Revisiting the Geochemical Classification of Zircon Source Rocks Using a Machine Learning Approach

Keita Itano[1,2] · Hikaru Sawada[3]

Abstract

Trace element fingerprints preserved in zircons offer clues to their origin and crystallization conditions. Numerous geochemical indicators have been established to evaluate the source rock characteristics from a geochemical perspective; however, multivariate trace element data have not been sufficiently investigated statistically. As substantial amounts of zircon data from a wide range of rock types have become accessible over the past few decades, it is now essential to reassess the utility of trace elements in discriminating source rock types. We employed a new zircon trace element dataset and established classification models to distinguish eight types of source rocks: igneous (acidic, intermediate, basic, kimberlite, carbonatite, and nepheline syenite), metamorphic, and hydrothermal. Whereas a conventional decision tree analysis was unable to correctly classify the new dataset, the random forest and support vector machine algorithms achieved high-precision classifications ($> 80\%$ precision, recall, and F1 score). This work confirms that trace element composition is a helpful tool for province studies and mineral exploration using detrital zircons. However, the compiled dataset with many missing values leaves room for improving the models. Trace elements, such as P and Sc, which cannot be measured by quadrupole inductively coupled plasma mass spectrometry, are vital for more accurate classification.

✉ Keita Itano
k.itano@gipc.akita-u.ac.jp; inv.itano@ugr.es

Hikaru Sawada
hsawada@sus.u-toyama.ac.jp

[1] Department of Mathematical Science and Electrical-Electronic-Computer Engineering, Akita University, Akita 010-8502, Japan

[2] Present Address: Department of Mineralogy and Petrology, University of Granada, Granada 18071, Spain

[3] Department of Earth System Science, School of Sustainable Design, Toyama 930-0087, Japan

1 Introduction

Sedimentary rocks and detrital minerals can be used for average sampling of crustal rocks exposed in the hinterland, and the approach using sediment materials complements outcrop-based geological studies to reveal the evolution of continental crusts (e.g., Allègre and Rousseau 1984). Although sedimentary rocks are derived from multiple sources through sedimentary recycling processes, whole-rock chemical or isotopic composition reflects the averages of various components, and cannot decipher geological records from a single sample. In contrast, detrital mineral grains from a single sedimentary rock sample individually record the crystallization timing and geochemical characteristics of the source rocks, providing more quantitative information about ancient geotectonic environments (Gaschnig 2019; Gehrels 2014).

Detrital zircon is ubiquitous in sedimentary rocks and has remarkable advantages in detrital studies, such as (i) chemical and physical stability against weathering processes (Ewing et al. 1995), (ii) U–Pb geochronology (Bowring and Schmitz 2003), (iii) geochemical tracers (Burnham and Berry 2017; Zhu et al. 2020), and (iv) isotopic tracers (Iizuka et al. 2017; Valley 2003). The applications of detrital zircons are wide-ranging, such as the evaluation of global continental growth (Rino et al. 2004; Sawada et al. 2018), reconstruction of paleo-tectonic history (Cawood et al. 2012; Grimes et al. 2007), and exploration of ore deposits (Pereira and Storey 2023; Wang et al. 2021). These applications have shown that it is crucial for the versatility to not only date a single grain but also estimate the source rock type.

Trace element chemistry is the key to deciphering the origin of zircons. A number of researchers have reported that the abundance and ratios of trace elements (e.g., P, Sc, Ti, Nb, rare earth elements [REE], Hf, Ta, Th, and U) are useful for interpreting the crystallization conditions of zircons (Burnham and Berry 2017; Grimes et al. 2007; Hoskin and Ireland 2000). The rationale is that trace element partitioning is remarkably sensitive to changes in coexisting mineral assemblages and accordingly reflects the differences in the source rock type. Many studies commonly report REE data; however, the selection of other trace elements based on specific research objectives is inconsistent. Despite the increase in trace element data over the last few decades, the full potential of multivariate data has not been realized, and far too little attention has been paid to statistical approaches to quantitatively assess their potential as geochemical indicators. Multivariate statistical analysis was first performed by Belousova et al. (2002), which demonstrated that geochemical fingerprints were effective in distinguishing different origins within igneous zircons based on decision tree classification using the CART (classification and regression tree) algorithm.

