



A machine learning method for distinguishing detrital zircon provenance

S. H. Zhong^{1,2} · Y. Liu^{1,2} · S. Z. Li^{1,2} · I. N. Bindeman³ · P. A. Cawood⁴ · R. Seltmann^{1,5} · J. H. Niu^{1,2} · G. H. Guo^{1,2} · J. Q. Liu^{1,2}

Received: 28 November 2022 / Accepted: 2 May 2023 / Published online: 20 May 2023
© The Author(s) 2023

Abstract

Zircon geochemistry provides a sensitive monitor of its parental magma composition. However, due to the complexity of the uptake of trace elements during zircon growth, identifying source magmas remains challenging, particularly for detrital grains whose petrological context is lost. We use a machine learning-based approach to explore the classifiers for zircon provenance, based on 3794 published, high-quality zircon trace element analyses compiled from I-, S-, and A-type granites. Three supervised machine learning algorithms, namely, Support Vector Machine (SVM), Random Forest (RF), and Multilayer Perceptron (MLP) were used and trained with 11 features, including 7 trace elements (Ce, Eu, Ho, Nb, Ta, Th, and U) and 4 derived trace element ratios (Th/U, U/Yb, Ce/Ce*, and Eu/Eu*). Our results show that all three trained machine learning methods perform very well with accuracy varying from 0.86 to 0.89, and that input–output relationships captured by different ML methods are nearly consistent and can be explained by the known petrological processes. The application of our trained machine learning classifiers to detrital zircon studies will enhance the interpretability of zircon assemblages of different origins. It also helps develop interpretations, approaches, and tools that will benefit, for example, the study of continental crust evolution and mineral exploration.

Keywords Detrital zircon · I-type granite · S-type granite · A-type granite · Tectonic setting · Mineral exploration · Machine learning

Communicated by Daniela Rubatto.

S. H. Zhong and Y. Liu contributed equally to this work.

✉ S. H. Zhong
zhongshihua@ouc.edu.cn

¹ Key Lab of Submarine Geosciences and Prospecting Techniques, Frontiers Science Center for Deep Ocean Multispheres and Earth System, MOE and College of Marine, Geosciences, Ocean University of China, Qingdao 266100, China

² Laboratory for Marine Mineral Resources, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China

³ Department of Earth Sciences, University of Oregon, Eugene, OR 97403-1272, USA

⁴ School of Earth, Atmosphere and Environment, Monash University, Melbourne, VIC 3800, Australia

⁵ Department of Earth Sciences, Center for Russian and Central EurAsian Mineral Studies, Natural History Museum, London SW7 5BD, UK

Introduction

Zircons as an indicator for tectonic provenance and mineralization

Due to the physio-chemical resilience, detrital zircons may undergo multiple episodes of sedimentation, magmatism, and/or metamorphism, yet retain information on the age and chemistry of original parental magmas (Grimes et al. 2007; Cawood et al. 2013; Bindeman et al. 2018). This resilience has provided significant insight into the long-term evolution of the continental crust (Cawood et al. 2012, 2013), as well as enabled the development of provenance tools for mineral exploration (Belousova et al. 2002; Nardi et al. 2013). However, in many circumstances, to obtain the above information, we first need to identify the source rocks of detrital zircons.

The most common source rocks for detrital zircon of magmatic origin are I-, S-, and A-type granitoids. I- and S-type granitoids, which were first proposed by Chappel

and White (1974), mainly occur in convergent plate margins (generally arc settings for I-type dominated granites and continent–continent collisional settings for S-type dominated granites; Fig. 1) and form by partial melting of igneous rocks and sedimentary rocks, respectively (Blevin and Chappell 1995). In contrast, A-type granitoids, which were first proposed by Loiselle and Wones (1979), have somewhat alkalic compositions and are linked to tectonic settings at the stage of post-orogenic collapse when convergent stresses become reversed or in intraplate and anorogenic regimes (Fig. 1; Eby 1992; Foden et al. 2015). Thus, correctly identifying the source rocks of detrital zircons is vital, which can provide a basis for interpreting the geodynamic settings in which the zircons formed. This effort is further enhanced by the different affinities of these three types of granitoids (and thus zircon) with metal mineralization. I-type granitoids (especially those with high whole-rock Sr/Y) are generally related to porphyry Cu-Mo-Au deposits (Wang et al. 2018), whereas S-type granitoids show a particular affinity for W-Sn mineralization (Chappell and White 2001) and A-type granitoids can generate Sn, Li, Nb, W and rare earth element (REE) deposits (Fig. 1; Vasyukova and Williams-Jones 2020; Zhao et al. 2021). Thus, the identification of source rocks for detrital zircon can be used as a regional discriminatory tool for fertility evaluation, especially at an early stage of exploration (Ballard et al. 2002; Lee et al. 2021); for example, the predominance of I-type detrital zircons indicates very low fertility in generating W-Sn deposits but high fertility for generating porphyry Cu deposits (e.g., Lu et al. 2016).

Complexity of zircon compositions

The geochemical features of I-, S-, and A-type rocks are closely related to melt compositions, temperature, and oxidation state (Whalen et al. 1987; Blevin and Chappell 1992; Chappell and White 1992; Eby 1992; Breiter et al. 2014; Foden et al. 2015). Elevated Nb and Ta are a common feature of within-plate, alkaline A-type granites (Collins et al. 1982); A-type magmas are also characterized by low H₂O,

oxygen fugacity, and high temperature (Collins et al. 1982; Whalen et al. 1987; Eby 1990; Li et al. 2012). In contrast, both I- and S-type granites are generally characterized by much lower high-field-strength elements (HFSEs, e.g., Nb and Ta) and form from melts with higher H₂O and lower temperature (Collins et al. 1982; Eby 1990; Foden et al. 2015). Moreover, I-type magmas are also characterized by much higher oxidation states than A- and S-type magmas (Blevin and Chappell 1992; Foden et al. 2015). These distinct features of I-, S-, and A-type granites should, thus, be reflected in the compositions of zircon that grows from them. Indeed, according to our compiled database, zircons from A-type rocks tend to display much higher Nb and Ta than the other two groups (Fig. 2a, b); zircons from I-type rocks, tend to display much higher Ce/Ce* and Eu/Eu* (and thus higher oxidation state) than zircons from A-type granites (Fig. 2c, d). Zircon from S-type rocks also displays much lower Eu/Eu* than the I-type population, although the range of Ce/Ce* for the two groups shows considerable overlap (Fig. 2d). This confirms the existence of differences in zircon chemistry for the three groups of granites. Thus, it is reasonable to assume that bivariate plots, comprising the above elements and/or ratios, would be useful in the identification of the provenance of zircon. However, as shown in Fig. 2, all the bivariate diagrams display noticeable overlaps among the three groups. This limits the possible use of such diagrams in identifying the source rocks for detrital zircons.

Zircon trace element overlaps on the bivariate plots result from the complexity in the uptake of trace elements during zircon growth (Storm et al. 2014; Grimes et al. 2015; Rubatto 2017). Trace elemental chemistry of zircons is not solely dependent on trace element contents of magma itself (Chapman et al. 2016; Zhong et al. 2021b), but also partition coefficients dependent on temperature (Rubatto and Hermann 2007), and kinetics of zircon crystallization from magma (Melnik and Bindeman 2018; Bindeman and Melnik 2022). Pressure, oxygen fugacity, and competition from other co-crystallizing accessory and major minerals also impact zircon compositions (Grimes et al. 2007; Claiborne

Fig. 1 Schematic diagram showing the relationships of three common granite types with tectonic environments and mineralization

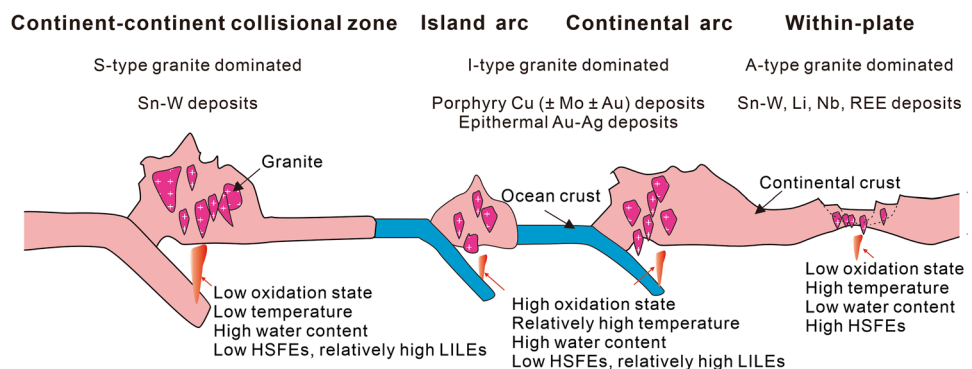
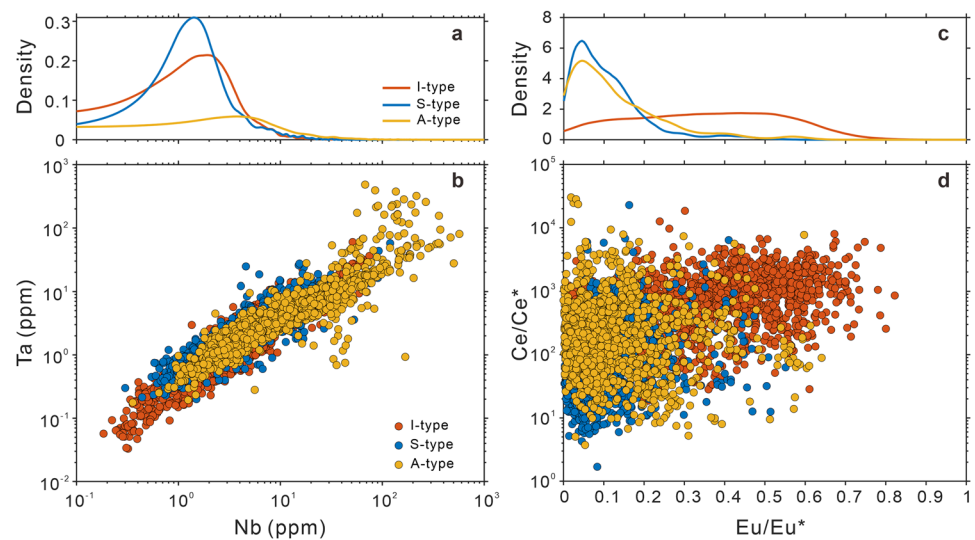


Fig. 2 Kernel density and binary diagrams for three types of zircons. **a** Kernel density diagram for Nb. **b** Nb versus Ta diagram. **c** Kernel density diagram for Eu/Eu*. **d** Eu/Eu* versus Ce/Ce* diagram



et al. 2018; Melnik and Bindeman 2018). However, differentiating the relative significance (and thus further deconvoluting the effects) of these variables is extremely challenging, especially for detrital grains that lack a direct link with the source rock from which they were derived (Grimes et al. 2007). Therefore, even if parental melt composition acts as a first-order control on the trace element composition of zircons that crystallized from it, the relationship between zircon compositions and their parental magma may not be as intuitive as expected. Deconvoluting the connection between zircon and its parental source magma is, thus, often difficult via conventional binary diagrams for elemental concentrations and ratios. This is the reason why, in this study, machine learning (ML) models are applied, as they enable more features that can impact zircon chemistry to be considered when developing robust provenance classifiers.

