Machine Learning Applied to Apatite Compositions for Determining Mineralization Potential

Yu-yu Zheng¹, Bo Xu¹²³, David R. Lentz⁴, Xiao-yan Yu¹, Zeng-qian Hou⁵ and Tao Wang⁵

¹ School of Gemology, China University of Geosciences Beijing, 29 Xueyuan Road, Beijing 100083, China.
² State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Beijing 100083, China.
³ Frontiers Science Center for Deep-time Digital Earth, China University of Geosciences (Beijing), Beijing 100083, China.
⁴ Department of Earth Sciences, University of New Brunswick, P.O. Box 4400, Fredericton, NB E3B 5A3, Canada.
⁵ Beijing SHRIMP Center, Institute of Geology, Chinese Academy of Geological Sciences, Beijing 100037, China.

Corresponding author: Bo Xu (xubo@outlook.com.cn)
ABSTRACT

Apatite major and trace element chemistry is a widely used tracer of mineralization, as it sensitively records the characteristics of the magmatic-hydrothermal system at the time of its crystallization. Previous studies have proposed useful indicators and binary discrimination diagrams to distinguish between apatites from mineralized and unmineralized rocks; however, their efficiency has been found to be somewhat limited in other systems and larger scale datasets. This work applied a machine learning (ML) method to classify the chemical compositions of apatites from both fertile and barren rocks, aiming to help determine the mineralization potential of unknown system. Approximately 13,328 apatite compositional analyses were compiled and labeled from 241 locations in 27 countries worldwide, and three apatite geochemical datasets were established for XGBoost ML model training. The classification results suggest that the developed models (accuracy: 0.851–0.992; F1 score: 0.839–0.993) are much more accurate and efficient than conventional methods (accuracy: 0.242–0.553). Feature importance analysis of the models demonstrates that Cl, F, S, V, Sr/Y, V/Y, Eu*, (La/Yb)_N, and La/Sm are important variables in apatite that discriminate fertile and barren host rocks and indicates that V/Y and Cl/F ratios and the S content, in particular, are crucial parameters to discriminating metal enrichment and mineralization potential. This study suggests that ML is a robust tool for processing high-dimensional geochemical data and presents a novel approach that can be applied to mineral exploration.

Keywords: Apatite; Major and Trace Element; Machine Learning; Mineralization Potential; XGBoost
INTRODUCTION

Apatite \( (\text{Ca}_5\text{PO}_4)_3[\text{F,Cl,OH}] \) is a ubiquitous accessory mineral in most igneous and metamorphic rocks and derived elasic sediments and is relatively resistant to weathering (O'Sullivan et al., 2020). It is considered to be an ideal indicator mineral, given its chemical composition sensitivity to the crystallization environment (Bruand et al., 2017; Mao et al., 2016). Trace elements and volatile chemistry and isotopic signature of apatites can characterize diverse crystallization environments, including magmatic systems (Cao et al., 2022; Gao et al., 2020; Li et al., 2021; Long et al., 2023; Palma et al., 2019; Qu et al., 2021; Tang et al., 2021; Xu et al., 2023; Zhang et al., 2021), low-grade metamorphic systems (Bea and Montero, 1999; El Korh et al., 2009; Henrichs et al., 2018; Nutman, 2007), and sedimentary environments (Joosu et al., 2016). Accordingly, the trace element chemistry of apatite is widely used to characterize the lithology of source rocks (Belousova et al., 2002), including tracing detrital provenance (Bruand et al., 2017; Dill, 1994; O'Sullivan et al., 2018; O'Sullivan et al., 2020), and used to constrain petrogenetic process (Chu et al., 2009; La Cruz et al., 2020; Sun et al., 2022; Tollari et al., 2008; Zafar et al., 2019), especially for revealing the origin and evolution of magma (Gao et al., 2020; Meng et al., 2021; O'Reilly and Griffin, 2000). Moreover, the major and trace element chemistry of apatite is applied to mineral exploration (Belousova et al., 2002; Cao et al., 2012; Mao et al., 2016; Sha and Chappell, 1999; Xu et al., 2015). A series of indicators, including \( \text{Sr}/\text{Y}, \text{Mn}, \text{Eu}/\text{Eu}^*, \text{Th}/\text{U}, \text{La}/\text{Sm}, \) and \( (\text{Ce}/\text{Yb})_N \) (Belousova et al., 2002), and several binary classification diagrams, such as \( \text{Sr} \) vs. \( \text{F} \) (\( \text{Mn}, \text{Y}, (\text{La}/\text{Yb})_N, \text{Eu}/\text{Eu}^* \)), \( \text{F}/\text{Cl} \) vs. \( \text{F} \) (Azadbakht et al., 2018; Cao et al., 2012; Zhong et al., 2018), \( \text{Cl} \) vs. \( \text{Eu}/\text{Eu}^* \) (Mao et al., 2016), \( \text{V}/\text{Y} \) vs. \( \text{REE}+\text{Y}, \text{Cl} \) vs. \( \text{SO}_3 \), and \( ^{87}\text{Sr}/^{86}\text{Sr} \) vs. \( \text{Cl}/\text{F} \), are commonly used to diagnose the metallogenic fertility of magmatic rocks (Xu et al., 2021). Unfortunately, as interest in apatite has recently increased and numerous major
and trace element data have been reported (Adlakha et al., 2018; Bruand et al., 2019; Cao et al., 2022; Chakhmouradian et al., 2017; Chen and Zhang, 2018; Chen et al., 2019; Gao et al., 2020; Glorie et al., 2019; Henrichs et al., 2018; Hoshino et al., 2017; La Cruz et al., 2020; Li et al., 2021; Liu et al., 2021; Long et al., 2023; Lupulescu et al., 2017; Meng et al., 2021; Mercer et al., 2020; Palma et al., 2019; Qu et al., 2021; Sun et al., 2022; Tang et al., 2021; Xie et al., 2018; Xu et al., 2023; Yang et al., 2018; Zafar et al., 2019; Zhang et al., 2021), it is challenging to validate these individual indicators and binary discrimination techniques due to a large area overlap of compositional spots, suggesting that those traditional low-dimensional classifiers that seemed to work well in specific systems might be invalid in other systems or datasets of larger scales. Consequently, a novel data processing method that can handle high-dimensional compositional data is imperative for identifying robust indices to aid in exploring various systems for new mineral resources.

