Environmental Science and Pollution Research (2022) 29:87490–87508
https://doi.org/10.1007/s11356-022-21890-8

RESEARCH ARTICLE

Machine learning-based prediction of toxic metals concentration in an acid mine drainage environment, northern Tunisia

Mariem Trifi[1] · Anis Gasmi[2,3] · Cristina Carbone[4] · Juraj Majzlan[5] · Nesrine Nasri[6,7] · Mohja Dermech[8] · Abdelkrim Charef[1] · Hamza Elfil[2]

Received: 8 April 2022 / Accepted: 2 July 2022 / Published online: 9 July 2022

Abstract

In northern Tunisia, Sidi Driss sulfide ore valorization had produced a large waste amount. The long tailings exposure period and in situ minerals interactions produced an acid mine drainage (AMD) which contributed to a strong increase in the mobility and migration of huge heavy metal (HM) quantities to the surrounding soils. In this work, the soil mineral proportions, grain sizes, physicochemical properties, SO_4^{2-} and S contents, and Machine Learning (ML) algorithms such as the Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) models were used to predict the soil HM quantities transferred from Sidi-Driss mine drainage to surrounding soils. The results showed that the HM concentrations had significantly increased with the increase of decomposition and oxidation of galena, marcasite, pyrite, and sphalerite-marcasite and Fe-oxide-hydroxides quantities and the sulfate dissolution (marked with SO_4^{2-} ions increase) that produced the decreased soil pH. Compared to SVM, and ANN models outputs, the RF model that revealed higher R^2_{val}, RPD, RPIQ, and lower error indices had satisfactorily predicted the soil HM accumulation coming from the AMD environment.

Keywords Machine Learning (ML) · Acid mine drainage (AMD) · Heavy metals (HMs) · Mine tailings · Mineralogical compositions

Introduction

Tailings that are widely distributed around the globe (Carmo et al. 2020) are considered as a reservoir and a vector of HM mobility and transport and which caused the biocenosis (fauna, flora, and human health), lithosphere (agricultural soils), hydrosphere (surface, and groundwater), and atmosphere (atmospheric dust) degradation (Jambor et al. 2003;

Responsible Editor: Marcus Schulz

Highlights
- Soil heavy metals were predicted using RF, SVM, and ANN.
- Soil Zn, Pb, Mn, Cu, and Cd are closely correlated with mineralogical compositions, dissolution sulfates (marked with increasing SO_4^{2-} ions, and decreasing soil pH values).
- Soil Fe and Mn concentrations are closely related to the iron minerals.
- RF outperforms SVM and ANN to predict soil HMs.

✉ Mariem Trifi
mariem.trifi@certe.rnrt.tn

[1] Georesources Laboratory, Water Research and Technology Center (CERTE), Borj-Cedria Technopole, B.P. 273, Soliman 8020, Tunisia

[2] Laboratory Desalination and Natural Water Valorization (LaDVEN), Water Research and Technology Center (CERTE), Borj-Cédria Technopole, B.P. 273, Soliman 8020, Tunisia

[3] Center for Remote Sensing Application (CRSA), Mohammed VI Polytechnic University (UM6P), 43150, Ben Guerir, Morocco

[4] Departiment of Earth, Environment and Life Sciences (DISTAV), University of Genoa, 26 Corso Europa, 16132 Genoa, Italy

[5] Institute of Geosciences, Friedrich-Schiller University, Burgweg 11, 07749 Jena, Germany

[6] Higher Institute of Environmental Technologies, Urban Planning and Construction, University of Carthage, Charguia II, 2035 Tunis, Tunisia

[7] Laboratory in Hydraulic and Environmental Modelling, National Engineering School of Tunis, University of Tunis, Tunis, Tunisia

[8] Mineral Resources and Environment Laboratory, LR01ES06, Sciences Faculty of Tunis, Tunis El Manar University, 1092 Tunis, Tunisia

Fengxia and Wenbao 2010; Frau and Marescotti 2011; Khalil et al. 2014; Tan et al. 2014; Ayari et al. 2016; Pourret et al. 2016; Dold 2006; Zhang et al. 2020). The total assessment of HM mobility (Kemper and Sommer 2002; Trifi et al. 2019), and their environmental fate require physico-chemical, mineralogical, and chemical analyses of a large number of tailings and soils samples (various locations and depths) (Kemper and Sommer 2003; Trifi et al. 2018 and 2019). This approach is time-consuming and expensive (Chang et al. 2001; Tan et al. 2014; Yaseen 2021).

However, the soil HM content prediction could be established by remote sensing (Malley and Williams 1997), spectroscopy (Choe et al. 2008), and algorithms such as multiple linear regression, generalized linear models, partial least squares regression, weighted geographic regression, and linear and kriging mixed models (Thompson et al. 2006; Kumar et al. 2012). The inconvenience is the limited prediction performance (Lark 1999; Wu et al. 2005; Tajik et al. 2012; Wang et al. 2014; Guo et al. 2015; Zhang et al. 2017).

Compared to other prediction models, Machine Learning ML methods provide increasingly accurate analysis of nonlinear and hierarchical relationships, insensitivity to noise characteristics, and resistance to overfitting, with reliable error measurement (Breiman, 2001; Zeraatpisheh et al. 2020; El Amri et al. 2022). Mojid et al. (2019) showed that ML algorithms facilitate the understanding of the nonlinear data of HM removal in comparison to the isotherm, statistical, mathematical, physical, and empirical models. Alvarez-Guerra et al. 2010; Tepanosyan et al. 2020 have considered the use of the ML models more than the classical statistical methods owing to solving the nonlinearity and non-stationarity phenomena. The ML algorithms were used for soil quantification such as soil organic matter, texture, and other properties (Gomez et al. 2019; Gasmi et al. 2014, 2022). Also, Guyon and Elisseeff (2003) and Wu et al. (2008) reported that predictive models for HMs based on the number of data sets, type data, catchment characteristics, optimization prior to model building, weight minimization, and the choice of the specific algorithm. Yaseen et al. (2018) applied ML for soil, water, and bodies HM simulation without the pre-knowledge for the data statistical distribution.

For example, based on pH, SO_4, HCO_3, TDS, EC, Mg, and Ca input variables, the Ni and Fe concentrations were predicted by Gholami et al. (2011) using a support vector machine (SVM) and back-propagation neural network (BPNN) algorithms. To predict the soil Cd concentrations, auxiliary land use information, soil organic matter, pH, and topographic data were inputted by Qiu et al. (2016) using the stepwise linear regression (SLR), classification, and regression tree (CART), and random forest (RF) models. Cd and Pb were estimated by adaptive neurofuzzy inference systems (ANFIS), artificial neural network (ANN), and multiple linear regression (MLR) models based on organic carbon, phosphorus, total Nitrogen, pH, and clay Bazoobandi et al. (2019). Omondi and Boitt (2020) applied RF to predict the spatial distribution of soil Cd, Pb, and Zn contents. Zhang et al. (2020) used the urban expansion, areas of different land-use types in a specific grid, urbanization history, soil properties, and ML models to predict the spatial distribution of soil HMs of the studied site. Lu et al. (2022) applied (RF) model to predict the removal efficiency of HMs using flocculant properties, flocculation conditions, and HM properties.

As we know, the input data to ML algorithms in HM predictions were, until now, the soil properties, topography, changing environmental conditions, and urban expansion. However, previous studies revealed that there were strong correlations between the HMs of polluted soils with mineralogical compositions that corresponded to soluble minerals or prone to oxidation of tailings (Jambor et al. 2000; Hammarstrom et al. 2005; Dold 2006; Carbone et al. 2013; Bouzahzah et al. 2014; Murray et al. 2014; Balci and Demirel 2018; Trifi et al. 2018 and 2019). Therefore, it is for the first time, this study has applied ML methods to linking soil HMs to mineralogical compositions of an AMD environment.

The objectives of this study were to (1) predict the HM accumulation in AMD environment using the mineral proportions, physico-chemical properties and grain size and the advanced ML approaches, (2) determine the important attributes for predicting soil HMs based on the predictor importance result, and (3) evaluate and compare the performance of the RF, SVM, and ANN models used to predict the soil HM concentrations.

Materials and methods

Site description

The Sidi-Driss mine is located in the northwest region of Tunisia, 150 km of Tunis, at 9°06′55″ E latitude and 37°03′14 ″ N longitude (Fig. 1a, b). The ore is in the Neogene filling of the Sidi-Driss-Tamra basin. It is limited in the southeast by a complex magmatic-saline structure of the "Oued Belif," in the east-southeast by Damous river, and in the west by the Mn-Fe Tamra mine (Fig. 1c) (Trifi et al. 2018 and 2019). The ore area is approximately 2 km^2, and its average elevation is around 150 m above sea level, with predominant rough topography. Sidi-Driss is in humid to subhumid Mediterranean region where temperatures range from − 1 to 47 °C. The maximum evapotranspiration level is 320 mm. The annual precipitation average is 1000 $mm{\cdot}year^{-1}$, and the mean wind speed is 2.7 $m{\cdot}s^{-1}$.

The host rocks are siliciclastic continental sediments with oxides-hydroxides (hematite and goethite), and abundant silicates (quartz, K-feldspar, plagioclase, mica, kaolinite, illite, and smectite), coming from fragments of (i) Numidian sandstone Unit, (ii) "Yellow Balls," and clayey limestones (alumino-silicate cement) of ed Diss Unit (Rouvier, 1977), (iii) the ferruginous breccia of "Oued Belif," (iv) rhyodacite, (v) pyroclasts, and (vi) gneiss from the magmato-salt complex (Dermech 1990; Dermech et al. 2022; Trifi et al. 2018 and 2019). The major ore minerals are sulfides (marcasite, galena, sphalerite, and pyrite), sulfates (barite, celestite, and gypsum) (Dermech 1990; Dermech et al. 2022), and the minor are low quantities of carbonates (6% calcite and siderite) (Trifi et al. 2018 and 2019). Many rare minerals such as arsenopyrite, phosphates, vanadates, arsenates, Bi-sulfosalts, anglesite ($PbSO_4$), cerusite ($PbCO_3$), jarosite and plumbojarosite ($PbFe_6\ (SO_4)_4\ (OH)_{12}$), wulfenite ($PbMoO_4$), coronadite ($PbMn_8O_{16}$), calamine ($Zn_4\ [Si_2O_7]\ (OH)_2, H_2O$) and smithsonite ($ZnCO_3$), chalcophanite ((Zn, Fe, Mn) $Mn_3O_73(H_2O)$), heterolite ($ZnMn_2O_4$), and stilpnosiderite ($Fe_2O_3\ (2H_2O)$) were also reported by Stefanov and Ouchev (1972), and Negra (1987).

Two mineralization types are distinguished at Sidi-Driss namely stratiform and vein (hydrothermal) ores of Late Miocene to Pliocene–Pleistocene age. The dominant sulfides are pyrite, sphalerite, marcasite, and galena (Dermech 1990). This ore is characterized by strong oxidation of galena (35%), and sphalerite (25%) because the most mineralized layer is above 60 m of the Damous river (Trifi et al. 2018).

Pb and Zn were mined at Sidi-Driss in an open-pit mine, (Fig. 2a) during three periods (1902–1925, 1925–1950, and 1969 until today). The mine wastes were deposited in the edge dump that is only 3 m far from the Damous river (Fig. 1c). One depression area immersed with meteoric water and the peripheral dry zone was observed at the surface of the dump (Fig. 1c, b). Its upper perimeter is 492 m, the upper surface is 1.61 ha (hectare), the basal perimeter is 610 m and the basal surface is 2.46 ha (Trifi et al. 2019).

Owing to the long-term waste exposition, this tailings dump became an ideal experimental site to identify the effects of minerals dissolution on the HM mobility because of (i) the ore and the tailings dump are close to the Damous river, (ii) the advanced degree of galena and sphalerite oxidation, and (iii) the moisture content of the remaining flotation water, (iv) and the highest sphalerite-marcasite amounts in the tailings. HM mobility is also controlled by the dilution and evaporation processes, the acid mine solution in the submerged area, runoff-percolation, and infiltration of water flowing into the dump and its drainage by the dump drains to Damous river.

Soil sampling and analytic methods

A total of 72 topsoils (0–20 cm) were sampled. Wet and ochreous sediments and yellow, ochreous, and rusty deposits were collected from the dump surface and its sides, respectively. All samples were homogenized, passed through a 2-mm stainless steel sieve, lyophilized (freeze-dried), ground, and kept in a cooler room for further mineralogical and geochemical analyses.

