Journal of Materials Research and Technology 29 (2024) 668–677

Contents lists available at ScienceDirect

Journal of Materials Research and Technology

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

Predicting gold accessibility from mineralogical characterization using machine learning algorithms

Fabrizzio Rodrigues Costa[a,*], Cleyton de Carvalho Carneiro[b], Carina Ulsen[c]

[a] Universidade de São Paulo, Escola Politécnica, Department of Mining and Petroleum Engineering, São Paulo, São Paulo State, Brazil
[b] Universidade de São Paulo, Escola Politécnica, InTRA – Integrated Technologies for Rock and Fluid Analysis, Santos, São Paulo State, Brazil
[c] Universidade de São Paulo, Escola Politécnica – Technological Characterization Laboratory, Department of Mining and Petroleum Engineering, São Paulo, São Paulo State, Brazil

ARTICLE INFO

Keywords:
Gold ore
Machine learning
Supervised models
Data analysis of applied mineralogy

ABSTRACT

Information about accessibility is of great relevance for gold recovery studies. Obtaining these variables from machine learning models can greatly assist in quickly determining accessibility. Few studies have been published relating the mineralogy of the gold ore process and the application of artificial intelligence, mainly algorithms in predicting variables related to gold recovery and extraction. Accessibility is an important variable for understanding the ability to recover gold from a cyanide solution, which can occur through fractures or some other means that provides access to the solution and consequent leaching of the gold grain. This study aims to present a model capable of predicting the accessibility variable using a data set with 168 characterization results from different ML methods, such as Linear Regression (LR), Random Forest (RF), Sequential Minimum Optimization for Support Vector Machine (SMOreg) and Gaussian Processes (GP). In this context, it was possible to establish that the random forest model performed best by presenting a coefficient of determination R^2 (0.77), MAE (11.76), and RMSE (14.48). It was also reported from the SHAP analysis that the Au_grade, exposed_a, and As_grade showed the highest contribution level towards the perdition process of the model.

1. Introduction

Process Mineralogy is an interdisciplinary approach aimed at linking the study of specific aspects of the ore bodies and plant products that can directly help in determining the mineralogical characteristics of the ore bodies, the potential for recovery, and the identification of their behavior in face of the beneficiation process. It provides subsidies for metallurgists, process engineers, and geologists in mine planning, development, and optimization of the ore's beneficiation process and hydrometallurgical operations [1–6].

The mineralogical characterization has benefited from the advancement of techniques for electronic microscopy, particularly in the automation of quantitative image analysis techniques (SEM-IA). This is a branch of mineralogy applied to the determination of quantitative mineralogy, mineral's association and liberation, grain size distribution, particle size, particles, and their inclusions, among other characteristics related to their morphology. In recent years, there has been an expansion in the use of the quantitative analysis technique, along with the improvement of equipment and the development of systems coupled with Energy Dispersive Spectroscopy (EDS) and Image Analysis (IA). A great advantage of automated image analysis methods is that they allow faster analysis, statistical robustness with the generation of a large amount of data, and reliability in the results, thus minimizing the analysis error [7–13].

In gold ores, the classic concept of mineral liberation described by Gaudin [14], must be adapted when based on specific properties of gold extraction and recovery. The term “accessible” or “accessibility” is more adequate and means the portion accessible is directly proportional to the ability to extract gold from a cyanide solution via fractures or microfractures. The gold that can be leached or recovered is the portion of gold in which there is some perimeter either exposed to alkaline cyanide solutions, on the surface or included in particles whose gold is accessible by microfractures or some medium in which the solution can flow [15]. Although Fig. 1A shows a free gold grain, Fig. 1B displays in the same mineral a locked gold grain and gold that may be extracted by fracture for the solution percolation, making it accessible. Different percentages

[*] Corresponding author. Department of Mining and Petroleum Engineering, Escola Politécnica, Universidade de São Paulo, Av. Prof. Mello Moraes, 2373, Butantã, São Paulo, SP, Brazil.
E-mail address: fab.costa@usp.br (F.R. Costa).

https://doi.org/10.1016/j.jmrt.2024.01.139
Received 30 September 2023; Received in revised form 30 November 2023; Accepted 15 January 2024
Available online 17 January 2024

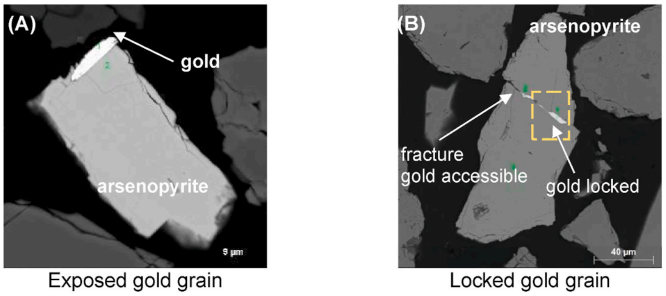


Fig. 1. Accessibility of gold grains: (A) Gold grain exposed and (B) Gold grain with minimum exposure and accessibility, liable to be leached and gold grain locked (yellow rectangle) in arsenopyrite particle. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

of accessibility can occur. In the two-dimensional representation, it is evident that the gold may be extracted by paths located in another part of the grain not visualized in the 2D image. Thus, information on gold accessibility by image analysis may be undersized or underestimated due to technique limitations.

The classic definition of mineral liberation shows that one of the species minerals in a population of different mineral species consists of the percentage of mineral, which occurs as free particles concerning the total amount of mineral in mixed or free particles. The association corresponds to the percentage of the mineral included in two or more phases about the total. In Fig. 2, different classes of particles in exposed area and perimeter are illustrated.

Artificial Intelligence (AI), particularly Machine Learning (ML) is a discipline of computer science that focuses on studying mathematical models and several different algorithms to make predictions utilizing knowledge and providing a feasible solution to dataset [16–18]. A dataset is the central part of any learnable decision-making system for automated classification, regression tasks, clustering, association rule learning, and reinforcement learning. Machine learning can adopt new methods according to its characteristics, simulate human learning methods, or combine the two to form new methods [19].

Through data interpretation, predictions are developed and obtained by connecting the data with the knowledge set and developing the learning algorithms [20,21]. ML usually provides systems with the ability to learn and enhance from experience without being specifically programmed automatically. The methods show the advantages particularly in geosciences, where they challenge grade and recovery in flotation [22] application to classify minerals automatically [23], technological advancement in the electronic industry [24], geometallurgy [25,26] and classification of drill core textures for process simulation [27]. The ML algorithms can be categorized into four primary types: supervised [28], unsupervised [29,30], semi-supervised [31] and reinforcement learning [32]. Learning concerns a set of procedures defined to adjust the parameters of an AI, so that it can learn a certain function.

In concern of mineral characterization, recent publications have been produced mainly focus on a framework based on ML to maximize the use of such classifications for decision-making to improve the grade or recovery, optimize the throughput, reduce the environmental footprint of the process or provide confidence in predictions of metal production at geometallurgical model [26]. ML has been concentrated on prediction by using specified learning algorithms to find underlying patterns in large amounts of complex data. The ML methods can be effective even when the data are gathered without a carefully controlled experimental design and in the presence of complicated nonlinear interactions [33].

