors (Kazapoe et al., 2023). These risks emanate from the industry's small-scale and artisanal mining (ASM) sector (Achina-Obeng and Aram, 2022). Some of these activities are carried out with rudimentary equipment, are labour-intensive, and often lack the regulation major mines are subjected to (Bansah et al., 2016). This may result in the exposure of the environment to mine waste or inputs which cause significant damage to water bodies and the environment (Agboola et al., 2020; Amuah et al., 2022a). Until recently, these activities have been particularly severe

* Corresponding author.
E-mail address: amuahyahans@gmail.com (E.E.Y. Amuah).

https://doi.org/10.1016/j.hazadv.2024.100480
Received 17 August 2024; Received in revised form 17 September 2024; Accepted 21 September 2024
Available online 22 September 2024

only in the Southern part of Ghana (Nunoo et al., 2022).

Gold mining in Northern Ghana is in its relative infancy, and generally not as intensive as is the case in the Southern mining district. However, the advent of mining communities in some Northern areas and reported issues of adverse effects on the environment have brought it under the spotlight. Studies in areas such as Nangodi, Sheaga and Lawra areas have highlighted the risk the activities of such miners have on the water bodies and soils of the local communities who rely on the water and the crops which grow on possibly compromised soils for their daily sustenance (Arhin et al., 2016; Moomen and Dewan, 2017; Isung, 2021; Akoto et al., 2023). Equally as threatening are geogenic contaminants such as fluoride and arsenic whose anomalous concentrations in soil and groundwater have been reported in the study area (Arhin et al., 2016; Arhin et al., 2017; Akoto et al., 2023).

This situation necessitates advanced environmental monitoring techniques, such as spatiotemporal and multivariate analysis of elements, which have proven effective in capturing complex environmental changes. Additionally, pollution and ecological indices have been successfully used to assess environmental quality. Building on these methods, this study incorporates machine learning models to enhance monitoring accuracy and efficiency. Machine learning offers robust analytical capabilities, identifying intricate patterns and relationships within large datasets (Kamm et al., 2023; Strielkowski et al., 2023). This integration allows for precise predictions and actionable insights, contributing to more effective environmental management and decision-making.

Combining machine learning with traditional techniques represents a significant advancement, ensuring thorough assessments and a proactive approach to combating environmental challenges (Bibri et al., 2024). To help in decoding the outcomes of the pollution indices, ML models were used. Machine Learning (ML) is a contribution of various algorithms that undertake predictions and tasks by learning from a given "problem-specific" data and training on it. Machine learning analyzes large amounts of data in a short time possibly providing precise prediction. Javed et al. (2023) critically provides an understanding of how machine learning can be utilized in improving speed and precision in the medical sector. ML has made significant contributions to the advancement of environmental science offering a wide range of innovative solutions for the management and mitigation of environmental issues plaguing the world. Through the analysis of vast amounts of data, ML models are employed in the prediction of temperature changes, precipitation patterns, and disastrous weather with incredible accuracy (Fister et al., 2023; Iglesias et al., 2024).

Multivariate Adaptive Regression Spline (MARS), a type of ML algorithm, is a non-parametric variant of regression that extends the ability of linear models to identify linear relationships to non-linear relationships. The technique was developed by Friedman (1991) and can automatically identify and model complex relationships without requiring the usual assumptions that come with regression (Zhang and Goh, 2016; Adiguzel and Cengiz, 2023). The technique operates through the division of the dataset into separate sections with each section fitted with its own piecewise linear regression. This is achieved with the help of basis functions, the points where these pieces connect are known as knots. Naser et al. (2022) was successful in employing the use of the MARS algorithm in the prediction of compressive strength of eco-friendly concrete and even found it superior to well-known ML models like Random Forest (RF) and Support Vector Machine (SVM). Gackowski et al. (2022) was also successful in predicting antitumor activity in anthrapyrazole derivatives. These studies highlight MARS as a suitable tool for us due to its high level of accuracy and flexibility across several fields.

The objective of this study was to assess the environmental condition of the soil in the Nangodi region concerning heavy metal pollution to determine (i) the concentration of heavy metal pollutants; (ii) sources, and interactions between these elements; (iii) spatial distribution; (iv) ecological risk related to elements, and (v) whether ML models can serve as reliable predictors of soil labelling such models as effective.

2. Methodology

2.1. Description of the study area

The Nangodi area (Fig. 1) is located within the Talensi-Nabdam district of the Upper East Region (Amosah and Lukman, 2023). According to the Ghana Statistical Service (2010), the wider area has a population of 94,650. It adjoins the Bolgatanga Municipality to the North, the West and East Mamprusi Districts to the South, the Kassena-Nanakana District to the West and the Bawku West District to the East (Ghana Statistical Services, 2010). The district is located between the latitudes of 10. 15° and 10. It is located at latitude 60 ° north of the equator and longitude 0. 31° and 10. 50. The area is predominantly an agricultural-based economy. A significant proportion of the households (i.e. 85.9%) in the area are engaged in the cultivation of crops, rearing of animals, and planting of trees among others. About 49.3% of those actively engaged in agricultural practices are males while 50.7% are females (Ministry of Food and Agriculture, n.d.). However, the relatively recent discovery of gold in the area has attracted a lot of unemployed youth and small-scale mining ventures to the area, whose activities are affecting the area (Tom-Dery et al., 2012; Kazapoe et al., 2021).

The area is characterized by a gentle to medium slope of 1%-5% while the upland slopes are about 10% (Ministry of Food and Agriculture, n.d.). The district is drained mainly by the Red and White Volta and their tributaries. The climate is warm and is characterized by two seasons namely the Wet and the Dry. The rainfall is irregular and received from May to October each year with an average annual rainfall of 950 mm. The dry season lasts for 7 months every year beginning in October and ending in April (Sekyi-Annan, 2019). The temperature can rise to 45°C in March and April and can drop to 12°C in December (Ghana Statistical Services, 2010). The district has three gazetted forest reserves that cover a total area of 455.21 km; Nyokoko (established in 1954), Tankwiddi East and The Red Volta were respectively established in the year 1956 (Ministry of Food and Agriculture, n.d.). The forest reserves unfortunately host enclaves of illegal mining pits which poses an environmental challenge in the area (Okyere et al., 2021).

The area falls within the Nangodi greenstone belt, which is one of six (6) Paleoproterozoic Birimian gold-bearing belts in Ghana (Dzikunoo et al., 2021). These Paleoproterozoic formations are found in the Baoulé Mossi domain of the West African Craton (Abouchami et al., 1990). The stratigraphic package of the area is composed of volcanics including andesites, spilites, greenstones, porphyries, meta-basalts, vesicular basalts, rhyolites and tuffs along with metasediments (mostly phyllites, greywackes and chlorite schist) (Kazapoe, 2014). The greenstones are commonly massive and are often chloritised with the intrusives mostly made up of granodiorites (Murray, 1960).

2.2. Soil sampling

For this study, a grid-based sampling method was adopted in order to cover the entire area and obtain a good representation of the soil samples. This approach was selected to address natural factors including topography and vegetation and anthropogenic factors including mining and agriculture concerning soil parameters. The study area was divided into equal-sized cells and soil samples were collected at regular intervals thus 552 samples were obtained. The grid size and the sampling points were computed according to the size of the study area and possible variation in the soil type, which includes agricultural lands, mines, and transitional area.

The samples were taken from the B-horizon by digging up to a depth of 20 cm using soil augers and shovels. Materials such as stones, roots, and other plant debris were sieved out in the field separately from the soil particles and were neatly wrapped in polythene bags and labelled

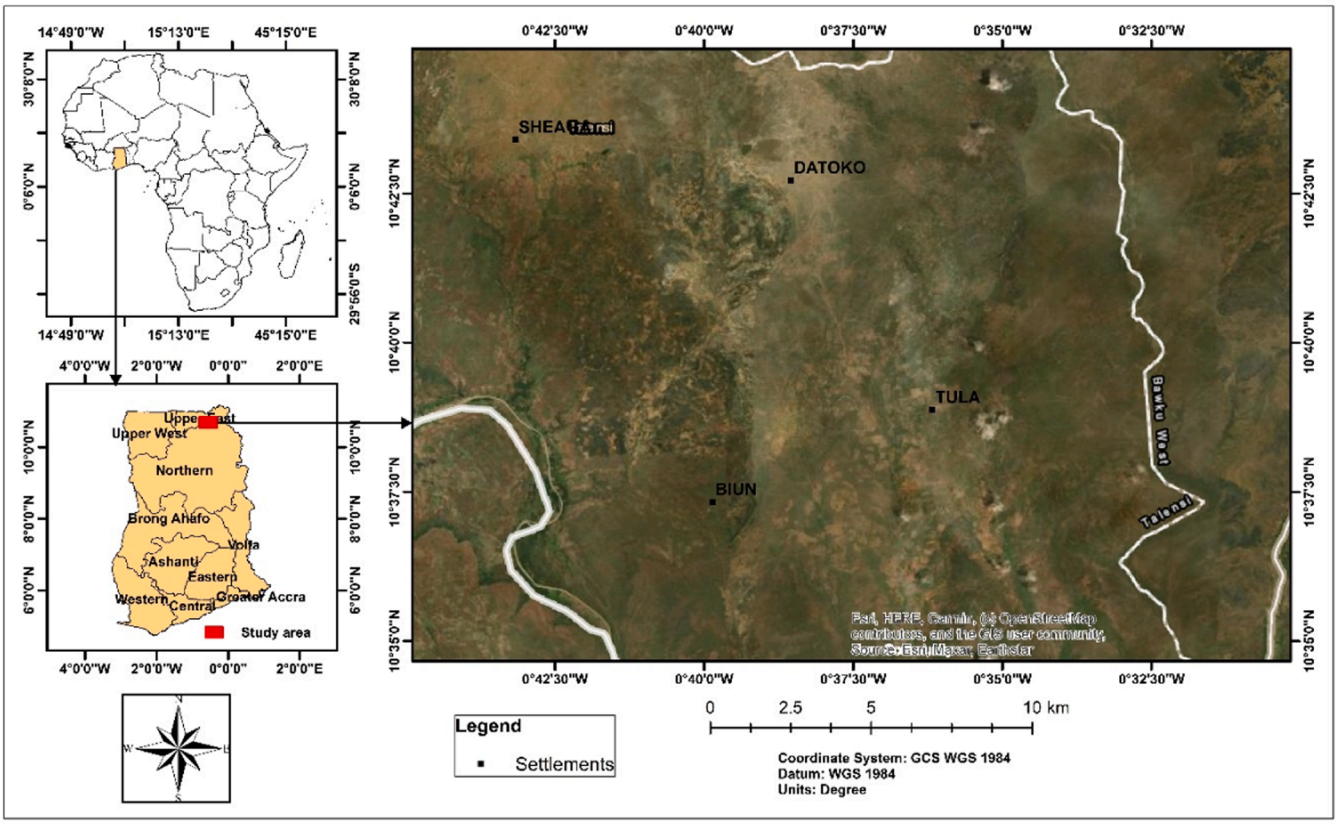

Fig. 1. Map of the study area showing the sampling locations.

for laboratory analysis.