The field of machine learning (ML) encompasses the use of computational algorithms to discern patterns within datasets, which can subsequently be applied to make predictions. ML offers a robust toolkit for decoding latent information within high-dimensional data. Over the past few years, there has been an explosion of interest in the applications of ML to solid Earth geoscience (Li et al., 2023). ML has been widely applied in earthquake phase detection and seismicity classification (Cianetti et al., 2021; Linville, 2022), geophysical data processing and image interpretation (Xiao et al., 2021), geophysical inversion (Cai et al., 2022), and multi-physical and multidisciplinary information integration. Given the complexity and diversity of geochemistry data, ML-based classification methods have emerged as a promising approach that outperforms conventional methods, especially in large-scale geological processes, such as in predicting mantle metasomatism worldwide (Qin et al., 2022), revealing source compositions of intraplate
basaltic rocks (Guo et al., 2021), identifying primary water concentrations in mantle pyroxene (Chen et al., 2021), determining the quartz-forming environments (Wang et al., 2021), and classifying the source rocks of detrital zircons (Zhong et al., 2023a, 2023b). In the field of mineral exploration, two studies tried to apply ML to characterize magma fertility based on zircon compositional data, aiming to identify porphyry copper mineralization potential (Zhou et al., 2022; Zou et al., 2022). Tan et al. (2023) employed partial least squares discriminant analysis (PLS-DA) to the apatite trace element dataset (4,298 data) to distinguish between apatites from different types of deposits and rocks. Their plots could not directly discriminate ore magmatic and hydrothermal apatites, but showed a great potential in classifying barren and ore magmatic apatites from granitoid-related deposits and highlighted the role of V, Eu, and Sr for classification.

Here, three global datasets of the major and/or trace element chemistry of apatites were compiled from both mineralized and unmineralized rock samples, and a series of XGBoost models were trained to determine the mineralization potential. The classification results compared with traditional binary diagrams demonstrated an improvement in accuracy and efficiency in discriminating whether apatite is derived from a fertile rock suite or a barren suite. In addition, the feature importance analysis suggested that V/Y and Cl/F ratios and the S content are crucial to metal enrichment and mineralization.
DATA COMPILATION AND LABELING

Data Compilation

All of the apatite compositional data used for modeling were collected and compiled from 241 locations in 27 countries worldwide (Figure 1) from preexisting literature. Each location included multiple samples and analyses. This raw dataset (Table S1) contains 13,382 rows of compositional data, including spot analyses and mean values for those references in which spot analyses were not given. Figure 2 shows an overview of the elements and geochemical parameters contained in this dataset. All the data and related sources can be found at https://github.com/YuyuJo/Supplementary-file-for-AM-9115R.

Data Labeling

Analyses of apatites collected from rock samples with obvious mineralization that formed in association with mineral deposits were labeled as “Mineralized.” The analyses of apatites in barren rocks were labeled as “Unmineralized.” Descriptions regarding the deposit and rock samples (including location, mineralization, and alteration information) can be easily found in the literatures when collecting apatite data. Based on these criteria, 9,104 and 4,278 analyses were labeled as “Mineralized” and “Unmineralized,” respectively. The deposit type of each data was also identified based on the classification in Mao et al. (2016). For data with “Mineralized” labels, their deposit types included porphyry (no. = 2,251), skarn (5,075), orogenic Au (875), carbonatite deposits (207), iron oxide Cu–Au (IOCG, 80), Kiruna type (IOA, 579), orogenic Ni–Cu ± platinum group element (28), and epithermal Au–Ag (9).

Always consult and cite the final, published document. See http://www.minsocam.org or GeoscienceWorld
Sub-Dataset Construction

The raw dataset was divided into three subsets to further differentiate the role of major and trace elements. The analyses of samples containing CaO, P$_2$O$_5$, SO$_3$, Cl, and F were selected as “Major” dataset, and the analyses of samples including trace elements were selected as “Trace” dataset. The analyses with both major and trace elements were set into the “Major and Trace” dataset.

To preprocess the collected data, the initial step involved handling the missing values, whereby any element with missing values >60% of the entire column was excluded. After this filtering, the “Major” dataset comprised 5,618 analyses (Table S2). The features therein included CaO, P$_2$O$_5$, SO$_3$, F, Cl, FeO, MnO, Na$_2$O, SiO$_2$, and Cl/F. The “Trace” dataset (Table S3) contained 9,979 data and included V, Mn, Rb, Sr, Y, Zr, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, and Lu. Additionally, certain geochemical parameters, which are considered significant for mineralization and magma evolution, were computed and added to the “Trace” dataset. These parameters included LREE, HREE, Sr/Y, V/Y, Ce/Nd, Eu*, Ce*$_N$, Eu$_N$/Eu*$_N$, Ce/Ce*, Eu/Eu*/Y, REE+Y, (La/Yb)$_N$, and La/Sm. The “Major and Trace” dataset (Table S4) included 2,448 analyses and 43 features (CaO, P$_2$O$_5$, SO$_3$, F, Cl, FeO, Cl/F, SiO$_2$, Na$_2$O, MgO, Rb, Sr, Y, Zr, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Th, U, Sr/Y, V/Y, Ce/Nd, Eu*, Ce*$_N$, Eu$_N$/Eu*$_N$, Ce/Ce*, Eu/Eu*/Y, REE+Y, (La/Yb)$_N$, La/Sm, LREE, HREE).
METHODS