A series of works have been developed on the application of ANN techniques.

The correlations of the variables obtained by the technological characterization of the gold ore, especially the accessibility variable, using the Self-Organizing Maps (SOM) in the formation of clusters and in the implementation as an alternative tool to impute the missing data of the low-grade gold ore were object of study by Ref. [34].

Through process testing and mineralogical characterization, the development of a methodology for integrating process properties into a spatial model using ML methods and comparing performance in terms of its accuracy was also a related topic [25].

Classification performance was the subject of a study in which a reliable ML classifier was evaluated to identify several heavy minerals based on EDS data. The results indicated that Random Forest can be used as the most effective classifier for heavy mineral classification [35].

[36] published a review equipping researchers and industrial professionals with structured knowledge of the state of machine learning applications in mineral processing. Variables from gold ore mineralogical characterization such as arsenic, gold, and sulfur content as well as mineral associations and grain size of gold exposure influence the accessibility of the leach solution [15,34]. The researchers took on the task of modeling and predicting chemical-mineralogical behavior using mechanistic or empirical models. In supervised learning, some algorithms also are well suited for empirical regression modeling of a multivariable operation.

In this context, the present work aims to use a dataset from mineralogical characterization to predict the variable accessibility from different ML methods, comparing their performance in terms of their accuracy. Algorithms such as K-Nearest Neighbor (k-NN), Random Forest (RF), Sequential Minimum Optimization for Support Vector Machine (SMOreg), and Gaussian Processes (GP) were the ML models chosen. Furthermore, variable importance in the modeling and prediction was examined. Precision is expressed as coefficient of determination, mean absolute error, and root mean square error (R^2, MAE, and RMSE, respectively).

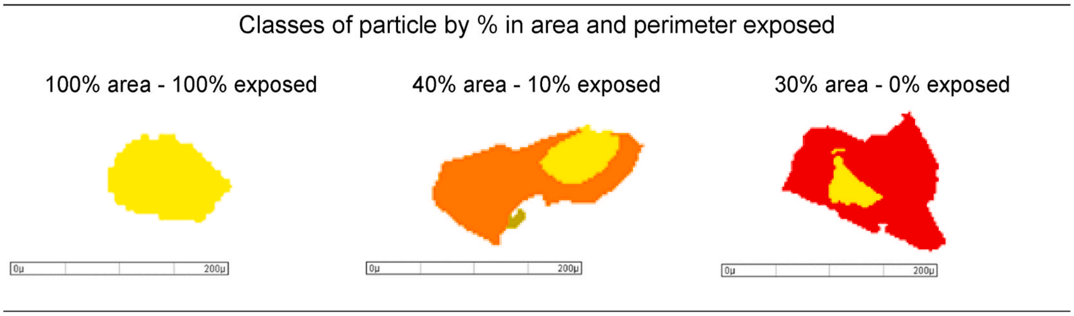


Fig. 2. Schematic representation of classes of particles by percentage in area and perimeter exposed (phase of interest in yellow). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

2. Conceptual background

2.1. Goodness-of-fit indicators

The following standard statistics metrics are used herein. In equations (1)–(3) the value x_I is the predicted value, y_I is the measured value. MAE means mean absolute error, RMSE means root mean squared error, R^2 coefficient of determination. Low values of MAE and RMSE, as well as high values of R^2, indicate a good fit of data.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|x_i - y_i| \tag{1}$$

x_i and y_i indicate actual and imputed values for n samples. MAE estimates the mean error of the predicted and actual values and evaluates continuous value imputation.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

RMSE measures the root mean square error for the predicted continuous variables concerning the actual variables. MAE and RMSE express the average error of the predictive model, concerning the original data (training and/or test).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)2}{\sum_{i=1}^{n}\left(x_i - \underline{y}_i\right)2}$$

$$\underline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3}$$

R^2 x_i and y_i indicate actual and imputed values for n samples and $\underline{y}_i$ is the mean value of x. It is a statistical measure that indicates how well the predicted values are close to the real data values

2.2. K-Nearest Neighbor (k-NN)

The k-Nearest Neighbors (k-NN) algorithm, non-parametric supervised learning method, is widely used for classification and regression problems in the industry [37]. The implementation of KNN regression is to calculate the average label attributes of the k known samples. Another approach uses an inverse distance weighted average label attributes of the k known samples [38]. The disadvantages are the computation of accurate distances as well as how to set K value [39].

However, before a classification can be made, the distance must be defined. Euclidean distance is most commonly used (Eq. (4)).

$$d = \sqrt{(x_1 - x_2)(x_1 - x_2)^T} \tag{4}$$

Therefore, the performance of these classification algorithms significantly depends on the k (Eq. (5)); the key parameter for k-NN. In this study, the value of k ranging from 1 to 50 was tested and an ideal value of k = 12 (smallest MAE) was found that best resulted in the prediction.

$$\hat{y} = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i \tag{5}$$

$N_k(x)$ is the neighborhood of x defined by closet k points Friedman, 2017.

2.3. Random forest (RF)

Random Forest (RF) is a nonparametric ensemble method developed by Ref. [40] and is used for both classification and regression analysis. RF is a modification of bagging that creates a collection of K-randomized regression trees and averages them. For classification problems, a set of decision tree classifiers is trained.

The training algorithm for RFs applies the general technique of bootstrap aggregating, or bagging, to tree learners. Bootstrap aggregating is used for training data creation by resampling the original data set randomly with replacement. This leads to more efficient model performance. While the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not that sensitive as long as the trees are not correlated [41].

Some advantages can be highlighted when using RF regression: bias, few hyperparameters input, and minimized risk of overfitting. It corrects the overfitting of the training set by constructing a multitude of Decision Trees (DT) and outputting the mean prediction (regression) of the individual trees. Therefore, each DT predicts the output independently, and then the predictions are averaged to generate the result. The equation (Eq. 6) summarizes the RF operator where x denotes input and $\hat{T}_k$ (x) is the estimation produced by the *kth* tree.

$$\hat{T} = \frac{1}{K} + \sum_{k=1}^{K} \hat{T}_k(x) \tag{6}$$

2.4. Sequential minimal optimization for support vector machine (SMOReg)

Sequential minimal optimization (SMO) an algorithm was developed by Ref. [42] to train SVM models. Models SVM can offer an advantage in generalization performance for solving pattern recognition, and complex regression problems and use Lagrange to solve the optimization problem is simply defined as a hyperplane between a set of positive data and a set of negative data.

It converts a very large quadratic programming (QP) optimization problem to the smallest possible QP problems which can be solved analytically. This feature of SMOreg provides a faster solution in numerical QP optimization than the chunking algorithm that is used conventionally to train the SVM [42].

SMOreg starts with the initial two Lagrange multipliers and continues until optimal values of these multipliers' values are found. One of the advantages of the SMOreg algorithm is that extra matrix storage is not needed in the training process of SVM. SMOreg algorithms work in two stages. In the first stage, the two Lagrange multipliers are solved with an analytic method, and in the other stage, the multipliers are chosen and optimized heuristically [43].