Advantages of machine learning

ML is a technology that is designed to automatically learn from experience and recognize complex patterns and relationships in data (Jordan and Mitchell 2015; Bergen et al. 2019). It has been one of today's most rapidly growing technical fields, lying at the intersection of computer science and statistics, and the core of artificial intelligence and data science (Jordan and Mitchell 2015). The growth of datasets of ever-increasing size has made ML an important tool in the integration of data and its application across the geosciences (e.g., Zuo et al. 2016; Petrelli and Perugini 2016; Petrelli et al. 2017; 2020).

ML methods are robust, fast, and allow exploration of a large function space (Bergen et al. 2019). ML enables the utilization of a large number of features as well as the ability to capture complex nonlinear relationships among large datasets (Jordan and Mitchell 2015; Bergen et al. 2019; Reichstein et al. 2019).

This is different from traditional geochemical classification strategies that are generally based on single elements (e.g., Burnham and Berry 2012; Trail et al. 2017) or some binary and/or triangular diagrams where fewer elements are utilized (e.g., Wang et al. 2012; Grimes et al. 2015). Thus, ML promises to achieve a much higher level of classification precision than the previous methods, especially for complex geological problems characterized by a large enough number of input variables (Petrelli and Perugini 2016). Moreover, ML learns the classification features by itself and does not need to be explicitly programmed; therefore, the internal, complex relationships within the data can be discovered algorithmically without the requirement for preexisting knowledge (Zhong et al. 2023).

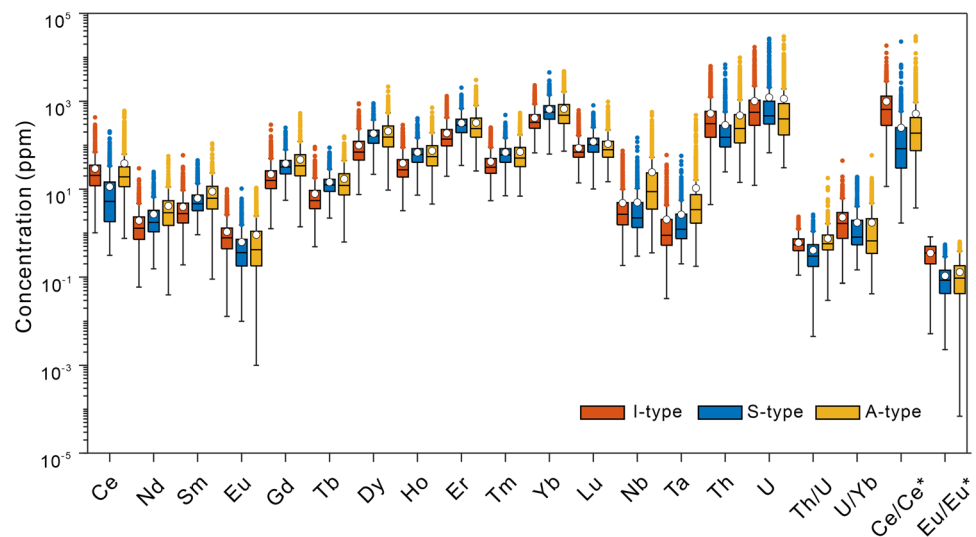
Therefore, in this study, we apply ML technology to relate the trace element geochemistry of zircon with the type of source magma. To build the ML classifiers applicable to the identification of zircon provenance, we first compiled a zircon trace element database comprising ~4000 published analyses for which source rocks are well known. Then we conducted ML modeling using this zircon dataset and three supervised ML algorithms, namely, Support Vector Machine (SVM), Random Forest (RF), and Multilayer Perceptron (MLP). We finally demonstrate the good performance of these novel ML methods in distinguishing zircon provenance by evaluating the outputs according to different metrics and by applying them to the recently published zircon database with known source rocks.

Data compilation and pre-processing

Data compilation

Zircon can be of magmatic, metamorphic or more controversially hydrothermal origin (e.g., Hoskin and Schaltegger

Fig. 3 Box and whisker plots of 20 zircon trace element concentrations and/or ratios for I-, S-, and A-type rocks. The height of the colored bars represents the interquartile range. The horizontal black lines within the colored bars are the median and the open circles with black edges represent the mean value. “Whiskers” of each box illustrate the maximum values lying within 1.5 times the interquartile range beyond the edges of the bars. The colored crosses represent the outliers deviating by more than $\pm 1.5 \sigma$. The data points outside the box are all higher than the average because of the use of log transformation



2003; Rubatto 2017). Only magmatic zircon can serve as an indicator for the magmas from which it crystallized, whereas hydrothermal and metamorphic zircon records fluid-infiltration and water–rock interaction, and metamorphic events, respectively (Rubatto et al. 2001; Schaltegger 2007). We compiled more than 20,000 magmatic zircon trace element analyses from published studies, for which the source rocks are known. For each zircon analysis, 20 features, including 12 REEs (Ce, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, and Lu), Nb, Ta, Th, U and 4 derived trace element ratios (Th/U, U/Yb, Ce/Ce* and Eu/Eu*), were initially compiled. Ce/Ce* and Eu/Eu* were calculated using the exponential power function, as described by Zhong et al. (2019). These ratios reflect the degree of Ce and Eu anomalies, which have been suggested to be related to magmatic oxidation state (Ballard et al. 2002; Smythe and Brenan 2016; Zhong et al. 2019).

Due to radioactive decay, the measured Th and U concentrations would be lower than those at the time of crystallization. This is especially true for the deep-time Hadean and Eoarchean zircons. Thus, both Th and U (and thus the derived Th/U and U/Yb) were corrected back to the time of crystallization. Furthermore, following Zhong et al. (2023), we did not consider two REEs—La and Pr—in our algorithms. This is because La and Pr are present at very low levels in natural zircons and are generally below the limit of detection. We did not compile features like Al, P, Ti, Sc, Hf, and Y, as well as isotopes (O, Hf, Zr, Li), which may also be useful in the identification of the origin of zircon (e.g., Bindeman 2008; Burnham and Berry 2017; Melnik and Bindeman 2018; Ackerson et al. 2021), since these values are often not reported.

The compiled 20 elements and/or ratios have many advantages in their application to ML modeling. First, they are routinely analyzed in many laboratories and are more commonly reported in literature studies. Second, they have

been shown to be useful in discriminating zircon provenance (Belousova et al. 2002; Grimes et al. 2007, 2015), despite some claims to the contrary (Hoskin and Ireland 2000; Coogan and Hinton 2006). Moreover, our statistical analysis work has indicated that although none of these selected elements and/or ratios can independently identify all three types of zircons, most can distinguish at least one zircon type from the rest (Fig. 3). For example, most zircons from I-type rocks can be distinguished from the other two types by higher Eu/Eu* and Ce/Ce*, most zircons from S-type rocks can be distinguished by lower Ce and Th/U, and most zircons from A-type rocks can be distinguished by much higher Nb and Ta (Fig. 3).

Data filtering

The use of the above composite dataset that comprises different sources of data requires quality control. Many studies showed that zircon compositions are highly susceptible to contamination by accessory mineral inclusions (Zhong et al. 2018, 2021a). To help to exclude such “artifacts”, we followed previously published studies (e.g., Tang et al. 2022) and used the selection criterion of La < 1 ppm. The remaining > 14,800 analyses might not all be autocrystic and some can be of hydrothermal, metamorphic or inherited origin. To exclude metamorphism-influenced zircons, zircon grains (~ 1100) from magmatic rocks with noticeable metamorphic overprint were filtered. To exclude hydrothermal zircons or analyses possibly experiencing chemical alteration, ~ 5600 analyses were discarded using the criterion of LREE-I > 30 (LREE-I = (Dy/Nd) + (Dy/Sm)), which was proposed by Bell et al. (2016). We also discarded ~ 1500 zircon analyses with discordant ages (> 10%) which are probably related to alteration and/or metamorphism and ~ 1000 analyses with noticeably older individual ages than the host rocks (and

thus probably be of inherited origin; Siégel et al. 2018). Moreover, not all of the above elements are reported in the publications, causing gaps in the compiled database. In this study, we excluded the analyses that contain missing values for Nb, Ta, Th, and U, but analyses with partial missing values of REE data were not excluded. In this latter case, the missing values can be easily extrapolated from other REE concentrations using the method of Zhong et al. (2019).

After the above filtering, 1442 zircon grains from I-type granites, 1051 from S-type granites, and 1301 from A-type granites are retained, with the sample locations shown in Fig. S1, Online Resource 1. It can be seen that 82% of the retained samples are sourced from Asia, 9% from North America, 4% from Australia, 3% from Africa, 1% from Europe, and 1% from South America. This database may not be optimal and should be updated in the future due to geographical bias, but it contains all the data available to authors at the time of manuscript writing. The full features for this database are presented in Online Resource 2.

Feature selection

For each zircon analysis, 20 features were compiled, but we did not use all these features during the ML training. This is because many studies show that the incorporation of more features during the ML modeling does not necessarily guarantee better performance (Yang et al. 2022). Rather, the presence of more features and more noise sometimes increases the likelihood of incorrect decisions (Salama and El-Gohary 2016; Wang et al. 2022). One common reason is multicollinearity, which indicates that several features are significantly correlated not only with the dependent feature but also with each other (Dormann et al. 2013; Garg and Tai 2013). Multicollinearity can not only prompt skewed or deluded results but also result in the inaccurate interpretation of the effects of explanatory features because the change of one feature would inevitably lead to the change of another feature (Shrestha 2020). In other words, the findings from a model with multicollinearity may not be trustworthy. Thus, to improve the model performance and make the model more explainable, it is important and valuable to focus on the feature selection.

In this study, we use the correlation matrix (Tay 2018) and variance inflation factor (VIF; Dormann et al. 2013) to identify multicollinearity among the 20 features mentioned above. The correlation matrix shows the correlation coefficient for all pairs of input features. The typical correlation value that has been suggested as the threshold ranges from 0.6 to 0.8 (Tay 2018). If the correlation coefficients among the independent features are higher than the suggested threshold, then it can be deduced that there is a multicollinearity problem among the features. VIF is the other widely used selection criterion (Dormann et al. 2013; Rawal and Ahmad 2021). It is used to measure how much the variance

of the estimated regression coefficient is inflated if the independent features are correlated (Shrestha 2020). If $VIF < 5$, there is no multicollinearity issue in the data. If $VIF \geq 5$ to 10, there is multicollinearity among the features, whereas $VIF > 10$ indicates the regression coefficients are feebly estimated with the presence of multicollinearity (Dormann et al. 2013; Rawal and Ahmad 2021). Figure 4 displays the heatmap of the correlation matrix and the VIF for the 20 features. Both of them confirm that only REEs (except Ce and Eu) are highly correlated and have a significant multicollinearity problem. The multicollinearity of zircon REEs illustrated by these diagrams is consistent with previous studies, which show that single REE concentrations can be well predicted by others (e.g., Zhong et al. 2019). According to the correlation matrix (Fig. 4a) and VIF values (Fig. 4b), Ho is mostly correlated with all other REEs (except Ce and Eu), indicating that Ho could likely represent the majority of features illustrated by these REEs. Therefore, we finally selected 11 features for the ML modeling: Ce, Eu, Ho, Nb, Ta, Th, U, Th/U, U/Yb, Ce/Ce*, and Eu/Eu*. Recalculation of VIF confirms that there is no multicollinearity relationship among these 11 features (Fig. S2, Online Resource 1).