The pH and electrical conductivity (EC) were measured using soil: distilled water ratios of 1:2.5, and 1:5, respectively after stirring for 2 h (Montoroi 1997; Rayment and Higginson 1992). The pH and EC were determined using a pH meter (LPH 230 T-type) and a conductivity meter ORION 150, respectively.

Mineralogical composition was determined using powder X-ray diffractometry (PXRD) with CuKα radiation (current 40 mA, voltage 40 kV). Each sample was scanned between 3 and 80° 2θ at a scan rate of 1°/min. Minerals were identified based on the sets of their peak positions using a standard database.

About 0.5 g of each sample was digested by 5 ml of a mixture of HF (40%), 1.5 ml $HClO_4$ (70%), 3.75 ml, HCl (37%), and 1.25 ml HNO_3 (65%) in a sand bath at 250 °C. Afterward, 100 ml of ultrapure water was added and the resulting solution was acidified to pH≈1.5 with ultrapure HNO_3 (Hageman and Briggs 2000). The Fe, Zn, Pb, Cd, Cu, Mn, and S concentrations were carried out at Georessources laboratory, (Water Research and Technology Center, Technopole of Borj Cedria, Tunisia) using atomic absorption spectrometry (AAS)-Perkin Elmer type and ion chromatography on a non-acidified solution. In the whole analysis, reagents and standard solutions were prepared with millipore deionized water. To check the accuracy of HM analysis, NIST 2709a was used in the digestion and analysis of the samples as part of the quality assurance/quality control (QA/QC) protocol. For trace element concentrations, 20% (chosen at random) of samples were replicated. Also, the samples which had either highest or lowest concentrations were replicated three times. To check the precision of measurement, blank samples were run after every five samples. The value below the detection limit and within 2% of the certified one was only accepted.

Statistical analysis

Statistical analysis was implemented using SPSS software (SPSS Inc., USA). Descriptive statistics like standard deviation (SD), mean, minimum, maximum, and quartile of input parameters including the mineralogical compositions, physicochemical properties, grain size, and HMs were used. For input variables, the sample size was taken as $n = 72$. Pearson correlation coefficients and principal component analysis (PCA). PCA is a descriptive method that reduces and transforms the dimensions of normalized variables into new principal components (PC) by varimax rotation was used in the current study. This technique produces multiple

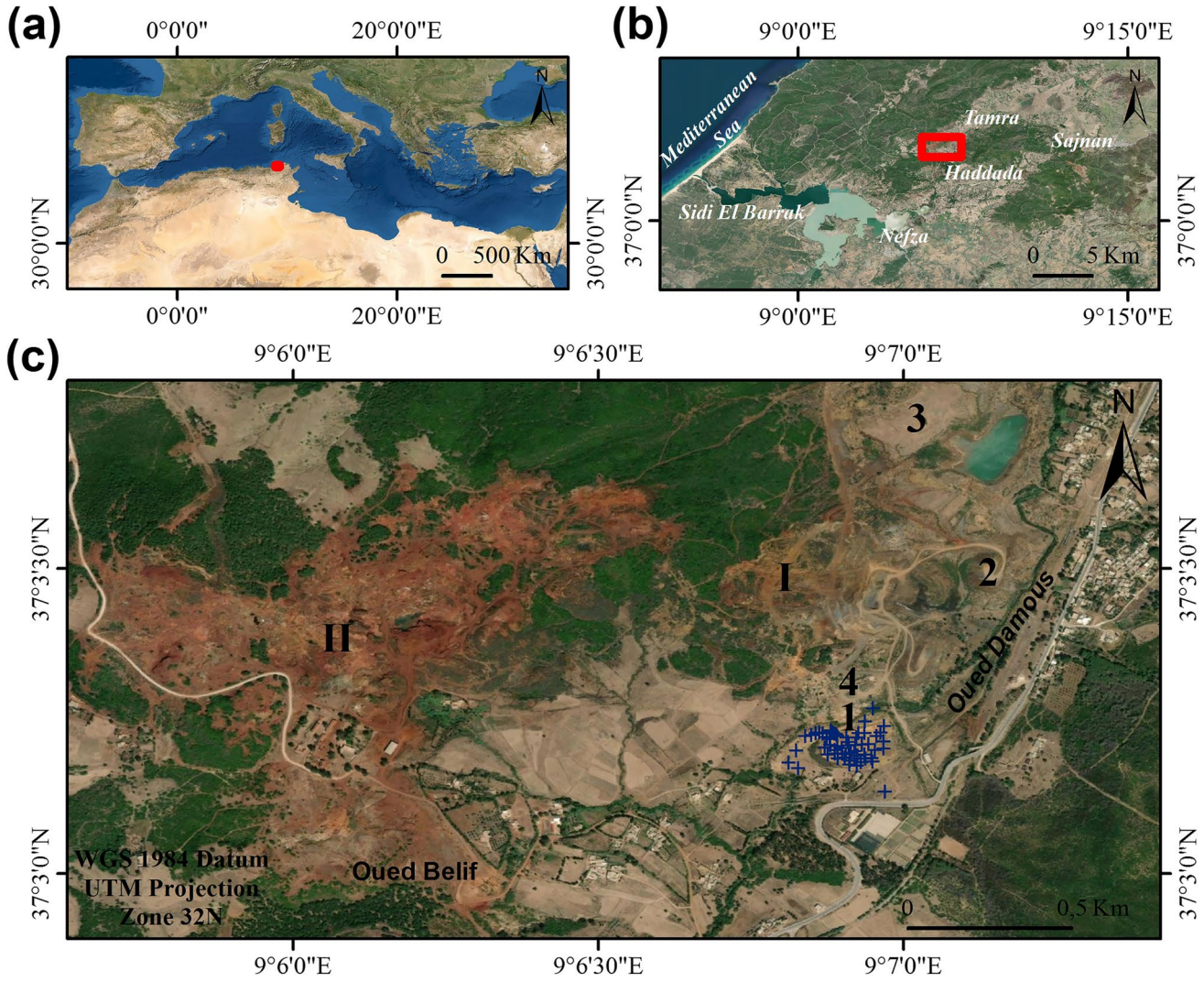

Fig. 1 **a**, **b** Geographic location of the study area in northwestern Tunisia (red), and **c** the soil sampling sites (blue dots). (I) Pb–Zn Sidi-Driss mineralization, (II) Fe Tamra mineralization, (1, 2, 3) the three dumps of Sidi-Driss mine wastes, and (4) the laundries

groups of results after removing minor contributing variables. The suitability of data reduction through PCA analysis is indicated by Bartlett's sphericity test, Kaiser-Meyer-Olken (KMO) test, eigenvalues, percent variance, and component matrix (before and after rotation).

ANOVA was used to access the variance between means of analyzed parameters and where a significant Fisher (*F*) value was determined.

Prediction models

The RF, ANN, and SVM models were established to predict toxic Zn, Pb, Fe, Mn, Cu, and Cd concentrations in the mine wastes and soil. The total datasets sample was subdivided into 75% training and 25% validation datasets. The response variables such as HM concentrations were sorted in ascending order. First, the three samples having the lowest element concentrations were successively placed in the training set and the next sample in the validation set. Then the procedure was pursued by alternately putting the following three samples in the training set and the next sample in the validation set. This subdivision ensured a similar distribution of HM concentrations in both training and validation sets (Gasmi et al. 2019, 2021). The training dataset was used to model the relationship between the toxic HM concentrations and the predictors that are the mineralogical compositions, physico-chemical properties, such as pH, EC, CEC, sediment grain sizes, SO_4^{2-}, and S contents, and the dataset validation to evaluate the performances of the predictive models.

We note that the RF, SVM, and ANN models are implemented using WEKA software (version 3.9.5, University

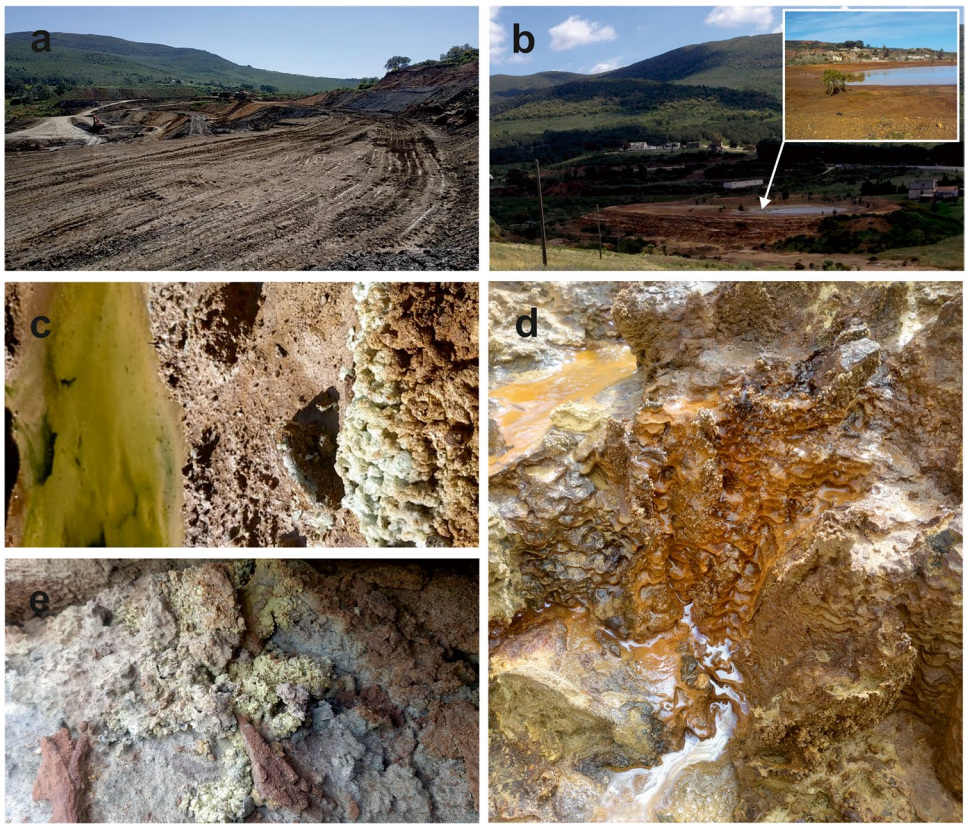

Fig. 2 **a** A panoramic view of the open pit mining activities of Pb–Zn-Sidi-Driss deposits, Northern Tunisia, **b** tailings dump with deep chronic ponds from rainfall in the surface dump, pH = 3.77, T° = 24.8 °C, EC = 1516 µS cm^{-1}, Eh = 182.3 mV draining the Damous river through three drains, **c** shallow transient ponds from rainfall in the side dump (pH = 2.25, T° = 20.6 °C, EC = 28,700 µS cm^{-1}, Eh = 278.5 mV) with precipitation of rozenite, **d** acid water in the side dump (pH = 1.8), **e** yellow-ochre-rust deposits from the side dump showed the presence of jarosite, gypsum (pH = 3, EC = 2271 µS.$cm^{-.1}$, Eh = 230 mV)

of Waikato, Hamilton, NZ), with trees of RF (Breiman 2001), SMOreg (Shevade et al. 1999), and multiple layer perceptron (MLP) (Werbos 1975) packages, respectively. The hyper-parameters of each classifier are optimized using multi-search-weka-package and the RegSMO improved package for the SVM regression, which is the default RegOptimizer.

Random forest

Random forest (RF) regression (Breiman 2001) is a ML model that makes output predictions by combining outcomes from a sequence of regression decision trees sequence. Each tree is constructed independently and depends on a random vector sampled from the input data, with all the trees in the forest having the same distribution (Williams et al. 2020). It includes three parameters named ntree, mtry, and nodesize. Ntree refers to the number of decision trees in the model, mtry the number of variables selected from a decision split for the next split, and nodesize the minimal number of samples allowed in a node (Breiman 2001).

The relative attribute/predictor importance in RF regression models was calculated via mean decrease impurity (MDI, or Gini importance) (Breiman 2002). Each split at a node should increase the homogeneity (purity) of the two descendant nodes resulting in a reduction of the MDI. The average decrease in the impurity for each variable over all the trees gives an accurate picture of the relative importance of the variables in the model. However, one should keep in mind that true attribute importance can be obscured by complex variable interactions (Breiman 2001).