2.3. Analytical procedures

Composite samples were prepared using a mixture of three homogenised samples taken at each sampling point. The samples were sent to the Ghana Geological Survey Authority (GGSA) for further treatment and analysis. After the samples were brought to the laboratory, the soil samples were left to dry at room temperature and in normal atmospheric conditions for 72 h to remove moisture from the samples. These were then dried in an oven and sieved through a 2 mm mesh to ensure no large particles and other debris were included in the analyses. To make the samples as uniform as possible, 50 g of the sieved soil from each sample was ground with a mechanical agate mortar and pestle, ensuring all the particles were of small size. The powdered samples were then kept in clean, airtight containers until the next step of the experiment. The elemental analysis of the soil samples was done by using an Energy Dispersive X-ray Fluorescence (ED-XRF) spectrometer. The XRF spectrometer was checked and standardized with standard samples to give precise results for the readings taken. For each sample, approximately 10 grams of soil was placed in a die and compacted into a pellet at a pressure of 10,000 kN for 1 min. The samples obtained were then placed in the sample holder of the XRF spectrometer for analysis. All the samples were prepared and analysed in triplicate to ensure the results were precise. The spectrometer was set to operate under the following conditions: The instrument parameters used were an excitation voltage of 50 kV, a current of 1 mA, and a counting time of 300 s per sample. The spectra of the emitted X-rays were recorded and processed to quantify the concentrations of the materials present, including some of the heavy metals. Some steps were taken in order to guarantee the accuracy and credibility of the analysis. Blank measurements were done at appropriate time intervals in order to check for contamination and instrumental interferences. Control samples were used in each batch to check the accuracy of the measurements due to sample repetition. To check the reliability and precision of the instrument, CRMs with established elemental concentrations were used alongside the soils. The XRF spectrometer was routinely run in the calibration mode with the CRMs to check its stability and efficiency. All the procedures, settings, and measurements were properly documented to create an organized record of the quality control measures.

2.4. Quality Assurance and Quality Control Analysis (QA/QC)

Certified Reference Materials (CRMs) were applied as the control samples in the course of the analysis to check the correctness of the measurements. One (1) control sample was inserted for every twenty (20) field samples based on international recommendations. The control samples were used to check the validity of the elemental analysis done on the samples of the soil. The measurement precision and the method reproducibility were checked by including the duplicate samples in the analysis. Consistent with normal protocol, about 5 per cent of the field samples were duplicated. Hence, 28 duplicate samples were incorporated after every 20 samples when the samples were being analysed. As both the original and duplicate samples were analyzed, the study was able to assess the variability of the analytical procedures. The elements show small average variations of 1.3–9.2% between duplicate sample pairs, which was considered accepted following Arhin et al. (2019).

2.5. Machine Learning Model

2.5.1. Pollution indices

The NIPI, PLI, Cdeg and mCdeg were used in this study. Table 1 provides details of these.

2.5.2. Multivariate Adaptive Regression Splines (MARS)

The process through which the MARS model is built involves two

Table 1
A table showing the details of Pollution indices used in this study.

Index	Equation
NIPI	$C_f = \frac{C_i}{C_s}$ $NIPI = \sqrt{\frac{(Cf)^2_{max} + Cf^2_{mean}}{2}}$ Where Cf represents the concentration factor, C_i represents the measure concentration, and C_s represents the evaluation standard (Wang et al., 2022).
PLI	$PLI = \sqrt[n]{Cf_1 \times Cf_2 ..... \times Cf_n}$ Where Cf represents the concentration factor and n represents the number of heavy metals considered (Tomlinson et al., 1980).
Cdeg	$Cdeg = \sum_{i=1}^{n} C_f^i$ Cf represents the concentration factor and n represents the number of heavy metals considered (Hakanson, 1980).
mCdeg	$mCdeg = \frac{1}{n} \sum_{i=1}^{n} C_f^i$ Cf represents the concentration factor and n represents the number of heavy metals considered (Abrahim and Parker, 2008)

phases, forward and backwards. The forward phase involves the addition of basis functions considering the interactions that exist between variables and knots (Ağyar et al., 2022). This usually results in the creation of an overfitted model (Zhang and Goh, 2016; Ağyar et al., 2022). The backward phase involves the pruning of basis functions that contribute very little to the overall performance of the model. This reduces the possibility of overfitting thus enhancing model generalization (Zhang and Goh, 2016).

The MARS model is expressed by the equation:

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{K_m} h_{km}\left(X_{v(k,m)}\right) \quad (1)$$

Where f(x) represents the predicted value of the dependent variable y based on the independent ones x, β_0 the intercept, a constant representing the baseline value of the dependent variable when all predictors are zero, β_m the coefficient associated with the m^{th} basis function to the overall model, $\prod_{k=1}^{K_m}$ denotes the product of K_m hinge functions for the m^{th} basis function, $h_{km}\left(X_{v(k,m)}\right)$ represents the hinge function, a piecewise linear function of the variable $X_{v(k,m)}$, and $X_{v(k,m)}$ the predictor variable indexed by $v(k,\ m)$ (Şengül et al., 2020).

After the first phase, the basis function that did not significantly contribute to model performance is removed using a method called pruning. This is achieved using the generalized cross-validation error (GCV). This is expressed by the Eq. (2):

$$GCV\ (\lambda) = \frac{\sum_{i=1}^{n}\left(y_i - y_{ip}\right)^2}{\left(1 - \frac{M(\lambda)}{n}\right)^2} \quad (2)$$

Where n represents the number of training cases, y_i is the observed values of the dependent variable, y_{ip} is the predicted values of the dependent variable, *and* $M(\lambda)$ is the penalty function for the model complexity (Şengül et al., 2020).

2.5.3. Evaluation metrics

To determine the accuracy of the MARS model, several well-known statistical evaluation metrics were employed. These metrics provide important insights as to how well the model fits, the accuracy of the predictions made by the models, and the errors generated by the model.

R squared/ Adjusted R squared ($R^2/Adj\ R^2$)

The coefficient of determination, also known as R^2 measures the proportion of the variances that exist within the dependent variable that is explained by the independent variables. This metric is expressed by the Eq. (3):

$$R^2 = 1 - \frac{SSres}{SStot} \quad (3)$$

Where SSres represents the residual sum of squares and SStot, the total sum of squares. A coefficient of determination of 1 indicates the model predicted perfectly and thus explains the variance within the dependent variable, 0 indicates an inability to predict accurately thus unable to explain the variance within the dependent variable (Kim, 2018). According to Hair et al. (2021), an R^2 of 0.75 is described as substantial, 0.50 as moderate, and 0.25 as weak.

Adj R^2 is a modified version of the R^2. It adjusts the R^2 value based on the number of predictors in the model. This helps in accounting for overfitting. It is expressed by equation (Eq. (4)):

$$Adj\ R^2 = 1 - \frac{n-1}{n-m-1}\left(1 - R^2\right) \quad (4)$$

Where n represents the number of observations, and m, the number of predictors (Trunfio et al., 2022).

2.5.4. Root Mean Square Error (RMSE)

RMSE determines the average errors generated between the predicted and observed values. It is denoted by the Eqn. 5:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y}_i - y_i)^2}{n}} \quad (5)$$

()Where y_i represents the observed values and $\widehat{y}_i$, the predicted values, n the number of observations. A lower RMSE value represents a more accurate model and vice versa (Chai and Draxler, 2014).

2.5.5. Mean Absolute Error (MAE)

MAE determines the average magnitude of absolute errors between the observed and predicted values. It is represented by the Eq. (6):

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

Where y_i represents the observed values and $\widehat{y}_i$, the predicted values, n the number of observations. Similar to RMSE, a lower MAE value represents a more accurate model and vice versa (Chai and Draxler, 2014).

2.5.6. Implementation

The MARS model was implemented using R studio employing the use of the earth package. Model selection was conducted to ensure all variables were suitable for the model training. PLI variable failed to meet the requirement for homoscedasticity after the Breusch-Pagan test was conducted. Cu variable was also removed after a test to determine the presence of multicollinearity was conducted. The train test split was 80:20. 80% of the data was used to train the models and 20% for testing. Parameter optimization was conducted using grid search which runs several iterations of the MARS models changing parameters at each iteration to determine which combination yields the best set of results. Three different MARS models were developed separately for three different dependent variables. Model 1 had Nemerow Integrated Pollution Index (NIPI) as the dependent variable, model 2, degree of contamination (Cdeg), and model 3, modified degree of contamination (mCdeg). Vanadium (V), Chromium (Cr), Cobalt (Co), Zinc (Zn), Strontium (Sr), Molybdenum (Mo), Barium (Ba), and Lead (Pb) were the independent variables for all models. The main aim of this study is to assess the environmental quality of the Nangodi region through the identification of complex interactions that may exist within the heavy metal concentrations in the area. MARS helps us achieve this goal through the identification of a set of simple linear functions that collectively results in the best predictive performance. The identification of these complex relationships may have been particularly challenging given the distribution of the dataset. By employing three separate MARS

models (NIPI, Cdeg, and mCdeg), we are able to identify distinct relationships that may exist with regard to each dependent variable providing us with a clearer and more interpretable set of results.