Treatment of class imbalance

Class imbalance is a major problem in ML where the instances of one of the two classes are in abundance (say majority class) while the instances belonging to the other class are low (say minority class) (Chawla 2009; Kotsiantis et al. 2006). Previous studies have demonstrated training on an imbalanced dataset is risky and detrimental to classification performance because of neglecting the minority class (e.g., del Río et al. 2014). In this study, the proportion of I-, S-, and A-type zircons is 39.3%, 27.1%, and 33.6%, respectively. Although this class imbalance is not extreme especially compared to that possible in other ML problems (e.g., anomaly detection, object recognition; Chandola et al. 2009; Elyan et al. 2020), a balanced class distribution is optional (Nathwani et al. 2022).

A common technique to solve this problem is oversampling the minority class or undersampling the majority class to produce a class-balanced database (e.g., Kubat and Kubat 2017; Alpaydin 2020). In this study, undersampling was used because our preliminary investigation showed that it worked better than oversampling according to the performance metrics including accuracy, precision, and recall. Specifically, we followed our recent work (Zhong et al. 2023) to use “TomekLinks” (an undersampling approach that aims to remove samples, which are nearest neighbors) to get a balanced dataset (Tomek 1976). After the treatment of class imbalance, we found the accuracy of the trained models could be improved by 2–5%.

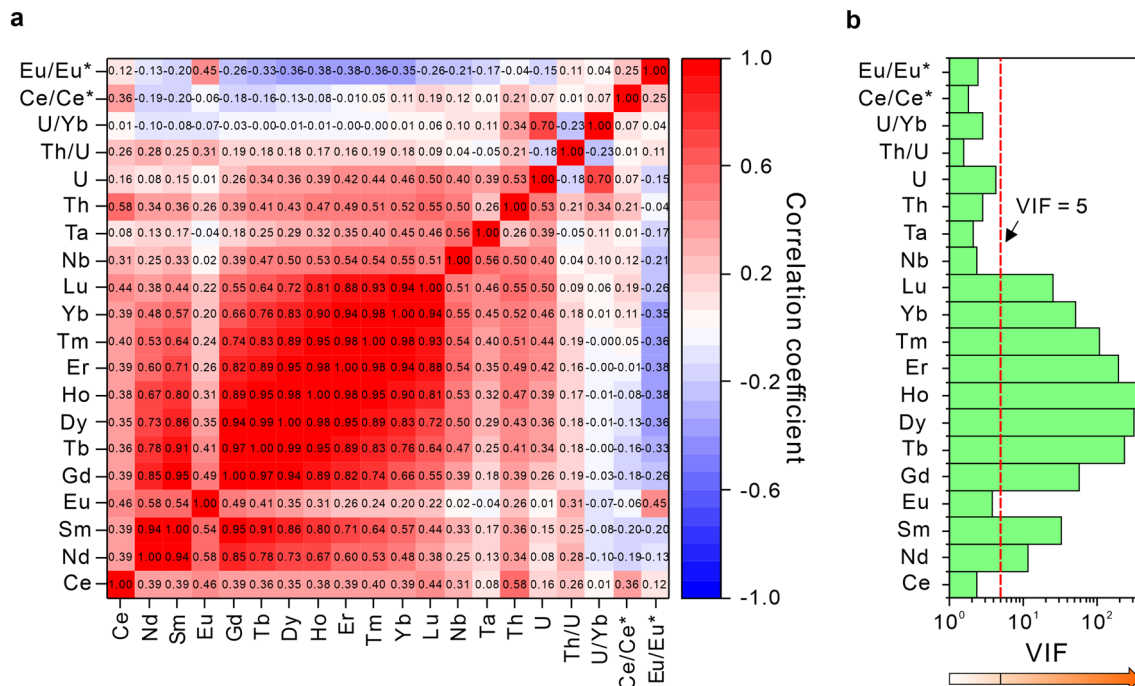


Fig. 4 Parameters used to identify the multicollinearity of the 20 zircon features. **a** Heatmap of the correlation matrix. The heatmap shows the darker the color, the stronger the correlation, with the red and blue colors showing positive and negative correlations, respectively. **b** Variance inflation factor (VIF) for the 20 zircon features. The color bar at the bottom shows the correspondence between VIF and

multicollinearity, with large VIF values showing more serious multicollinearity between features. Both the correlation matrix and the VIF values confirm that there is strong multicollinearity among REEs (except Eu and Ce); thus, some of them should be discarded before ML training (see Section "Feature selection" for more details)

Standardization and data splitting

Standardization of datasets is a common requirement for many ML estimators (Pedregosa et al. 2011; Petrelli et al. 2020). This helps to avoid the dominance of features in greater numeric ranges and reduces the calculation expense (Hsu et al. 2003). Following the method of previous studies (e.g., Wang et al. 2021), we first transformed all the zircon trace element data by applying a natural logarithmic scale to reduce the effect of outliers. Then a "StandardScaler" (centering of data by removing the mean value of each feature, and then scaling it by dividing non-constant features by their standard deviation) was conducted to get zero mean and unit variance distributions (e.g., Petrelli et al. 2020).

With the leave-out method (e.g., Petrelli and Perugini 2016), the whole dataset was randomly divided into a training set and a test set at a ratio of 8:2 while keeping the proportions of each class (Fig. 5). Then a tenfold cross-validation technique was performed on the training set to measure the performance of three ML methods. This involves splitting the training set into ten subsets, with nine subsets (real "training set") used to train the algorithm, and the remaining one subset (the validation set) used to validate the algorithm. This was repeated ten times until every subset had appeared

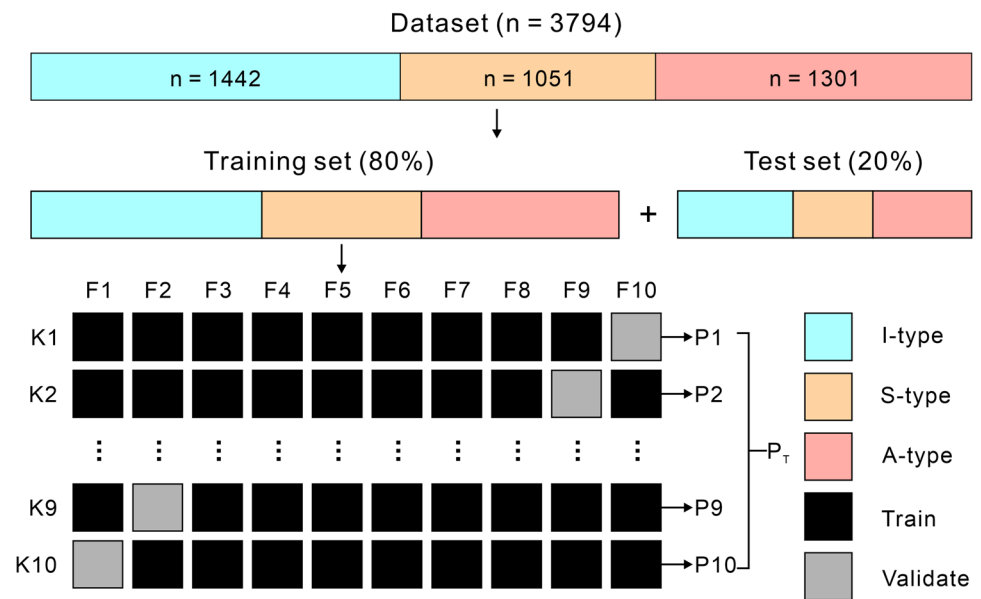
once as the validating set. The algorithm performance was finally indicated by an average of the metric scores. The benefit of this approach, as opposed to a single train/test set, is that it reduces the possibility of high bias that may arise from a single train/test set and helps to ensure algorithms generalize better on unseen data (Nathwani et al. 2022). It is noted that the resulting average accuracy based solely on the training set is likely, to some degree, an underestimate of the true accuracy of these classifiers when the algorithm is trained on all data and tested on unseen data. However, in most cases, this estimate is reliable, particularly if the amount of labeled data is sufficiently large and if the unseen data follow the same distribution as the labeled examples (Urueta-Hinojosa et al. 2020).

Machine learning methods

Machine learning model selection

Many ML algorithms have been used to solve classification problems in geosciences (Bergen et al. 2019). Among these methods, Support Vector Machine (SVM) and Random Forest (RF) are two of the most widely used algorithms (Petrelli

Fig. 5 Data splitting scheme and schematic illustration of a tenfold cross-validation workflow (modified after Pedregosa et al. 2011). The dataset is split into a training set and a test set. The training set is further divided tenfold. On this basis, ten times of cross-validation are performed, with one of them selected as a validation set for evaluation in each training. Performance metrics (P) are calculated for each fold and the mean metric (P_T) is calculated as the overall performance. The proportion of each class remains the same throughout the training



and Perugini 2016; Petrelli et al. 2020; Wang et al. 2021; Zou et al. 2022). In addition, our recent work shows that Multilayer Perceptron (MLP) also performs well in identifying the source rocks of detrital zircons (Zhong et al. 2023). Thus, in this study, these three algorithms were selected to explore the classifiers for zircon provenance. The basic features of SVM, RF, and MLP are summarized here.

SVM is a supervised classifier based on statistical learning theory and the principle of structural risk minimization (Cortes and Vapnik 1995). The basic idea of SVM is to create the optimal fitting hyperplane in the sample space or feature space that best separates classes (Vapnik 1999). To solve linearly inseparable problems, the input data in the low-dimensional space are mapped into the high-dimensional space using the kernel function, thereby transforming them into linearly separable problems in the high-dimensional space (Fig. 6a). More detailed mathematical principles for SVM can be found in Burges (1998) and Chang and Lin (2011).

RF is a powerful ensemble learning method proposed by Breiman (2001). The idea of ensemble learning is to combine several weak classifiers into a strong classifier. RF algorithm is an averaging algorithm based on randomized decision trees. It produces multiple decision trees, using a randomly selected subset of training samples and variables (Fig. 6b). The bagging algorithm generates training data for each tree by sampling with the replacement of several samples equal to the sample number in the source dataset (Breiman 2001). As an ensemble method, each decision tree votes for a category prediction, and the top-voted one is then used to make the final prediction. A more detailed introduction can be found in Breiman (1996).

MLP is one of the most popular feed-forward neural network architectures in use today. The architecture of MLP consists of three or more layers of nodes with unidirectional connections (an input layer, a single or more hidden layer(s), and an output layer, see Fig. 6c) to capture more complex relationships in the sample dataset. The nodes of the input layer are linked to the hidden layers via connections, with the weights for the connections depending on their importance that is learned during the training. The value at the output nodes is the result of the weighted sum of the hidden nodes. The performance and the error of the algorithm in predicting the real outcomes are then evaluated. A more detailed introduction can be found in Hinton (1989) and He et al. (2015).