It is widely used for predicting soil attributes and extracting the important predictors relative to linear regression (Zhang et al. 2017; Lagacherie et al. 2019). It can handle either categorical or continuous targets or environmental variables (Zhang et al. 2020). An RF model ultimately produces one single prediction, and the bias and variance values generated are usually low. The model is resistant to overfitting because each tree is trained on a unique bootstrap subsample of the original dataset (Zhang et al. 2020).

Support vector machine

Support vector machine (SVM) regression (Smola and Schölkopf 1998) is a kernel-based learning approach that projects data into a new hyperspace in which complex non-linear relationships can be represented. This approach considers the predictor interactions (Besalatpour et al. 2012). The use of the kernel functions is possible to derive a hyperplane as a decision function for non-linear problems, and then to apply a back-transformation in the non-linear space

(Boser et al. 1992). The regression hyperplane is determined by optimizing the distances from the nearby data points known as support vectors. There are four SVM kernels are linear, polynomial, radial basis function (RBF), and sigmoid (Gasmi et al. 2015).

The RBF is the most popular of kernel types used in SVM methods for classification, and regression problems (Gasmi et al. 2016, 2017). In this study, the RBF-kernel was chosen for the SVM regression process, as it provides a reasonable trade-off between the number of kernel parameters to be optimized and the adaptability and flexibility of non-linear data (Zhang et al. 2020). The SVM RBF-kernel has two hyperparameters associated with the regularization parameter C, and kernel width γ.

Artificial neural networks

Artificial neural networks (ANN) is a computational model (Werbos 1975) that was inspired by networks of biological neurons (Puri et al. 2016). It includes input and output layers that are interconnected. After that, each data was considered as a neuron working that transformed the input data into output values (Were et al. 2015; Zolfaghari et al. 2015). In the input layer, the number of neurons and the predictor number is equal, while the neurons in the hidden layer are determined by the training, and prediction errors calculation. The output layer included a single neuron representing the soil heavy metal concentrations (Zhang et al. 2020). The ANN model calculates the gradient of the case-wise error function concerning the network and minimizes the overall network error (Werbos 1975; Zhang et al. 2020).

Models performance analysis

The prediction performance of the ML models is evaluated using (i) for the training set: the determination coefficient (R^2_{train}), the root means square error ($RMSE_{train}$), the mean absolute error (MAE_{train}), and (ii) for the validation set: the determination coefficient (R^2_{val}), the root mean square error ($RMSE_{val}$), the mean absolute error (MAE_{val}), the ratio performance deviation (RPD), the ratio of the standard deviation RMSEP (Chang et al. 2001), and the ratio performance to interquartile (RPIQ). This last coefficient is the interquartile ratio of the RMSEP (Bellon-Maurel et al. 2010).

Results and discussion

Mineralogical, physiochemical, and geochemical characterizations of mine tailings, and soils

The semi-quantitative evaluation was based on the PXRD peak intensities (Fig. 3a, b). The major and minor minerals are galena (0–2%), marcasite (0–8%), hematite (0–2%), quartz (10–89%), barite (0–26%), phyllosilicates (4–49%), jarosite (0–16%), bassanite (0–49%), anglesite (0–13%), gypsum (0–62%), and goethite (0–46%) (Table 1). The pH, EC, and CEC ranges of mine wastes were 1.8–6.9, 180–17,500 $\mu S \cdot cm^{-1}$, and 4–33 meq 100 g^{-1}, respectively. The mean sand, silt, and clay percentages were 57%, 23%, and 20%, respectively (Table 1).

The total field samples dataset showed that the mean Zn, Pb, Fe, Mn, Cu, and Cd concentrations were 103.5 $g \cdot kg^{-1}$, 5.2 $g \cdot kg^{-1}$, 103.5 $g \cdot kg^{-1}$, 11.7 $g \cdot kg^{-1}$, 0.058 $g \cdot kg^{-1}$, and 0.008 $g \cdot kg^{-1}$, respectively (Table 2). The CV variation of soil HM contents that ranged from 41.3 to 95.5% was considered moderate. However, the CV of Cd concentration was high, (around 101.2%).

The effect of the AMD environment on soil HM dynamics

The (i) long exposure period of mine wastes (1922–2021), (ii) humid to subhumid climate, (iii) position of the Damous river to the ore (60 m) and the edge of the tailings dump (≈ 3 m), (iv) advanced oxidation degree of sulfide minerals such as galena (35%) and sphalerite (25%) (Stefanov and Ouchev 1972), (v) wet wastes rich in sphalerite and -marcasite, and (vi) low neutralization power (6% of carbonates) are the origin of the destabilization conditions that degraded over time the sulfides (Dermech 1990; Trifi et al. 2018). They also favored partial or the total alteration and dissolution of initial material, such as the host rock "Yellow Balls" (YB), pyrite and sphalerite-marcasite mixture, the iron oxides precipitation (mostly goethite), and the neoformation of basanite (Fig. 4), jarosite, gypsum, and anglesite (Fig. 3a, b). These formed minerals, as a consequence of AMD witnesses (Trifi et al. 2018) controlled the HM mobility or retention (HM leaching).

Indeed, sulfides were oxidized and decomposed and Fe oxides were solubilized. These processes liberated SO_4^{2-}, Fe^{2+}, Fe^{3+}, and H_3O^+. The acidic environment increase promoted the precipitation of the secondary iron oxide-hydroxide (goethite) (Figs. 2e and 4) (Aubertin et al. 2002; Dold 2014; Trifi et al. 2018 and 2019). The strong increase of the sulfated ion quantities (saturated environment) initiated the gypsum, bassanite, and jarosite precipitation. These typical minerals of AMD (Dold and Fontbote 2001; Hudson-Edwards and Wright 2011; Carbone et al. 2013; Murray et al. 2014) were detected only by PXRD (Eq. 1 and Fig. 3a, b). By precipitating, these last neo-formed minerals had trapped the available Pb, Zn, Fe, Cd, Cu, and SO_4^{2-} in these tailings in dry periods and released them by dissolution in surrounding areas in rainy periods (Jambor et al. 2000; Gieré et al. 2003; Hammarstrom et al. 2005; Carbone et al. 2013; Murray et al. 2014; Trifi et al. 2018 and 2019).

$$3Fe^{3+} + M + 2(RO_4)^{2-} + 12H_2O \rightarrow MFe_3(RO_4)_2(OH)_6(\text{jarosite}) + 6H_3O^+ \quad (1)$$

where M are monovalent or divalent cations (Na, K, Ag, Tl, NH_4, H_3O, Pb) and RO_4 are tetrahedral anions (SO_4), (PO_4), or (AsO_4) (Dutrizac and Jambor 2000; Murray et al. 2014).

Statistic results

In this statistical study, the Pearson correlation test, PCA, and ANOVA were adopted.

Independent variables and HMs in pairs

By statistical correlation coefficients, the relationship among mineralogical compositions, physico-chemical properties, S and SO_4^{2-} contents, and grain size as independent variables (Table 3) and between HMs in pairs was assessed (Table 4).

Hight correlation coefficients were found between barite/jarosite (0.70), barite/marcasite (0.75), S/marcasite (0.76), S/galena (0.67), S/barite (0.76), SO_4^{2-}/gypsum (0.80), SO_4^{2-}/jarosite (0.60), and SO_4^{2-}/quartz (−0.76). However, the gypsum/goethite, jarosite/marcasite, jarosite/galena, SO_4^{2-}/goethite, SO_4^{2-}/bassanite, and S/jarosite had moderate correlation coefficients values (between 0.43 and 0.59) (Table 3).

The positive correlation coefficients among Zn/Mn, Zn/Fe, Mn/Cu, and Zn/Cu were 0.72, 0.67, 0.64, and 0.61, respectively, were also high. However, the Fe/Mn and Fe/Cu (r=0.60) correlations were positives and moderates (Table 4).

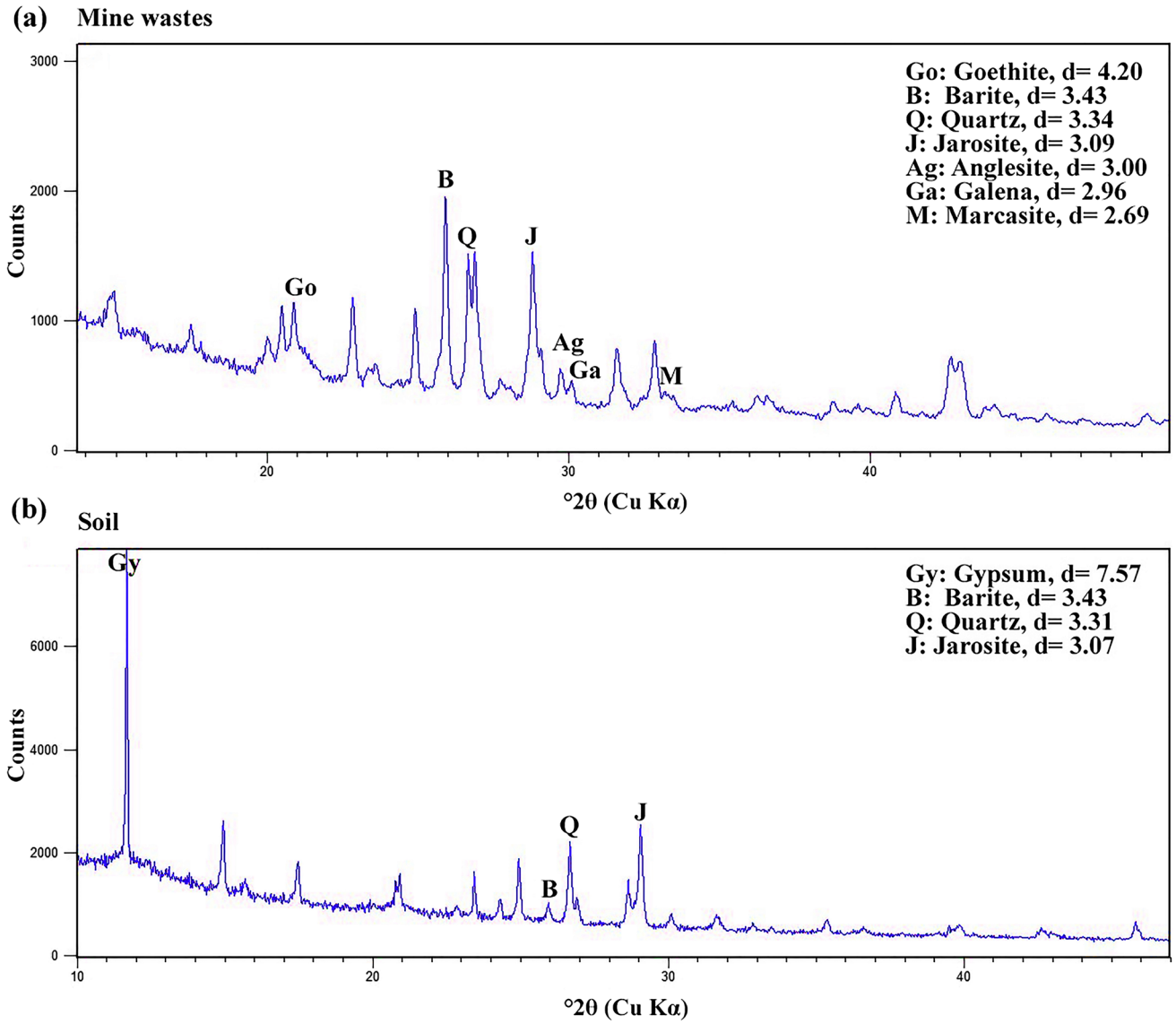

Fig. 3 XRD results of representative **a** mine wastes and **b** soil sample examples

Table 1 Summary statistics of mineralogical compositions, physico-chemical, chemical, and grain size properties of field samples