2.6. Statistical data analysis

The R software was used to perform geostatistical analysis with the help of the "compositions" and "robcompositions" packages. Descriptive measures, including minimum, maximum, mean, median, standard deviation (SD), coefficient of variation (CV%), kurtosis, and skewness, were determined to understand the distribution of the elemental concentrations. For this purpose, the analysis concerned nine potential toxic elements (PTEs): V, Cr, Co, Cu, Zn, Sr, Mo, Ba, and Pb, and their interconnections with other geochemical factors, which were estimated based on Pearson correlation coefficients (r) and their corresponding p-values for both positive and negative relationships. PCA was used to determine the sources of the elemental concentration in the soil samples. PCA is not only useful in simplifying the analysis of large datasets through the identification of the important components but also demonstrates the relationships between the variables, reduces the number of factors to consider and pinpoints the factors which have the most impact on the quality of the soil (Kazapoe et al., 2021; Abu et al., 2021). Before performing PCA, the variables were ranked using normal score transformation to select the components with the highest signal.

3. Results

3.1. Spatial distribution of elements cross the study area

Table 2 presents the general distribution of elemental values across the study area. Mo (144.66%), Cu (102.45%), Cr (94.74%), Sr (92.19%), Co (87.13%), and Ba (78.50%) displayed very high CV%. Comparatively, Pb (43.71%) showed moderate variability. Positive skewness exhibited by all the elements suggests that the concentrations of elements are significantly higher than their average values at most parts across the area. Similarly, elements such as Co, Zn, Mo, and Ba record high kurtosis values (> 3), signifying that their concentrations have more extreme values in comparison to a normal distribution.

In terms of the general distribution of the PTEs, V ranges from 9.45-528 mg/L with an average value of 120.86 mg/L (Fig. 2A). These values exceed the average distribution of V set as the contaminant level when compared to values from the European Union (2002), the United States Environmental Protection Agency (USEPA) (2002) and other parts of Europe. Cr averaged a concentration of 242.42 mg/L across the area which exceeded the contaminant level reported by Kazapoe and Arhin (2021) for the Birimian of Ghana (80.74 mg/L). This value is similarly high as compared to the standards set by the EU (75 mg/L) and USEPA (11 mg/L). Co distribution in the soils within the study area ranged from 3.3 to 311.8 mg/L with an average value of 30.92 mg/L. Also, 40% of the

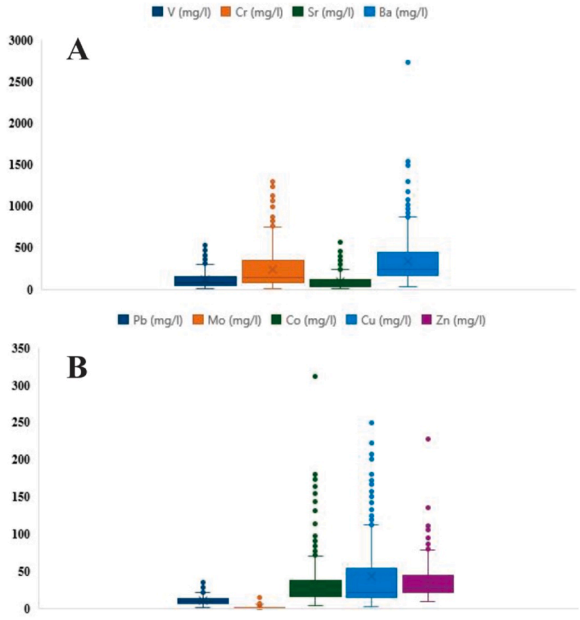

Fig. 2. Box and whisker plots of the elemental distribution across the study area for (A) V, Cr, Sr and Ba (B) Pb, Mo, Co, Cu and Zn.

samples fall above the contaminant threshold of 24 mg/L suggested by Crommentuijn et al. (2000). Similarly, Cu recorded figures (3.25–249.9 mg/L) which far exceeded the average global concentration of 30 mg/L. 12% of the samples fell above the average of 100 mg/L set by McLaughlin et al. (2000). The results of Zn varied between 9.8 and 228.2 mg/L (35.42 mg/L). The results show that 109 (19.74%) of the samples were above the set global contaminant level Zn (30 mg/L). Sr and Mo recorded average values of 98.17 mg/L and 0.95 mg/L respectively which were below the average crustal values (Sr = 375 mg/L and Mo = 1.5 mg/L) reported by Taylor (1964). Ba showed values between 40.5 and 2735 mg/L. This indicates that all of the samples were above the recommendation (9 mg/L) by Crommentuijn et al. (2000). Pb concentration (Fig. 2B) ranged from 1.7 to 35.8 mg/L (Avg. 10.07 mg/L). The results for Pb show that all the samples fell below the guideline value for contaminants globally.

3.2. Relationships among the pollutants and trace elements

The eigenvalues and rotated sum of squares loadings from the PCA are shown in Table 3. The first three components (PCs) explain 81.30 % of the total variance. PC1 loadings contributed 42.12 % of the total variance, PC2 contributed 23.39% while PC3 loadings contributed 15.79 % (Table 3). The factor analysis using R–Mode factor analysis

Table 2
Statistical summary of the results from the Nangodi area.

Elements	V	Cr	Co	Cu	Zn	Sr	Mo	Ba	Pb
Min	9.45	11.6	3.3	3.25	9.8	6.9	0.4	40.5	1.7
Max	528	1294	311.8	249.9	228.2	567	14.2	2735	35.8
Mean	120.86	242.42	30.92	43.8	35.42	98.17	0.95	337.62	10.07
Standard Deviation (SD)	106.11	229.65	26.94	43.77	21.58	90.5	1.38	265.02	4.4
Skewness	1.6	1.65	3.8	1.86	2.51	2.1	5.94	2.64	1.26
Kurtosis	1.85	2.79	26.04	3.2	12.76	4.83	42.46	13.81	3.49
Coefficient of Variation (CV%)	87.79	94.74	87.13	102.45	60.93	92.19	144.66	78.50	43.71
*World**		100		30	50				42.5
*EU**		75			1				
*USEPA**		11		270	1100				200
Kazapoe and Arhin (2021)		80.74	6.44						6.99
McLaughlin et al. (2000)	200	100		100					150
Crommentuijn et al. (2000)	1.1	3.8	24	3.5				9	55

World (Vinogradov, 1959), EU (EU, 2002) and USEPA (USEPA, 2002).

Table 3
The variance contribution of each principal component in the soil sample analysis, highlighting their significance in explaining the dataset's variation.

Total Variance Explained									
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.46	43.27	43.27	3.46	43.27	43.27	3.37	42.12	42.12
2	1.82	22.81	66.08	1.83	22.81	66.08	1.87	23.39	65.51
3	1.22	15.22	81.30	1.22	15.22	81.30	1.26	15.79	81.30
4	0.62	7.75	89.05						
5	0.47	5.82	94.87						
6	0.20	2.51	97.38						
7	0.14	1.70	99.07						
8	0.07	0.93	100.00						

Extraction Method: Principal Component Analysis.

produced the rotated components (Table 4) depicting various component elements of the PC. PC1 holds V, Co, Cu and Zn, PC2 holds Sr and Ba while PC2 has Mo and Pb (Fig. 3; Table 4).

3.3. Heavy metal contamination in the Nangodi area

A summary of the results of the indices used to assess the level of contamination or pollution in the soil samples of the study area is shown in Table 5. The Nemerow Pollution Index reflects a mixed picture where the majority of the samples (46.74%) fall under the Slight Polluted class while 26.81% and 5.98% are classed in the Moderate and Heavy Polluted categories respectively. The central part of the area around Datoko is classed as clean under the NIPI while Shaega in the west and other areas around the eastern margins of the study area are categorised as moderate to heavily polluted respectively (Fig. 4A). These areas are known centres of galamsey activities, which may explain the values derived. The calculation for the Cdeg shows that 36.96% and 41.49%, representing the majority of the samples, are classed as "moderate" to "considerable" respectively, while 13.41% of the samples are classed as "low" in terms of contamination. The results for the Cdeg further underscore the effect of illegal mining on Shaega and the eastern part of the study area which are classed as having considerable pollution (Fig. 4B). The Pollution Load Index analysis shows that nearly half of the samples (43.84%) indicate a progressive deterioration of the quality of the area and the ecosystem. PLI shows that Shaega and Buin rank as showing a progressive deterioration of the quality of the area and ecosystem (Fig. 4C). Correspondingly, Table 6 shows that most of the samples (39.67% and 4%) fall under the "moderate" and "high" classes respectively depicting a moderate level of contamination among the soil samples in the area in terms of the mCdeg Index. Mirroring the other indices, the mCdeg shows that Shaega and Buin are moderately polluted while the eastern part around Tula records some spots which are categorised as high (Fig. 4D). This is mainly due to the activities of the illegal miners in the area A summary of the results of the indices used to assess the level of contamination or pollution in the soil samples of the

Table 4
Rotated Component Matrix from the Principal Component Analysis (PCA) of Soil Samples.

	Component		
	1	2	3
V (mg/l)	**0.94**	0.01	0.08
Co (mg/l)	**0.83**	0.23	0.20
Cu (mg/l)	**0.97**	0.09	-0.01
Zn (mg/l)	**0.83**	0.31	-0.13
Sr (mg/l)	-0.19	**0.84**	-0.23
Mo (mg/l)	0.04	0.04	**0.89**
Ba (mg/l)	-0.23	**0.88**	-0.10
Pb (mg/l)	-0.43	0.44	**0.55**

Extraction Method: Principal Component Analysis.
a. 3 components extracted.