Validation and tuning parameters

In this study, we use the following metrics to evaluate the performance of each classifier: accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic (ROC) curve (AUC). The accuracy, precision, recall, and F1-score are calculated based on the confusion matrix, in which true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are presented (Fig. 7). Accuracy is the ratio of the number of correct predictions to the total number of samples. Precision is the ratio between the number of correct predictions and all the samples predicted in this class. Recall is the ratio of the number of correct predictions to the total samples of this class. F1-score is the weighted harmonic mean of precision and recall. The AUC provides a single measure of the overall model accuracy that is threshold independent (Narkhede 2018). An AUC value of

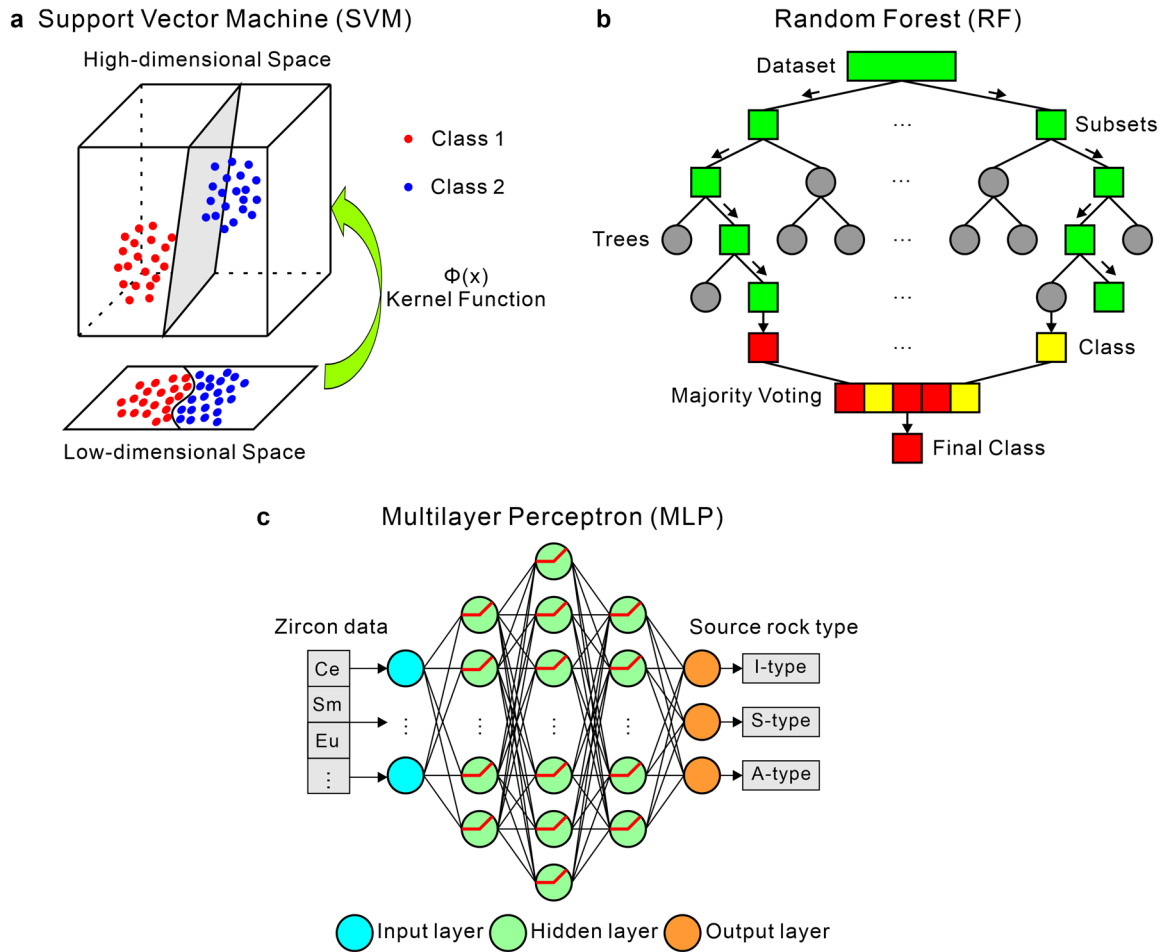


Fig. 6 Schematic illustrations of the three supervised ML methods used in this study

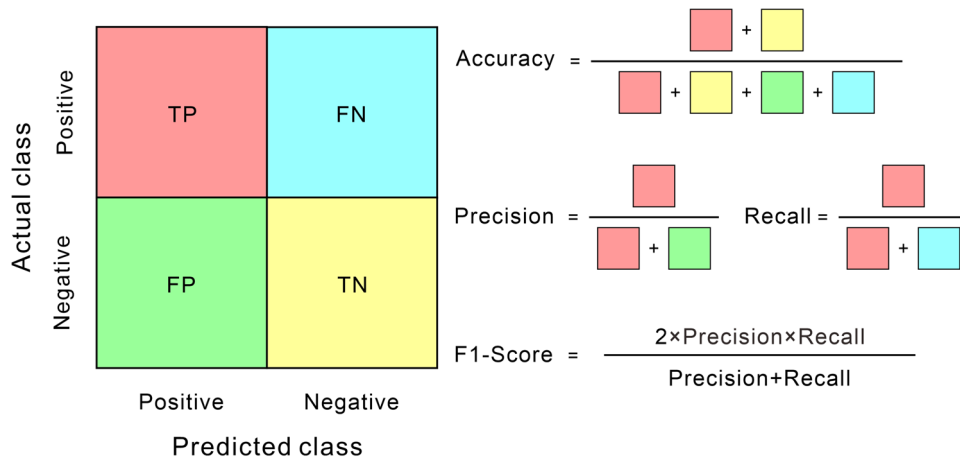


Fig. 7 Schematic diagram of the calculation of the four metrics (accuracy, precision, recall, and F1-score) based on the confusion matrix (left). TP (true positive) is the number of positive samples predicted correctly. FP (false positive) is the number of positive samples predicted incorrectly. TN (true negative) is the number of nega-

ive samples predicted correctly. FN (false negative) is the number of negative samples predicted incorrectly. Accuracy, precision, recall, and F1-score range from 0–1; theoretically, if the samples from the test set are all predicted correctly, accuracy, precision, recall, and F1-score would be 1

1 indicates perfect prediction whereas 0.5 indicates prediction performance is as good as random.

Many ML algorithms are parameterized, and their parameters should be tuned to achieve the best result with a particular dataset (Wang et al. 2021; Zou et al. 2022). Hyperparameters are parameters that are not directly learned from estimators (Nathwani et al. 2022). For the sake of the optimization of the three ML methods, a “grid search” technique based on the tenfold cross-validation method was used to determine the best hyperparameters (Hsu et al. 2003). For MLP, SVM, and RF, there are three, two, and five hyperparameters to be optimized, respectively (Figs. S3–S5, Online Resource 1). The best combination suggested by grid search was chosen as the optimal hyperparameter values (Table S1, Online Resource). The methods using the optimal hyperparameter values were then retained and evaluated using the test set to provide the final, optimal performance metrics.

Model interpretation

ML is becoming increasingly used in the solid Earth geoscience. However, ML is commonly considered a black-box method, meaning that humans cannot fully understand the underlying reasoning behind the predictions/classifications (Bergen et al. 2019). This is different from the traditional methods, in which the results obtained are generally explainable and transparent. Therefore, the results predicted by ML are not always fully endorsed by the community. Fortunately, model interpretability approaches, such as SHAP (Shapely Additive exPlanations; Lundberg and Lee 2017) and LIME (Local Interpretable Model-Agnostic Explainer; Ribeiro et al. 2016), have been developed, and these overcome the mentioned limitation by allowing interpretation of the estimated results and analysis of the importance and dependency of features. In this study, we use SHAP, which is based on game theory (Lundberg and Lee 2017), to investigate how the ML methods have learned input–output relationships. As described in previous studies (e.g., Lundberg and Lee 2017), the SHAP value is a measure of the contribution of each feature to the output that interprets the estimated results. However, it should be noted that SHAP values reflect the importance of a feature to the model, which can be different from the direct importance of this feature in nature, as already emphasized in many studies (Nathwani et al. 2022).

Results

Performance metrics of three trained models

Average performance metrics for each classifier from the tenfold cross-validation process with individual metrics for each fold are summarized in Tables S2 to S4, Online

Resource 1. The results of tenfold cross-validation show that three classifiers have higher stability. Three ML classifiers are applied to the test set, and their performance indicators are reported in Fig. 8 and Table S5, Online Resource 1. The SVM method performs well, as indicated by high accuracy (0.89), precision (0.82–0.93), recall (0.87–0.90), and F1-score (0.84–0.92) scores (Table S5, Online Resource 1). For the RF and MLP methods, all the metrics scores are also high (mostly > 0.80) with accuracy being 0.86 and 0.89, respectively (Table S5, Online Resource 1). The three methods are also characterized by high AUC values (0.966 for SVM, 0.966 for RF and 0.972 for MLP) (Fig. S6, Online Resource 1). All these metric scores indicate that three supervised ML methods are robust in predicting zircon types.

In each method, the prediction performance for the zircon populations from I- and A-type source rocks is always better than that from S-type source rocks. For example, for the I- and A-type zircon, the precision scores are almost all above 0.90 (average 0.92); in contrast, they range from 0.78 to 0.82 (average 0.80) for S-type zircons. This indicates that S-type zircons are slightly more difficult to be correctly distinguished compared to I- and A-type zircons.

SHAP values

Mean absolute SHAP values (feature importance scores) and the relative importance of features are shown for SVM (Fig. 9a–c), RF (Fig. 9d–f), and MLP (Fig. 9g–i), respectively. The five most important features in distinguishing I-type zircons are Eu/Eu*, Ce/Ce*, Eu, Ho, and Ta for the SVM method (Fig. 9a); Eu/Eu*, Ce/Ce*, Nb, Ta, Ce for the RF method (Fig. 9d); and Ce/Ce*, Nb, Ta, Th, and Ce for the MLP method (Fig. 9g). It can be seen that in each method, Ce/Ce* plays an important role in distinguishing zircons from I-type rocks. Three methods also display quite similar feature importance patterns in distinguishing S- and A-type zircons. Particularly, Ce and Ce/Ce* are ranked as the two most important features in identifying S-type zircons (Fig. 9b, e and h), whereas Nb is ranked as the most important feature in identifying A-type zircons (Fig. 9c, f and i).