2.5. Gaussian process regression (GP)

GP is a nonlinear, nonparametric regression tool, useful for interpolating between data points scattered in a high-dimensional input space. It is based on Bayesian probability theory and has very close connections to other regression techniques, such as kernel ridge regression (KRR) and linear regression with radial basis functions [44]. It can capture a wide variety of relations between inputs and outputs by utilizing a theoretically infinite number of parameters and letting the data determine the level of complexity through the means of Bayesian inference [45,46].

GPR provides a solution to the modeling problem such that the locality of the interpolation may be explicitly and quantitatively controlled by encoding it in the a priori assumption of smoothness of the underlying function. Gaussian process regression can serve as a useful tool for performing inference both passively describing a given data set as best as possible, allowing one to also predict future data as well as actively, learning while choosing input points to produce the highest possible outputs. There are two equivalent approaches to deriving the GPR framework: the weight-space and the function-space views, each highlighting somewhat different aspects of the fitting process.

2.6. Shapley additive exPlanations

The SHAPley Additive exPlanations (SHAP) is a visualization tool used to making different a machine-learning model more explainable by visualizing its output. It can be used for explaining the prediction of any model by computing the contribution of each feature to the output prediction. This is done by SHAP assigning a score to each variable (SHAP value), which indicates how important the variable was [47,48].

Local accuracy, missingness, and consistency are properties to satisfy SHAP. Local accuracy means the explanation model should match the original model. The missingness property enforces that missing variables in the dataset are attributed no importance [49]. The consistency property says that if a model changes so that the marginal contribution of a feature value increases or stays the same (regardless of other features), the Shapley value also increases or stays the same (Eq. (7)).

SHAP specifies the explanation as:

$$g(z') = \varnothing_0 + \sum_{j=1}^{m} \varnothing_j z_j' \tag{7}$$

where g is the explanation model. $z' \in \{0,1\}m$ is the coalition vector. M is the maximum coalition size and $\varnothing_j \in \mathbb{R}$ is the feature attribution for a feature j, the Shapley values.

Two methods can be used to approximate SHAP values Kernel SHAP and TreeSHAP: KernelSHAP estimates for instance x the contributions of each feature value to the prediction and TreeSHAP is a variant of SHAP for tree-based machine learning models such as decision trees, random forests, and gradient boosted trees. TreeSHAP was introduced as a fast, model-specific alternative to KernelSHAP, but it turned out that it can produce unintuitive feature attributions [50].

SHAP feature is an alternative to permutation feature importance. There is a difference between both measures: Permutation feature importance is based on the decrease in model performance and SHAP is based on the magnitude of feature attributions. The fast computation makes it possible to compute the many Shapley values needed for the global model interpretations. The global interpretation methods include feature importance, feature dependence, interactions, clustering, and summary plots.

3. Materials and methods

3.1. Framework

The structure consists of the acquisition of a set of chemical-mineralogical data of gold ore obtained by X-ray based automated image analysis (SEM-IA), regression analysis, and efficiency test by ML of four different algorithms. The samples were characterized at the Technological Characterization Laboratory (LCT) of the University of São Paulo (USP), Brazil. Fig. 3 shows the flowchart of activities and procedures performed according to the methodology used.

The application of the ML algorithms presented here was performed using WEKA [51]. For the database training and testing stages, multiple combinations were applied to test and identify the best hyperparameters that fit each model. Table 1 show the main adjusted parameters.

3.2. Dataset

The dataset is composed of 168 samples obtained by image analysis

Table 1
Description of the hyperparameters used for each machine learning model.

K-Nearest Neighbor (K-NN)		**Sequential Minimum Optimization (SMOReg)**	
Number of neighbors	12	Batch Size	100
Batch Size	100	Complexity Parameter	5
Distance weighting	No distance weighting	Filter type	Normalize training data
Mean Squared	True	Kernel	Puk
Search Algorithm	Euclidean Distance	Reg Optimizer	Reg SMO Improved
Random Forest (RF)		**Gaussian Processes (GP)**	
Break Ties Randomly	True	Batch Size	100
Execution Slots	3	Filter type	Normalize training data
Max Depth	6	Kernel	Puk
Iterations	110	noise	1
Seed	6	Seed	8

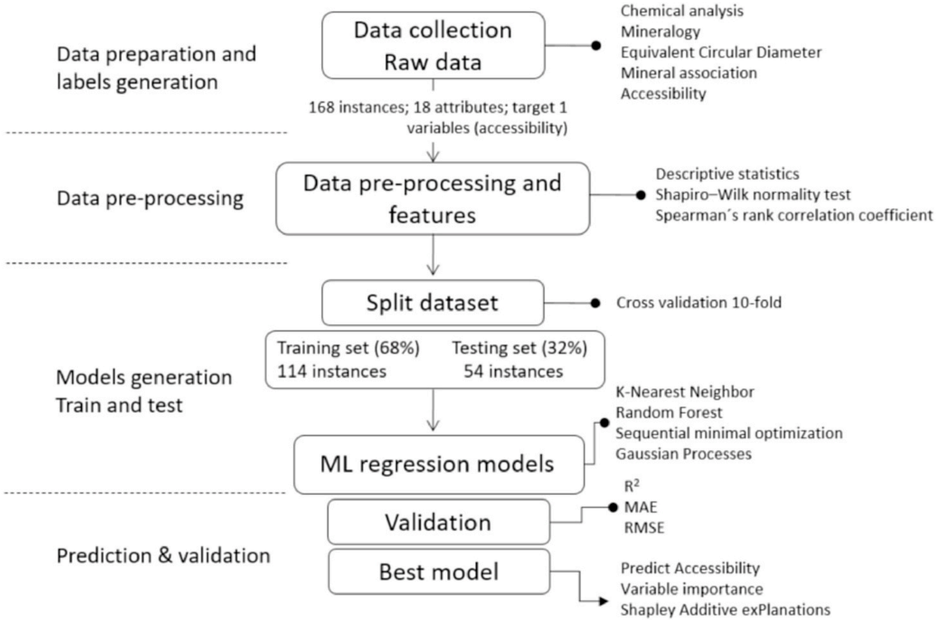


Fig. 3. Flowchart of activities performed until prediction of accessibility and variable importance.

(SEM-AI) of sulphide refractory gold ore samples and was provided by a low-grade gold-producing company (<0.6 g/t) in the state of Minas Gerais, Brazil. The deposit was hosted in carbonaceous seriticitic phyllite with the intercalation of phyllosilicate essentially composed of chlorite and millimeter quartzite lenses and venules. Sulphides, in general, are represented by pyrite, arsenopyrite, and sparse occurrences of pyrrhotite, sphalerite, chalcopyrite, and galena. Gold grains occur essentially associated with sulfides, mainly pyrite and arsenopyrite. Few free gold grains were observed.