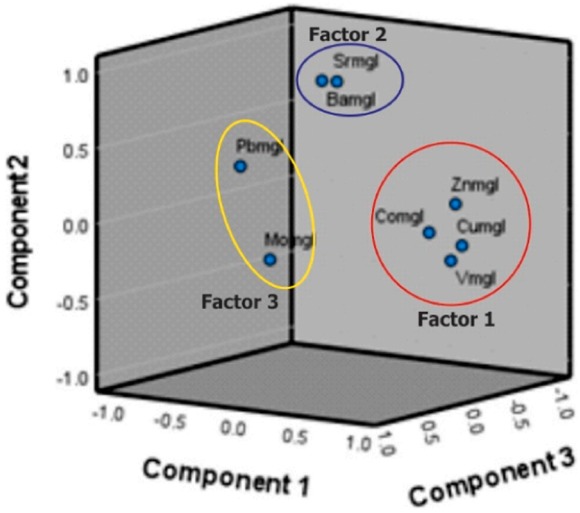

Fig. 3. Plots of R-mode factor analysis of the PTEs.

study area is shown in Table 5. The Nemerow Pollution Index reflects a mixed picture where the majority of the samples (46.74%) fall under the Slight Polluted class while 26.81% and 5.98% are classed in the Moderate and Heavy Polluted categories respectively. The central part of the area around Datoko is classed as clean under the NIPI while Shaega in the west and other areas around the eastern margins of the study area are categorised as moderate to heavily polluted respectively (Fig. 4A). The calculation for the Cdeg shows that 36.96% and 41.49%, representing the majority of the samples, are classed as "moderate" to "considerable" respectively, while 13.41% of the samples are classed as "low" in terms of contamination. The Pollution Load Index analysis shows that nearly half of the samples (43.84%) indicate a progressive deterioration of the quality of the area and the ecosystem. PLI shows that Shaega and Buin rank as showing a progressive deterioration of the quality of the area and ecosystem (Fig. 4C). Correspondingly, Table 6 shows that most of the samples (39.67% and 4%) fall under the "moderate" and "high" classes respectively depicting a moderate level of contamination among the soil samples in the area in terms of the mCdeg Index.

Classification of the dataset was conducted using existing literature as guidelines (Hakanson, 1980; Tomlinson et al., 1980; Abrahim and Parker, 2008; Wang et el., 2022). Computation of the classification was conducted using Microsoft Excel.

3.4. Machine Learning Results

Table 6 shows the performance of all three models after training and testing, GCV, the degree of interactions, and the number of terms are also listed. GCV is used for model validation, it assesses how well the

Table 5
The table showing summary of the results of the indices.

Index	Min	Max	Mean	SD	Skewness	Kurtosis	Class	%
NIPI	0.55	5.69	1.70	0.75	0.86	0.80	NIPI≤0.7 – Clean	8 (1.45%)
							0.7<NIPI≤ 1- Warning limit	105 (19.02%)
							1<NIPI≤2 – Slight pollution	258 (46.74%)
							2<NIPI≤3 – Moderate	148 (26.81%)
							NIPI>3 - Heavy	33 (5.98%)
Cdeg	3.78	68.11	17.63	9.34	0.90	0.91	Cdeg < 8 - low degree	74 (13.41%)
							8<Cdeg<16 - moderate	204 (36.96%)
							16<Cdeg< 32 – considerable	229 (41.49%)
								45 (8.15%)
PLI	0.27	2.85	0.97	0.34	1.05	2.11	<1	310 (56.16%)
							=1	0
							>1	242 (43.84%)
mCdeg	0.42	7.57	1.96	1.04	0.90	0.91	mCdeg<1.5(nil to very low) 1.5≤mCdeg<2(low)	234 (42.39%)
							2≤mCdeg<4(moderate)	75 (13.59%)
							4≤ mCdeg<8(high)	219 (39.67%)
							8≤mCdeg<16(very high)	24 (4.35%)
							16≤ mCdeg<32 (extremely	0
							high)	0
							mCdeg≥ 32 (ultra-high)	0

results from the model will generalize on unseen data (Bottegal and Pillonetto, 2018). It achieves this by adjusting the number of terms in the model penalizing complexity to prevent overfitting. The degree of interactions specifies the maximum number of interactions allowed between the predictor variables and the basis functions. The low GCV values across all three models indicate a high level of generalizability by the models (Oduro et al., 2015). This essentially means that the models perform well on unseen data.

The study sought to examine the relationship between heavy metals (Co, Cr, V, Mo, Sr, Ba, Pb, Zn) on water quality indices (NIPI, Cdeg, mCdeg) using MARS algorithms. To this end, three MARS models were developed one for each dependent variable. The results from Table 6 suggest that model 1 (NIPI) can explain a significant portion of the variance specifically 96.65%. The RMSE and MAE associated with model 1 are relatively low at 0.8227 and 0.4287 respectively, indicating high prediction accuracy. Models 2 (Cdeg) and 3 (mCdeg) produced similar outputs as well in terms of explanation of the variance. Model 2 explains 98.63% of the variance while model 3 explains 98.44% of the variance. The RMSE and MAE for model 2 were slightly higher than that of model 1, 1.0416 and 0.6181 indicating a slightly less accurate model. Model 2 produced less RMSE and MAE as compared to the other models, 0.1225 and 0.0670 respectively making it the most accurate model. Although model 3 produced the least errors, R^2 value was only slightly less than model 3′s. Model 1 had the lowest R^2 value, it also had the least number of interaction terms with only Cr, Co, and V being used by the model, all other independent variables were pruned this was not the case for models 2 and 3. The adjusted R^2 values for all 3 models indicate slight R^2 change, this indicates the possibility of overfitting for all 3 models is highly unlikely. The high R^2 values from all 3 models mirror findings from existing research like Hlokoe et al. (2022) where MARS models were used to examine the effects of biometric traits on the body weight of Nguni cows and Zhang and Goh (2016) where MARS models were used in the prediction of pile drivability thus supporting MARS algorithms as a highly accurate predictive model across varying variables.

Comparing our MARS model with other ML models used in the prediction of pollution indices, our MARS models outperformed the ML models used by Wang et al. (2022) which used ML models like Partial Least Squares Regression, Support Vector Machines, and Gaussian Process Regression in the prediction of pollution indices like NIPI and Potential Ecological Risk Index (RI). This is consistent with Naser et al. (2022) which found MARS models superior to other ML models like Random Forest (RF) and Support Vector Machine (SVM) in the prediction of compressive strength of eco-friendly concrete.

The results from Table 7 indicate that the model's intercept which is a representation of the baseline level of the dependent variable NIPI is 9.7502264. The coefficient associated with BF1 suggests a negative contribution to NIPI when Cr values are less than 492.5. Therefore, below the threshold of 492.5, as Cr decreases, NIPI is expected to decrease. Above the same threshold, a positive coefficient is observed specifically 0.0209554 for BF2 implying that above this threshold, an increase in NIPI is expected as Cr increases. BF3 describes the combined effects of V and Cr values above the thresholds of 263.7 and 492.5 respectively. A positive coefficient is observed although quite low suggesting a slight increase in NIPI if both threshold conditions are met. BF4 and BF5 also indicate slight increases in NIPI upon meeting their threshold conditions due to the low positive coefficient associated with them.

The results from Table 8 indicate a value of 41.984806 for the dependent variable Cdeg when all other variables are 0. BF1 and BF2 suggest a negative impact on Cdeg when V is below 67.2 and a positive above the same threshold. Similar effects are observed for BF3 to BF10 with a negative impact being observed below their varying thresholds and a positive one observed above their thresholds. The interaction term associated with BF11 describes the combined effect of Cr values below the threshold limit of 760 and that of Zn above 28. A positive coefficient is observed suggesting a slight increase in Cdeg when both threshold conditions are met.

From Table 9 the intercept value for mCdeg when all other terms are 0 is 4.4746867. BF1 and BF2 capture the effects of V at the threshold limit of 67.2. Above the threshold a positive impact is observed, below it, a negative one is observed. For BF3 and BF4 a similar effect is observed for Cr but for the threshold of 760. A trend Is observed for the rest of the interaction terms for BF5 to BF12. A negative impact is observed below their various threshold limits and a positive one above.

Fig. 5 contains the subplots of each predictor variable relating to the response variable NIPI. The black line represents the model's prediction, and the blue points are the observed data points. As part of its results, the MARS model also provides data in terms of variable importance. In terms of variable importance Co was the most influential followed by Cr, V, Mo, Zn, and finally, Ba. Sr and Pb were unused by the model meaning that it played no significant role in model output. This is consistent with Table 7, the trend observed for Co and Cr is positive indicating a strong relationship. The trend captured for Ba, Zn, and Mo is flat indicating a weak relationship. It can be observed that the prediction for Cr was the most accurate as well. Data points from the Cr subplot closely follow the regression line, this effect is observed in decreasing intensity for Co, and V indicating less prediction accuracy for the 2 variables. In terms of feature importance, Cr was the most important variable followed by Co