Besides, the SHAP summary plots also indicate that the relationship between input and output is captured consistently in almost all methods. In each method, high Ce/Ce* inputs (red) produce high SHAP values for I-type zircons and, therefore, have a strong positive influence on the model output, whereas for S- and A-type zircons, low Ce/Ce* inputs (blue) produce high SHAP values and, therefore, have a strong positive influence on the model output (Fig. 9). This is consistent with that derived from the statistical analysis result, where I-type zircons are visually characterized by noticeably higher Ce/Ce* than other types of zircons

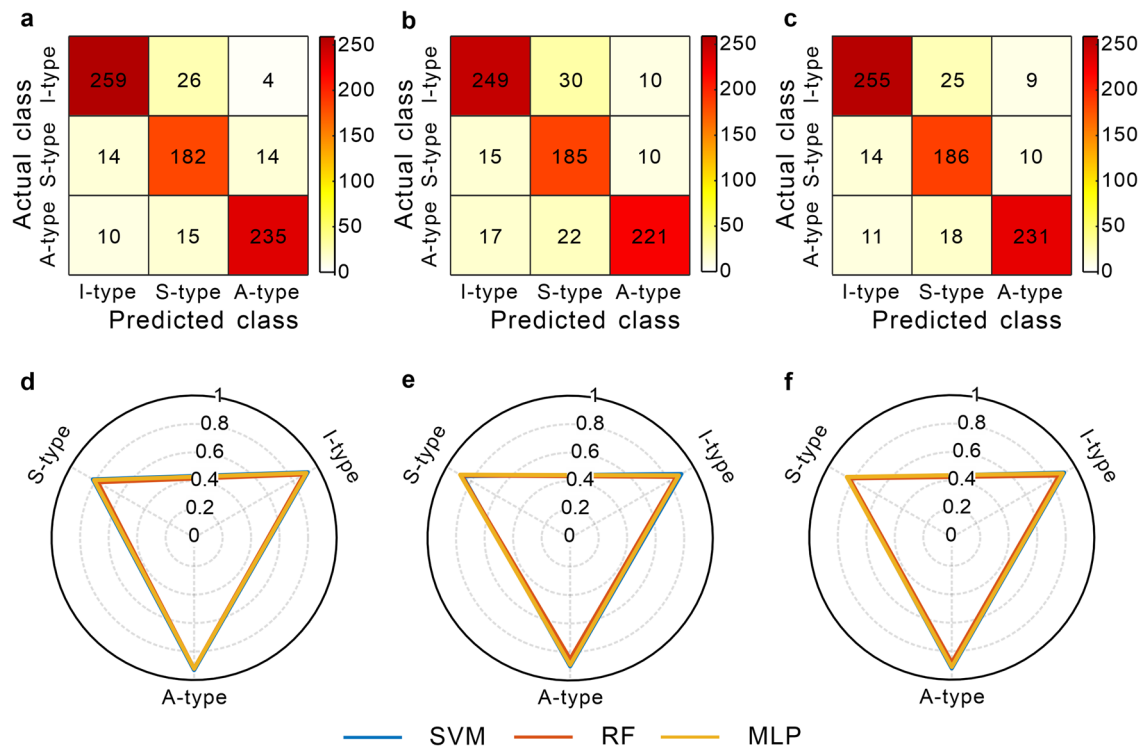


Fig. 8 Performance for the test set in different ML methods. **a–c** Confusion matrix for SVM, RF, and MLP, respectively. **d–f** Polar plot of precision, recall, and F1-score, respectively. It can be seen that most

of the metric scores are >0.85 , confirming that the three trained models perform very well

(Fig. 3). This is also true for Eu/Eu^* . Low Eu/Eu^* (blue) has a strong positive influence on the model output for the S- and A-type zircons, whereas it shows a strong negative influence on the model output for the I-type zircons (Fig. 9). Higher Nb and Ta (red) produce high SHAP values for A-type zircons and, therefore, have a strong positive influence on the model output, whereas for I- and S-type zircons, negative Nb and Ta inputs (blue) generally produce high SHAP values and, therefore, have a strong positive influence on the model output (Fig. 9). For other features, like Ho, Th, U, U/Yb, and Th/U, the three ML methods also display nearly same input–output relationships for each zircon population. To first order, the nearly consistent feature importance pattern and input–output relationship captured by different ML methods confirm the feasibility of zircon trace elements in the identification of zircon provenance and the plausibility of the ML results.

Discussion

Independent tests of model performance

In this study, all three trained ML methods showed high classification accuracy (0.89 for SVM, 0.86 for RF, and 0.89

for MLP). However, published studies show that a method's stated performance may sometimes not accurately reflect its performance post-deployment because of, for example, overfitting (Reunanen 2003) and black-box effects (Rudin 2019) of the used ML methods. Additionally, in this study, although all the metric scores were derived from the test set, which was never encountered by the algorithm during the training process and thus can reflect the “real” performance of each method, in practice, this pre-processing method may still result in overestimation of the metric scores. This is because the test set zircon data may be from the same locations and even from the same granite samples that the methods have been trained on. To further manifest the robustness of three trained classification techniques, they were tested on three independent datasets (Online Resource 3) containing recently published zircon data from known granite provenance that do not appear in the training and test dataset at all. These three datasets include 109 I-type zircon analyses from Qu et al. (2022) and Tang et al. (2022), 55 S-type analyses from Xu et al. (2022), and 128 A-type analyses from Zhang et al. (2022) datasets.

All three trained classification methods produced a good performance for the three independent datasets, with the dominant zircon types identified by the three methods consistent with the provenance illustrated in the

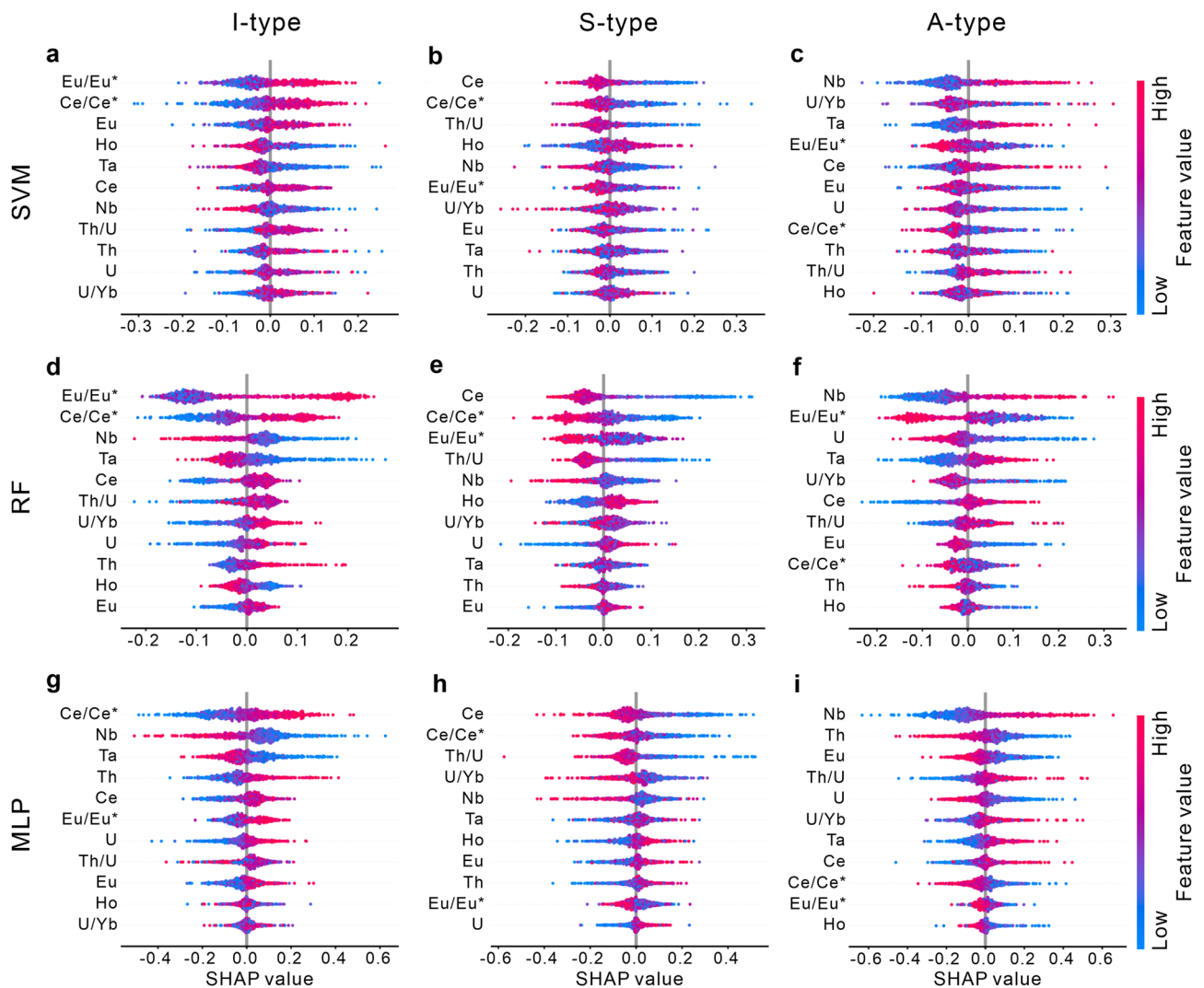


Fig. 9 SHAP summary plots for three different methods in this study. **a–c** SVM-based SHAP plot for I-, S-, and A-type zircons, respectively. **d–f** RF-based SHAP plot for I-, S-, and A-type zircons, respectively. **g–i** MLP-based SHAP plot for I-, S-, and A-type zircons, respectively. Features are sorted by the sum of their SHAP value magnitudes across all samples in the test set. The order from the top vertically indicates the importance of the feature. Red color indicates a high value and blue color indicates a low value of the feature. The horizontal axis denotes the impact of the value of the feature on the

output. The density achieved by the dots indicates their intensity. For example, for the SVM method, Eu/Eu* is the most important feature in identifying I-type zircons, whereas Ce for S-type zircons and Nb for A-type zircons; moreover, high Ce/Ce* values (red symbols) contribute toward an I-type prediction (positive SHAP value), whereas low Ce/Ce* values (blue symbols) contribute toward an S- and A-type prediction (negative SHAP value). See Section “SHAP values” for more details

publications according to other methods (Fig. 10). However, all the trained methods performed better on the I-type zircon dataset, with only a percentage of 2–6% of I-type grains being misclassified into the A- and/or S-type populations (Fig. 10a–c). For the S-type zircon database, the classification performance is weaker with accuracy being 0.76–0.80; a percentage of 15–22% of grains were misclassified into the I-type population (Fig. 10d–f). For the A-type zircon database, the classification performance is also weaker than that for the I-type zircon database with accuracy being 0.71 to 0.76; a percentage of 13% to 23%

of grains from the A-type zircon database were misclassified into the S-type population (Fig. 10g–i). These are consistent with the results given by the confusion matrix (Fig. 8a–c) and reflect that in cases where the wrong classification exists, S-type grains are more likely to be misclassified into the I-type population, whereas some A-type zircons are more likely misclassified into the S-type population. For the I-type zircons, the proportion of the wrong classification is negligible. The data falling into unexpected populations may reflect magma contamination, unusual fractionation processes, and/or random variations

between isolated melt pockets formed late in the crystallization sequence (Grimes et al. 2015). Thus, the misclassification for a small group of a few grains seems to be inevitable. Enlarging the training database and involving complementary methods (e.g., isotopic analyses) will help to fully characterize the likely parental melt sources of out-of-context grains.