Chemical analysis was carried out by fire assay to dosage of Au content, As by ICP OES and S by the pyrolysis method in an induction furnace with determination by infrared cell. Quantitative mineralogy, composition, forms of occurrence, and association of gold were assessed at 0.50–0.020 mm, carried out on sink products of heavy liquid separation in polished sections by SEM-IA using the MLA/FEI software coupled to a FEI Quanta 600 FEG scanning electron microscope. The automated search of the gold grains in polished sections of 30 mm in diameter relative to the heavy product with an analysis time of approximately 2.5 h section under the conditions of 300X using sparse phase liberation (SPL) analysis. A specific database for the present study, containing data on specific weight, chemical composition and EDS spectrum, was created for all minerals present (mineral reference). In total, 551 polished sections were analyzed.

A total of 18 input variables were treated, resulting from image analysis and chemical analysis, in which allow tracing of the most relevant components that influence the accessibility indices. These variables were used to build machine learning models. The variables inputted are listed in Table 2 and the range investigated for each variable.

Descriptive statistical analyses were performed using the standard library packages of Python. Significant differences ($P < 0.05$) between data sets are indicated where appropriate. The Shapiro–Wilk normality test was used to determine whether a data set followed a normal distribution [52].

Table 2
Characterization of 18 variables inputted for ML modeling.

#	Input variable	Description	unit	Range investigated
1	As_grade	Arsenic grade	ppm	280–5691
2	S_grade	Sulfur grade	%	0.30–2.00
3	Au_grade	Gold grade	g/t	0.04–1.36
4	qtz_m	Mineralogical distribution of quartz	%	11.3–78.3
5	mica_m	Mineralogical distribution of mica	%	16.6–49.9
6	clt_m	Mineralogical distribution of chlorite	%	0.1–3.3
7	alb_m	Mineralogical distribution of albite	%	1.0–13.8
8	heavy_m	Mineralogical distribution of heavy mineral[a]	%	0.7–3.8
9	carb_m	Mineralogical distribution of carbonates[a]	%	0.0–14.6
10	py_m	Mineralogical distribution of pyrite	%	0.0–3.6
11	aspy_m	Mineralogical distribution of arsenopyrite	%	0.0–2.0
12	othersulph_m	Mineralogical distribution of other sulphides[a]	%	0.1–5.5
13	D_{50}	Equivalent diameter D_{50} in 2D	μm	2.0–42.0
14	exposed_a	Exposed perimeter association	%	0.0–91.7
15	py_a	Grain gold association with pyrite	%	0.0–94.8
16	aspy_a	Grain gold association with arsenopyrite	%	0.0–100
17	othermin _a	Grain gold association with other minerals	%	0.0–77.0
18	accessibility	Accessibility gold grain	%	0.4–100

[a] Heavy minerals: ilmenite. rutile. goethite; carbonates: siderite. ankerite; other sulphides: galena. sphalerite. chalcopyrite.

For the study of correlations, Spearman's rank correlation coefficient was used [53] as it is a usual alternative to estimate linear correlations in situations where there is joint non-normality between variables. Spearman's rank is a nonparametric rank statistic proposed by Charles Spearman as a measure of the strength of an association between two variables (Eq. (8)). Spearman correlation coefficient can be computed as follows:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{8}$$

where ρ means Spearman's rank correlation coefficient, d_i, the difference between the two ranks of each observation n: number of observations. The Spearman rank correlation can take a value from +1 to −1. By convention, classify Spearman's correlation coefficients as weak: 0.0 to 0.3 or −0.3 to 0.0, moderate: 0.3 to 0.7 or −0.7 to −0.3. strong: >0.7 or < −0.7.

4. Results and discussion

4.1. Database description

One of the elements necessary for an accurate of machine learning application is model information diversity. The diversity of the database reflects the incorporation of measurement information characterizing relations across different elements and variables, representing the analyzed sample space, which must contain sufficient and necessary information for prediction to be effective.

The statistical analysis is performed to find the mean, median, standard deviation, coefficient of variation, variance and kurtosis of all the 18-input chemical-mineralogical variables that were considered for the model development. To test the normality of the data set, the Shapiro–Wilk test was applied (Table 3). The test rejects the hypothesis of normality when the p-value is less than or equal to 0.05. Failing the normality test allows you to state with 95 % confidence the data does not fit the normal distribution. Passing the normality test only allows you to state no significant departure from normality was found.

Kurtosis data is a measure that describes how heavily the tails of a distribution differ from the tails of a normal distribution. The results reinforce the non-normality of the data. Values positive suggests heavy tails (leptokurtic) and negative values mean that there are light tails (platykurtic). The tail heaviness or lightness is in comparison with the normal distribution and it suggests whether the data distribution is flatter or less flat than the normal distribution.

The results show that the p-values are below 0.05 indicating non-normality of the data. The py_m variable was the only one that came close, indicating a value of 0.04. Most of the variables approached 0.0. SEM-IA mineralogy analysis was performed to obtain the detailed quantitative mineralogical composition of samples (Fig. 4).

The mineralogical compositions show the considerable presence of quartz, mica, and albite, which together represent, on average, 85 % of the total sample composition. Sulfide minerals such as sphalerite, galena, and chalcopyrite were not represented in the table due to erratic occurrence being grouped into othersulph_m.

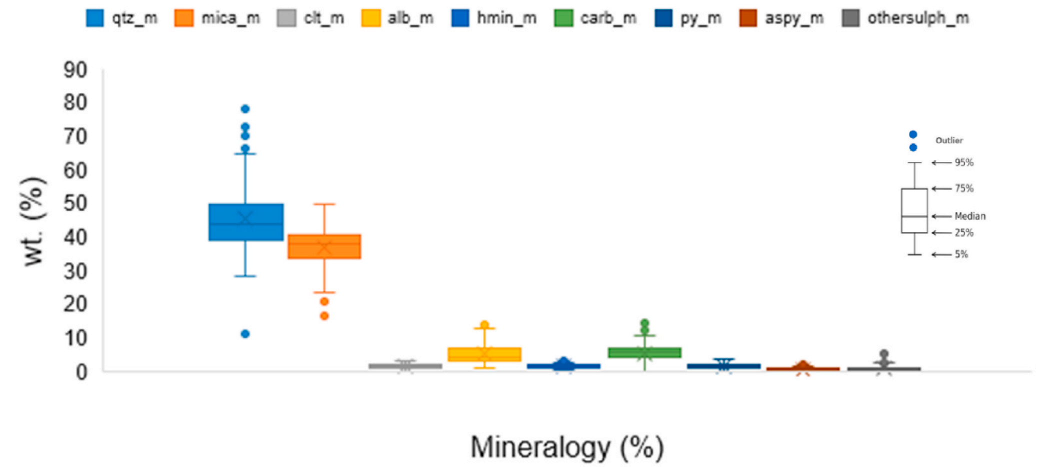
The main gold associations, assessed by SEM-IA are shown in Fig. 5 and expressed in terms of perimeter of contact with other minerals. The gold with exposed perimeter (exposed_a) on bearing particles of pyrite (pyrite_a), arsenopyrite (aspy_a), and a grouping of galena, sphalerite, chalcopyrite represented by other sulphides (othersulph_a). The occurrence of gold is mainly associated with arsenopyrite and pyrite and in very low proportions with silicates, carbonates, and heavy minerals.