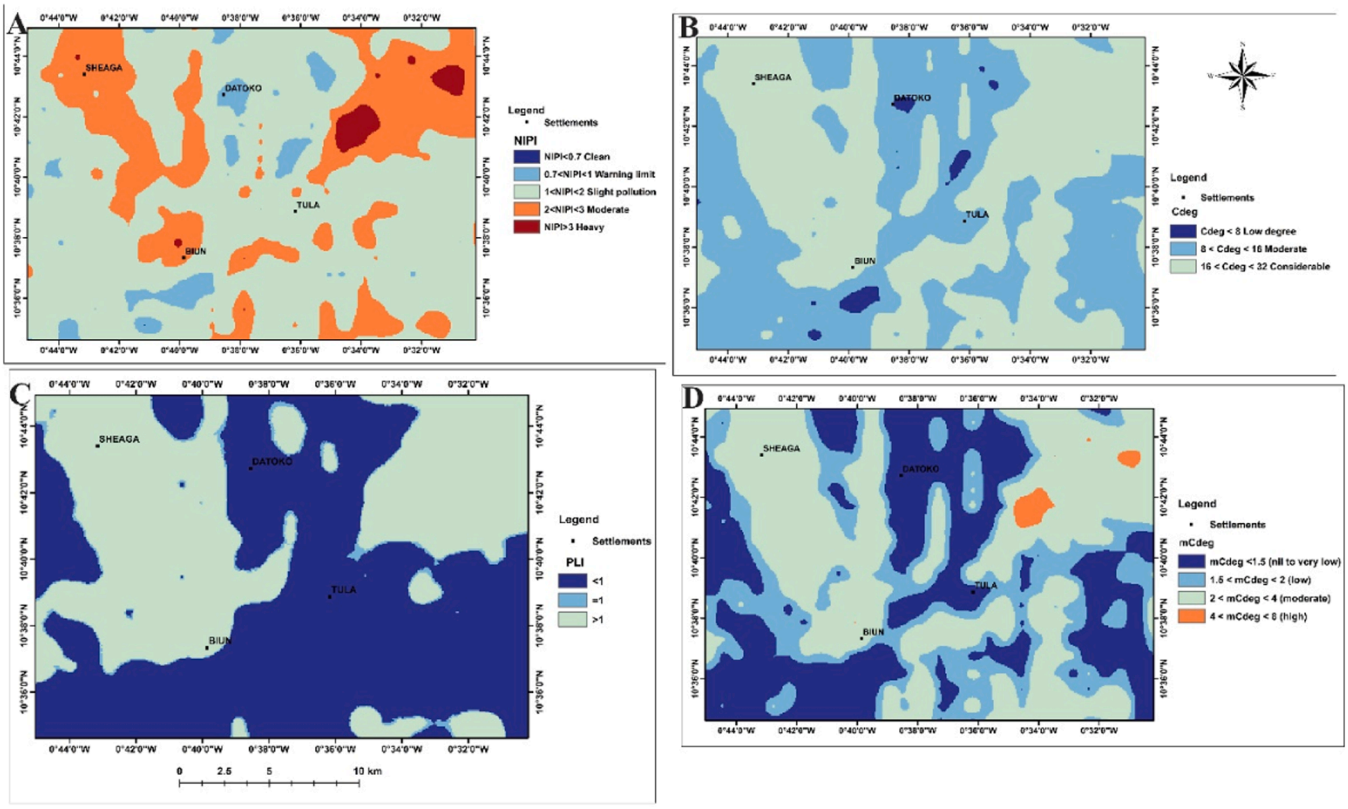

Fig. 4. Spatial distribution of (A) NIPI (B) Cdeg (C) PLI (D) mCdeg.

and V respectively, this is expected due to the positive correlation between Cr, Co and NIPI.

Regarding feature importance for Cdeg, Co played the most important role, followed by Cr, V, Zn, Mo, and Ba, Sr and Pb were unused by the model. This differs from model 1 (NIPI) where Cr was the most important variable. Another noticeable difference was the addition of Zn, Mo, and Ba, these variables were unused by model 1, however, although used in model 2, their input was really low given the level of positive correlation between them and the dependent variable Cdeg.

The most important predictor variable was Co once again followed by Cr, V, Mo, Zn, and Ba. This is expected given how close Cdeg and mCdeg are, Sr and Pb were also unused by the model. Predictor variables on the lower end of feature importance demonstrate weak relationships between the dependent variable mCdeg thus Mo, Zn, and Ba did not play a critical role in terms of model output.

Considering the Figs. 5–7, Co, Cr, and V are the most important drivers in terms of model output. This finding differs from the findings from Wang et al. (2022) which found Cu, Pb, and Zn to be the most important contributors in predicting pollution indices in mining areas within China.

4. Discussions

According to Amuah et al. (2022b) and Ahado et al. (2024), a high coefficient of Variation (CV%), as recorded in the study area, suggests a high degree of variability in elemental distribution and can be used to infer the level of anthropogenic effect on elemental concentration or distribution. The relatively moderate CV% shown by Pb suggests that while there are fluctuations in Pb concentrations across the area, they are not as pronounced as the other elements. The variability in the bulk of the elements can be attributed to the gold anomalies recorded across the area as seen in Table 2, occurring as localized hotspots both as primary sources and from contamination from illegal mining in the area. This is corroborated by the skewness and kurtosis values of the elements. The combination of high skewness and kurtosis values for most of the elements depict uneven distribution patterns and localized contamination hotspots across the area, suggesting a combined effect of mineralized anomalies and influences from illegal mining activities.

Table 6
Model performance for all three MARS models used in the study.

Model	R^2 (Adj)R^2		RMSE		MAE		GCV	Degree of int	No. of terms
	Training	Testing	Training	Testing	Training	Testing			
1(NIPI)	0.9812 (0.9808)	0.9665 (0.9638)	0.6339	0.8227	0.3723	0.4287	0.428	1	6
2 (Cdeg)	0.9921 (0.9920)	0.9863 (0.9852)	0.8419	1.0416	0.5702	0.6181	0.810	1	12
3 (mCdeg)	0.9918 (0.9916)	0.9844 (0.9832)	0.0958	0.1225	0.0607	0.0670	0.010	1	13

From the results provided in Table 7, the regression equation for the dependent variable NIPI is:
NIPI = 9.7502264 - 0.0191000 * max (0, 492.5 - Cr) + 0.0209554 * max (0, Cr - 492.5) + 0.0000481 * max (0, V - 263.7) * max (0, 492.5 - Cr) + 0.0001471 * max (0, 492.5 - Cr) * max (0, Co - 20.5) + 0.0001924 * max (0, Cr - 492.5) * max (0, Co - 16.5)

Table 7
Model 1 (NIPI) indicates the basis function, its related equations and coefficients.

Basis function (BF)	Equation	Coefficients	Std. Error	Pr(>|t|)
	(Intercept)	9.7502264	7.649e-02	0.000e+00
BF1	max (0, 492.5-Cr)	-0.0191000	2.267e-04	1.579e-271
BF2	max (0, Cr-492.5)	0.0209554	3.589e-04	5.878e-208
BF3	max (0, V-263.7) * max (0, 492.5-Cr)	0.0000481	3.019e-06	2.332e-45
BF4	max (0, 492.5-Cr) * max (0, Co-20.5)	0.0001471	5.137e-06	3.711e-102
BF5	max (0, Cr-492.5) * max (0, Co-16.5)	0.0001924	1.719e-05	9.839e-26

From Table 8, the regression equation for Cdeg is:
Cdeg = 41.984806 - 0.017853 * max(0, 67.2 - V) + 0.029273 * max(0, V - 67.2) - 0.028772 * max(0, 760 - Cr) + 0.029502 * max(0, Cr - 760) - 0.151267 * max(0, 92.1 - Co) + 0.083055 * max(0, Co - 92.1) - 0.596140 * max(0, 4.2 - Mo) + 0.661543 * max(0, Mo - 4.2) - 0.002323 * max(0, 590.5 - Ba) + 0.002591 * max (0, Ba - 590.5) + 0.000108 * max(0, 760 - Cr) * max(0, Zn - 28).

The values recorded for V are comparable to findings by Reyes et al. (2020) in Taltal, northern Chile which ranges between 58.00–663 mg/L. Additionally, these values are similar to findings of Kazapoe et al. (2021) in the Birimian of Southwestern Ghana which averaged 92.06 mg/L. The study by Kazapoe et al. (2021) was conducted in an area with a similar geological setting which is also populated with mining firms. These relatively high values can be attributed to the underlying geology, which has been described as composed of granitic intrusions high in alkaline (Kesse, 1985; Kazapoe, 2014). The recorded values for Cr far exceeded

Table 8
Results for model 2 (Cdeg) indicating the basis function, its related equations and coefficients.

Basis Function (BF)	Equation	Coefficients	Std. Error	Pr(>|t|)
	(Intercept)	41.984806	3.927e-01	2.308e-311
BF1	max (0, 67.2-V)	-0.017853	3.502e-03	5.144e-07
BF2	max (0, V-67.2)	0.029273	8.732e-04	6.506e-122
BF3	max (0, 760-Cr)	-0.028772	2.221e-04	0.000e+00
BF4	max (0, Cr-760)	0.029502	9.810e-04	1.056e-107
BF5	max (0, 92.1-Co)	-0.151267	4.058e-03	1.258e-136
BF6	max (0, Co-92.1)	0.083055	3.773e-03	2.029e-72
BF7	max (0, 4.2-Mo)	-0.596140	6.622e-02	7.274e-18
BF8	max (0, Mo-4.2)	0.661543	6.403e-02	1.714e-22
BF9	max (0, 590.5-Ba)	-0.002323	2.886e-04	8.175e-15
BF10	max (0, Ba-590.5)	0.002591	3.288e-04	2.753e-14
BF11	max (0, 760-Cr) * max (0, Zn-28)	0.000108	6.493e-06	1.945e-48

From Table 9, the regression equation for mCdeg is:
mCdeg = 4.4746867 - 0.0011756 * max(0, 67.2 - V) + 0.0033762 * max(0, V - 67.2) - 0.0031351 * max(0, 760 - Cr) + 0.0034487 * max(0, Cr - 760) - 0.0161624 * max(0, 92.1 - Co) + 0.0092752 * max(0, Co - 92.1) - 0.0069885 * max(0, 38.3 - Zn) + 0.0054938 * max(0, Zn - 38.3) - 0.0785176 * max(0, 1.1 - Mo) + 0.0680869 * max(0, Mo - 1.1) - 0.0002002 * max(0, 640.2 - Ba) + 0.0002763 * max(0, Ba - 640.2).

Table 9
Results for model 3 (mCdeg) indicating the basis function, its related equations and coefficients.