Petrogenetic implications

The feature importance obtained from SHAP values can be used to interpret key petrogenetic processes of the source magmas from which zircons crystallize. As mentioned above, analysis of those importance scores indicates that the features that tend to display high concentrations in I-type

zircons but lower in S-type zircons are Ce/Ce* and Eu/Eu* (Fig. 9), all of which are related, in this case, to magma oxygen fugacity (Zhong et al. 2019). As already mentioned, these relationships are consistent with the well-known key difference between the two types: oxidation state (Blevin and Chappell 1992). The relatively reduced nature of S-type granites compared to I-type granites has been ascribed to the presence of graphite within the source rocks (Flood and Shaw 1975). A-type zircons are also characterized by lower Ce/Ce*, and Eu/Eu* than I-type zircons (Figs. 3 and 9), which is also consistent with previous work that suggested that the closest matching of A-type whole-rock compositional trends was achieved by the closed system (i.e., unbuffered) fractional crystallization at reduced oxidation states (Foden et al. 2015). It is noted that zircon Eu/Eu* is also

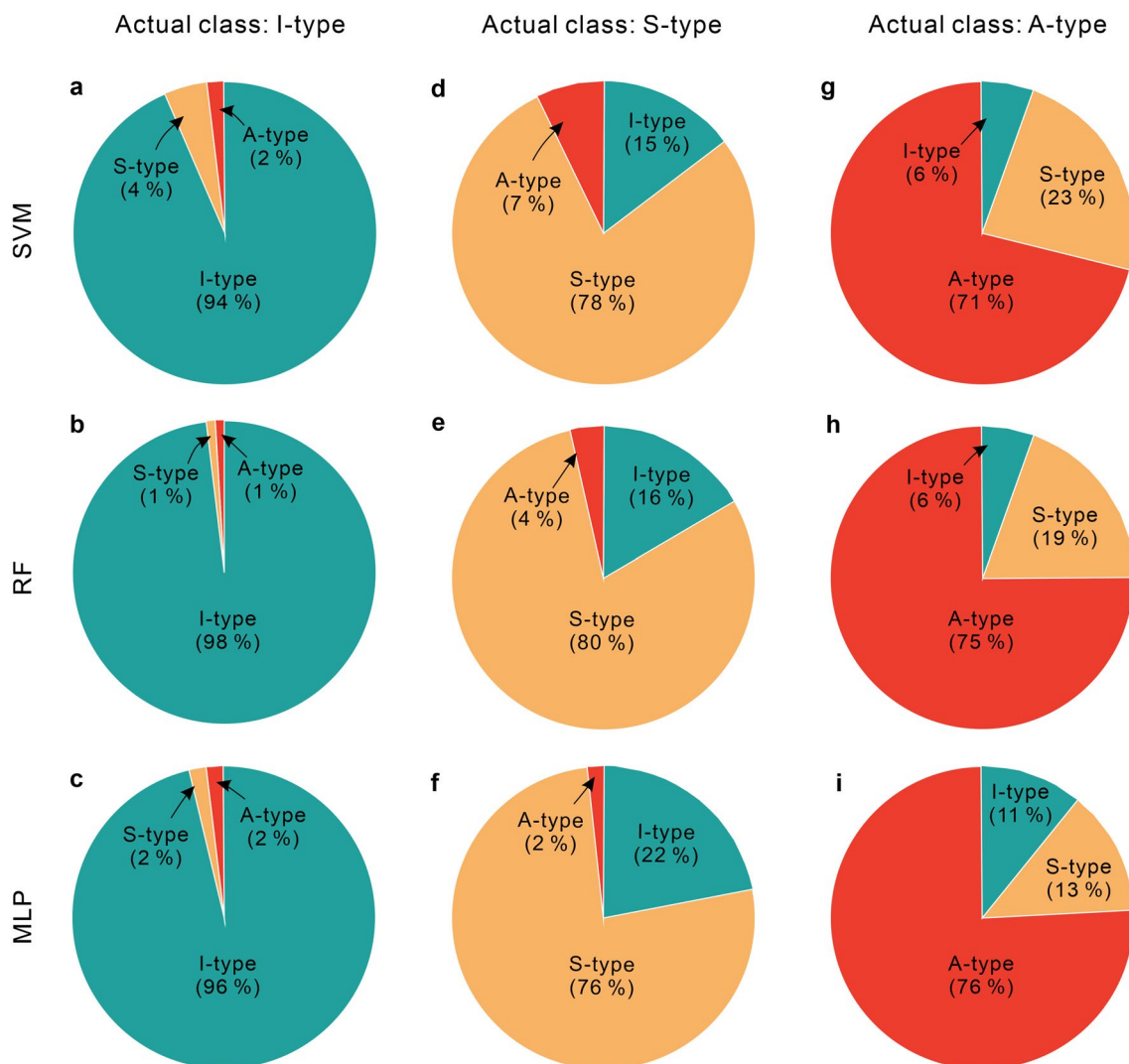


Fig. 10 Pie charts showing the classification results using three trained ML classifiers for independent tests. **a–c** Results for the I-type zircons from Qu et al. (2022) and Tang et al. (2022) ($n=109$). **d–f**

Results for the S-type zircons from Xu et al. (2022) ($n=55$). **g–i** Results for the A-type zircons from Zhang et al. (2022) ($n=128$)

positively correlated to the water content of magmas (Zhong et al. 2018, 2019). A-type granites are generally derived from magmas with low water contents, which, thus, may enhance the low Eu/Eu^* and also probably the Eu feature of zircons derived from such magmas. Figure 9 shows that different from the low Ce/Ce^* and Eu/Eu^* features, A-type zircons tend to display high Ce contents and S-type zircons generally tend to display high Eu contents. The uncoupling of Ce and Eu with the indicators of magma oxidation state is interesting and needs further research.

The SHAP pattern suggesting A-type zircons tend to have high Nb and Ta contents whereas the others tend to be characterized by lower contents (Fig. 9), is also consistent with the corresponding source rock features mentioned previously (see Section “Complexity of zircon compositions”). For example, elevated Nb and Ta, as well as low Ti, are a noticeable feature of within-plate, alkaline A-type granites (Collins et al. 1982; Eby 1990), which can be explained by the crystallization of ilmenite, with little influence of rutile or titanite (Li et al. 2012). This is due to ilmenite, rutile, and titanite being the main Ti-enriched minerals, with ilmenite stable at high temperatures and low pressure (Liou et al. 1998). Different from rutile and titanite, which have high concentrations of Nb and Ta (Rudnick et al. 2000; Liang et al. 2009), ilmenite is characterized by lower Nb and Ta (Cole and Stewart 2009). Thus, ilmenite crystallization will lead to the depletion of Ti without significant decreases in Nb and Ta (Li et al. 2012). Our ML methods indicate Nb as the most important feature in identifying A-type zircons, further supporting the above arguments and the high-temperature characteristics of A-type granites. In contrast, S-type zircons tend to be characterized by lower Nb and Ta contents because S-type granites are derived mainly from the remelting of sedimentary rocks or upper crustal materials, which are strongly depleted in high-field strength elements (e.g., Nb and Ta) (e.g., Tang et al. 2012; Wang et al. 2012). The low Nb and Ta contents of I-type zircons could not be explained by the above mechanism; rather, they are likely related to the crystallization of titanite (and probably also rutile), which are common in I-type rocks (Clemens et al. 2011).

The relationships between other elements (Ho, Th, U, Th/U, and U/Yb) and the granite genetic types cannot be explained intuitively as those mentioned above. Further research is, thus, merited. Nonetheless, since all the key geochemical characteristics illustrated by the SHAP value analysis are in agreement with previous petrological studies, this confirms our ML methods are credible and robust.

Cautions on the application

This work confirms the feasibility of using ML methods for identifying zircon provenance. These methods are designed

for detrital grains for which the petrological context is missing. Meanwhile, they will also be useful for inherited, xenocrystic populations and grains from geochemically altered host rocks, as long as it can be established that they are of granitic magma origin and the analyzed areas do not undergo alteration and metamorphism. In practice, the identification of hydrothermal and metamorphic zircon populations can be facilitated by, for example, backscattered electron (BSE) and cathodoluminescence (CL) images, and isotopic data.

One limitation of these ML classifiers is that they can only make predictions on the classes they were trained on, i.e., all zircons will be classified as I-type, S-type or I-type even if they were sourced from something completely different. They cannot identify zircon suites from magmatic rocks except for granitoids. Indeed, zircons that grew from mafic rocks are also reported (Borisova et al. 2020). However, zircons of mafic magma origin are not very common; thus, the possibility to be preserved in detrital zircon populations should be low. Therefore, as is usually done (Zhu et al. 2020), we generally do not need to consider detrital grains of mafic origin, unless there is evidence to indicate that zircon populations from mafic rocks may not be negligible.

Another possible limitation is that these ML classifiers may be less robust when applied to detrital zircons derived from Europe, South America, Australia, and Africa. This is because the proportion of zircons from these four areas only accounts for less than 10% of the training dataset, whereas more than 90% are from Asia and North America. Thus, as we mentioned earlier, the training database should be updated in the future, which may improve the performance of these ML classifiers. Nonetheless, according to our knowledge, no geographically geochemical distinctions have been reported for I-, S-, and A-type granites. Therefore, despite the existence of geographical bias for the training zircon dataset, no evidence indicates that the performance of three ML classifiers will be significantly compromised.

Lastly, we re-emphasize the need to carefully filter zircon data when using these ML classifiers. For example, according to our previous studies, mineral inclusion contamination is quite common during zircon analyses (Zhong et al. 2018); this creates many “artifacts” with deceptive genetic information and thus should be excluded. In this study, we have provided a detailed filtering strategy for the trained database (e.g., $\text{La} < 1$ ppm, $\text{LREE-I} > 30$), which can also be applied to treating the input grains. Additionally, as with many geochemical tools, successful applications of ML classifiers discussed here to out-of-context zircon will be most effective when considered alongside additional constraints, especially the geological background from which the detrital zircons are collected.

To make our ML models more easily used, three trained ML models are packaged as an installation-free .exe file,

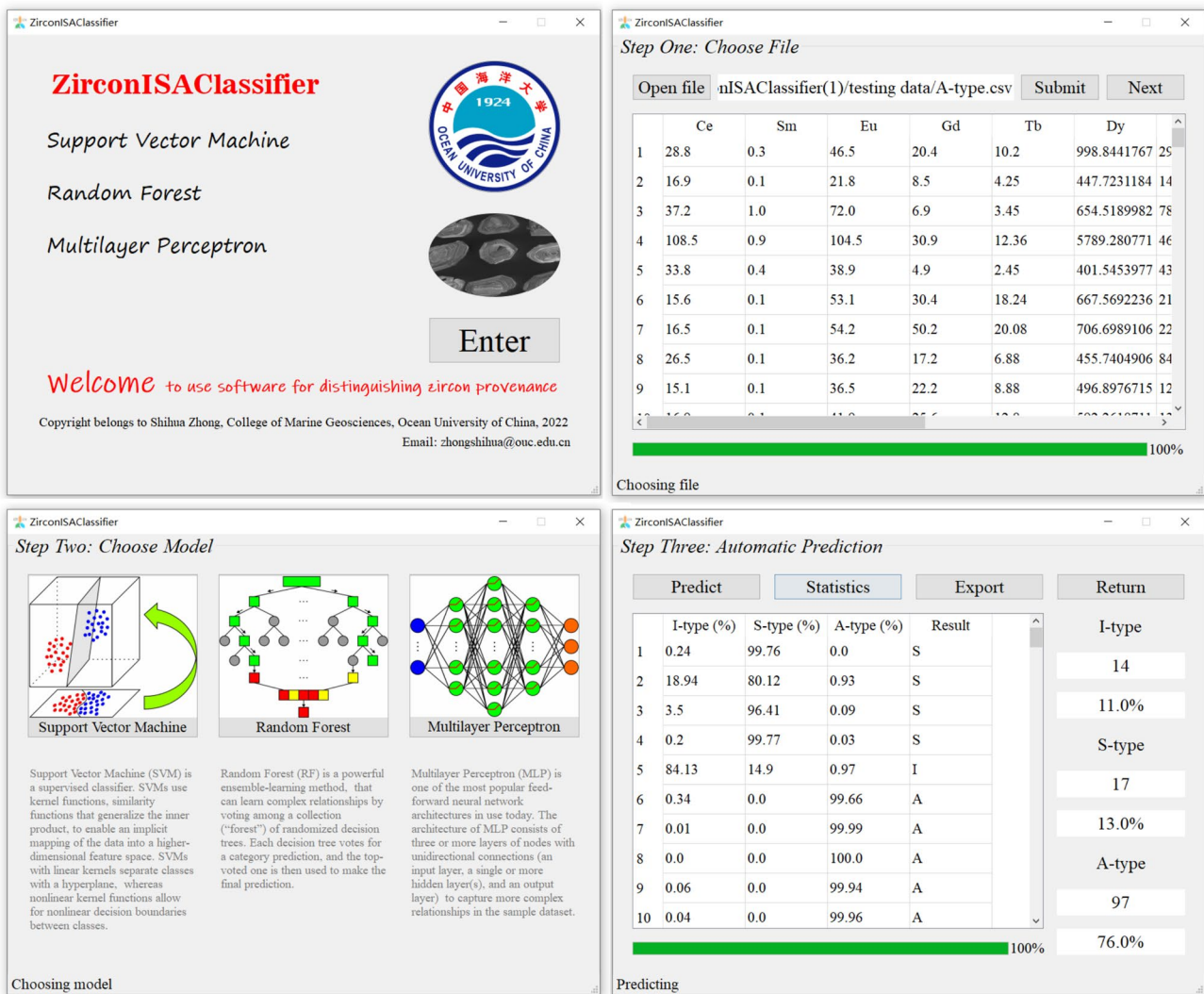


Fig. 11 The user interface of ZirconIASClassifier, a software that is designed in this study to predict the source rocks of zircons (see Text S1, Online Resource 1 for more details)