	Min	Q1	Mean	Q3	Max	SD	IQ	CV	Sk	K
Gy	0.000	0.000	8.634	9.343	61.860	16.111	9.343	186.587	2.087	3.384
Phy	4.000	9.000	15.871	20.250	49.000	9.673	11.250	60.949	1.492	2.283
Q	9.540	31.730	41.176	50.928	88.530	17.689	19.198	42.959	.449	0.528
B	0.000	0.593	4.531	6.525	25.640	4.981	5.933	109.916	1.862	4.915
J	0.000	0.703	2.877	4.140	15.800	2.911	3.438	101.186	1.843	5.282
Ga	0.000	0.000	0.447	0.835	2.350	0.668	0.835	149.523	1.372	0.680
M	0.000	0.000	1.592	2.653	8.450	1.742	2.653	109.422	1.450	3.178
Go	0.000	8.308	22.897	32.965	45.820	13.983	24.658	61.071	− .624	− 1.048
Bas	0.000	0.000	0.913	0.000	49.080	5.923	0.000	648.562	7.850	64.004
Ag	0.000	0.000	0.754	0.000	13.090	2.507	0.000	332.459	3.549	12.518
H	0.000	0.000	0.024	0.000	1.760	0.207	0.000	848.528	8.485	72.000
S	0.000	0.000	2.039	3.113	9.570	2.123	3.113	104.121	1.163	1.591
SO_4^{2-}	0.000	2.743	9.946	11.555	41.630	11.122	8.813	111.828	1.559	1.497
pH	1.800	3.295	4.017	4.585	6.900	1.183	1.290	29.440	.799	0.241
CE	180.000	811.000	1953.979	2372.500	17,500.000	2369.151	1561.500	121.248	4.573	26.996
CEC	4.000	12.000	17.199	21.330	33.330	6.541	9.330	38.030	.297	− 0.315
Sa	2.090	27.000	43.190	57.000	91.000	21.544	30.000	49.882	.029	− 0.394
Si	1.000	29.000	40.919	56.000	86.000	20.600	27.000	50.343	.067	− 0.588
C	0.000	9.000	15.887	20.248	49.000	9.824	11.248	61.839	1.334	2.045

Gy gypsum, *Phy* phyllosilicates, *Q* quartz, *B* barite, *J* jarosite, *Ga* galena, *M* marcasite, *Go* goethite, *Bas* bassanite, *Ag* anglesite, *H* hematite (in %), *S* sulfur and SO_4^{2-} sulfate ions (in g kg^{-1}), *pH* hydrogen potential, *EC* electric conductivity (in µS cm^{-1}), *CEC* cation exchange capacity (in meq 100 g^{-1}), *Sa* sand, *Si* silt, *C* clay (in %)

Table 2 Summary statistics of HM contents in field samples

	Min	Q1	Mean	Q3	Max	SD	IQ	CV	SK	K
Zn	0.380	2.905	11.722	19.200	33.950	9.508	16.295	81.114	0.501	− 0.928
Pb	1.750	4.545	13.943	22.275	59.200	11.862	17.730	85.075	1.360	2.109
Fe	2.600	39.500	103.469	158.000	226.000	61.478	118.500	59.417	0.015	− 1.308
Mn	0.074	1.105	5.189	8.263	20.040	4.953	7.158	95.454	1.204	0.997
Cu	0.011	0.035	0.058	0.075	0.102	0.024	0.041	41.267	− 0.196	− 1.065
Cd	0.000	0.004	0.008	0.009	0.051	0.008	0.005	101.219	2.768	9.875

Q1 first quartile, *Q3* third quartile, *SD* standard deviation, *IQ* interquartile, *CV* coefficient of variation (%), *Sk* skewness, *Zn*, *Pb*, *Fe*, *Mn*, *Cu*, and *Cd* HM contents (in g.kg^{-1})

HMs and independent variables

By combining all Pearson correlation analysis results, we deduced that the (Table 5):

- Zinc was significantly ($p < 0.05$) correlated with gypsum, barite, galena, marcasite, goethite, S, and clay texture (correlation coefficients range 0.25–0.59 and $p < 0.05$). These good relationships were mainly due to the total dissolution of mixed sphalerite-marcasite and sphalerite concentrates (Trifi et al. 2018),
- Lead concentration was significantly ($p < 0.05$) correlated with the jarosite, galena, marcasite, S, SO_4^{2-}, EC, quartz, barite, and anglesite. The *R* ranges between 0.35 and 0.64 showed that mineral dissolution favors the Pb mobility. The moderate correlation between pH/Pb indicated that the acidic pH also enhanced the Pb mobility. The low galena and marcasite concentrations in the tailings evidenced the advanced alteration of the initial materials (Fig. 3a, b),
- Iron was significantly ($p < 0.05$) correlated with quartz, galena, marcasite, goethite, S, EC, CEC, silt, and clay. These correlations confirmed the decomposition and oxidation of the sulfides and the precipitation of secondary goethite,
- Manganese was significantly ($p < 0.05$) correlated with gypsum, marcasite, goethite, S, SO_4^{2-}, pH, CEC, silt,

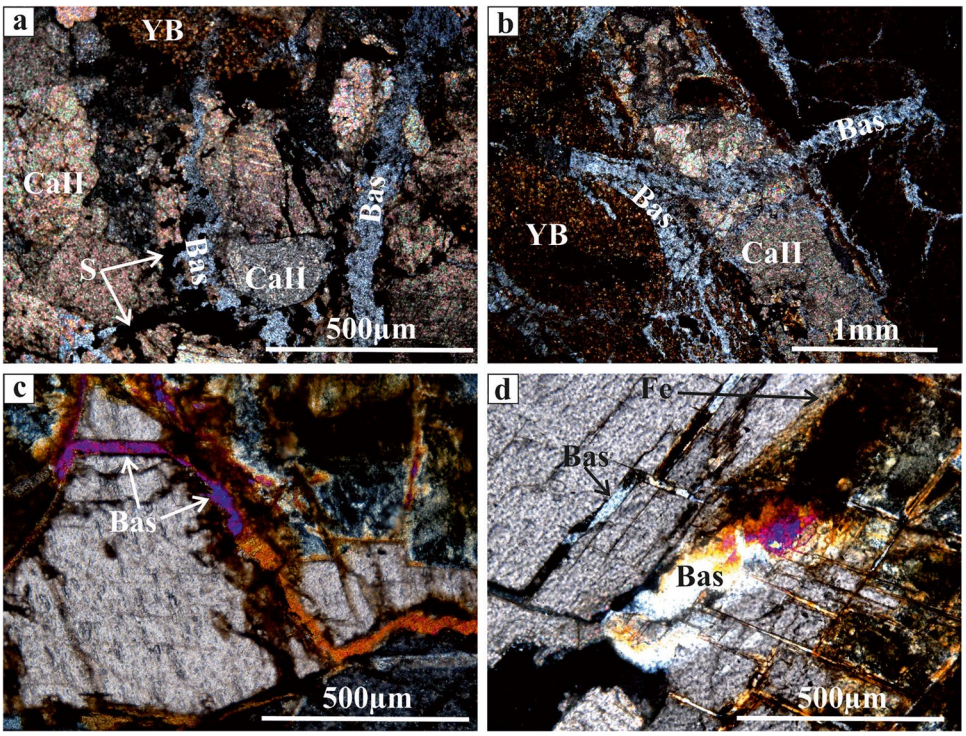

Fig. 4 Polarized cross light photomicrographs showing in **a**, **b** the alteration forms of initial materials such as yellow balls (YB), calcite II (Ca-II), and sulfides (S), and precipitation of neoformed bassanite sulfate in microfractures, **c**, **d** the oxidation of sulfides (S) and precipitation of iron oxide-hydroxide (Fe), and bassanite (Bas) in microfractures and cleavages

and clay. These correlation levels confirmed the previous results where found that Mn was a minor element in siderite, calcite, and marcasite and had higher concentrations in the Tamra mine breccias (Trifi et al. 2018),

- Copper was significantly ($p < 0.05$) correlated with the marcasite, goethite, bassanite, calcite, S, EC, CEC, sand, and silt,
- Cadmium was significantly ($p < 0.05$) correlated with the bassanite, and EC. In the initial materials, the Cd was often present in the sphalerite-marcasite mixed concentrate (Stefanov and Ouchev 1972), and in sphalerite and galena (Dermech 1990; Trifi et al. 2018) that were not detected in the friable fraction of actual tailings (Fig. 3a, b), suggesting its mobility.

Previous studies revealed significant correlations, among the minerals and their majors (e.g., Zn, Pb, Fe) and trace (e.g., Mn, Cu, Cd, As) elements (Stefanov and Ouchev 1972; Dermech 1990, Trifi et al. 2018), the EC, CEC, pH and sediment texture of mine wastes that confirmed once again that mineralogical composition, dissolution, and oxidation processes induced an essential variation of their physicochemical characteristics and closely controlled the Pb, Zn, Mn, Fe, Cu, and Cd mobility levels (Jambor et al. 2000; Hammarstrom et al. 2005; Dold 2006; Carbone et al. 2013; Bouzahzah et al. 2014; Murray et al. 2014; Balci and Demirel 2018; Trifi et al. 2018 and 2019). Therefore, Salonen and Korkka-Niemi (2007) and Hu and Cheng (2013) also reported that soil Cr and Ni are majorly sourced from parent materials. Moreover, Garcıa-Sanchez et al. (1999) reported that Fe-rich soils containing high levels of clay minerals can adsorb considerable amounts of As. Zhang et al. (2020) showed a strong correlation between As, Ni, and Cr were significantly ($p < 0.05$) correlated with the Fe-Al-oxides and quartz. Gholami et al. (2011) showed a strong relationships among pH, SO_4, HCO_3, TDS, EC, Mg, Ca, Ni, and Fe concentrations. Also, Rooki et al. (2011) revealed high values of the correlation coefficients between heavy metals and pH, SO_4, and Mg^{2+} concentrations.

To reduce and transform the dimensions of normalized variables into new principal components (PC) (Fig. 5), the PCA with varimax rotation was applied. Therefore, data reduction through PCA analysis was achieved after removing minor contributing variables (such as sand, silt, and quartz) and items that have representation qualities lower than 0.5. In this study, the suitability of the PCA is indicated by Bartlett's sphericity test (0.000), Kaiser-Meyer-Olken (KMO = 0.763) test, eigenvalues, percent variance, and component matrix (after rotation).

In this study, the reason is to understand the interrelationship among mineralogical compositions, grain size, and HM predictors in the soil, and the source identification of the pollutants. Three correlated components were extracted by PCA witch a total explained variance was 75% (Table 6). Factor 1 seems to be an ax of mineralogy and, sulfur and sulfates contents that revealed strong relationships among jarosite, galena, S, and SO_4 contents. Factor 2 showed a close correlation among pollutants

Table 3 Pearson correlation coefficients between the different independent variables (mineralogical compositions, physico-chemical, chemical, and grain size properties of field samples)

	Gy	Phy	Q	B	J	Ga	M	Go	Bas	Ag	H	S	SO_4^{2-}	pH	CE	CEC	Sa	Si	C
Gy	1																		
Phy	0.171	1																	
Q	**−0.575****	**−0.250***	1																
B	−0.081	−0.210	**−0.407****	1															
J	0.213	0.064	**−0.610****	**0.709****	1														
Ga	0.054	−0.055	**−0.357****	**0.461****	**0.495****	1													
M	−0.172	**−0.240***	**−0.301***	**0.756****	**0.535****	**0.440****	1												
Go	**−0.537****	**−0.481****	0.040	0.091	−0.204	0.088	0.211	1											
Bas	−0.006	0.073	**−0.258***	0.041	**0.242***	−0.054	−0.042	**−0.256***	1										
Ag	−0.043	−0.116	**−0.251***	**0.239***	**0.265***	0.015	**0.242***	−0.040	**0.340****	1									
H	**0.395****	−0.110	−0.168	−0.031	−0.035	−0.010	−0.061	−0.139	−0.018	−0.036	1								
S	−0.124	−0.214	**−0.359****	**0.766****	**0.595****	**0.676****	**0.959****	0.200	−0.052	0.204	−0.053	1							
SO_4^{2-}	**0.800****	0.100	**−0.769****	**0.350****	**0.606****	0.220	0.153	**−0.527****	**0.433****	**0.265***	**0.297***	0.195	1						
pH	**−0.375****	−0.029	**0.506****	**−0.373****	**−0.411****	−0.170	−0.206	0.217	**−0.269***	**−0.272***	−0.184	−0.223	**−0.583****	1					
CE	**0.301***	0.111	**−0.507****	**0.257***	**0.487****	0.140	0.068	**−0.393****	**0.825****	**0.308****	0.182	0.100	**0.704****	**−0.458****	1				
CEC	0.040	−0.144	−0.016	−0.047	−0.126	0.015	0.142	0.190	−0.218	0.012	−0.071	0.121	−0.081	0.141	**−0.246***	1			
Sa	−0.176	**−0.291***	**0.318****	−0.039	−0.175	−0.182	0.003	0.039	−0.061	0.168	−0.034	−0.054	−0.181	0.165	−0.176	**−0.235***	1		
Si	0.100	−0.152	−0.230	0.174	0.189	0.230	0.131	0.175	0.030	−0.125	0.087	0.179	0.154	−0.182	0.139	**0.303****	**−0.892****	1	
C	0.179	**0.956****	−0.215	**−0.279***	−0.012	−0.081	**−0.281***	**−0.455****	0.069	−0.107	−0.110	**−0.256***	0.075	0.020	0.091	−0.119	**−0.321****	−0.142	1