ML Algorithms

XGBoost is an ML system based on gradient tree boosting, which was originally proposed by Friedman (2001). It has gained widespread recognition in numerous ML and data mining challenges due to its ability to solve real-world-scale problems using minimal resources (Chen and Guestrin, 2016; Python et al., 2021). XGBoost is a distributed gradient boosting library that has been optimized for high efficiency and flexibility. Its flexibility is exemplified by its ability to handle sparse data with multiple possible causes, including missing values and frequent zeros. In addition, its parallel and distributed computing capabilities facilitate faster learning, enabling quicker model exploration. The highly scalable end-to-end tree boosting system allows for efficient scaling to larger datasets with minimal cluster resources (Chen and Guestrin, 2016). Moreover, the tree structure of XGBoost enables the identification of important features and enhances the interpretability of results (Azodi et al., 2020; Qin et al., 2022), which is beneficial in elucidating the relationship between apatite composition and mineralization and exploring the geochemical implications.

XGBoost is an ML algorithm that operates under a gradient boosting framework. Its training methodology is additive, with each new tree added to fit the residuals of the prior predictions. The results of all the trees are summed up to obtain the final predictions. Given a dataset $D = \{(x_i, y_i) \} (|D| = n, x_i \in R^m, y_i \in R)$ with n examples and m features, the output of a tree ensemble model that uses K additive functions is predicted as a sum of k times scores:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(X_i), \ f_k \in \mathcal{F},$$ (1)
where $\mathcal{F} = \{ f(x) = w_q(x) \}$ ($q: \mathbb{R}^m \rightarrow T$, $w \in \mathbb{R}^T$) represents the space of regression trees, the function $q$ denotes the structure of each tree that maps an example to the corresponding leaf index, $T$ is the number of leaves in the tree, each $f_k$ corresponds to an independent tree structure $q$ and leaf weights $w$, and $w_i$ represents the score on the $i$-th leaf (Chen and Guestrin, 2016).

The following regularized objective is constructed and minimized to evaluate the quality of a tree structure $q$:

$$
\mathcal{L}(\varphi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (2)
$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda \| w \|^2$. The regularization term penalizes the complexity of the model and helps to smooth the final learned weights to avoid overfitting. The parameter $\gamma$ controls the degree of regularization, while $\lambda$ controls the strength of the penalty. In this equation, $l$ is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target $y_i$.

**Model Construction Processes**

A four-step modeling process was employed to construct a classification model that best fitted the apatite compositional data (Figure 3).

**Data Preprocess and Splitting.** All three sub-datasets were used to train the classification models. Taking the “Trace” dataset as an example, the elemental data were used as input without any transformation. The inputed dataset was first split into “Features” and “Class” subsets, which...
were uniformed as 0 (Unmineralized) and 1 (Mineralized) using the “LabelEncoder” function. Maintaining the original proportion of each class, both subsets were randomly split into training (80%) and test (20%) sets.

**XGBoost Modeling.** To avoid overfitting, a fivefold cross-validation (Kohavi, 1995) was employed to train the model. The training set was divided into five folds of equal sizes, where four subsets were used to train the ML model, and the left-out fold was used for validation and classification evaluation. This process was repeated five times, with each validation fold being different, and the output score represented the mean value of all five predictions.

**Model Hyperparameter Tuning.** A fivefold cross-validation approach was utilized in conjunction with a grid search strategy to optimize the XGBoost model. This strategy exhaustively generated candidates from a grid of parameter values and selected the candidate with the highest output scores, as evaluated by a predefined metric. Specifically, the goal of the grid search procedure was to identify the optimal combination of hyperparameters (eta, gamma, max depth, and alpha) and to generate 3,600 candidates from which the optimal model was selected.

**Model Validation and Evaluation.** Predictions were obtained by applying the test set to the above optimal XGBoost model. To clearly observe the classification results, the predictions were generally displayed as a confusion matrix (Stehman, 1997), of which the rows represent the true number of each class (from labeled dataset) and the columns display the predicted number of each class. The commonly used classification metrics for evaluating the model performance can be calculated based on the confusion matrix. Here, the accuracy and the F1 score were used as the evaluation indicators of the model. For the convenience of description, true “Mineralized,” which was also predicted as “Mineralized,” is abbreviated as “MM”; true “Mineralized,” which
was falsely predicted as “Unmineralized,” is abbreviated as “MU”; true “Unmineralized,” which
was also predicted as “Unmineralized,” is abbreviated as “UU”; and true “Unmineralized,”
which was falsely predicted as “Mineralized,” is abbreviated as “UM.”