named ZirconIASClassifier, that can be run on Windows platforms (Fig. 11). Using this software, users can directly get the classification results for their zircon datasets. More details about this software can be found in the Supplementary Material (Text S1, Online Resource 1). This software, as well as the associated ML code (written in Python) to reproduce the results in this paper, is available at <https://github.com/ShihuaZhong/CTMP2023ZirconIASClassifier>.

Conclusion

Using high-quality, published zircon trace element data from known source rocks (I-, S-, and A-type granites) and three supervised machine learning algorithms (Support Vector Machine, Random Forest, and Multilayer

Perceptron), we have developed a novel approach to identify zircon types. The results demonstrate that trace element compositions of zircons can well mirror the nature of plutons from which they crystallized, and that the trained machine learning methods are robust in discriminating zircons from I-, S-, and A-type granites. The feature importance analysis indicates that the features sensitive to magma oxidation state (e.g., Ce/Ce* and Eu/Eu*) are most important in distinguishing zircons from I- and S-type granites, whereas Nb ranks among the most important in identifying zircons from A-type granites. All these are consistent well with known petrological processes during the formation of three groups of granites. To make the trained zircon classifiers more accessible, specialized software for classifying zircon types is developed. The application of our classifiers to detrital zircon studies will

improve the accuracy and efficiency for identifying zircon assemblages of different origins, and help develop interpretations, approaches, and tools that will benefit continental crust evolution studies and mineral exploration.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00410-023-02017-9>.

Acknowledgements We appreciate helpful reviews and constructive suggestions from Associate Editor Daniela Rubatto, Maurizio Petrelli, Coralie Siegel, and an anonymous reviewer. This work was financially supported by the Marine S&T Fund of Shandong Province for the National Laboratory for Marine Science and Technology (Qingdao) (No. 2022QNLM050302); Fundamental Research Funds for the Central Universities (202172002); the National Natural Science Foundation (42203066); and the Natural Science Foundation of Shandong Province (ZR2020QD027); Australian Research Council (FL160100168). RS acknowledges funding under Natural Environment Research Council Grant NE/P017452/1 “From arc magmas to ores (FAMOS): A mineral systems approach”.

Data availability The authors confirm that the data supporting this study is available within the online supplementary materials.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ackerson M, Trail D, Buettner J (2021) Emergence of peraluminous crustal magmas and implications for the early Earth. *Geochem Perspect Lett* 17:50–54
- Alpaydin E (2020) Introduction to machine learning. MIT Press
- Ballard JR, Palin MJ, Campbell IH (2002) Relative oxidation states of magmas inferred from Ce(IV)/Ce(III) in zircon: application to porphyry copper deposits of northern Chile. *Contrib Mineral Pet* 144(3):347–364
- Bell EA, Boehnke P, Harrison TM (2016) Recovering the primary geochemistry of Jack Hills zircons through quantitative estimates of chemical alteration. *Geochim Cosmochim Acta* 191:187–202
- Belousova E, Griffin W, O'Reilly SY, Fisher N (2002) Igneous zircon: trace element composition as an indicator of source rock type. *Contrib Mineral Petr* 143(5):602–622
- Bergen KJ, Johnson PA, de Hoop MV, Beroza GC (2019) Machine learning for data-driven discovery in solid Earth geoscience. *Science* 363(6433):eaau0323
- Bindeman I (2008) Oxygen isotopes in mantle and crustal magmas as revealed by single crystal analysis. *Rev Mineral Geochem* 69(1):445–478
- Bindeman IN, Melnik OE (2022) The rises and falls of zirconium isotopes during zircon crystallization. *Geochem Perspect Lett* 24:17–21
- Bindeman IN, Schmitt AK, Lundstrom CC, Hervig RL (2018) Stability of zircon and its isotopic ratios in high-temperature fluids: long-Term (4 months) isotope exchange experiment at 850°C and 50 MPa. *Front Earth Sci* 6:59
- Blevin PL, Chappell BW (1992) The role of magma sources oxidation states and fractionation in determining the granite metallogeny of eastern Australia. *T R Soc Edin Earth* 83(1–2):305–316
- Blevin PL, Chappell BW (1995) Chemistry origin and evolution of mineralized granites in the Lachlan fold belt Australia; the metallogeny of I- and S-type granites. *Econ Geol* 90(6):1604–1619
- Borisova AY, Bindeman IN, Toplis MJ, Zagrtdenov NR, Guignard J, Safonov OG, Bychkov AY, Shcheka S, Melnik OE, Marchelli M, Fehrenbach J (2020) Zircon survival in shallow asthenosphere and deep lithosphere. *Am Mineral* 105(11):1662–1671
- Breiman L (1996) Bagging predictors. *Mach Learn* 24(2):123–140
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Breiter K, Ackerman L, Ďurišova J, Svojtka M, Novák M (2014) Trace element composition of quartz from different types of pegmatites: a case study from the Moldanubian Zone of the Bohemian Massif (Czech Republic). *Mineral Mag* 78(3):703–722
- Burges CJC (1998) A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc* 2(2):121–167
- Burnham AD, Berry AJ (2012) An experimental study of trace element partitioning between zircon and melt as a function of oxygen fugacity. *Geochim Cosmochim Acta* 95:196–212
- Burnham AD, Berry AJ (2017) Formation of Hadean granites by melting of igneous crust. *Nat Geosci* 10(6):457–461
- Cawood PA, Hawkesworth CJ, Dhuime B (2012) Detrital zircon record and tectonic setting. *Geology* 40(10):875–878
- Cawood PA, Hawkesworth CJ, Dhuime B (2013) The continental record and the generation of continental crust. *Geol Soc Am Bull* 125(1–2):14–32
- Chandola V, Banerjee A, Kumar V (2009) Anomaly detection. *ACM Comput Surv* 41(3):1–58
- Chang CC, Lin CJ (2011) Libsvm. *ACM Trans Intell Syst Tech* 2(3):1–27
- Chapman JB, Gehrels GE, Ducea MN, Giesler N, Pullen A (2016) A new method for estimating parent rock trace element concentrations from zircon. *Chem Geol* 439:59–70
- Chappell BW, White AJR (1974) Two contrasting granite types Pacific. *Geology* 8:173–174
- Chappell BW, White AJR (2001) Two contrasting granite types: 25 years later. *Aust J Earth Sci* 48(4):489–499
- Chappell BW, White AJR (1992) I- and S-type granites in the Lachlan Fold Belt. *Earth Environ Sci Trans R Soc* 83(1–2):1–26
- Chawla NV (2009) Data mining for imbalanced datasets: an overview. In: Maimon O, Rokach L (eds) *Data min knowl disc handbook*. Springer US, pp 875–886
- Claiborne LL, Miller CF, Gualda GAR, Carley TL, Covey AK, Wooden JL, Fleming MA (2018) Zircon as Magma Monitor. In: Moser DE, Corfu F, Darling JR, Reddy SM, Tait K (eds) *Microstructural Geochronology*. <https://doi.org/10.1002/9781119227250.ch1>
- Clemens JD, Stevens G, Farina F (2011) The enigmatic sources of I-type granites: the peritectic connexion. *Lithos* 126(3):174–181
- Cole RB, Stewart BW (2009) Continental margin volcanism at sites of spreading ridge subduction: examples from southern Alaska and western California. *Tectonophysics* 464(1):118–136
- Collins W, Beams S, White A, Chappell B (1982) Nature and origin of A-type granites with particular reference to southeastern Australia. *Contrib Mineral Petr* 80(2):189–200
- Coogan LA, Hinton RW (2006) Do the trace element compositions of detrital zircons require Hadean continental crust? *Geology* 34(8):633–636
- Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3):273–297