Basis Function (BF)	Equation	Coefficients	Std. Error	Pr(>|t|)
	(Intercept)	4.4746867	3.858e-02	0.000e+00
BF1	max (0, 67.2-V)	-0.0011756	4.363e-04	7.332e-03
BF2	max (0, V-67.2)	0.0033762	9.560e-05	6.735e-129
BF3	max (0, 760-Cr)	-0.0031351	2.506e-05	0.000e+00
BF4	max (0, Cr-760)	0.0034487	1.094e-04	1.129e-113
BF5	max (0, 92.1-Co)	-0.0161624	4.925e-04	6.224e-119
BF6	max (0, Co-92.1)	0.0092752	4.342e-04	1.876e-69
BF7	max (0, 38.3-Zn)	-0.0069885	9.805e-04	4.378e-12
BF8	max (0, Zn-38.3)	0.0054938	4.048e-04	3.791e-35
BF9	max (0, 1.1-Mo)	-0.0785176	2.176e-02	3.447e-04
BF10	max (0, Mo-1.1)	0.0680869	4.254e-03	1.609e-45
BF11	max (0, 640.2-Ba)	-0.0002002	3.575e-05	3.801e-08
BF12	max (0, Ba-640.2)	0.0002763	3.996e-05	1.697e-11

what was reported by Kazapoe et al. (2021) in the Southwestern part of Ghana (17.00–331.00 mg/L). According to the World Health Organization (2000), the ingestion of crops which have high Cr concentrations due to uptake from the soil may result in allergic eczematous and acute irritative dermatoses, chrome ulcers, and other health complications.

The range of values identified for Pb is consistent with findings by This was similar to the findings of Darko et al. (2019) in Gbani, Ghana and Kazapoe et al. (2021) in Southwestern Ghana where Pb levels were 3.6–63.20 mg/L and 5.00 - 71.00 mg/L (Avg. 7.85 mg/L), respectively. Machiwa (2010) also recorded similar mean values in the Lake Victoria Basin of Tanzania.

In terms of the PCA; PC1, which accounts for the greatest variance among the dataset and may therefore suggest the most dominant process in terms of control on the distribution of heavy metals in the area, may be linked to an external process that affects the distribution of some metals in the soil, as indicated by the high CV% for elements like Cu (102.45%) and V (87.79%). The source of Cu could be agrochemical applications such as fertilizers and biocides (Lamichhane et al., 2018) or metallurgical activities associated with gold mining (Kazapoe, 2023). PC2 implies a geological or natural source, for instance, through the process of weathering of rocks containing minerals. The presence of Ba and Sr, which can be associated with fertilizers or pesticides, suggests a mixed source with some contribution from low-level agricultural pollution. The third factor (PC3) suggests human disturbances, especially from mining activities. The high CV% for Mo (144. 66%) and its association with gold in the Birimian terrain may also imply that mining waste can alter the soil chemistry. The relatively low standard deviation of Pb (43. 71%) suggests a minor anthropogenic impact, following the findings of Arhin et al. (2019) and Akoto et al. (2023), who link Pb to the local gold-bearing rocks.

The results in Table 4 and Fig. 3 suggest lithological control of the elements present in the samples, particularly those associated with mafic rocks. The elemental associations in the main cluster of Fig. 3 additionally show the elemental assemblages typically associated with gold mineralisation in Ghana (Kazapoe et al., 2022).

The pollution indices further indicated the pollution status of the area. According to the Nemerow Pollution Index (NIPI), the area around

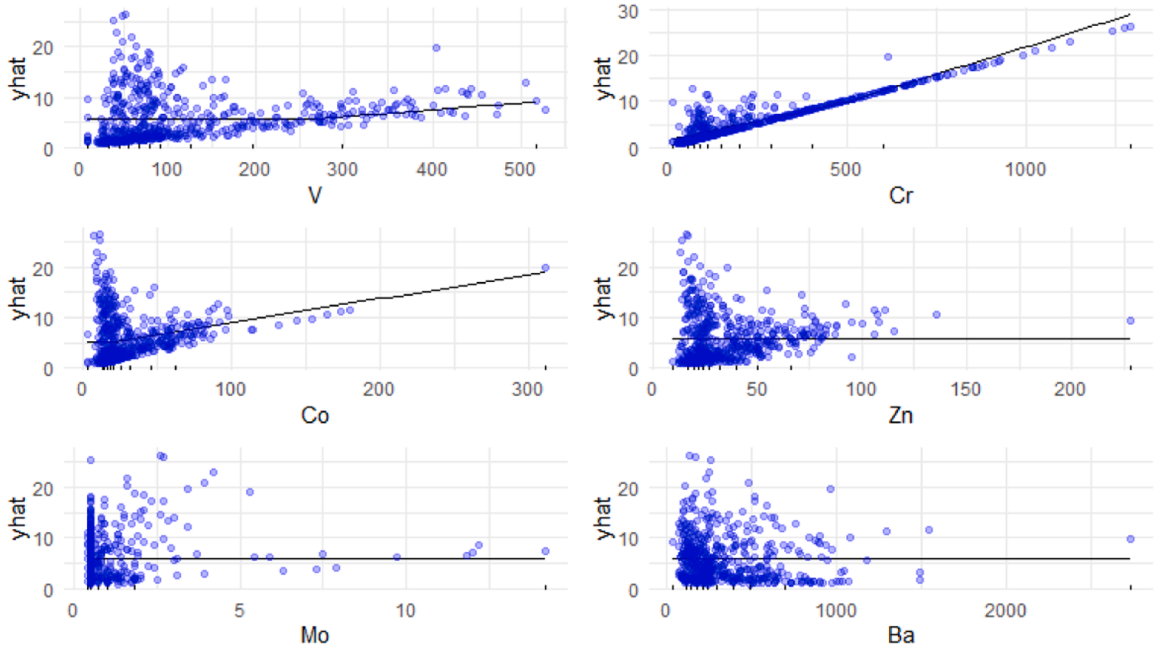

Fig. 5. Partial dependence plot for each predictor variable against NIPI.

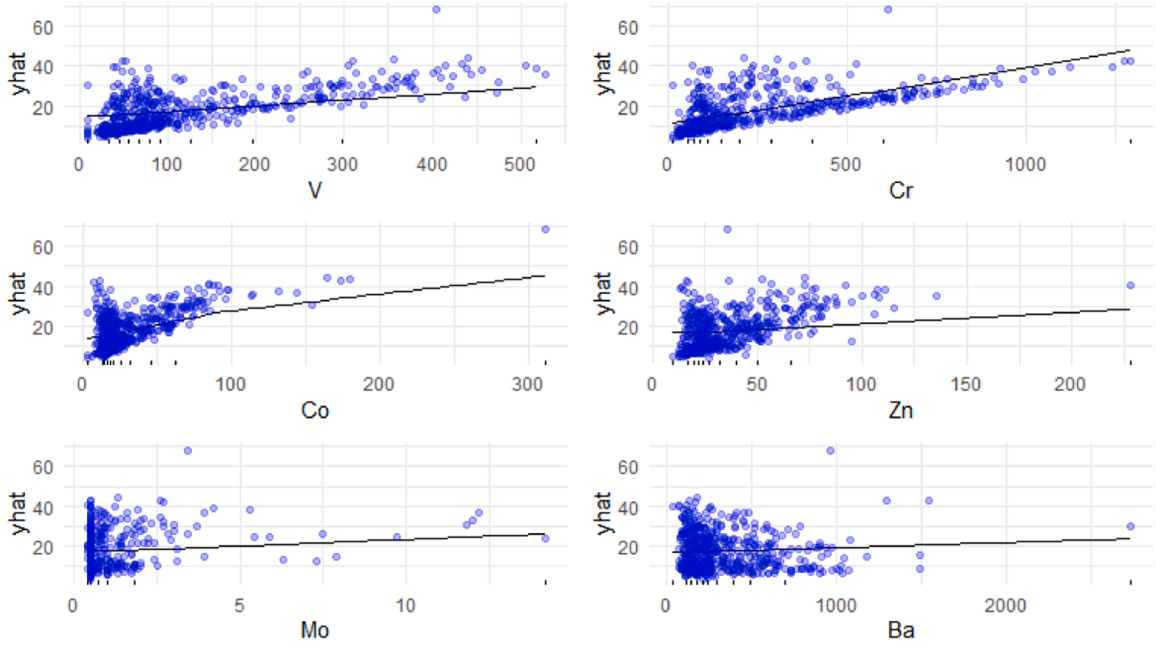

Fig. 6. Partial dependence plot for each predictor variable against Cdeg (yhat = Cdeg).

Datoko is relatively clean while Shaega in the western part and some areas in the eastern part are moderately to highly polluted. These areas are known centres of galamsey activities, which may explain the values derived. The results for the Cdeg further underscore the effect of illegal mining on Shaega and the eastern part of the study area which are classed as having considerable pollution (Fig. 4B). Mirroring the other indices, the mCdeg shows that Shaega and Buin are moderately polluted while the eastern part around Tula records some spots which are categorised as high (Fig. 4D). This is mainly due to the activities of the illegal miners in the area.

The MARS models identified several complex non-linear relationships between the independent variables (V, Cr, Co, Zn, Mo, and Ba) and the dependent variables (NIPI, Cdeg, and mCdeg). Cr, V, and Co consistently emerged as the vital role players amongst the independent variables with significant non-linear relationships. This differs from findings in existing literature. Wang and Xu (2022) used ML models like support vector machines (SVM), partial least squares (PSLR), and gaussian process regression (GPR) in the prediction of pollution indices like NIPI and ecological risk index (RI. Their findings found Cu, Pb, and Zn to be the most significant contributors to their model output. The common element between both studies is Zn, which was found to be vital to model by Wang and Xu (2022) but not in our case.

Cr across all 3 models plays a dominant role with splits at multiple thresholds across all 3 models suggesting a strong relationship between the chemical element and all 3 dependent variables subject to their individual thresholds.