Accuracy is a metric that measures the number of correctly predicted cases relative to the total
number of samples used. It is calculated as the ratio of the number of correct predictions to the
sum of all the utilized samples, which can be expressed as

\[
\text{Accuracy} = \frac{MM + UU}{MM + MU + UU + UM} \tag{3}
\]

The F1 score is a measure of classification accuracy that combines precision and recall. Specifically, it is the harmonic mean of precision and recall and is expressed as

\[
F1 \text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}
\]

Precision is a measure of the accuracy of predictions, and it indicates the probability that a sample is truly positive among all samples predicted to be positive. Taking class “Mineralized” as an example, precision can be calculated as follows:

\[
\text{Precision} = \frac{MM}{MM + UM} \tag{5}
\]

Recall is a measure of how well the classifier identifies the actual positive cases, and it indicates the probability that a sample predicted to be positive is actually positive. Taking class “Mineralized” as an example, recall can be calculated as follows

\[
\text{Recall} = \frac{MM}{MM + MU} \tag{6}
\]
RESULTS

Classification Results and Feature Importance

A total of 14 XGBoost models were trained using three apatite compositional datasets based on different feature selection. Five models were trained using the “Major and Trace” dataset, and the number of selected features was 43, 35, 22, 12, and 6, respectively (Table 1). Two models were trained using the “Major” dataset. All the ten major elements were applied to train model M-1, while four selected elements to model M-2. The “Trace” dataset was thought to be important to identify mineralization and was thus employed to train seven XGBoost models. The related feature numbers were sequentially set as 33, 28, 21, 14, 7, 3, and 2. The classification results of these XGBoost models are displayed as confusion matrices. Figure 4 shows the results of the four representative models. Table 1 presents the F1 scores and accuracies of 14 models, which were calculated based on the classification confusion matrices.

The relative importance of all the features used in each model was obtained from the XGBoost algorithm to determine the elements of apatite that are highly relevant to mineralization (Table 1). Among all the models for the “Trace” dataset, V, Sr, Y, Eu, Ce, and Rb appeared most frequently in the top ten feature relative importance. Vanadium was particularly important in the rank. Of all five models in which V was selected, the relative importance of V was highest in four and ranked second in the remaining one. Some geochemical parameters also contributed to the rank, including Sr/Y, V/Y, Eu*, (La/Yb)\text{N}, and La/Sm. In the two models for the “Major” dataset, the content of SO\textsubscript{3} showed the highest relative importance. However, the proportion of each feature was quite uniform. The features that played a key role in the models for the “Major
and Trace” dataset were somehow familiar to those in the models for the “Trace” dataset. In addition, Cl, F, and Cl/F were also noteworthy.

**Feature Selection and ML Model Performance Evaluation**

The variance of classification metrics presented in Table 1 has been shown to be related to the input dataset type and the number of selected features in the model. The highest F1 scores of the models are 0.9939 and 0.9933 obtained by models T-1 and M-T-1, respectively, while that of models trained by the “Major” dataset is 0.9259 obtained by model M-1. The accuracies of T-1 and M-T-1 are 0.9900 and 0.9918, respectively, while that of M-1 is 0.9386. This indicates that the models trained using the “Trace” and “Major and Trace” datasets achieved better performance than those using the “Major” dataset.

The classification results in Table 1 also suggest a positive correlation between feature number and model performance. As displayed in Figure 5, the XGBoost models achieved higher scores when they are trained on more elements and geochemical parameters, which was also observed in other research (Qin et al., 2022). For instance, the accuracy and F1 score increased from 0.9146 and 0.8507 for model T-7 (no. of features = 2) to 0.9682 and 0.9474 for model T-5 (no. of features = 7) and 0.9939 and 0.9900 for model T-3 (no. of features = 33).

Furthermore, feature selection also showed a salient effect on the model performance, as evidenced by models M-T-4 and T-3. The features selected to train model M-T-4 were almost the same as the crucial elements and geochemical parameters summarized in Section 4.1, leading to the result that although the model was trained only on 12 features, its score is slightly higher.
than that of M-T-2 and M-T-3 that were trained on 35 and 22 features, respectively. This effect
was even more pronounced in models trained using “Trace” datasets. Model T-4 was trained on a
feature dataset consisting entirely of 14 rare-earth elements, but scored worse than model T-5
(no. of features = 7). This indicates that REEs might not play a positive role in training the
XGBoost model for discriminating between fertile and barren apatites in general. As shown in
Figure 5, the performance of model T-3 trained on 21 features without REEs was better than that
of T-2 trained with 28 features with REEs.

On the whole, 10 of the 12 models could correctly classify more than 90% samples of the test set
(accuracy > 0.9), indicating the excellent performance of the models in this study for
distinguishing between “Mineralized” and “Unmineralized” apatites. Among all 14 models,
model M-T-1 achieved the highest scores for both training and test sets. In the results of this
model, all samples in the training set were correctly classified (accuracy = 1), and more than
99% samples of the test set were correctly classified (accuracy = 0.9918). Elemental data
obtained in practice might not be enough as those used for model M-T-1; however, model M-T-4
can achieve a similar performance with an accuracy and an F1 score of 0.9878 and 0.900,
respectively, when only using 9 elements (12 features). This demonstrates that the classification
models in this study can work in a variety of situations. However, the XGBoost model
performance sharply declined when the number of selected features decreased to 2 (Figure 5).
Given the overall classification results, it was clear that the XGBoost models in this study can
achieve excellent performance when proper feature selection was performed and can be applied
in various scenarios.
DISCUSSION