The Spearman correlation matrix "r" of the variables obtained by image analysis (SEM-AI) and chemical analysis show moderate correlations (0.3–0.7) between accessibility, variable studied for prediction, arsenic content (As_grade), sulfur content (S_grade), both showing a

Table 3
Descriptive statistical analysis of the process mineralogy variables used as input for models and the hypothesis test (p-value) by Shapiro-Wilk.

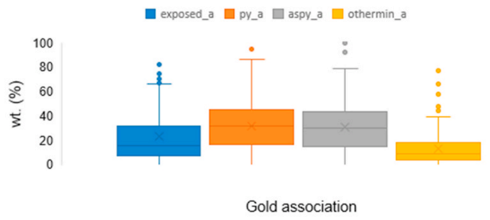
#	Input variable	Mean	Median	standard deviation	CV	Variance	Kurtosis	p-value
1	As_grade	2313.6	2156.5	1229.5	0.53	1511565	−0.69	0.00
2	S_grade	1.17	1.10	0.43	0.37	0.18	−0.89	3.5e-05
3	Au_grade	0.61	0.55	0.35	0.54	0.11	−0.88	5.8e-06
4	qtz_m	45.5	43.8	9.09	0.20	82.58	1.85	2.7e-06
5	mica_m	36.9	37.9	6.22	0.17	38.66	1.16	6.7e-06
6	clt_m	1.55	1.60	0.72	0.46	0.51	−0.87	0.00
7	alb_m	5.20	4.20	2.98	0.57	8.85	−0.33	3.4e-07
8	heavy_m	1.71	1.60	0.55	0.32	0.30	1.73	9.8e-07
9	carb_m	5.46	5.80	2.66	0.49	7.09	0.88	0.00
10	py_m	1.59	1.50	0.84	0.53	0.70	−0.47	0.04
11	aspy_m	0.61	0.40	0.52	0.86	0.27	−0.19	9.8e-11
12	othersulph_m	1.18	0.90	0.97	0.82	0.94	4.30	2.9e-14
13	D_{50}	14.4	10.0	10.2	0.71	104.3	0.21	1.2e-10
14	exposed_a	24.9	16.9	22.0	0.88	484.6	0.65	3.3e-11
15	py_a	31.9	31.8	20.6	0.65	425,2	−0.11	0.00
16	aspy_a	32.1	30.7	21.1	0.66	444.5	0.08	0.00
17	othermin_a	13.3	9.10	14.3	1.07	203.4	5.57	9.9e-15
18	accessibility	60.7	68.2	28.5	0.47	813.5	−1.03	1.1e-07

Heavy_m: ilmenite, rutile, goethite; carbonates: siderite, ankerite; othersulph_m: galena, sphalerite and chalcopyrite. Othermin _a: quartz, albite, heavy minerals, galena, sphalerite and chalcopyrite.



qtz: quartz; mic: mica; clt: chlorite, alb: albite, hmin: heavy minerals (ilmenite, rutile) and goethite; carb: carbonates siderite and ankerite; py: pyrite; aspy: arsenopyrite. Othersulph: other sulphides such as sphalerite, galena, and chalcopyrite.

Fig. 4. The major composition of the sample was determined by SEM-IA mineralogy analysis (wt%).



Exposed: exposed perimeter; grain gold association with py: pyrite; aspy: arsenopyrite. Othermin: quartz, mica, albite, ilmenite, rutile, siderite, anquerita, sphalerite, galena, and chalcopyrite.

Fig. 5. The gold association by perimeter of contact is determined by SEM-IA (wt%). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

correlation of 0.65 and 0.74, respectively and gold exposure (exposed_a) of 0.7. As expected, accessibility is directly correlated with gold content and exposed perimeter since the increase or decrease in accessibility is related to the amount or the occurrence of gold grains. Spearman assessments also indicated that there is a good correlation between gold and arsenic content (0.73), shown also with the mineral occurrence of arsenopyrite (aspy_m).

Negative correlations occur with mica, heavy minerals (rutile, ilmenite, and goethite), and a fact that may be associated with the low occurrence of gold grains in these minerals. Fig. 6 shows the correlation matrix (Spearman rank correlation) between 18 features.

4.2. *Predictive performance*

The predictive performance of four machine learning model methods for the training, validation, and test data presented in Table 3 were compared and evaluated in terms of better predictive performance for the accessibility variable.

To evaluate the performance of the algorithm was employed a 10-fold cross-validation to obtain a generalized model that does not overfit on training data. In 10-fold cross-validation on the training set, the original samples were divided randomly into ten equal-sized subsets, of which a unique fold was picked as a validation set for testing the model and the remaining nine subsets were used as training data. This process

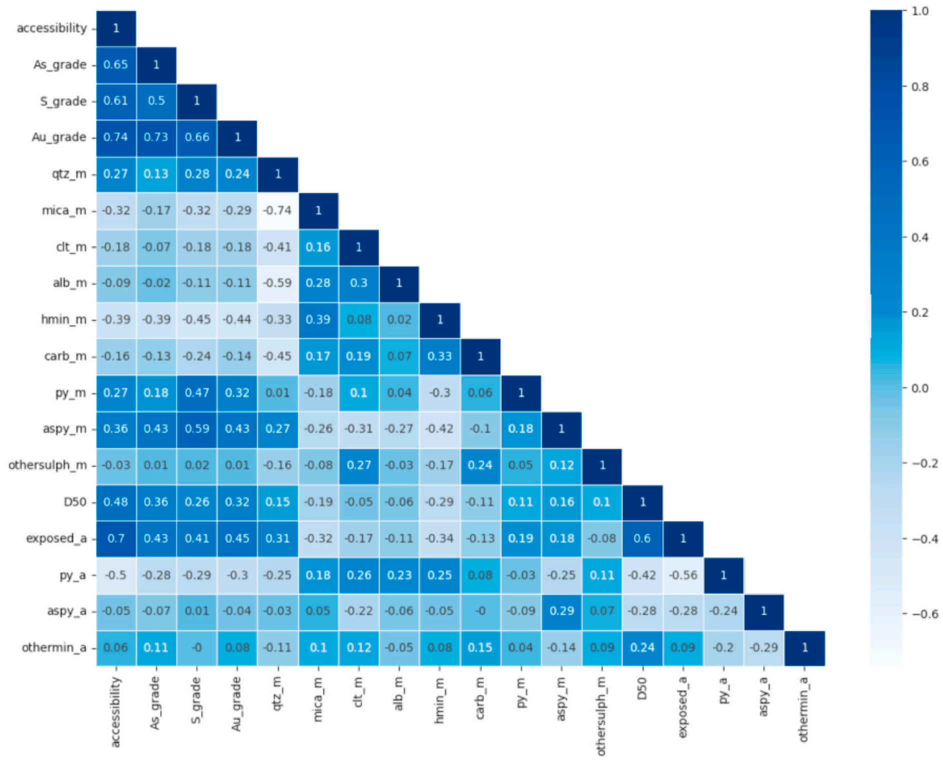


Fig. 6. Spearman's rank correlation map shows the correlation between the input variables.

was repeated ten times, with each of the ten subsets used exactly once as the validation data [54].