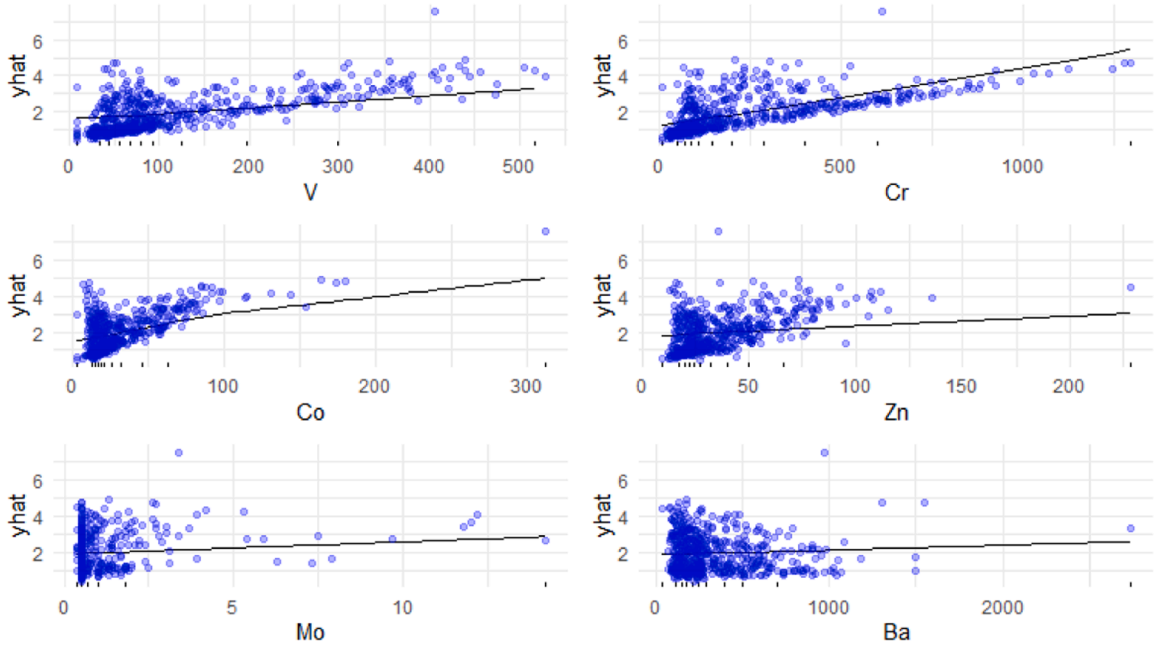

Fig. 7. Partial dependence plot for each predictor variable against mCdeg.

Results from Table 7 suggest V acts as a moderator to Cr enhancing its effect on NIPI past specifics thresholds. In Tables 8 and 9 however, V has a more active role directly impacting the dependent variables Cdeg and mCdeg subject to their individual thresholds.

A similar effect is observed for Co where in model 1 is acts as an amplifier to Cr's ability to influence NIPI subject to the threshold but in model 2 and 3 a more direct interaction between Co and the two dependent variables Cdeg and mCdeg is observed at specific thresholds.

Mo, Ba, and Zn also play notable roles but only in models 2 and 3. No notable relationships are observed between the 3 elements and model 1. Mo has relationships with Cdeg and mCdeg but these interactions are subtle at best.

Ba on the other hand although smaller in magnitude exhibits significant threshold effect towards both Cdeg and mCdeg.

Zn has a direct although subtle interaction with mCdeg highlighting a difference between the two dependent variables even though they share similar features. No such direct interaction is observed in Cdeg but rather a slight interaction between Cr and Zn for model 2.

The interactions identified by the MARS models highlight not only interactions between the dependent variables and independent ones but also between independent variables suggesting that changes in one element's concentration may influence the effects of another towards the dependent variables pointing to a highly dynamic system.

5. Conclusion

This paper examines the environmental quality of the Nangodi region in terms of the concentrations of heavy metals in the soil, the sources and interactions between these elements, the spatial distribution, the ecological risk associated with these elements, and the effectiveness of ML models as reliable predictors of soil.. Considering the performance of the MARS algorithm in this study, the very high predictive ability observed between the variables considered both dependent and independent highlights NIPI, Cdeg, and mCdeg as viable tools for predicting soil quality. Based on the comprehensive analysis, Cr, Co, and V are chosen as important predictors for the dependent variables; they have high CV%, skewness, and kurtosis, which show a large range of dispersion and non-normal distribution. This indicates scattered hotspots, which may have been shaped by gold deposits and illicit mining operations. This is further corroborated by the PCA which produced three principal components which cumulatively accounted for 81.30% of the total variance. PC1 (V, Co, Cu, and Zn), PC2 (Sr and Ba), and PC3 (Mo and Pb) suggest lithological controls on the elemental concentrations. The pollution indices reflected the state of soil contamination at different levels. The Nemerow Pollution Index (NIPI) indicated that the Shaega and certain parts of the eastern region of the study area and Datoko had moderate to high levels of pollution, which is consistent with areas with high galamsey activities. Cdeg analysis further corroborated these findings. The Pollution Load Index (PLI) also showed a decreasing trend in the quality of soil with 43.84% of samples showing evidence of environmental disturbance, especially in the Shaega, Buin and the eastern region close to Tula which are the most influenced by mining.

Based on the above findings, the authors recommend the following measures that if adopted, may assist in reducing the impacts of soil contamination, enhancing public health and environmental stewardship in the study area.

- There is a need to enhance the implementation of environmental laws to address the challenges of galamsey in areas like Shaega, Buin, and the eastern parts of Tula.
- There is a need to embark on immediate soil remediation measures in the affected areas to restore soil fertility and prevent further pollution of the environment.
- Establish a consistent monitoring system to assess the soil's condition and detect potential contaminations. This will enable the implementation of early intervention measures and the reduction of extreme pollution. To raise awareness among the local communities regarding the detrimental effects of illegal mining on their health and the environment. A routine monitoring system, as recommended in this study, is highly feasible when implemented, as it will mitigate the consequences of soil contamination on mining communities. A well-organised monitoring schedule would enable the identification of pollution surges in advance, thereby reducing the environmental impact. It is advisable to conduct soil sampling on a bi-annual basis, at the conclusion of the dry and rainy season. The implementation of more stringent monitoring protocols is likely to be restricted by the necessity of funding these exercises. These periods are significant

because they can reveal varying pollution patterns, such as the concentration of pollutants during periods of drought or precipitation. In order to ensure that the results of the monitoring are accessible and comprehensible to the residents whose health is being impacted by the pollution, this routine monitoring must be supplemented with other aspects of data management and community engagement.Educate and equip them with the right tools and techniques to embrace sustainable mining. Encourage and sustain other sources of income to lessen the reliance on galamsey. This can range from agricultural projects, eco-tourism, and other income-generating activities that are environmentally friendly.

- Further studies should be conducted to determine the effects of soil pollution on crop yields and ecosystem services in the long run to formulate effective measures for soil protection.

CRediT authorship contribution statement

Daniel Kwayisi: Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Raymond Webrah Kazapoe:** Writing – original draft, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Seidu Alidu:** Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Samuel Dzidefo Sagoe:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Aliyu Ohiani Umaru:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Ebenezer Ebo Yahans Amuah:** Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Prosper Kpiebaya:** Writing – original draft, Validation, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Funding

This research did not receive any grant from any funding agency, commercial or profit sectors.

Compliance with ethical standards

Conflict of interest
The authors declare that they have no competing interests.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.hazadv.2024.100480.