Gy gypsum, *Phy* phyllosilicates, *Q* quartz, *B* barite, *J* jarosite, *Ga* galena, *M* marcasite, *Go* goethite, *Baq* bassanite, *Ag* anglesite, *H* hematite (in %), *S* sulfur and SO_4^{2-} sulfate ions (in $g.kg^{-1}$). *pH* hydrogen potential, *EC* electric conductivity (in µS cm^{-1}), *CEC* cation exchange capacity (in meq 100 g^{-1}), *Sa* sand, *Si* silt, *C* clay (in %), $n = 72$

*Significant at 5% probability

**Significant at 1% probability

Table 4 Pearson correlation coefficients between different HM contents of field samples

	Zn	Pb	Fe	Mn	Cu	Cd
Zn	1					
Pb	**0.479****	1				
Fe	**0.673****	**0.257***	1			
Mn	**0.726****	0.204	**0.609****	1		
Cu	**0.617****	0.144	**0.606****	**0.641****	1	
Cd	0.151	0.224	0.029	**0.295***	0.095	1

$n = 72$

*Significant at 5% probability

**Significant at 1% probability

as Zn, Mn, Fe, and Cu. Factor 3 correspond to clay fraction. This PCA method was widely tested in previous studies (e.g., Bhuiyan et al. 2011; Ogwueleka 2014; Rahman et al. 2014; Chowdhury and Maiti 2016).

ANOVA was used to access the variance between means of analyzed pollutants and here significant Fisher (F) value was observed. ANOVA revealed that Zn, Mn, Fe, Pb, and Cu and associated group of mineralogical compositions, S and SO_4 contents, and clay fraction showed a significant difference in mean ($p < 0.001$) except Cu ($p < 0.05$), and Cd ($p > 0.05$) (Table 7). This is showed that mineralogical composition plays an important role in defining metal pollution in soil. This ANOVA test was popularly applied in previous studies (e.g., Chowdhury and Maiti 2016; Chowdhury et al. 2021; Zarei et al. 2014).

Therefore, the founded strong correlations among the minerals, physico-chemical properties, and soil grain sizes and each soil toxic HMs were used as important predictors and inputs for the ML-predicted models.

Performance of HM prediction models based on all and important variables

To evaluate the prediction model performance first only the highly correlated variables (on important independent variables) and second, all independent variables were considered.

Table 5 Pearson correlation coefficients between different independent variables (mineralogical compositions, physico-chemical properties, and soil grain size) and soil HMs

	Zn	Pb	Fe	Mn	Cu	Cd
Gy	−0.263	−0.070	−0.16	−0.316	−0.162	0.024
Phy	−0.252	−0.166	−0.22	−0.277	−0.001	−0.043
Q	−0.202	**−.473****	−0.308	−0.076	−0.221	−0.165
B	**0.344****	**0.779****	**0.403****	0.150	0.096	0.085
J	0.154	**0.644****	0.198	−0.066	0.014	0.127
Ga	**0.325****	**0.489****	**0.308****	0.156	0.114	0.173
M	**0.442****	**0.609****	**0.473****	**0.286***	**0.277***	0.142
Go	**0.588****	0.158	**0.558****	**0.619****	**0.523****	0.003
Bas	−0.179	0.189	−0.201	−0.132	−0.242	**0.269***
Ag	0.073	**0.354****	0.074	0.104	−0.034	0.083
H	−0.052	−0.089	−0.071	−0.121	−0.172	−0.020
S	**0.465****	**0.654****	**0.485****	**0.284***	**0.263***	0.171
SO4	−0.153	**0.348****	−0.055	−0.263	−0.195	0.164
pH	0.067	−0.410	−0.124	**0.242***	0.055	0.033
CE	−0.218	**0.355****	−0.235	−0.223	−0.305	**0.269***
CEC	0.203	−0.043	**0.269***	**0.304****	**0.373****	0.125
Sa	−0.065	−0.001	−0.211	−0.140	−0.434	−0.082
Si	0.196	0.100	**0.342****	**0.266***	**0.444****	0.098
C	−0.269	−0.205	−0.254	−0.252	0.019	−0.027

Gy gypsum, *Phy* phyllosilicates, *Q* quartz, *B* barite, *J* jarosite, *Ga* galena, *M* marcasite, *Go* goethite, *Bas* bassanite, *Ag* anglesite, *H* hematite (in %), *S* sulfur and SO_4^{2-} sulfate ions (in g kg^{-1}), *pH* hydrogen potential, *CE* electric conductivity (in µS cm^{-1}), *CEC* cation exchange capacity (in meq 100 g^{-1}). Sand, silt, and clay (in %). *Zn, Pb, Fe, Mn, Cu, Cd* heavy metal contents (in g. kg^{-1}), $n = 72$

*Significant at 5% probability

**Significant at 1% probability

Based on important independent variables (imp) in predicted soil HMs, the RF_{imp}, SVM_{imp}, and ANN_{imp} were considerably different (Table 8). The important independent variables were extracted from the Pearson correlation analysis (Table 5) where (i) eight to estimate the Zn contents, (ii) nine to estimate the Pb, Fe, and Cu contents, and (iii) ten to estimate the Mn and Cd concentrations. The obtained R^2_{val} of the different toxic HMs using ANN_{imp}, and SVM_{imp} models were less or equal to 0.38. Hence, the ANN_{imp} and SVM_{imp} models generated poorer accuracies in predicting the toxic HM content compared with RF_{imp} models as they yielded lower R^2_{val}, RPD, and RPIQ and higher error indices (MAE_{val} and $RMSE_{val}$).

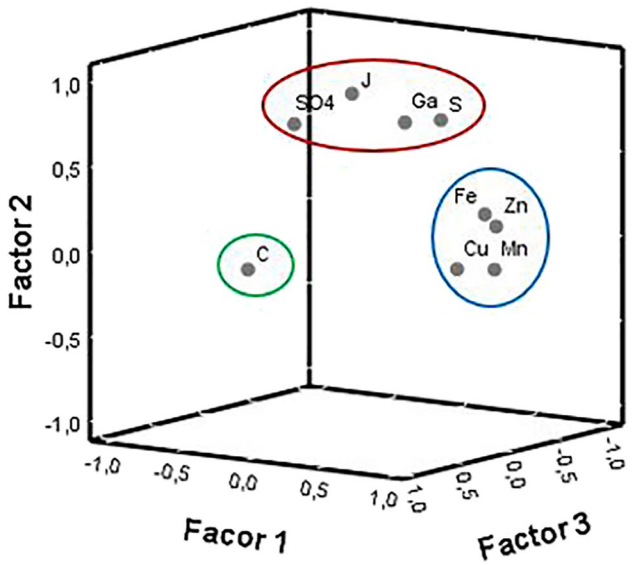

Fig. 5 Principal component analysis (PCA) with varimax rotation of all variables. J, jarosite; Ga, galena; M, marcasite, S, sulfur; SO_4^{2-} sulfate ions, and C, clay

For the soil Pb, Cu, Mn, and Fe the R^2_{val} of RF ranged between 0.67 (Pb) and 0.53 (Fe) (Table 8). Thus, the RF_{imp} models achieved acceptable success, indicating the ability of RF in predicting the HM dynamics. For the other soil HM contents, the RF less well predicted the toxic metal dynamism as the R^2_{val} of Zn, and Cd were 0.44 and − 0.15, respectively.

Owing to the complexity of the field settings, it is necessary to consider all variables (all) used to test soil HM content in the AMD environment.

The ANN_{all} and SVM_{all} models showed poorer accuracies in predicting the toxic HM contents, compared with RF_{all} models (Table 9). Thus, based on all or important independent predictors, RF outperforms SVM and ANN and has the ability to satisfactory predict the Zn, Pb, Fe, Mn, Cu, and Cd contents of this AMD environment. This outcome is similar to previous studies which showed that RF algorithm performs well for the HM prediction (e.g., Schwarz et al. 2013; Qiu et al. 2016; Zhang et al. 2020). Added to that, Schwarz et al. (2013) estimated the soil Pb contamination in Baltimore (Maryland, USA), and obtained an overall accuracy of 72% for the RF model. Qiu et al. (2016) showed that Cd concentration in Fuyang soil (eastern China) deduced by predictive models and exhibited acceptable overall accuracies 72.2% for stepwise linear regression (SLR), 70.4% for classification and regression tree (CART), and 75.9% for random forest (RF). The results performed by RF model were closest to the observed values within the SLR and CART models. The good performance of the RF model was attributable to its ability to handle the non-linear and hierarchical relationships between Cd and environmental variables (Qiu et al. 2016). In the same case, Hong et al. (2019) estimated Pb and Zn concentrations in peri-urban agricultural soils through reflectance spectroscopy and concluded that the accuracy of RF was generally better than that of PLSR. According to Tan et al. (2019), the RF can make full use of

Table 6 Total variance explained and extracted components of PCA. Extraction method: principal component analysis

Total variance explained									
Component	Initial eigen value			Extraction sums of squared loadings			Rotation sums of squared loading		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total	% de la variance	Cumulative %
1	3569	39,657	39,657	3569	39,657	39,657	3172	35,246	35,246
2	2214	24,603	64,260	2214	24,603	64,260	2455	27,280	62,526
3	1031	11,456	75,716	1031	11,456	75,716	1187	13,189	75,716
4	,719	7984	83,700						
5	,387	4296	87,996						
6	,339	3770	91,766						
7	,296	3288	95,053						
8	,240	2667	97,720						
9	,205	2280	100						

the input spectral data from a hyperspectral sensor in the estimation of four kinds of HMs, such as Zn ($R^2 = 0.90$), Cr ($R^2 = 0.91$), As ($R^2 = 0.99$), and Pb ($R^2 = 0.97$). However, Zhang et al. (2020) suggest that SVM and RF perform much better than ANN in explaining HM spatial variations in urban soil. Wang et al. (2020) explored the application of RF models in predicting the soil concentration and HM spatial distribution (Pb, Cd, Cr, As, Hg, and Zn) compared with land use regression (LUR) models and demonstrates that the RF models performed better and could be captured effectively the complex non-linear relationships. Four mathematical models, namely partial least squares regression (PLSR), adaptive neural fuzzy inference system (ANFIS), RF, and generalized regression neural network (GRNN), were used by Xu et al. (2021) to estimate the concentration of Hg, Cr, and Cu in agricultural soils based on spectroscopy data, and obtain an acceptable accuracy. The RF models could accurately predict the HM removal efficiency via chitosan-based flocculants ($R^2 = 0.93$) according to flocculant properties, flocculation conditions, and HM properties (Lu et al. 2022). The difference in the predictive accuracy between these previous studies may be linked to the extent of the study area, sampling density, different target variables, or the quality and quantity of the auxiliary data.