The performance comparison based on RMSE, MAE, and R^2 for training, validation, and testing, for regression is summarized in Table 4. Random Forest algorithm outperformed all other regression algorithms based on lower RMSE and MAE tests. The mean absolute error on test data for the RF model was 11.76; RMSE of 14.48 and 0.77 of R^2, indicating good predictability. The k-NN model had the worst training error in comparison with an R^2 of 0.65. RF and GP achieved good performance on the training data.

Comparing the error associated with the regression, calculated by the difference between the original and the predicted value. Fig. 7 shows, in box form, the interquartile difference of the set of samples tested. The low interquartile difference (Q3-Q1) shows a low data dispersion in the range of 50 % of the data for the RF algorithm. The RF models produced the lowest mean CV error, whereas the other models produced much higher mean CV error.

The performance of predict vs. original accessibility data based on RF model on the training and test data is given in Fig. 8. The prediction indicates that, for the training tests, the data set presented an excellent coefficient (0.97) between the predicted and original values. However, the tested data showed a moderate coefficient of determination (R^2; 0.77).

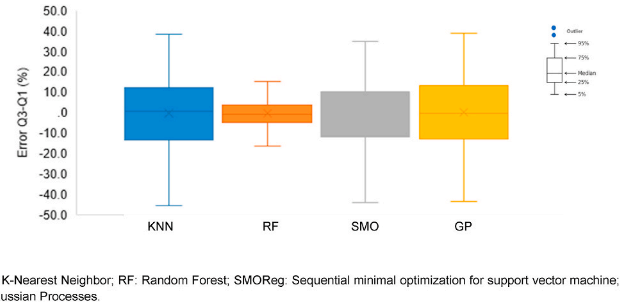


Fig. 7. Box showing the regression error data between original and predicted values
*K-NN: K-Nearest Neighbor; RF: Random Forest; SMOReg: Sequential minimal optimization for support vector machine; GP: Gaussian Processes.

4.3. Variable importance

SHAP analysis investigates the impact of each variable on your prediction model by showing its significance level. As shown in Fig. 9A, each variable had a different level of contribution to the prediction result of the accessibility variable.

SHAP assessments can determine the multiple correlations between variables rank them based on their importance and represent the

Table 4
Scoring of 5 model regression performances based on MAE, R^2 and RMSE after training dataset, cross-validation (10 folds), and testing.

Algorithm	Train			Validation			Test		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
K-NN	0.71	15.95	12.64	0.64	17.50	13.88	0.65	17.51	13.47
RF	0.97	5.61	4.59	0.71	15.59	12.83	0.77	14.48	11.76
SMOReg	0.72	15.08	11.80	0.65	16.95	13.78	0.69	16.50	13.44
GP	0.91	10.27	8.40	0.63	18.04	14.65	0.67	18.73	14.08

K-NN: K-Nearest Neighbor; RF: Random Forest; SMOReg: Sequential minimal optimization for support vector machine; GP: Gaussian Processes.

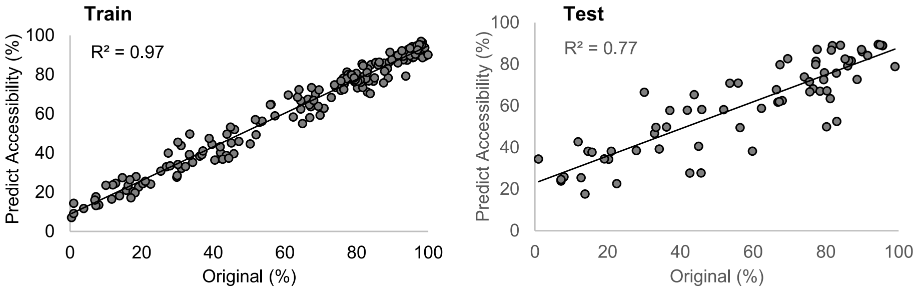


Fig. 8. Random Forest Original vs. predicted accessibility variable for train and test.

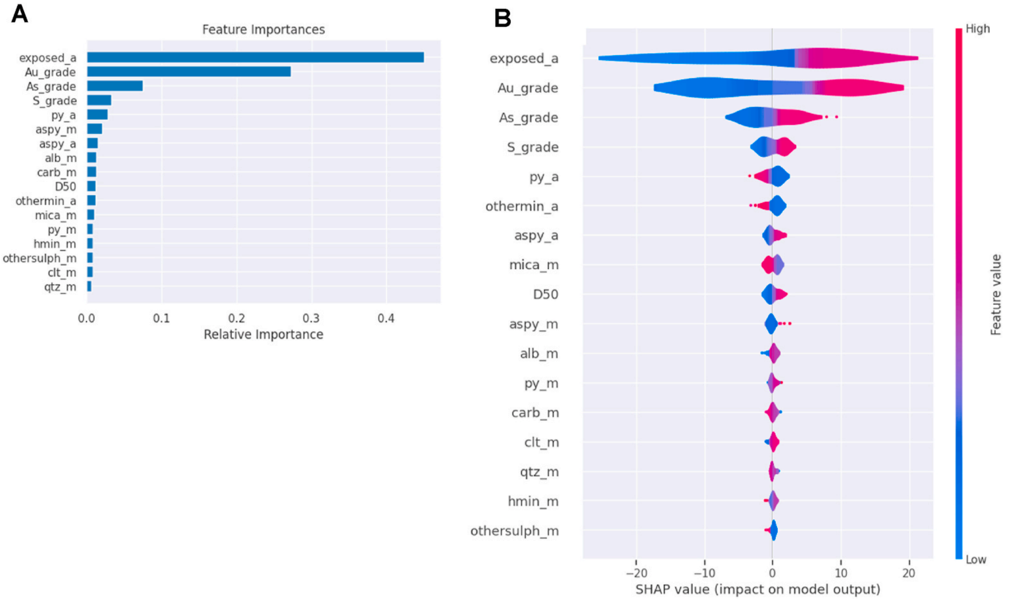


Fig. 9. Ranking variables based on their mean SHAP value and their relationship for accessibility prediction. (A) Feature importance plot using Random Forest; (B) Impact of the input parameters for the accessibility model's output from SHAP analysis.

average magnitude of their impacts. The assessment of the importance ranking of the SHAP variable revealed a pattern similar to the Spearman correlation assessment (Fig. 6), showing that the variables Au_grade, exposed_a, D50, and S_grade indicate a significant impact on the model result. All variables are shown in the order of global feature importance, the first being the most important and the last being the least important Fig. 9B.

The exposed_a, Au_grade, and As_grade variables have a high positive contribution when their values are high and a low negative contribution on low values. The variables mica_m, othersulph_m, hmin_m, aspy_m, alb_m, othermin_a, carb_m, py_m, qtz_m, and clt_m have almost no contribution to the prediction of whether their values are high, or low.

The results obtained through the feature importance show that the exposed_a variable has great weight in predicting accessibility, as shown in Fig. 9A.