Limitation of Conventional Apatite Fertility Indicators

Previous research has suggested that magmas characterized by high water content, high Sr/Y ratio, and high oxidation state play a vital role in the genesis of porphyry Cu deposits (Lu et al., 2015; Richards, 2011; 2015). Recent investigations have indicated that chlorine and sulfur are crucial components of ore-forming fluids due to their ability to form complexes with ore metals, including Cu, Au, Pb, Zn, Fe, and Mo (Hsu et al., 2019; Xu et al., 2021; 2022). These geochemical signatures of magma could be inherited by apatite crystallized from such fertile magmas. Accordingly, various apatite fertility indicators, such as Sr/Y, \((\text{Ce/Yb})_N\), Cl/F, V/Y, and \((\text{Ce/Pb})_N\), have been proposed to distinguish between fertile and barren suites (Belousova et al., 2002; Mao et al., 2016; Xu et al., 2021). In this study, the performance of several traditional apatite fertility indicators was evaluated using the raw dataset (Figure 6). However, the efficacy of these indicators in predicting mineralization at a global scale was found to be limited, despite their effectiveness in specific metallogenic systems, such as porphyry deposits; for instance, Xu et al. (2021) proposed three indicators in apatite that worked effectively in differentiating fertile and barren porphyries. However, it only showed the best accuracy of 0.553 (Figure 6a) when applied to the dataset in this study. More precisely, the Cl/F ratio-based classification (Figure 6a) yielded a true-positive rate (TPR) of 0.421 for fertile apatite and a true-negative rate (TNR) of 0.580 for barren apatite. The biplot of V vs. Y (Figure 6b) has an accuracy, TPR, and TNR of 0.261, 0.866, and 0.026, respectively, indicating its ability to identify fertile apatite but not barren apatite. Comparative scores are 0.423, 0.007, and 0.919 for Cl/F vs. \((\text{Ce/Pb})_N\) biplot (Figure 6c), 0.242, 0.181, and 0.640 for V/Y vs. \((\text{Ce/Pb})_N\) biplot (Figure 6d), and 0.299, 0.172,
and 0.750 for V/Y vs. Cl/F biplot (Figure 6e), indicating that the (Ce/Pb)$_N$ ratio might perform better in determining barren apatite.

The traditional discrimination diagrams exhibit low accuracies (from 0.242 to 0.553) on a global scale dataset and could result in inconclusive findings and imprecise mineralization targets when applied to mineral exploration. As an increasing amount of geochemical data pertaining to apatite becomes publicly accessible, the limitations of individual geochemical indicators are becoming progressively conspicuous. One major limitation is the lack of transferability of a geochemical indicator that accurately identifies fertile rocks in one system to another. Additionally, traditional methods that combine limited indicators could not comprehensively introduce the underlying pattern of multiple elements and efficiently assess the metallogenic fertility.

Accordingly, an ML model that can process high-dimensional geochemical data is considered to be a robust mineral exploration tool. The XGBoost models in this study are apparently more accurate and efficient than traditional elemental biplots with accuracies ranging from 0.8507 to 0.9918, suggesting a higher success rate during prospecting and exploration. In addition, ML can simultaneously integrate all apatite trace element features and directly capture the relationships between geochemical data and mineralization. The advantage of this approach is the applicability of results to any geological environment, while the disadvantage is that it required a systematic and comprehensive apatite geochemical dataset from worldwide occurrences. With the growth in the volume of geochemical data on apatite from various deposit types, ML models trained on such datasets are likely to become more sophisticated and accurate. This is because ML
algorithms excel at identifying complex patterns and relationships in large datasets, which can lead to more precise discrimination of mineralization. As a result, hopes are high for the robustness of this ML approach in mineral exploration. With the continuous expansion of geochemical databases and the ongoing refinement of ML algorithms, further improvements in the performance of these models are expected.

**Model Application and Limitation**

To further substantiate the reliability of our model and elucidate its potential applications, a set of unlabelled apatite compositional data reported by Xu et al. (2021) was employed as a validation dataset (Table S5). These apatites were extracted from rock samples collected in 12 distinct localities spanning Iran and western China (Tibet and Yunnan) which include both barren localities (Liuhe, Nanmuqie, Renduoxiang, Songgui, Wolong) and porphyry deposits (Beiya, Chongmuda, Jiama, Machangqing, Masjed Dagheri, Qulong, Zhunuo). In an effort to strike a balance between feature quantity and model performance, we utilized the cost-effective model M-T-5 to predict the fertility of the host rocks of these apatites and the corresponding mineralization potential in the respective regions. To render these data points visually interpretable, Principal Component Analysis (Smith, 2002) was employed to reduce their dimensionality to two dimensions. Figure 7a illustrates that their distribution primarily aligns with the clusters of the “Major & Trace” subset, signifying the comprehensive spectrum covered by our established database.

After applying model M-T-5 on the validation dataset and then organizing the resulting probability values and prediction results (Figure 7b-f and Table 2), the robust performance of our model was reconfirmed. As evident from Figure 7b and Table 2, apatites originating from fertile
rock occurrences in all seven instances were accurately identified, with five of them exhibiting a
100% likelihood of mineralization, while the remaining two displayed probabilities exceeding
70%. Additionally, four barren rock samples were predicted with high probabilities, whereas the
prediction results for apatite from Renduoxiang displayed suboptimal performance.

These cases emphasize the optimistic perspective for utilizing ML models and the compositional
data of apatite to predict the mineralization potential in this region. However, it is crucial to note
that our current database exhibits a degree of data imbalance, with a predominant proportion of
apatite data from porphyry and skarn deposits. This imbalance may render our models more
sensitive to these two systems. Hence, it is imperative to conduct a thorough examination of data
distribution before deploying the models, and data clusters that substantially deviate from our
database's distribution should be used judiciously. This highlights the need for larger-scale and
more diverse datasets, which relies on contributions of more geological researchers.