References

Abouchami, W., Boher, M., Albarede, F., 1990. 2.1 mafic magmatism in West Africa: an early stage of crustal accreation. In: Seventh international conference on geochronology, cosmochronology and isotope geology.
Abrahim, G.M.S., Parker, R.J., 2008. Assessment of heavy metal enrichment factors and the degree of contamination in marine sediments from Tamaki Estuary, Auckland, New Zealand. Environ. Monit. Assess. 136 (1), 227–238.
Abu, M., Kalimenze, J., Mvile, B.N., Kazapoe, R.W., 2021. Sources and pollution assessment of trace elements in soils of the central, Dodoma region, East Africa: implication for public health monitoring. Environ. Technol. Innov. 23, 101705.
Achina-Obeng, R., Aram, S.A., 2022. Informal artisanal and small-scale gold mining (ASGM) in Ghana: assessing environmental impacts, reasons for engagement, and mitigation strategies. Resour. Policy. 78, 102907.
Adiguzel, M.B., Cengiz, M.A., 2023. Model selection in multivariate adaptive regressions splines (MARS) using alternative information criteria. Heliyon. 9 (9).
Agboola, O., Babatunde, D.E., Fayomi, O.S.I., Sadiku, E.R., Popoola, P., Moropeng, L., Mamudu, O.A., 2020. A review on the impact of mining operation: Monitoring, assessment and management. Results Eng. 8, 100181.
Ağyar, O., Tırınk, C., Önder, H., Şen, U., Piwczyński, D., Yavuz, E., 2022. Use of multivariate adaptive regression splines algorithm to predict body weight from body measurements of anatolian buffaloes in türkiye. Animals 12 (21), 2923.
Ahado, S.K., Agyeman, P.C., Borůvka, L., Kanianska, R., Nwaogu, C., 2024. Using geostatistics and machine learning models to analyze the influence of soil nutrients and terrain attributes on lead prediction in forest soils. Model. Earth. Syst. Environ. 10 (2), 2099–2112.
Akoto, O., Yakubu, S., Ofori, L.A., Bortey-Sam, N., Boadi, N.O., Horgah, J., Sackey, L.N., 2023. Multivariate studies and heavy metal pollution in soil from gold mining area. Heliyon. 9 (1).
Amosah, J., Lukman, T., 2023. From adaptation to resilience: the capability of women smallholder farmers in The Nabdam district of the upper east region. Int. J. Manag. Entrepr. Res. 5 (7), 483–502.
Amuah, E.E.Y., Fei-Baffoe, B., Sackey, L.N.A., Dankwa, P., Nang, D.B., Kazapoe, R.W., 2022a. Remediation of mined soil using shea nut shell (Vitellaria paradoxa) as an amendment material. J. Environ. Chem. Eng. 10 (6), 108598.
Amuah, E.E.Y., Fei-Baffoe, B., Sackey, L.N.A., Douti, N.B., Kazapoe, R.W., 2022b. Understanding the distribution, source-pattern and geochemical controls of soils in an artisanal mine site during a ban on illegal mining activities: Is a ban an absolute solution? Soil Secur. 9, 100078.
Arhin, E., Boansi, A.O., Zango, M.S., 2016. Trace elements distributions at Datoko-Shega artisanal mining site, northern Ghana. Environ. Geochem. Health 38, 203–218.
Arhin, E., Kazapoe, R., Zango, M.S., 2017. The Hidden Dangers of Unknowingly Ingesting Harmful Trace Elements from Food Crops and Their Health Implications. a case study at Talensi District in the Upper East Region, Ghana.
Arhin, E., Zhang, C., Kazapoe, R., 2019. Medical geological study of disease-causing elements in Wassa area of Southwest Ghana. Environ. Geochem. Health 41 (6), 2859–2874.
Bansah, K.J., Yalley, A.B., Dumakor-Dupey, N., 2016. The hazardous nature of small scale underground mining in Ghana. J. Sustain. Min. 15 (1), 8–25.
Bibri, S.E., Krogstie, J., Kaboli, A., Alahi, A., 2024. Smarter eco-cities and their leading-edge artificial intelligence of things solutions for environmental sustainability: A comprehensive systematic review. Environ. Sci. Ecotechnol. 19, 100330.
Bottegal, G., Pillonetto, G., 2018. The generalized cross validation filter. Automatica 90, 130–137.
Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geosci. Model. Dev. 7 (3), 1247–1250.
Dzikunoo, E.A., Kazapoe, R.W., Agbetsoamedo, J.E., 2021. An integrated structural and geophysical approach to defining the structures of part of the Nangodi greenstone belt, northeastern Ghana. J. Afr. Earth Sci. 180, 104238.
European Union, 2002. European Union. Heavy Metals in Wastes. Eur. Commission Environ. Retrieved from h.t.t.p://ec.e uropa.eu/environment/waste/studies/pdf/heavy metals report.pdf.
Fister, D., Pérez-Aracil, J., Peláez-Rodríguez, C., Del Ser, J., Salcedo-Sanz, S., 2023. Accurate long-term air temperature prediction with machine learning models and data reduction techniques. Appl. Soft. Comput. 136, 110118.
Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19 (1), 1–67.
Gackowski, M., Szewczyk-Golec, K., Pluskota, R., Koba, M., Mądra-Gackowska, K., Woźniak, A., 2022. Application of multivariate adaptive regression splines (MARSplines) for predicting antitumor activity of anthrapyrazole derivatives. Int. J. Mol. Sci. 23 (9), 5132.
Ghana Statistical Services (GSS), 2010. Nabdam District. Provisional Results of the 2010 Population and Housing Census, pp. 33,826Retrieved from. https://unstats.un.org/unsd/demographicsocial/census/documents/Ghana/Provisional_results.pdf. Accessed July 7, 2024.
Hair Jr, J.F., Hult, G.T.M., Ringle, C.M., Sarstedt, M., Danks, N.P., Ray, S., ... & Ray, S. (2021). Evaluation of the structural model. Partial least squares structural equation modeling (PLS-SEM) using R: a workbook, 115–138.
Hakanson, L., 1980. An ecological risk index for aquatic pollution control. A sedimentological approach. Water. Res. 14 (8), 975–1001.
Hlokoe, V.R., Mokoena, K., Tyasi, T.L., 2022. Using multivariate adaptive regression splines and classification and regression tree data mining algorithms to predict body weight of Nguni cows. J. Appl. Anim. Res. 50 (1), 534–539.
Iglesias, J., Cuesta, I., Salueña, C., Solé, J., Prevatt, D.O., Fabregat, A., 2024. Predictive modeling of severe weather impact on individuals and populations using Machine Learning. Int. J. Disaster Risk Reduc. 105, 104398.
Isung, C.B., 2021. The socio-economic implications of artisanal and small-scale mining on mining communities in northern Ghana. Open Access Lib. J. 8 (03), 1.
Javed, H., Muqeet, H.A., Javed, T., Rehman, A.U., Sadiq, R., 2023. Ethical Frameworks for Machine Learning in Sensitive Healthcare Applications. IEEE Access.
Kamm, S., Veekati, S.S., Müller, T., Jazdi, N., Weyrich, M., 2023. A survey on machine learning based analysis of heterogeneous data in industrial automation. Comput. Ind. 149, 103930.

Kazapoe, R.W., 2014. Geological Structural Interpretation of the Bongo and Nabdam districts of the Bole Nangodi belt, North-Eastern Ghana, from Integrated Geophysical and Field Mapping Datasets (MPhil Thesis). University of Ghana.
Kazapoe, R.W., 2023. Assessing the Lithium Potential of the Paleoproterozoic Rocks of the West African Craton; the Case so Far. Geosystem Eng. 26 (6), 257–271.
Kazapoe, R.W., Amuah, E.E.Y., Dankwa, P., 2022. Sources and pollution assessment of trace elements in soils of some selected mining areas of southwestern Ghana. Environ. Technol. Innov. 26, 102329.
Kazapoe, R.W., Amuah, E.E.Y., Dankwa, P., Ibrahim, K., Mville, B.N., Abubakari, S., Bawa, N., 2021. Compositional and source patterns of potentially toxic elements (PTEs) in soils in southwestern Ghana using robust compositional contamination index (RCCI) and k-means cluster analysis. Environ. Challenges 5, 100248.
Kazapoe, R., Arhin, E., 2021. Determination of local background and baseline values of elements within the soils of the Birimian Terrain of the Wassa Area of Southwest Ghana. Geol. Ecol. Landsc. 5 (3), 199–208.
Kim, H.Y., 2018. Statistical notes for clinical researchers: simple linear regression 2–evaluation of regression line. Restor. Dent. Endod. 43 (3).
Lamichhane, J.R., Osdaghi, E., Behlau, F., Köhl, J., Jones, J.B., Aubertot, J.N., 2018. Thirteen decades of antimicrobial copper compounds applied in agriculture. a review. Agron. Sustain. Dev. 38 (3), 28.
Ministry of Food and Agriculture. (n.d.). Nabdam District. Retrieved from http://mofa.gov.gh/site/sports/district-directorates/upper-east-region/269-talensi-nabdam.
Moomen, A.W., Dewan, A., 2017. Assessing the spatial relationships between mining and land degradation: evidence from Ghana. Int. J. Min. Reclam. Environ. 31 (7), 505–518.
Murray, R.J., 1960. The geology of the Zuarungu 12° field sheet. Geol. Surv. Ghana, Bull. 25.
Naser, A.H., Badr, A.H., Henedy, S.N., Ostrowski, K.A., Imran, H., 2022. Application of Multivariate Adaptive Regression Splines (MARS) approach in prediction of compressive strength of eco-friendly concrete. Case Stud. Constr. Mater. 17, e01262.
Nunoo, S., Manu, J., Owusu-Akyaw, F.K., Nyame, F.K., 2022. Impact of artisanal small-scale (gold and diamond) mining activities on the Offin, Oda and Pra rivers in Southern Ghana, West Africa: a scientific response to public concern. Heliyon. 8 (12).
Oduro, S.D., Metia, S., Duc, H., Hong, G., Ha, Q.P., 2015. Multivariate adaptive regression splines models for vehicular emission prediction. Visualiz. Eng. 3, 1–12.
Okyere, M., Ayitey, J.Z., Ajabuin, B.A., 2021. Large scale mining in Ghana: a review of the implications on the host communities. Journal of Degraded and Mining Lands Management 9 (1), 3193.
Sekyi-Annan, E., 2019. Performance Evaluation of Reservoir-Based Irrigation Schemes in the Upper East Region of Ghana (Doctoral dissertation. Universitäts-und Landesbibliothek Bonn.
Şengül, T., Çelik, Ş., Şengül, Ö., 2020. Use of Multivariate Adaptive Regression Splines (MARS) for predicting parameters of breast meat in quails. JAPS: J. Animal Plant Sci. 30 (4).
Strielkowski, W., Vlasov, A., Selivanov, K., Muraviev, K., Shakhnov, V., 2023. Prospects and challenges of the machine learning and data-driven methods for the predictive analysis of power systems: a review. Energies. (Basel) 16 (10), 4025.
Tom-Dery, D., Dagben, Z.J., Cobbina, S.J., 2012. Effect of illegal small-scale mining operations on vegetation cover of arid northern Ghana. Res. J. Environ. Earth Sci. 4 (6), 674–679.
Tomlinson, D.L., Wilson, J.G., Harris, C.R., Jeffrey, D.W., 1980. Problems in the assessment of heavy-metal levels in estuaries and the formation of a pollution index. Helgoländer Meeresuntersuchungen 33, 566–575.
Trunfio, T.A., Scala, A., Giglio, C., Rossi, G., Borrelli, A., Romano, M., Improta, G., 2022. Multiple regression model to analyze the total LOS for patients undergoing laparoscopic appendectomy. BMC. Med. Inform. Decis. Mak. 22 (1), 141.
United States Environmental Protection Agency, 2002. Supplemental Guidance for Developing Soil Screening Levels For Superfund Sites. Office of Solid Waste and Emergency Response, Washington, D.C. Retrieved from. http://www.epa.gov/superfund/health/conmedia/soil/index.htm.
Vinogradov, A.P., 1959. The Geochemistry of Rare and Dispersed Chemical Elements in Soils, 2nd ed. Consultants Bureau Enterprises, New York, p. 209. Revised and enlarged.
Wang, Y., Zhao, Y., Xu, S., 2022. Application of VNIR and machine learning technologies to predict heavy metals in soil and pollution indices in mining areas. J. Soils. Sediments. 22 (10), 2777–2791.
Zhang, W., Goh, A.T., 2016. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geosci. Front. 7 (1), 45–52.