Recent studies have applied machine learning techniques to classification problems in geochemical fields (Itano et al. 2020; Petrelli and Perugini 2016). Although difficulties still exist in the statistical analysis of compositional data (Aitchison 1982), appropriate preprocessing enables us to overcome these obstacles and solidify the advantages of the machine learning approach. However, its application to zircon trace element data is extremely limited (Ziyi et al. 2022; Zhong et al. 2023).

In this study, we update geostatistical methods and the zircon dataset that has been significantly augmented in the last two decades, and statistically reassess the potential of trace element composition as an indicator of source rock type. Two machine learning

techniques, random forest and support vector machine, were applied to the newly compiled zircon data, and their predictive performance was compared with that of the traditional CART algorithm used in a previous study (Belousova et al. 2002). Both techniques have been increasingly used in Earth science applications, particularly for remote sensing and geophysical data (Belgiu and Drăguţ 2016). Application to petrological data has only recently begun (Ueki et al. 2018; Zhao et al. 2019). We address three classification tasks to distinguish between (i) igneous, metamorphic, and hydrothermal origins; (ii) acidic, intermediate, basic, kimberlite, carbonatite, and nepheline syenite; and (iii) the S–I–A–M types of granitic rocks.

2 Data Compilation

We collected zircon trace element data from various source rocks (n = 7,999 from 90 papers; see Table 1). All citations of the data used are summarized in Supplementary Information S1. The database, including detailed lithology, sampling locality, and chemical composition, is available in Supplementary Information S1.

The first objective of this data collation was to systematically compare zircon data from igneous, metamorphic, and hydrothermal rocks. This study also aimed to further examine various types of igneous rocks. For simplification, the categories of igneous rock types were reclassified into acidic ($SiO_2 > 63$ wt%), intermediate (63 wt% $\geq SiO_2 \geq 52$ wt%), basic ($SiO_2 > 52$ wt%), kimberlite, carbonatite, and nepheline-bearing syenite (Ne-syenite) based on the whole-rock chemical composition and the original descriptions on rock names and mineral assemblages if the former was not present. The acidic rock data were further subdivided into S- (sedimentary protolith, n = 328), I- (igneous protolith, n = 1,358), A- (anorogenic, n = 156), and M-type granites (direct mantle source, n = 212), which is the geochemical classification of granitic rocks of different origins (Chappell and White 2001; Bonin 2007; Whalen 1985). The classification of the granite types followed descriptions in the literature.

For analytical accuracy, we only compiled trace element data measured using secondary ion mass spectrometry (SIMS) and laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS). The number of measured elements varies from study to study (43 elements in total: Li, Be, B, F, Na, Al, P, S, Cl, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Ga, Sr, Y, Nb, Mo, Sn, Ba, La, Ce, Pr, Nd, Sm, Ru, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Hf, Ta, Pb, Th, and U) because of differences in analytical methods and their scopes. Some elements, such as Li, Be, P, and Sc, are difficult to measure by quadrupole-type ICP-MS due to their low concentrations and mass interferences, leading to numerous missing data (Fig. 1).

A common problem with compiled trace element data is missing data. Missing data for trace elements occur simply because they have not been measured or because of an instrument sensitivity threshold, that is, the limit of detection (LOD). The former corresponds to data missing completely at random (MCAR), whereas the latter is considered as missing not at random (MNAR) (Do et al. 2018; Doucet et al. 2022). Because most of the missing data for Y, Nb, REE, Hf, Th, and U are MCAR-type data, we adopted a data removal approach to consistently treat data from various references and to avoid the risk of potential bias in the imputed data. On the other hand,

Table 1 Rock types with the corresponding numbers of samples in the dataset

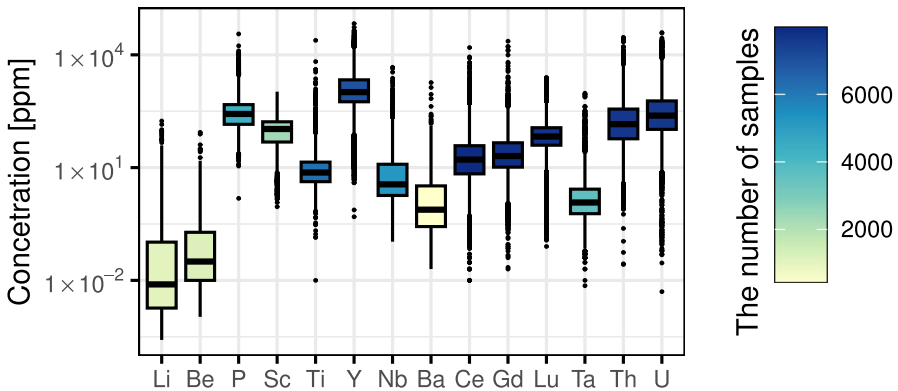| Igneous					Metamorphic	Hydrothermal
7,388					378	233
Acidic	Intermediate	Basic	Kimberlite	Carbonatite	Ne-syenite	
4,485	1,428	681	538	206	50	

Fig. 1 Representative trace element data for igneous zircons with the numbers of samples

unmeasured REE data were imputed based on the linearity of CI chondrite-normalized values except for Ce and Eu because of their consistent chemical affinities.