For Zn, Fe, and Cu, the applied RF_{all} models even perform better than RF_{imp} as they produced higher R^2_{val}, RPD, and RPIQ, and lower MAE_{val} and $RMSE_{val}$ (Tables 8 and 9). On one hand, the RF_{all} and RF_{imp} models yielded similar results for soil Mn prediction with an R^2_{val}, RPD, and RPIQ of 0.64, 1.72, and 2.15, respectively. On the other hand, the RF_{imp} outperforms RF_{all} in the prediction of soil Pb as it produced higher R^2_{val}, RPD, and RPIQ and lower error indices (MAE, $RMSE_{val}$) (Tables 8 and 9). Similarly, Zhang et al. (2020) also showed that the RF models based on important independent variables yielded satisfactory results in the HM prediction, although the R^2 and Nash–Sutcliffe model efficiency (NSE) of several HMs were lower than those of RF models based on all independent variables. For As and Cr,

Table 7 Analysis of variance (ANOVA) between HM concentrations, mineralogical compositions, S and SO_4 contents, and clay fraction in soils

	Zn	Pb	Fe	Mn	Cu	Cd
F value	5.570	15.490	4.928	3.640	2.394	0.786
p value	0.000	0.000	0.001	0.006	0.047	0.563

Table 8 Performance indices of the RF, SVM, and ANN models-predicted HM soils and mine wastes based on important predictors

	Model	R^2_{train}	$RMSE_{train}$	MAE_{train}	R^2_{val}	$RMSE_{val}$	MAE_{val}	RPD	RPIQ
Zn	RF	0.940	2.291	1.701	0.444	7.264	5.556	1.380	2.104
	SVM	0.254	8.056	6.538	0.187	8.785	6.837	1.141	1.740
	ANN	−0.669	12.054	9.066	−0.397	11.520	9.142	0.871	1.327
Pb	RF	0.928	2.897	2.017	0.671	8.103	3.642	1.794	2.285
	SVM	0.279	9.195	6.640	0.286	11.939	7.057	1.218	1.551
	ANN	−0.861	14.777	10.283	−0.741	18.647	12.269	0.780	0.993
Fe	RF	0.942	14.485	10.477	0.534	42.757	29.375	1.507	2.581
	SVM	0.394	47.037	40.045	0.312	51.945	43.969	1.241	2.124
	ANN	−0.059	62.145	54.913	−0.124	66.393	57.353	0.971	1.662
Mn	RF	0.926	1.310	1.026	0.642	3.108	2.793	1.721	2.158
	SVM	0.177	4.370	3.023	0.099	4.933	3.399	1.084	1.360
	ANN	−0.685	6.252	4.299	−0.734	6.844	4.697	0.781	0.980
Cu	RF	0.927	0.006	0.004	0.658	0.013	0.010	1.761	2.765
	SVM	0.281	0.020	0.016	0.386	0.018	0.016	1.313	2.062
	ANN	−0.081	0.024	0.021	−0.158	0.025	0.022	0.956	1.502
Cd	RF	0.839	0.002	0.001	−0.157	0.026	0.010	0.957	0.182
	SVM	−0.050	0.006	0.004	−0.092	0.026	0.009	0.985	0.188
	ANN	−0.008	0.006	0.004	-0.059	0.025	0.009	1.000	0.190

RF random forest, *ANN* artificial neural networks, *SVM* support vector machines, R^2_{train} coefficient of determination of training set, $RMSE_{train}$ root mean square error of training set (in g kg^{-1}), MAE_{train} mean absolute error of training set, R^2_{val} coefficient of determination of validation set, $RMSE_{val}$ root mean square error of validation set (in g. kg^{-1}), MAE_{val} mean absolute error of validation set, *RPD* the ratio of the performance to the deviation on validation set, *RPIQ* the ratio of performance to interquartile on validation set, *Zn, Pb, Fe, Mn, Cu, Cd* heavy metal contents (in g kg^{-1})

the RF models based on important independent variables models, even perform better than RF models based on all independent variables as their R^2 values were high and error indices were low (Zhang et al., 2020).

Implication of the soil HM prediction

First, PCA and Pearson correlation analyses were used to extract the independent parameters. Despite the high correlation coefficients and significances (Fig. 5 and Table 3) among the considered parameters, the impact of the interactions within the tailings material remained very complex. The obtained results could be improved, and the predictions strengthened if all the auxiliary mineralogical compositions, physico-chemical parameters, texture, and HM concentrations are introduced in the models.

The RF_{all} models produced robust predictions of HMs. The R^2 values of the RF_{all} model based on all variables were generally higher than those of the RF_{imp}, ANN_{imp}, and SVM_{imp} models based on important independent variables. In addition, the RF_{all} model performance of soil Zn, Pb, Fe, Mn, Cd, and Cu was improved by increasing the number of cross-correlated predictors, despite the RF model having greater power in predicting pollutant concentrations, compared with other models. Within this RF model, it is necessary to identify the parameters that strengthened this predictive power. Based on the mean decrease variance of the different parameters, the calculated attribute/predictor showed the importance of each soil HM predictor (Fig. 6).

The soil Zn, Fe, and Cu concentrations and mineralogical compositions (goethite, barite, and S) were identified with the highest importance as a factor group. In addition, the marcasite was an important factor for Zn. Jarosite was an important factor for Fe and Cu. The impacts of other factors on Zn, Fe, and Cu concentrations were moderate or weak. The Pb and S were ranked as the most important variables, followed by marcasite, bassanite, and anglesite. The importance of other variables was moderate or negligible. The most important factor of Mn was goethite and the moderately important factors including the pH, S, CEC, and silt. The bassanite, anglesite, pH, and goethite were the top-ranked factors of Cd.

The adopted classical analytical methods for the HM quantification by using harsh chemical use train re-precipitation of metals and destroys the physico-chemical properties of soils. However, spectroscopy, remote sensing, X-ray fluorescence (XRF), and X-ray diffraction (XRD) were used as a non-destructive analytical methods. Hence, in this study, coupling of ML models and mineralogical data as an output of on non-destructive method (XRD) solves problems of mentioned quantification analytical techniques and HM prediction with low costs and time consumption. In this study,

Table 9 Performance indices of the RF, SVM and ANN models-predicted HMs in soils and mine wastes based on all predictors

	Model	R^2_{train}	$RMSE_{train}$	MAE_{train}	R^2_{val}	$RMSE_{val}$	MAE_{val}	RPD	RPIQ
Zn	RF	0.946	2.177	1.714	0.597	6.185	5.110	1.621	2.471
	SVM	0.240	8.136	6.148	0.171	8.874	6.700	1.130	1.722
	ANN	−0.039	9.510	8.076	−0.072	10.091	8.368	0.994	1.515
Pb	RF	0.924	2.989	2.117	0.625	8.652	4.494	1.681	2.140
	SVM	0.330	8.863	6.178	0.206	12.591	7.645	1.155	1.470
	ANN	−0.861	14.777	10.283	−0.741	18.647	12.269	0.780	0.993
Fe	RF	0.945	14.193	10.426	0.575	40.810	28.831	1.579	2.704
	SVM	0.404	46.639	39.190	0.306	52.178	43.828	1.235	2.115
	ANN	−0.026	61.184	54.187	−0.001	62.678	54.419	1.028	1.761
Mn	RF	0.924	1.325	1.005	0.643	3.106	2.560	1.722	2.159
	SVM	0.213	4.274	2.950	0.146	4.804	3.285	1.113	1.396
	ANN	−0.241	5.366	3.803	−0.297	5.919	4.128	0.904	1.133
Cu	RF	0.942	0.005	0.004	0.752	0.011	0.009	2.068	3.247
	SVM	0.309	0.019	0.015	0.391	0.018	0.016	1.318	2.071
	ANN	−0.004	0.023	0.020	−0.002	0.023	0.020	1.028	1.614
Cd	RF	0.878	0.002	0.001	−0.052	0.011	0.006	1.003	0.421
	SVM	−0.003	0.006	0.004	−0.181	0.012	0.005	0.947	0.397
	ANN	−0.839	0.009	0.006	−0.519	0.013	0.008	0.835	0.350

RF random forest, *ANN* artificial neural networks, *SVM* support vector machines, R^2_{train} coefficient of determination of training set, $RMSE_{train}$ root mean square error of training set (in g. kg^{-1}), MAE_{train} mean absolute error of training set, R^2_{val} coefficient of determination of validation set, $RMSE_{val}$ root mean square error of validation set (in g. kg^{-1}), MAE_{val} mean absolute error of validation set, *RPD* the ratio of the performance to the deviation on validation set, *RPIQ* the ratio of performance to interquartile on validation set, *Zn*, *Pb*, *Fe*, *Mn*, *Cu*, *Cd* heavy metal contents (in g. kg^{-1})

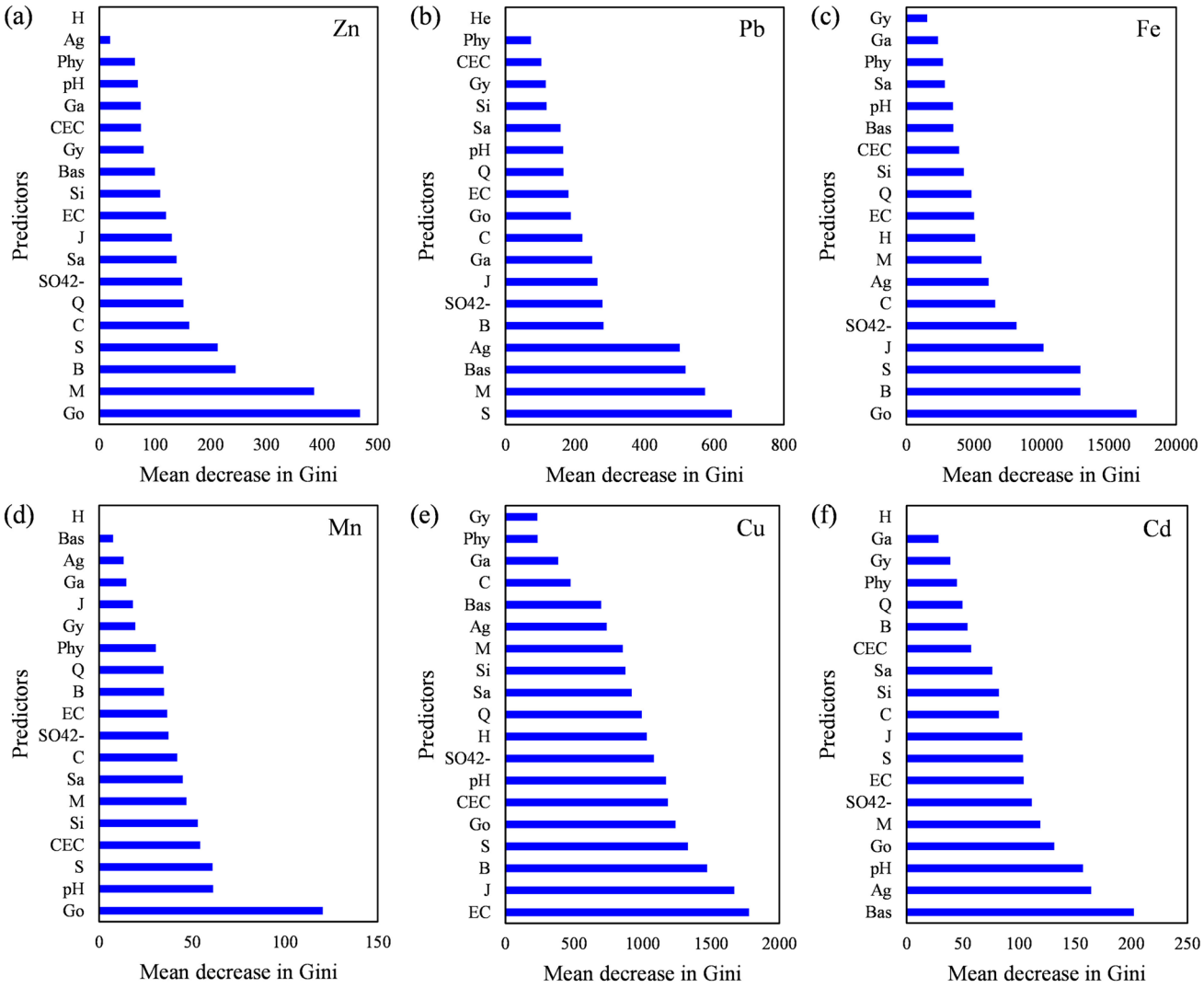

Fig. 6 Relative attribute importance ranked by the Random Forest (RF) regression models obtained via mean decrease impurity (MDI) and number of nodes using that attribute to predict soil heavy metal contents **a** Zn, **b** Pb, **c** Fe, **d** Mn, **e** Cu, and) Cd. The mean decrease in Gini measure was an average of 100 runs of RF. Gy, gypsum; Phy, phyllosilicates; Q, quartz; B, barite; J, jarosite; Ga, galena; M, marcasite; Go, goethite; Bas, bassanite; Ag, anglesite; H, hematite; S, sulfur; SO_4^{2-}, sulfate ions; pH, hydrogen potential; EC, electric conductivity; CEC, cation exchange capacity; Sa, sand; Si, silt; C, clay

the selection of the HM modeling in Tunisia as the economic disadvantage countries, where the lack of facility to monitor the HM quantity with an individual interval of time to report the local authorities to bring the appropriate policies and minimize AMD impacts loaded with HMs on the agricultural soil and water resources.