**Geochemical Explanation for ML Model**

The feature importance ranks of 14 XGBoost models reveal that several indicators are highly
relevant with mineralization based on our dataset, including Cl, F, S, V, Sr/Y, V/Y, Eu*,
(La/Yb)N, and La/Sm. This is in consistent with the conclusions of previous studies (Lu et al.,
2015; Richards, 2011; 2015; Xu et al., 2021; 2022). Several observations of higher Cl and S
contents in fertile than barren apatites (Chelle-Michou and Chiaradia, 2017; Xu et al., 2021; Zhu
et al., 2018) and fluid inclusion studies (Sillitoe, 2010) highlighted the significance of chlorine
and sulfur in supporting the transport and deposition of ore metals at magmatic hydrothermal
systems (Duan et al., 2021; Wang et al., 2021a; Zheng et al., 2021). These two elements form
ligands with ore metals, such as Cu, Au, Pb, Zn, Fe, and Mo, allowing their transport to the site
of ore deposition and are involved in causing hydrothermal alteration. Additionally, the V/Y ratio was proved to be high in apatite crystallized from ore-forming magma (Xu et al., 2021). The presence of elevated V contents in the host magma indicates high levels of dissolved H$_2$O in the melt, which was also recognized as a crucial factor for mineralization (Chiaradia, 2014; Lu et al., 2015). The mechanism therein is that amphibole has a more wide-ranging crystallization sequence than titanomagnetite in high H$_2$O content melt environment, leading to the retention of more V in the residual melt and the efficient extraction of Y from the melt into amphibole. Therefore, apatite crystallized from such magmas exhibited small negative Eu anomalies and high V/Y, reflecting early amphibole fractionation and suppression of plagioclase crystallization in hydrous melts (Davidson et al., 2007). Considering the feature importance analysis of this study and the suggestions of previous studies, Cl- and S-enriched hydrous magma seem to be crucial factors in metal enrichment and mineralization.

**IMPLICATIONS**

The XGBoost models developed in this study exhibit strong discriminatory power in distinguishing apatite samples from mineralized fertile suites and those from barren suites, with high accuracy and efficiency. This indicates that ML, when integrated with conventional geological and geochemical techniques, can offer a cost-effective and efficient approach to evaluate mineralization. Furthermore, this methodology holds potential for identifying fertile and barren magmas in other systems by other minerals, including quartz, zircon, and titanite. The versatility of ML models trained on different target variables can extend beyond solid Earth science to other fields.
In conclusion, this study demonstrates the efficacy of ML methods in capturing the intricate relationship between 43-dimensional apatite geochemical data and mineralization. The findings underscore the significant feasibility of ML in analyzing and processing high-dimensional data in solid Earth sciences, which could help elucidate underlying geological events.

However, the application of ML methods also has potential pitfalls. Firstly, ML algorithms do not include an inferential component such as an adequate assessment of uncertainty (Frenzel, 2023). Secondly, care of the breadth and representativeness of the data should be taken. Within different deposits of same type, factors such as the alteration degree of host rock and the mineral assemblage of apatite can significantly influence its composition. Furthermore, variations in the composition of rock samples from different locations in a deposit are noteworthy. Hence, the volume and representativeness of the data are of paramount importance. A small number of samples cannot adequately represent the characteristics of the entire deposit. Instead, a relatively large sample size is likely necessary to reasonably encompass the extent of the observed variability. In our database, there exists an imbalance in the quantity of "Mineralized" apatite data originating from different types of deposit. For instance, data from porphyry and skarn deposits are more abundant, while data from other deposit types are less represented. Consequently, this disparity may result in our model performing more effectively when applied to these two deposit types. Thirdly, interpretational pitfalls must be acknowledged. Owing to the influence of data imbalance and volume, it remains to be verified whether parameters identified by our models as highly sensitive to mineralization are equally effective across all ore systems.
ACKNOWLEDGMENTS

This research was funded by National Natural Science Foundation of China (42222304, 42073038), and the “Deep-time Digital Earth” Science and Technology Leading Talents Team Funds for the Central Universities for the Frontiers Science Center for Deep-time Digital Earth, China University of Geosciences (Beijing) (Fundamental Research Funds for the Central Universities; grant number: 2652023001), Young Talent Support Project of CAST, the Fundamental Research Funds for the Central Universities (Grant no. 265QZ2021012) and International Geoscience Programme (IGCP-741, IGCP-662). This is the XXth contribution of B.X. for National Mineral Rock and Fossil Specimens Resource Center. The authors would like to express their sincere gratitude to Jiali Lei for her valuable contributions in enhancing the model's performance. Furthermore, we extend our appreciation to Jiaxing Yu and Wenxuan Wu for their assistance in data collection. Their efforts significantly contributed to the successful completion of this study. Finally, we heartfeltly thank the Laboratory of the Jewelry College (China University of Geosciences, Beijing) for their assistance.

REFERENCES


Geochemistry, Geophysics, Geosystems, 19(4), 1309–1326.


FIGURE CAPTIONS

FIGURE 1. Representative locations of apatite samples of which analyses were collected in this study. The green points indicate that the analyses were collected from apatites that formed in mineralized rocks and were labeled as “Mineralized.” The orange points indicate that the analyses were collected from apatites in unmineralized rock samples and were labeled as “Unmineralized.”

FIGURE 2. Boxplots of major and trace elements and geochemical parameter data of apatite samples worldwide, expressed in wt.% (a) and ppm (b). The box represents the interquartile range (IQR), with the upper (75%) and lower (25%) quartiles demarcated. The outer whiskers extended to 1.5 times the IQR. A horizontal line within the colored box represents the median (50%). The black square symbols and circle symbols indicate the mean and outliers, respectively.

FIGURE 3. Workflow for the XGBoost modeling in this research. Step I: The labeled apatite dataset (“Major and Trace,” “Major,” or “Trace” dataset) is read as input and preprocessed and then randomly split into training set (80%) and test set (20%) by the holdout method. Step II: The training set is applied to train the XGBoost model using the fivefold cross-validation method. Step III: The optimal hyperparameters are determined by grid search techniques with the fivefold cross-validation method. Step IV: The best model obtained by Step III is applied to the test set. The classification results will be used to evaluate the model performance.
FIGURE 4. Confusion matrices (left) and feature importance ranks (right) of four representative XGBoost models. The confusion matrices display the prediction results for each class.