Due to the high level of importance of the exposed_a and Au_grade variable in predicting accessibility, a correlation was generated that relates the original gold content data and the predicted values of the accessibility variable resulting from the application of the RF algorithm

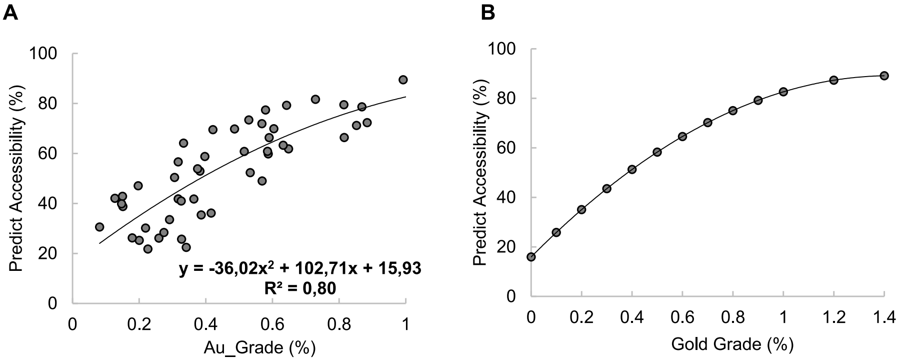


Fig. 10. Random Forest Original vs. predicted accessibility variable for training and test.

that had the best performance among the 4 analyzed algorithms. The correlation plot showed a good fit represented by R^2 (0.80) and a polynomial equation of the curve was generated (Fig. 10A). The equation was applied for gold grades between 0.1 and 1.4 ppm (Fig. 10B) generating a theoretical curve modeled by the RF algorithm.

The representation of the equation and the curve translates, in a theoretical and adjusted context, the prediction of the variable of interest (accessibility) by the gold grade. The theoretical curve using gold grades can provide a good prediction or tendency of accessibility percentage as a function of the gold grade of the sample, especially when using the variable gold grade, which is the variable predominant and essential in a database of auriferous ores.

The main observed features or features importance (exposed_a, Au_grade and As_grade) are directly related to the accessibility variable with a degree of dependence also evidenced by the Spearman's correlation. The accessibility of a gold particle is denoted by a lesser or greater degree of occurrence depending on its exposure to leaching fluids, the gold and arsenic content since the gold grain is largely associated with arsenopyrite. Therefore, accessibility is quantitatively linked to gold and arsenic contents and their exposure.

5. Conclusions

The main objective of this work was to predict accessibility variables through models using machine learning tools comparing four machine learning methods. The use of WEKA freeware allowed the production of fast and accurate machine learning models based on a mineralogical and chemical database. The performance of four ML algorithms (K-NN, RF, SMOReg, and GP) applied to a database allowed drawing the following conclusions.

(a) The machine learning approaches employed satisfactorily predicted the accessibility in gold grains, that is, the ability of a leachate fluid to access a portion of gold.

The random forest model outperforms the models in predicting accessibility. It presented a coefficient of determination R^2 (0.77), MAE (11.76), and RMSE (14.48).

(b) The validation K-fold (cross-validation 10-folds) confirms good precision in the model approach also better performance of the RF model towards the required outcome as opposed to the other three models.
(c) It was also reported from the SHAP analysis that the Au_grade, exposed_a, and As_grade showed the highest contribution level towards the prediction process of the model.
(d) Spearman's correlation coefficient is employed to check for collinearity and can be used to accurately capture the statistical dependence of input parameters. When compared with the importance variables, the variables were similar to the Spearman coefficient.

It remains for future studies to increase the systematic acquisition of analytical data to increase the number of samples since the resulting model is more suitable for visualizing trends and understanding the spatial properties of the properties of the modeled process.

In conclusion, machine-learning algorithms have proven to be very useful and versatile in many situations, but they also have some disadvantages that must be considered when creating and using these models, especially concerning inaccurate or biased data, since the algorithm may generate undesirable results.

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.

Acknowledgments

We would like to thank the technical team from the Technological Characterization Laboratory of the Polytechnic School at the University of São Paulo (USP) for their analytical support. The authors would also like to thank the anonymous referee for reviewing the manuscript and providing valuable comments and suggestions.

References

[1] Schach E, Buchmann M, Tolosana-Delgado R, Leißner T, Kern M, Van Den Boogaart G, Rudolph M, Peuker UA. Multidimensional characterization of separation processes – Part 1: introducing kernel methods and entropy in the context of mineral processing using SEM-based image analysis. Miner Eng 2019; 137:78–86. https://doi.org/10.1016/j.mineng.2019.03.026.

[2] Shouwastra RP, Smith AJ. Devepments in mineralogical techniques – what about mineralogists? Miner Eng 2011;24:1224–8. https://doi.org/10.1016/j.mineng.2011.02.002.

[3] Chryssoulis SL, Mcmullen J. Mineralogical investigation of gold ores. In: Adams Mike D, Wills BA, editors. Advances in gold ore processing, developments in mineral processing, vol. 15. Elsevier; 2005. p. 21–71.

[4] Chryssoulis SL, Mcintyre NS, Mycroft JR. Determination of gold in natural sulphide minerals using X-ray photoelectron spectroscopy. Canadian Mineralogist; 1993.

[5] Marsden J, House I. The chemistry of gold extraction. London: Ellis Horwood; 1992.

[6] Henley KJ. Ore-Dressing mineralogy - a review of techniques, applications and recent developments. Geological Society of South Africa 1983;7(7):175–200.

[7] Haberlah D, Owen M, Botha P, Gottlieb P. SEM-EDS based protocol for subsurface drilling mineral identification and petrological classification. 10th International Congress for applied mineralogy 2011;34:265–73.

[8] Lotter NO. Modern Process Mineralogy: an integrated multi-disciplined approach to flowsheeting. Miner Eng 2011;24(12):1229–37. https://doi.org/10.1016/j.mineng.2011.03.004.

[9] Goodall WR. Characterisation of mineralogy and gold deportment for complex tailings deposits using QEMSCAN. Miner Eng 2008;21:518–23. https://doi.org/10.1016/j.mineng.2008.02.022.

[10] Gu Y. Automated scanning electron microscope based mineral liberation analysis. An introduction to JKMRC/FEI Mineral Liberation Analyser. J Miner Mater Char Eng 2003;2(1):33–41.

[11] Petruk W. Applied mineralogy in the mining industry. Amsterdam; New York: Elsevier Science BV; 2000.

[12] Gottlieb P, Wilkie G, Shuterland D, Ho-Tun E, Suters S, Parera K, Jenkins B, Spencer S, Butcher A, Rayner J. Using quantitative electron microscopy for process mineralogy applications. Jom 2000;52(4):24–5.

[13] Jones MP. Applied mineralogy - a quantitative approach. United States: Graham & Trotman; 1987.

[14] Gaudin AM. Principles of mineral dressing. New Delhi: Tata McGraw Hill; 1939. p. 554.

[15] Costa FR, Nery GP, Carneiro CC, Kahn H, Ulsen C. Mineral characterization of low-grade ore to support geometallurgy. J Mater Res Technol 2022;21:2841–52. https://doi.org/10.1016/j.jmrt.2022.10.085.