There is still room, however, to improve the accuracy of this approach with high R^2 value in the validation set, and low error indexes, the additional data as input in relation to local rainfall, evapotranspiration, wind data, could be a solution. The complete balance must also include (i) the evolution of physic-chemical properties and (ii) the distribution of pollutants of the surface (stagnant and stream water) and aquifer water. In addition, remote sensing data have great potential to reveal spatial patterns of soil properties (Mulder et al., 2011). Therefore, in future research projecting the combination of the night-light data and thermal infrared data, soil attributes, and land use data can potentially improve again the HM prediction model.

Conclusions

Based on mineralogical compositions, physico-chemical conditions, soil grain size data, and three ML approaches, the soil HM sources and their proportions in an AMD environment were predicted. The data confirmed a close relationship among

HM contents, mineralogy, and soil pH in the Sidi-Driss tailings. The Zn, Pb, Mn, Cu, and Cd contents increased significantly with the galena, marcasite, pyrite, and sphalerite-marcasite dissolution and oxidation, the sulfate dissolution (marked with SO_4^{2-} ions increase) that contributed to the soil pH decline.

Based on independent soil variables that were the mineralogical compositions, physico-chemical properties, and grain size, the RF could be used as a practical model to provide the behavior of mine wastes and soil toxic HM contents. It performed much better than SVM and ANN, as it produced higher R^2_{val} and lower error indexes.

Finally, the RF model performance was enhanced by considering the environmental variables that were closely linked to the minerals dissolution and oxidation processes, acid soil pH values, EC, and CEC, grain sizes, and HM mobility, and retention in mine wastes and surrounding agricultural soils.

Author contribution Conceptualization, methodology: A.G. and M.T.; soil sampling: M.T., A.C., and M.D.; laboratory analysis: M.T.; data pretreatments, software, algorithms: A.G. and M.T.; results analysis: all authors; field expertise: M.T., A.G., A.C., and M.D.; writing: all authors. The manuscript was improved by the contributions of all the co-authors. All authors have read and agreed to the published version of the manuscript.

Funding This study was supported by the Water Research and Technologies Center of Borj-Cedria Technopark (CERTE, Tunisia), Georesources Laboratory. This project is carried out under the MOBIDOC scheme, funded by The Ministry of Higher Education and Scientific Research through the PromESsE project and managed by the ANPR.

Data availability Not applicable.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent to publish Not applicable.

Competing interests The authors declare no competing interests.

References

Alvarez-Guerra M, Ballabio D, Amigo JM, Bro R, Viguri JR (2010) Development of models for predicting toxicity from sediment chemistry by partial least squares-discriminant analysis and counter-propagation artificial neural networks. Environ. Pollut. 158:607e614. https://doi.org/10.1016/j.envpol.2009.08.007

Aubertin M, Fala O, Bussière B, Martin V, Campos D, Gamacherchette A, Chouteau M, Chapuis R (2002) Analyse des écoulements de l'eau en conditions non saturées dans les haldes à stériles. Symposium 2002 sur l'Environnement et les Mines. CIM, Rouyn-Noranda, CD-ROM

Ayari J, Agnan Y, Charef A (2016) Spatial assessment and source identification of trace metal pollution in stream sediments of Oued El Maadene basin, northern Tunisia. Environ Monit Assess 188:397

Balci N, Demirel C (2018) Prediction of acid mine drainage (AMD) and metal release sources at the Küre Copper Mine Site, Kastamonu, NW Turkey. Mine Water Environ 37:56–74. https://doi.org/10.1007/s10230-017-0470-4

Bazoobandi A, Emamgholizadeh S, Ghorbani H (2019) Estimating the amount of cadmium and lead in the polluted soil using artificial intelligence models. Eur J Environ Civil Eng 1e19. https://doi.org/10.1080/19648189.2019.1686429.

Bellon-Maurel V, Fernandez-Ahumada E, Palagos B, Roger JM, McBratney A (2010) Prediction of soil attributes by NIR spectroscopy. a critical review of chemometric indicators commonly used for assessing the quality of the prediction. TrAC, Trends Anal Chem 29:1073–1081

Besalatpour A, Hajabbasi MA, Ayoubi A, Gharipour A, Jazi AY (2012) Prediction of soil physical properties by optimized support vector machines. Int Agrophys 109e115

Bhuiyan MAH, Suruvi NI, Dampare SB, Islam MA, Quraishi SB, Ganyaglo S, Suzuki S (2011) Investigation of the possible sources of heavy metal contamination in lagoon and canal water in the tannery industrial area in Dhaka Bangladesh. Environ Monit Assess 175(1–4):633–649

Boser BE, Guyon IM, Vapnik VN (1992) A Training Algorithm for Optimal Margin Classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM. Pennsylvania, Pittsburgh, pp. 144e152. USA

Bouzahzah H, Benzaazoua M, Bussiere B, Plante B (2014) Prediction of Acid Mine Drainage: Importance of Mineralogy and the Test Protocols for Static and Kinetic Tests. Mine Water Environ 33:54–65. https://doi.org/10.1007/s10230-013-0249-1

Breiman L (2001) Random Forests. Mach Learn 45(1):5–32

Breiman L (2002) Manual on setting up, using, and understanding random forests v3.1. Technical report, http://oz. Berkeley. edu/user/breiman

Carbone C, Dinelli E, Marescotti P, Gasparotto G, Lucchetti G (2013) The role of AMD secondary minerals in controlling environmental pollution: Indications from bulk leaching tests. J Geochem Explor 132:188–200

Carmo FF, Lanchotti AO, Kamino LHY (2020) Mining Waste Challenges: Environmental Risks of Gigatons of Mud, Dust and Sediment in Megadiverse Regions in Brazil. Sustainability 12(20):8466. https://doi.org/10.3390/su12208466

Chang C-W, Laird DA, Mausbach MJ, Hurburgh CR Jr (2001) Near-infrared reflectance spectroscopy—principal components regression analysis of soil properties. Soil Sci Soc Am J 65:480–490

Choe E, Meer FVD, Ruitenbeek FV, Werff HVD, Smeth BD, Kim KW (2008) Mapping of heavy metal pollution in stream sediments using combined geochemistry, field spectroscopy, and hyperspectral remote sensing: a case study of the Rodalquilar mining area SE Spain. Remote Sens Environ 112(7):3222–3233

Chowdhury A, Naz A, Maiti SK (2021) Bioaccumulation of potentially toxic elements in three mangrove species and human health risk due to their ethnobotanical uses. Environ Sci Pollut Res 28:33042–33059. https://doi.org/10.1007/s11356-021-12566-w

Chowdhury A, Maiti SK (2016) Identification of metal tolerant plant species in mangrove ecosystem by using community study and multivariate analysis: a case study from Indian Sunderban. Environ Earth Sci 75:744. https://doi.org/10.1007/s12665-016-5391-1

Dermech M (1990) Le complexe de l'Oued Bélif-Sidi-Driss (Tunisie septentrionale). Hydrothermalisme et métallogénie. PhD thesis, Paris 7 University, France, 336p

Dermech M, Trifi M, Charef A, Boulègue J (2022) Base metal mineralization and peri-mediterranean Miocene magmatism: Geologic, Petrographic and Mineralogic arguments of the genesis model of Sidi Driss-Argoub Ressas (Fe)-Zn-Pb (North Tunisia). In press

Dold B, Fontbote L (2001) Element cycling and secondary mineralogy in porphyry copper tailings as a function of climate, primary mineralogy, and mineral processing. J Geochem Explor 74:3–55
Dold B (2006) Element flows associated with marine shore mine tailings deposits. Environ Sci Technol 40:752–758
Dold B (2014) Submarine tailings disposal (STD). Minerals 4:642–666
Dutrizac JE, Jambor JL (2000) Jarosites and their application in hydrometallurgy. In Sulphate Minerals: Crystallography, Geochemistry, and Environmental Significance; Alpers C.N, Jambor J.L & Nordstrom D.K Eds, Reviews in Mineralogy and Geochemistry, 40, Mineralogical Society of America: Chantilly, VA, USA 405–443
El Amri A, M'nassri S, Nasri N, Nsir H, Majdoub R (2022) Nitrate concentration analysis and prediction in a shallow aquifer in central-eastern Tunisia using artificial neural network and time series modelling. Environ Sci Pollut Res 10.1007/s11356-021-18174-y
Fengxia GU, Wenbao L (2010) Applications of remote sensing and GIS to the assessment of riparian zones for environmental restoration in agricultural watersheds. Geo-Spatial Inform Sci 13:263–268
Frau F, Marescotti P (2011) Mineralogical and geochemical techniques to investigate the relationships between minerals and contaminants in supergenic environments: an update review. Neues Jahrbuch Fur Mineralogie-Abhandlungen 188(1):1–9
Garcıa-Sanchez A, Alastuey A, Querol X (1999) Heavy metal adsorption by different minerals: application to the remediation of polluted soils. Sci. Total Environ 242:179188
Gasmi A, Gomez C, Zouari H et al (2014) Using Vis-NIR hyperspectral HYPERION data for bare soil properties mapping over Mediterranean area : plain of the Oued Milyan, Tunisia. Eur Acad Res II:11721–11739
Gasmi A, Gomez C, Zouari H, Antoine Masse A, Ducrot D (2016) PCA and SVM as geo-computational methods for geological mapping in the southern of Tunisia, using ASTER remote sensing data set. Arab J Geosci 9:753. https://doi.org/10.1007/s12517-016-2791-1
Gasmi A, Masse A, Ducrot D, Zouari H (2017) Télédétection et photogrammétrie pour l'étude de la dynamique de l'occupation du sol dans le bassin versant de l'oued Chiba (Cap-Bon, Tunisie). Revue Française De Photogrammétrie Et De Télédétection 215, 43–51. https://doi.org/10.52638/rfpt.2017.344.
Gasmi A, Gomez C, Lagacherie P, Zouari H (2019) Surface soil clay content mapping at large scales using multispectral (VNIR–SWIR) ASTER data. Int J Remote Sens 40(4):1506–1533. https://doi.org/10.1080/01431161.2018.1528018
Gasmi A, Gomez C, Lagacherie P, Zouari H, Laamrani A, Chehbouni A (2021) Mean spectral reflectance from bare soil pixels along a Landsat-TM time series to increase both the prediction accuracy of soil clay content and mapping coverage. Geoderma 388:114864. https://doi.org/10.1016/j.geoderma.2020.114864
Gasmi A, Gomez C, Chehbouni A, Dhiba D, Elfil H (2022) Satellite Multi-Sensor Data Fusion for Soil Clay Mapping Based on the Spectral Index and Spectral Bands Approaches. Remote Sens 14(5):1103. https://doi.org/10.3390/rs14051103
Gasmi A, Zouari H, Masse A, Ducrot D (2015) Potential of the Support Vector Machine (SVMs) for clay and calcium carbonate content classification from hyperspectral remote sensing. Int J Innov Appl Stud 13:497–506
Gomez C, Dharumarajan S, Féret J-B, Lagacherie P, Ruiz L, Sekhar M (2019) Use of Sentinel-2 Time-Series Images for Classification and Uncertainty Analysis of Inherent Biophysical Property: Case of Soil Texture Mapping. Remote Sens 11(5):565. https://doi.org/10.3390/rs11050565
Gholami R, Kamkar-Rouhani A, Doulati A, F., Maleki, Sh., (2011) Prediction of toxic metals concentration using artificial intelligence techniques. Appl Water Sci 1:125–134
Gierè R, Sidenko NV, Lazareva EV (2003) The role of secondary minerals in controlling the migration of arsenic and metals from high-sulfide wastes (Berikul gold mine, Siberia). Appl Geochem 18:1347–1359
Guo PT, Li MF, Luo W, Tang QF, Liu ZW, Lin ZM (2015) Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. Geoderma 237e238:49e59
Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J. Mach. Earn. Res. 3:1157e1182
Hageman PL, Briggs PH (2000) A simple field leach for rapid screening and qualitative characterization of mine-waste material on abandoned mine lands. In: Proceedings from the Fifth International Conference on Acid Rock Drainage, Denver, Colorado, May 21–24, 2000, vol. II. Society for Mining, Metallurgy and Exploration, Inc., pp. 1463–1475
Hudson-Edwards KA, Wright K (2011) Computer simulations of the interactions of the (0 1 2) and (0 0 1) surfaces of jarosite with Al, Cd, Cu^{2+} and Zn. Geochim Cosmochim Acta 75:52–62
Hu Y, Cheng H (2013) Application of stochastic models in identification and apportionment of heavy metal pollution sources in the surface soils of a largescale region. Environ Sci Technol 47:3752e3760
Hammarstrom JM, Seal RR, Meier AL, Kornfeld JM (2005) Secondary Sulfate Minerals Associated With Acid Drainage in the Eastern Us: Recycling of Metals and Acidity in Surficial Environments. https://lib.ugent.be/catalog/ebk01:4330000001250268.
Hong Y, Shen R, Cheng H, Chen Y, Zhang Y, Liu Y, Zhou M, Yu L, Liu Y, Liu Y (2019) Estimating lead and zinc concentrations in peri-urban agricultural soils through eflectance spectroscopy: Effects of fractional-order derivative and random forest. Sci Total Environ 651:1969–1982
Jambor JL, Nordstrom DK, Alpers CN (2000) Metalsulphate salts from sulphide mineral oxidation. Rev Mineral Geochem 40:303–350
Jambor JL, Blowes DW, Ritchie AIM (2003) In: Jambor, Blowes, Ritchie (Eds.), Environmental Problem Aspects of Mine Wastes. Short Course Series, 31. Vancouver: Mineralogical Association of Canada, p. 430
Kemper T, Sommer S (2002) Estimate of Heavy Metal Contamination in Soils after a Mining Accident Using Reflectance Spectroscopy. Environ Sci Technol 36:2742–2747
Kemper T, Sommer S (2003) Mapping and monitoring of residual heavy metal contamination and acidification risk after the aznalcóllar mining accident (Andalusia, Spain) using field and airborne hyperspectral data. Proceedings of the 3rd EARSeL Workshop on Imaging Spectroscopy, Herrsching, Germany, 13–16, pp. 333–343
Khalil A, Hanich L, Hakkou R, Lepage M (2014) Gis- Based environmental database for assessing the mine pollution: a case study of an abandoned mine site in Morocco. J Geochem Explor 144:468–477
Kumar S, Lal R, Liu D (2012) A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma 189e190:627e634
Lagacherie P, Arrouays D, Bourennane H, Gomez C, Martin M, Saby NPA (2019) How far can the uncertainty on a Digital Soil Map be known?: a numerical experiment using pseudo values of clay content obtained from Vis-SWIR hyperspectral imagery. Geoderma 337:1320–1328. https://doi.org/10.1016/j.geoderma.2018.08.024
Lark RM (1999) Soil eland form relationships at within-field scales: an investigation using continuous classification. Geoderma 92:141e165
Lu C, Xu Z, Dong B, Zhang Y, Wang M, Zeng Y, Zhang C (2022) Machine learning for the prediction of heavy metal removal by chitosan-based flocculants. Carbohyd Polym 285:119240
Montoroi JP (1997) Electric conductivity of soil solution and aqueous extract. Etudes Et Gestion Des Sols 4:279–298