FIGURE 5. Correlation between feature selection and XGBoost model performance. n = number of selected features.


FIGURE 7. Distribution patterns of the validation dataset (a) and prediction results generated by model M-T-5; (a) distribution patterns displayed after dimension reduction using PCA for the "Major & Trace" dataset and validation dataset; (b) prediction results compilation for each occurrence; (c-f) probabilities of mineralization for the Songgui, Liuhe, Machangqing, and Jiama.
APPENDIX

The data sets and the code for the machine learning models developed in this study are available at https://github.com/YuyuJo/Supplementary-file-for-AM-9115R.
<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>No. of Data</th>
<th>No. of Fea.</th>
<th>F1&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Accuracy&lt;sup&gt;c&lt;/sup&gt;</th>
<th>F1</th>
<th>Accuracy</th>
<th>Top 10 feature importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-T-1</td>
<td>“Major &amp; Trace”</td>
<td>2448</td>
<td>43</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9933</td>
<td>0.9918</td>
<td>Y, Rb, Cl, Eu, Ce/Ce*, Tm, V/Y, Zr, Sr/Y, Ce</td>
</tr>
<tr>
<td>M-T-2</td>
<td>“Major &amp; Trace”</td>
<td>2448</td>
<td>35</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9850</td>
<td>0.9816</td>
<td>Y, V/Y, Rb, Cl, FeO, REE+Y, Sr/Y, Sr, Tb, Cl/F</td>
</tr>
<tr>
<td>M-T-3</td>
<td>“Major &amp; Trace”</td>
<td>2448</td>
<td>22</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9884</td>
<td>0.9857</td>
<td>Rb, FeO, Y, Cl, V/Y, Eu, Sr, Eu&lt;sub&gt;N&lt;/sub&gt;/Eu*&lt;sub&gt;N&lt;/sub&gt;, (La/Yb)&lt;sub&gt;N&lt;/sub&gt;, Na&lt;sub&gt;2&lt;/sub&gt;O</td>
</tr>
<tr>
<td>M-T-4</td>
<td>“Major &amp; Trace”</td>
<td>2448</td>
<td>12</td>
<td>0.9996</td>
<td>0.9995</td>
<td>0.9900</td>
<td>0.9878</td>
<td>V/Y, Cl/F, Y, Sr/Y, F, (La/Yb)&lt;sub&gt;N&lt;/sub&gt;, Sr, Eu, La/Sm</td>
</tr>
<tr>
<td>M-T-5</td>
<td>“Major &amp; Trace”</td>
<td>2448</td>
<td>6</td>
<td>0.9987</td>
<td>0.9985</td>
<td>0.9733</td>
<td>0.9673</td>
<td>V/Y, Cl/F, (La/Yb)&lt;sub&gt;N&lt;/sub&gt;, Sr/Y, Eu, SO&lt;sub&gt;3&lt;/sub&gt;</td>
</tr>
<tr>
<td>M-1</td>
<td>“Major”</td>
<td>5618</td>
<td>10</td>
<td>0.9921</td>
<td>0.9933</td>
<td>0.9259</td>
<td>0.9386</td>
<td>SO&lt;sub&gt;3&lt;/sub&gt;, Cl/F, Na:O, MnO, F, Cl, SiO&lt;sub&gt;2&lt;/sub&gt;, FeO, P&lt;sub&gt;2&lt;/sub&gt;O&lt;sub&gt;5&lt;/sub&gt;, CaO</td>
</tr>
<tr>
<td>M-2</td>
<td>“Major”</td>
<td>5618</td>
<td>4</td>
<td>0.9308</td>
<td>0.9424</td>
<td>0.8391</td>
<td>0.8639</td>
<td>SO&lt;sub&gt;3&lt;/sub&gt;, F, Cl/F, Cl</td>
</tr>
<tr>
<td>T-1</td>
<td>“Trace”</td>
<td>9979</td>
<td>33</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9939</td>
<td>0.9900</td>
<td>Eu*&lt;sub&gt;N&lt;/sub&gt;, V, Tb, Rb, Mn, Ce/Ce*, (La/Yb)&lt;sub&gt;N&lt;/sub&gt;, Sr/Y, Sr, Ce</td>
</tr>
<tr>
<td>T-2</td>
<td>“Trace”</td>
<td>9979</td>
<td>28</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9849</td>
<td>0.9750</td>
<td>V, Sr/Y, Eu/Eu*/Y, Ce/Ce*, La/Sm, (La/Yb)&lt;sub&gt;N&lt;/sub&gt;, Tb, Pr, Dy, Sr</td>
</tr>
<tr>
<td>T-3</td>
<td>“Trace”</td>
<td>9979</td>
<td>21</td>
<td>0.9998</td>
<td>0.9996</td>
<td>0.9912</td>
<td>0.9855</td>
<td>V, Ce/Ce*, Eu*, Zr, Sr, Ce*&lt;sub&gt;N&lt;/sub&gt;, V/Y, Rb, Mn, (La/Yb)&lt;sub&gt;N&lt;/sub&gt;</td>
</tr>
<tr>
<td>T-4</td>
<td>“Trace”</td>
<td>9979</td>
<td>14</td>
<td>0.9955</td>
<td>0.9926</td>
<td>0.9593</td>
<td>0.9319</td>
<td>Nd, Yb, La, Dy, Tb, Ce, Gd, Tm, Lu, Er</td>
</tr>
<tr>
<td>T-5</td>
<td>“Trace”</td>
<td>9979</td>
<td>7</td>
<td>0.9919</td>
<td>0.9866</td>
<td>0.9682</td>
<td>0.9474</td>
<td>V, Sr/Y, Sr, Y, Ce, V/Y, Eu</td>
</tr>
<tr>
<td>T-6</td>
<td>“Trace”</td>
<td>9979</td>
<td>3</td>
<td>0.9601</td>
<td>0.9345</td>
<td>0.9543</td>
<td>0.9243</td>
<td>V, Sr, Y</td>
</tr>
<tr>
<td>T-7</td>
<td>“Trace”</td>
<td>9979</td>
<td>2</td>
<td>0.9215</td>
<td>0.8626</td>
<td>0.9146</td>
<td>0.8507</td>
<td>Eu, Ce</td>
</tr>
</tbody>
</table>