- del Río S, López V, Benítez JM, Herrera F (2014) On the use of MapReduce for imbalanced big data using Random Forest Information. *Sciences* 285:112–137
- Dormann CF, Elith J, Bacher S, Buchmann C et al (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36(1):27–46
- Eby GN (1990) The A-type granitoids: a review of their occurrence and chemical characteristics and speculations on their petrogenesis. *Lithos* 26(1):115–134
- Eby GN (1992) Chemical subdivision of the A-type granitoids: Petrogenetic and tectonic implications. *Geology* 20(7):641–644
- Elyan E, Jamieson L, Ali-Gombe A (2020) Deep learning for symbols detection and classification in engineering drawings. *Neural Netw* 129:91–102
- Flood R, Shaw S (1975) A cordierite-bearing granite suite from the New England Batholith NSW Australia. *Contrib Mineral Pet* 52(3):157–164
- Foden J, Sossi PA, Wawryk CM (2015) Fe isotopes and the contrasting petrogenesis of A- I- and S-type granite. *Lithos* 212–215:32–44
- Garg A, Tai K (2013) Comparison of statistical and machine learning methods in modelling of data with multicollinearity. *Int J Model Identif* 18(4):295–312
- Grimes CB, John BE, Kelemen PB, Mazdab FK et al (2007) Trace element chemistry of zircons from oceanic crust: a method for distinguishing detrital zircon provenance. *Geology* 35(7):643–646
- Grimes CB, Wooden JL, Cheadle MJ, John BE (2015) “Fingerprinting” tectono-magmatic provenance using trace elements in igneous zircon. *Contrib Mineral Pet* 170(5):1–26
- He KM, Zhang XY, Ren SQ, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE International Conference on computer vision (ICCV)*, pp 1026–1034
- Hinton GE (1989) Connectionist learning procedures. *Artif Intell* 40(1):185–234
- Hoskin PWO, Ireland TR (2000) Rare earth element chemistry of zircon and its use as a provenance indicator. *Geology* 28(7):627
- Hoskin PWO, Schaltegger U (2003) The composition of zircon and igneous and metamorphic petrogenesis. *Rev Mineral Geochem* 53(1):27–62
- Hsu CW, Chang CC, Lin CJ (2003) A practical guide to support vector classification
- Jordan MI, Mitchell TM (2015) Machine learning: Trends perspectives and prospects. *Science* 349(6245):255–260
- Kotsiantis S, Kanellopoulos D, Pintelas P (2006) Handling imbalanced datasets: a review. *GESTS Intern Trans Comput Sci Eng* 30(1):25–36
- Kubat M, Kubat JA (2017) An introduction to machine learning, vol 2. Springer International Publishing, Cham, pp 321–329
- Lee RG, Plouffe A, Ferbey T, Hart CJ et al (2021) Recognizing porphyry copper potential from till zircon composition: A case study from the Highland Valley Porphyry district south-central British Columbia. *Econ Geol* 116(4):1035–1045
- Li H, Ling MX, Li CY, Zhang H et al (2012) A-type granite belts of two chemical subgroups in central eastern China: indication of ridge subduction. *Lithos* 150:26–36
- Liang J, Ding X, Sun X, Zhang Z et al (2009) Nb/Ta fractionation observed in eclogites from the Chinese continental scientific drilling project. *Chem Geol* 268(1–2):27–40
- Liou J, Zhang R, Ernst W, Liu J et al (1998) Mineral parageneses in the Piampaludo eclogitic body Gruppo di Voltri western Ligurian Alps Schweizerische. *Mineral Petrogr Mitt* 78(2):317–335
- Loiselle MC, Wones DR (1979) Characteristics and origin of anorogenic granites. *Geol Soc Am Abstr Progr* 11:48
- Lu YJ, Loucks R, Fiorentini M, McCuaig T et al (2016) Zircon compositions as a pathfinder for porphyry Cu ± Mo ± Au deposits. *Soc Econ Geol* 19:329–347
- Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on neural information processing systems* edited, pp 4768–4777, Curran Associates Inc, Long Beach, California, USA
- Melnik OE, Bindeman IN (2018) Modeling of trace elemental zoning patterns in accessory minerals with emphasis on the origin of micrometer-scale oscillatory zoning in zircon. *Am Mineral* 103(3):355–368
- Nardi LVS, Formoso MLL, Müller IF, Fontana E et al (2013) Zircon/rock partition coefficients of REEs, Y, Th, U, Nb, and Ta in granitic rocks: Uses for provenance and mineral exploration purposes. *Chem Geol* 335:1–7
- Narkhede S (2018) Understanding auc-roc curve. *Towards Data Sci* 26:220–227
- Nathwani CL, Wilkinson JJ, Fry G, Armstrong RN et al (2022) Machine learning for geochemical exploration: classifying metallogenic fertility in arc magmas and insights into porphyry copper deposit formation. *Miner Deposita* 57(7):1143–1166
- Pedregosa F et al (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
- Petrelli M, Perugini D (2016) Solving petrological problems through machine learning: the study case of tectonic discrimination using geochemical and isotopic data. *Contrib Mineral Pet* 171(10):81
- Petrelli M, Bizzarri R, Morgavi D, Baldanza A et al (2017) Combining machine learning techniques microanalyses and large geochemical datasets for tephrochronological studies in complex volcanic areas: new age constraints for the Pleistocene magmatism of central Italy. *Quat Geochronol* 40:33–44
- Petrelli M, Caricchi L, Perugini D (2020) Machine learning thermobarometry: application to clinopyroxene-bearing magmas. *J Geophys Res-Sol Earth* 125(9):e2020JB20130
- Qu P, Niu HC, Weng Q, Li NB et al (2022) Apatite and zircon geochemistry for discriminating ore-forming intrusions in the Luming giant porphyry Mo deposit Northeastern China. *Ore Geol Rev* 143:104771
- Rawal K, Ahmad A (2021) Feature selection for electrical demand forecasting and analysis of pearson coefficient. In: *2021 IEEE 4th International Electrical and Energy Conference (CIEEC)* edited, pp 1–6
- Reichstein M, Camps-Valls G, Stevens B, Jung M et al (2019) Deep learning and process understanding for data-driven Earth system science. *Nature* 566(7743):195–204
- Reunanen J (2003) Overfitting in making comparisons between variable selection methods. *J Mach Learn Res* 3(Mar):1371–1382
- Ribeiro MT, Singh S, Guestrin C (2016) “Why Should I Trust You?”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining* edited, pp 1135–1144
- Rubatto D (2017) Zircon: the metamorphic mineral. *Rev Mineral Geochem* 83(1):261–295
- Rubatto D, Hermann J (2007) Experimental zircon/melt and zircon/garnet trace element partitioning and implications for the geochronology of crustal rocks. *Chem Geol* 241(1):38–61
- Rubatto D, Williams IS, Buick IS (2001) Zircon and monazite response to prograde metamorphism in the Reynolds Range central Australia. *Contrib Mineral Pet* 140(4):458–468
- Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1(5):206–215
- Rudnick RL, Barth M, Horn II, McDonough WF (2000) Rutile-bearing refractory eclogites: missing link between continents and depleted mantle. *Science* 287(5451):278–281

- Salama DM, El-Gohary NM (2016) Semantic text classification for supporting automated compliance checking in construction. *J Comput Civil Eng* 30(1):04014106
- Schaltegger U (2007) Hydrothermal zircon. *Elements* 3(1):51–79
- Shrestha N (2020) Detecting multicollinearity in regression analysis. *Am J Appl Math and Stat* 8(2):39–42
- Siégel C, Bryan SE, Allen CM, Gust DA (2018) Use and abuse of zircon-based thermometers: a critical review and a recommended approach to identify antecrystic zircons. *Earth-Sci Rev* 176:87–116
- Smythe DJ, Brenan JM (2016) Magmatic oxygen fugacity estimated using zircon-melt partitioning of cerium. *Earth Planet Sci Lett* 453:260–266
- Storm S, Schmitt AK, Shane P, Lindsay JM (2014) Zircon trace element chemistry at sub-micrometer resolution for Tarawera volcano New Zealand and implications for rhyolite magma evolution. *Contrib Mineral Pet* 167(4):1000
- Tang DM, Qin KZ, Sun H, Su BX et al (2012) The role of crustal contamination in the formation of Ni–Cu sulfide deposits in Eastern Tianshan Xinjiang Northwest China: evidence from trace element geochemistry Re–Os Sr–Nd zircon Hf–O and sulfur isotopes. *J Asian Earth Sci* 49:145–160
- Tang L, Chen PL, Santosh M, Zhang ST et al (2022) Geology and genesis of auriferous porphyritic monzogranite and its correlation with the Qiyugou porphyry-breccia system in East Qinling Central China. *Ore Geol Rev* 142:104709
- Tay R (2018) Correlation variance inflation and multicollinearity in regression model. *J Eastern Asia Soc Transp Stud* 12:2006–2015
- Tomek I (1976) Two Modifications of CNN. *IEEE T Syst Man and Cy SMC* 6(11):769–772
- Trail D, Tailby N, Wang Y, Mark Harrison T et al (2017) Aluminum in zircon as evidence for peraluminous and metaluminous melts from the Hadean to present. *Geochem Geophys Geosy* 18(4):1580–1593
- Urueta-Hinojosa D, Velázquez P, Gutiérrez-Andrade M, De-los-Cobos-Silva S et al (2020) A comparative clustering model that considers false positives and false negatives in some socioeconomic applications Fuzzy. *Econ Rev* 25(2):45–67
- Vapnik VN (1999) An overview of statistical learning theory. *IEEE Trans Neural Netw* 10(5):988–999
- Vasyukova O, Williams-Jones A (2020) Partial melting fractional crystallisation liquid immiscibility and hydrothermal mobilization—a ‘recipe’ for the formation of economic A-type granite-hosted HFSE deposits. *Lithos* 356:105300
- Wang Q, Zhu DC, Zhao ZD, Guan Q et al (2012) Magmatic zircons from I- S- and A-type granitoids in Tibet: Trace element characteristics and their application to detrital zircon provenance study. *J Asian Earth Sci* 53:59–66
- Wang R, Weinberg RF, Collins WJ, Richards JP et al (2018) Origin of postcollisional magmas and formation of porphyry Cu deposits in southern Tibet. *Earth-Sci Rev* 181:122–143
- Wang Y, Qiu KF, Müller A, Hou ZL et al (2021) Machine learning prediction of quartz forming-environments. *J Geophys Res-Sol Earth* 126(8):e2021JB021925
- Wang Z, Fok KW, Thing VL (2022) Machine learning for encrypted malicious traffic detection: approaches datasets and comparative study. *Comput Secur* 113:102542
- Whalen JB, Currie KL, Chappell BW (1987) A-type granites: geochemical characteristics discrimination and petrogenesis. *Contrib Mineral Petr* 95(4):407–419
- Xu G, Li Z, Yang X, Liu L (2022) The role of Jiningian Pluton in Yanshanian metallogenic events in the Dahutang Tungsten deposit: evidence from whole rock and zircon geochemistry. *Minerals* 12(4):428
- Yang Q, Su Y, Hu T, Jin S et al (2022) Allometry-based estimation of forest aboveground biomass combining LiDAR canopy height attributes and optical spectral indexes. *For Ecosyst* 9:100059
- Zhang Q, Wang Q, Li G, Sun X et al (2022) Crucial control on magmatic-hydrothermal Sn deposit in the Tengchong block SW China: evidence from magma differentiation and zircon geochemistry. *Geosci Front* 13(4):101401
- Zhao X, Li NB, Huizenga JM, Yan S et al (2021) Rare earth element enrichment in the ion-adsorption deposits associated granites at Mesozoic extensional tectonic setting in South China. *Ore Geol Rev* 137:104317
- Zhong S, Feng C, Seltmann R, Li D et al (2018) Can magmatic zircon be distinguished from hydrothermal zircon by trace element composition? The effect of mineral inclusions on zircon trace element composition. *Lithos* 314–315:646–657
- Zhong S, Seltmann R, Qu H, Song Y (2019) Characterization of the zircon Ce anomaly for estimation of oxidation state of magmas: a revised Ce/Ce* method. *Miner Petrol* 113(6):755–763
- Zhong S, Li S, Feng C, Liu Y et al (2021a) Porphyry copper and skarn fertility of the northern Qinghai-Tibet Plateau collisional granitoids. *Earth-Sci Rev* 214:103524
- Zhong S, Li S, Seltmann R, Lai Z et al (2021b) The influence of fractionation of REE-enriched minerals on the zircon partition coefficients. *Geosci Front* 12(3):101094
- Zhong S, Li S, Liu Y, Cawood PA et al (2023) I-type and S-type granites in the Earth’s earliest continental crust. *Commun Earth Environ* 4(1):61
- Zhu Z, Campbell IH, Allen CM, Burnham AD (2020) S-type granites: their origin and distribution through time as determined from detrital zircons. *Earth Planet Sci Lett* 536:116140
- Zou S, Chen X, Brzozowski MJ, Leng CB et al (2022) Application of machine learning to characterizing magma fertility in porphyry Cu deposits. *J Geophys Res-Sol Earth* 127(8):e2022JB024584
- Zuo RG, Carranza EJM, Wang J (2016) Spatial analysis and visualization of exploration geochemical data. *Earth-Sci Rev* 158:9–18

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.