[16] Chaube S, Goverapet Srinivasan S, Rai B. Applied machine learning for predicting the lanthanide-ligand binding affinities. Sci Rep 2020;10(1):14322. https://doi.org/10.1038/s41598-020-71255-9.

[17] Jain D, Chaube S, Khullar P, Srinivasan SG, Rai B. Bulk and surface DFT investigations of inorganic halide perovskites screened using machine learning and materials property databases. Phys Chem Chem Phys 2019;21(35):19423–36. https://doi.org/10.1039/C9CP03240A.

[18] Kaushik J, Dodagoudar GR. Explainable machine learning model for liquefaction potential assessment of soils using XGBoost-SHAP. Soil Dynam Earthq Eng 2023; 165:1–22. https://doi.org/10.1016/j.soildyn.2022.107662.

[19] Zheng L, Wang C, Chen X, Song Y, Meng Z, Zhang R. Evolutionary machine learning builds smart education big data platform: data-driven higher education. Appl Soft Comput 2023;136:1–10. https://doi.org/10.1016/j.asoc.2023.110114.

[20] Oliver S, Willingham D. Maximise orebody value through the automation of resource model development using machine learning. In: Perth, W.A. (Ed.). The third AusIMM international geometallurgy conference 2016. Australia, pp. 295–301..

[21] Suthaharan S. Machine learning models and algorithms for big data classification: thinking with examples for effective learning. Integr Series Inform. Syst. 2016;;36: 359.

[22] Gomez-Flores A, Ilyas S, Heyes GW, Hyunjung K. A critical review of artificial intelligence in mineral concentration. Miner Eng 2022;189:1–16. https://doi.org/10.1016/j.mineng.2022.107884.

[23] Li C, Wang D, Kong L. Application of machine learning tecniques in mineral classification for scanning electron microscopy – energy dispersive X-ray spectroscopy (SEM-EDS) Imagens. Journal of pretroleuim science and engineering 201;200:2-13..

[24] Daware S, Chandel S, Rai B. A machine learning framework for urban mining: a case study on recovery of copper from printed circuit boards. Miner Eng 2022;180:1–8. https://doi.org/10.1016/j.mineng.2022.107479.

[25] Lishchuk V, Lund C, Ghorbani Y. Evaluation and comparison of different machine-learning methods to integrate sparse process data into a spatial model in geometallurgy. Miner Eng 2019;134:156–65. https://doi.org/10.1016/j.mineng.2019.01.032.

[26] Koch PH, Lund C, Rosenkranz J. Automated drill core mineralogical characterization method for texture classification and modal mineralogy estimation for geometallurgy. Miner Eng 2019;136:99–109. https://doi.org/10.1016/j.mineng.2019.03.008.

[27] Tiu G. Classification of drill core textures for process simulation in geometallurgy. Aitik Mine, New Boliden: Luleå University of Technology; 2017.

[28] Haykin S. Redes neurais artificiais. Princípios e prática. 2. Porto Alegre: Bookman; 2001.

[29] Kohonen T. Self-Organizing maps. second ed. Berlin: Springer; 1997.

[30] Fraser SJ, Dickson BL. A new method for data integration and integrated data interpretation: Self-organizing maps: 5th Decennial Inter- national Conference on Mineral Exploration. Expanded Abstracts; 2007. p. 907–10.

[31] Taha K. Semi-supervised and un-supervised clustering: a review and experimental evaluation. Inf Syst 2023;114:1–34. https://doi.org/10.1016/j.is.2023.102178.

[32] Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. J Artif Intell Res 1996;4:237–85.

[33] Russell SJ, Norvig P. Artificial intelligence: a modern approach (AIMA). third ed. Prentice Hall; 2009.

[34] Costa FR, Carneiro CC, Ulsen C. Imputation of gold recovery data from low grade gold ore using artificial neural network. Minerals 2023;13(3):340. https://doi.org/10.3390/min13030340.

[35] Hao H, Guo R, Gua Q, Hu X. Machine learning application to automatically classify heavy minerals in river sand by using SEM/EDS data. Miner Eng 2019;143. https://doi.org/10.1016/j.mineng.2019.105899.

[36] McCoy JT, Auret L. Machine learning applications in minerals processing: a review. Miner Eng 2019;132:95–109.

[37] Taunk K, De S, Verma S, Swetapadma A. A brief review of nearest neighbor algorithm for learning and classification. International conference on intelligent computing and control systems (ICCS). IEEE; 2019. https://doi.org/10.1109/ICCS45141.2019.9065747.

[38] Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE Trans Inf Theor 1963;3(1):21–7.

[39] Vapnik V, Golowich S, Smola A. Support vector method for function approximation. Regression estimation and signal processing. Adv Neural Inf Process Syst 1997;9(9):281–7.

[40] Breiman L. Random forests. Mach Learn 2001;45:5–32. https://doi.org/10.1007/9781441993267_5.

[41] Ghamisi P, Plaza J, Chen Y, Li J, Plaza AJ. Advanced spectral classifiers for hyperspectral images: a review. IEEE Geosci. Remote Sens. Mag 2017;5:8–32. https://doi.org/10.1109/MGRS.2016.2616418.

[42] Platt JC. Fast training of support vector machines using sequential minimal optimization. In: Schcolkopf B, Burges C, Smola A, editors. Advances in kernel methods: support vector machines. Cambridge, MA: MIT Press; 1998.

[43] Kayadelen C, Altay G, Önal S, Önal y. Sequential minimal optimization for local scour around bridge piers. Mar Georesour Geotechnol 2022;40(4):462–72. https://doi.org/10.1080/1064119X.2021.1907635.

[44] Gershmana SJ, Blei DM. A tutorial on Bayesian nonparametric models. J Math Psychol 2012;56(1):1–12. https://doi.org/10.1016/j.jmp.2011.08.004.

[45] Rasmussen CE, Williams CKI. Gaussian Processes for machine learning. the MIT Press; 2006.

[46] Williams CKI. Computation with infinite neural networks. Neural Comput 1998;10(5):1203–16.

[47] García MV, Aznarte JL. Shapley additive explanations for NO2 forecasting, Ecol. Inform 2020;56:1–12.

[48] Wen Z, Zhou C, Pan J, Nie T, Zhou C, Lu Z. Deep learning-based ash content prediction of coal flotation concentrate using convolutional neural network. Miner Eng 2021;174:1–14. https://doi.org/10.1016/j.mineng.2021.107251.

[49] Liu X, Aldrich C. Explaining anomalies in coal proximity and coal processing data with Shapley and tree-based models. Fuel 2023;335:1–16.

[50] Lundberg SM, Lee SI. Consistent feature attribution for tree ensembles. Proceedings of the 34 th international conference on machine learning. Sydney, Australia: JMLR: W&CP; 2017.

[51] Frank E, Hall MA, Witten IH. The WEKA workbench. Online appendix. Data mining: practical machine learning tools and techniques". fourth ed. Burlington: Morgan Kaufmann; 2016.

[52] Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965;52:591–611.

[53] Spearman C. General Intelligence, objectively determined and measured. Am J Psychol 1904;15:201–93.

[54] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Int Joint Conf Artif Intell 1995;14(2):1137–45.