<sup>a</sup>Number of elements and geochemical parameters used in the model. <sup>b</sup>Calculated by Equation 4. <sup>c</sup>Calculated by Equation 3.
TABLE 2. Prediction Results of Model M-T-5 Applied to Unlabeled Validation Set

<table>
<thead>
<tr>
<th>Occurrence</th>
<th>Sample source given by Xu et al. (2021)</th>
<th>Prediction results of apatite composition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Probability of barren rock</td>
</tr>
<tr>
<td>Beiya</td>
<td>Porphyry-skarn Cu-Au</td>
<td>30%</td>
</tr>
<tr>
<td>Chongmuda</td>
<td>Porphyry-hydrothermal Cu-Mo</td>
<td>0</td>
</tr>
<tr>
<td>Jiana</td>
<td>Porphyry-skarn Cu-Au</td>
<td>0</td>
</tr>
<tr>
<td>Machangqing</td>
<td>Porphyry Cu-Au</td>
<td>25%</td>
</tr>
<tr>
<td>Masjed</td>
<td>Porphyry Cu-Mo</td>
<td>0</td>
</tr>
<tr>
<td>Daghi</td>
<td>Porphyry Cu-Mo</td>
<td>0</td>
</tr>
<tr>
<td>Qulong</td>
<td>Porphyry Cu-Mo</td>
<td>0</td>
</tr>
<tr>
<td>Zhunuo</td>
<td>Porphyry Cu</td>
<td>0</td>
</tr>
<tr>
<td>Renduoxiang</td>
<td>Barren porphyry</td>
<td>50%</td>
</tr>
<tr>
<td>Liuhe</td>
<td>Barren granodiorite</td>
<td>74%</td>
</tr>
<tr>
<td>Nanmuqie</td>
<td>Barren porphyry</td>
<td>89%</td>
</tr>
<tr>
<td>Songgui</td>
<td>Barren porphyry</td>
<td>98%</td>
</tr>
<tr>
<td>Wolong</td>
<td>Barren granodiorite</td>
<td>100%</td>
</tr>
</tbody>
</table>
FIGURE 1
FIGURE 2
FIGURE 3

**STEP I. Data Pre-process**
- Labeled Dataset (total: 9979)
  - Minimized: 6194
  - Unminimized: 1785

  Pre-process

  Training Set (80%; n = 7963)
  Test Set (20%; n = 1996)

**STEP II. XGBoost Modeling**
- Divide into 5 groups randomly
- Five-fold Cross Validation

**XGBoost Model(q, γ, max_depth, α)**
- Average Validation:
  - F1 score
  - Precision
  - Recall
  - Accuracy

**STEP III. Parameter Optimization**
- Eta (α) value from [0.05, 0.1, 0.15, 0.2, 0.25, 0.3]
- Inverse gamma (γ) value from [0.1, 2] with step 0.1
- Max depth value from [3, 4, 5, 6, 7]
- Alpha (α) value from [0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4]

- Select the best model
  - (η = 0.25, γ = 0.1, max depth = 7, α = 0.1)

**STEP IV. Validation**
- Testing Set
- The Best XGBoost Model
- Predictions
FIGURE 4

(a) Model M-1 (features: n = 10)

<table>
<thead>
<tr>
<th>Mineralized</th>
<th>Unmineralized</th>
</tr>
</thead>
<tbody>
<tr>
<td>431</td>
<td>41</td>
</tr>
<tr>
<td>28</td>
<td>624</td>
</tr>
</tbody>
</table>

(b) Model T-1 (features: n = 33)

<table>
<thead>
<tr>
<th>Mineralized</th>
<th>Unmineralized</th>
</tr>
</thead>
<tbody>
<tr>
<td>1634</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>342</td>
</tr>
</tbody>
</table>

(c) Model M-T-1 (features: n = 43)

<table>
<thead>
<tr>
<th>Mineralized</th>
<th>Unmineralized</th>
</tr>
</thead>
<tbody>
<tr>
<td>296</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>190</td>
</tr>
</tbody>
</table>

(d) Model M-T-4 (features: n = 12)

<table>
<thead>
<tr>
<th>Mineralized</th>
<th>Unmineralized</th>
</tr>
</thead>
<tbody>
<tr>
<td>297</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>187</td>
</tr>
</tbody>
</table>
FIGURE 5

"Major" Dataset

Accuracy

F1-score

"Major & Trace" Dataset

"Trace" Dataset

Model Name

M-2, M-1, M-T-5, M-T-4, M-T-3, M-T-2, M-T-1, T-7, T-6, T-5, T-3, T-2, T-1
FIGURE 6
**FIGURE 7**

(a) Scatter plot showing PC1 (28.50%) vs. PC2 (20.02%) with subsets and a validation set.

(b) Bar graph comparing Fertile and Barren occurrences with Unmineralized and Mineralized categories.

(c) Histogram for Songgui porphyry "Unmineralized" (n = 48).

(d) Histogram for Liuhe Granodiorite "Unmineralized" (n = 38).

(e) Histogram for Machangqing porphyry Cu-Au "Mineralized" (n = 48).

(f) Histogram for Jiama porphyry-skarn Cu-Au "Mineralized" (n = 84).