Malley DF, Williams PC (1997) Use of near-infrared reflectance spectroscopy in prediction of heavy metals in freshwater sediment by their association with organic matter. Environ Sci Technol 31(12):3461–3467

Mojid MA, Hossain ABMZ, Ashraf MA (2019) Artificial neural network model to predict transport parameters of reactive solutes from basic soil properties. Environ Pollut 255:113355. https://doi.org/10.1016/j.envpol.2019.113355

Mulder VL, de Bruin S, Schaepman ME, Mayr TR (2011) The use of remote sensing in soil and terrain mapping d a review. Geoderma 162:1e19

Murray J, Kirschbaum A, Dold B, Mendes Guimaraes E, Pannunzio Miner E (2014) Jarosite versus Soluble Iron-Sulfate Formation and Their Role in Acid Mine Drainage Formation at the Pan de Azúcar Mine Tailings (Zn-Pb-Ag), NW Argentina. J Miner 4:477–502

Negra L (1987) Pétrologie, minéralogie et géochimie des minéralisations et des roches encaissantes des basins associés aux structures tectoniques et magmatiques de l'Oued Bélif et du Jebel Haddada (Nord des Nefza, Tunisie septentrionale). Unpublished PhD thesis, Paris Sud University, France, 223p

Puri M, Pathak Y, Sutariya VK, Tipparaju S, Moreno W (2016) (Eds.). Artificial Neural Network for Drug Design, Delivery and Disposition. Academic Press, Boston, pp. 3–13. https://doi.org/10.1016/B978-0-12-801559-9.00001-6

Ogwueleka TC (2014) Assessment of the water quality and identification of pollution sources of Kaduna River in Niger State (Nigeria) using exploratory data analysis. Water Environ J 28(1):31–37

Omondi E, Boitt M (2020) Modeling the spatial distribution of soil heavy metals using random forest modelda case study of nairobi and thirirka rivers' confluence. J. Geogr. Inf. Syst. 12:597e619. https://doi.org/10.4236/jgis.2020.126035

Pourret O, Lange B, Bonhoure J, Colinet G, Decree S, Mahy G, Séleck M, Shutcha M, Pierre-Faucon M (2016) Assessment of soil metal distribution and environmental impact of mining in Katanga (Democratic Republic of Congo). Appl Geochem 64:43–55

Qiu L, Wang K, Long W, Wang K, Hu W, Amable GS (2016) A comparative assessment of the influences of human impacts on soil Cd concentrations based on stepwise linear regression, classification and regression tree, and random forest models. PLoS One 11:e0151131. https://doi.org/10.1371/journal.pone.0151131

Rahman MS, Saha N, Molla AH (2014) Potential ecological risk assessment of heavy metal contamination in sediment and water body around Dhaka export processing zone Bangladesh. Environ Earth Sci 71(5):2293–2308

Rayment GE, Higginson FR (1992) Handbook of Soil and Water Chemical Methods. Inkata Press, Melbourne, p 488p

Rooki R, Doulati Ardejani F, Aryafar A, Bani Asadi A (2011) Prediction of heavy metals in acid mine drainage using artificial neural network from the Shur River of the Sarcheshmeh porphyry copper mine, Southeast Iran. Environmental Earth Science 64:1303–1316. https://doi.org/10.1007/s12665-011-0948-5

Salonen V-P, Korkka-Niemi K (2007) Influence of parent sediments on the concentration of heavy metals in urban and suburban soils in Turku Finland. Appl Geochem 22:906e918

Schwarz K, Weathers KC, Pickett STA, Lathrop RG, Pouyat RV, Cadenasso ML (2013) A comparison of three empirically based, spatially explicit predictive models of residential soil Pb concentrations in Baltimore, Maryland, USA: understanding the variability within cities. Environ Geochem Health 35:495–510

Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK (1999) Improvements to the SMO Algorithm for SVM Regression. In: IEEE Transactions on Neural Networks 11, 5

Smola AJ, Schölkopf B (1998) A tutorial on support vector regression. Stat Comput 14:199–222

Stefanov St Hr, Ouchev A (1972) Gisement plombo-zincifère de Sidi Driss. Rapport géol. Avec estimation de réserves. Unpublished internal report. Office National des Mines de Tunisie

Tajik S, Ayoubi S, Nourbakhsh F (2012) Prediction of soil enzymes activity by digital terrain analysis: comparing artificial neural network and multiple linear regression models. Environ Eng Sci 29:798e806

Tan K, Ye Y, Cao Q, Du P (2014) Estimation of arsenic contamination in reclaimed agricultural soils using reflectance spectroscopy and ANFIS model. IEEE J Select Topics Appl Earth Observ Remote Sens 7(6):2540–2546

Tan K, Ma W, Wu F (2019) Random forest–based estimation of heavy metal concentration in agricultural soils with hyperspectral sensor data. Environ Monit Assess 191:446. https://doi.org/10.1007/s10661-019-7510-4

Tepanosyan G, Sahakyan L, Maghakyan N, Saghatelyan A (2020) Combination of compositional data analysis and machine learning approaches to identify sources and geochemical associations of potentially toxic elements in soil and assess the associated human health risk in a mining city. Environ Pollut 261:114210. https://doi.org/10.1016/j.envpol.2020.114210

Trifi M, Dermech M, Charef A, Azouzi R, Hjiri B (2018) Extraction procedures of toxic and mobile heavy metal fraction from complex mineralogical tailings affected by acid mine drainage. Arab J Geosci 11:328. https://doi.org/10.1007/s12517-018-3612-5

Trifi M, Dermech M, Charef A, Azouzi R, Chalghoum A, Hjiri B, Ben Sassi M (2019) Trend evolution of physicochemical parameters and metals mobility in acidic and complex mine tailings long exposed to severe Mediterranean climatic conditions: Sidi Driss tailings case (NW-Tunisia), published in African Earth Science journal 158:103509. https://doi.org/10.1016/j.jafrearsci.2019.05.017

Thompson JA, Pena-Yewtukhiw EM, Grove JH (2006) Soil elandscape modeling across a physiographic region: topographic patterns and model transportability. Geoderma 133:57e70

Wang J, Cui L, Gao W, Shi T, Chen Y, Gao Y (2014) Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy. Geoderma 216(4):1–9

Wang H, Yilihamu Q, Yuan M, Bai H, Xu H, Wu J (2020) Prediction models of soil heavy metal(loid)s concentration for agricultural land in Dongli: a comparison of regression and random forest. Ecol Indicat 119 (February). https://doi.org/10.1016/j.ecolind.2020.106801

Williams B, Halloin C, Löbel W, Finklea F, Lipke E, Zweigerdt R, Cremaschi S (2020) Data-Driven Model Development for Cardiomyocyte Production Experimental Failure Prediction, Editor(s): SauroPierucci Flavio Manenti, Giulia Luisa Bozzano, Davide Manca. Computer Aided Chem Eng Elsevier 48:1639–1644

Werbos PJ (1975) Experimental implications of the reinterpretation of quantum mechanics. Nuovo Cim B 29:169–177. https://doi.org/10.1007/BF02732237

Were K, Bui DT, Dick ØB, Singh BR (2015) A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol Ind 52:394–403

Wu Y, Chen J, Wu X, Tian Q, Ji J, Qin Z (2005) Possibilities of reflectance spectroscopy for the assessment of contaminant elements in suburban soils. Appl Geochem 20(6):1051–1059

Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Philip SY (2008) Top 10 algorithms in data mining. Knowl. Inf. Syst. 14:1e37

Xu X, Chen S, Ren L, Han C, Lv D, Zhang Y, Ai F (2021) Estimation of Heavy Metals in Agricultural Soils Using Vis-NIR Spectroscopy with Fractional-Order Derivative and Generalized Regression Neural Network. Remote Sens 13(14):2718. https://doi.org/10.3390/rs13142718

Yaseen ZM (2021) An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions. Chemosphere 277:130126. https://doi.org/10.1016/j.chemosphere.2021.130126

Yaseen ZM, Sulaiman SO, Deo RC, Chau K-W (2018) An enhanced extreme learning machine model for river flow forecasting: state-of-the-art, practical applications in water resource engineering area and future research direction. J. Hydrol. 569:387e408. https://doi.org/10.1016/j.jhydrol.2018.11.069

Zarei I, Pourkhabbaz A, KhuzestaniRB (2014) An assessment of metal contamination risk in sedimentsof Hara Biosphere Reserve, southern Iran with a focuson application of pollution indicators. Environ Monit Assess. https://doi.org/10.1007/s10661-014-3839-x

Zeraatpisheh M, Jafari A, Bagheri Bodaghabadi M, Ayoubi S, Taghizadeh- Mehrjardi R, Toomanian N, Kerry R, Xu M (2020) Conventional and digital soil mapping in Iran: past, present, and future. CATENA 188:104424

Zhang H, Wu P, Yin A, Yang X, Zhang M, Gao C (2017) Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: a comparison of multiple linear regressions and the random forest model. Sciences of Total Environment 592:704–713

Zhang H, Yin S, Chen Y, Shao S, Wu J, Fan M, Chen F, Gao C (2020) Machine learning-based source identification and spatial prediction of heavy metals in soil in a rapid urbanization area, eastern China. J Clean Prod 273:122–858

Zolfaghari Z, Mosaddeghi MR, Ayoubi S (2015) ANN-based pedotransfer and soil spatial prediction functions for predicting Atterberg consistency limits and indices from easily available properties at the watershed scale in western Iran. Soil Use Manag. 31:142e154