3 Methodology

3.1 Random Forest

A random forest (RF) is an extension of a decision tree classifier trained using the CART algorithm and an ensemble model, where predictions from hundreds of decision trees are combined to determine the final output (Fig. 2a). Random forest overcomes the issue of susceptibility to overfitting of the single-tree classifier (Olden et al. 2008) while retaining its advantage of easy interpretability (Kampichler et al. 2010). Random forest achieves this by implementing two processes of randomness: bootstrap sampling and random selection of variables for each decision tree (Breiman 2001). The random variation between trees generated by these processes prevents overfitting and improves generalizability.
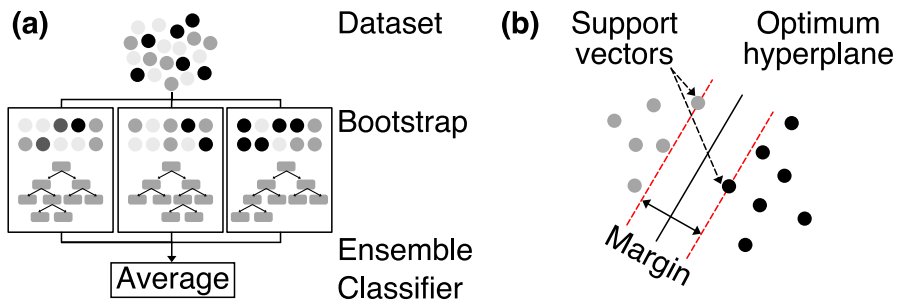

Fig. 2 Schematic images of random forest and support vector machine; see text for detailed explanations

One benefit of random forest is the ability to assess the relative importance of features in the classification results (Kampichler et al. 2010). The variable importance, which is how much the Gini index for each variable decreases at each split, allows us to interpret the impact of the variables on the classification. The Gini index is calculated as the probability of a specific variable classified incorrectly, which is expressed as

$$G(a) = 1 - \sum_{k=1}^{K} p_k^2,$$ (1)

where $G(a)$ is the Gini index of node a, K is the number of classes, and p_k is the probability that the sample belongs to class k.

3.2 Support Vector Machines

A support vector machine (SVM) is a supervised learning algorithm for classification problems (Vapnik 1999). For k-dimensional data, the SVM algorithm determines a $k-1$-dimensional hyperplane that classifies the data points according to which side of the margin the data points fall. The SVM finds support vectors (the closest data point to the hyperplane) and maximizes the margin, which is the gap between the support vectors and hyperplane (Fig. 2b). The optimized hyperplane is determined by using an optimization approach that utilizes Lagrange multipliers and quadratic programming methods (Pal and Mather 2004).

The SVM algorithm can also address nonlinear classification by introducing kernel tricks, where a kernel function adds an additional dimension to the data (Noble 2006). In this study, a radial basis function was used, and the selection of kernel parameters was optimized using a cross-validation method (Cherkassky and Mulier 2007). Furthermore, SVM can be extended to multi-class classification by integrating one-versus-one classifiers for each pair of classes (Mathur and Foody 2008).

3.3 Performance Metrics

The classified dataset used in this study has skewed class proportions (Table 1). Therefore, we applied nested 10-fold cross-validation to downsample the majority and assess the generalization performance (Fig. 3). Stratified 10-fold cross-validation was performed after the majority class was split into 10 folds. Prediction performance was evaluated using the following metrics: precision, recall, and F1 score. The precision is the ratio of correct predictions to the total number of predictions for a particular class. Recall is the ratio of correct predictions to the true total number for a particular class. In other words, precision and recall measure the correctness and completeness of the predictions, respectively. The F1 score is defined as the harmonic mean of precision and recall and is used as an overall indicator.

We used macro-average computation of the average of the metrics obtained for each class for taking the class imbalance into account (Manning 2009). The macro-average of the metric X_i is expressed by the following equation
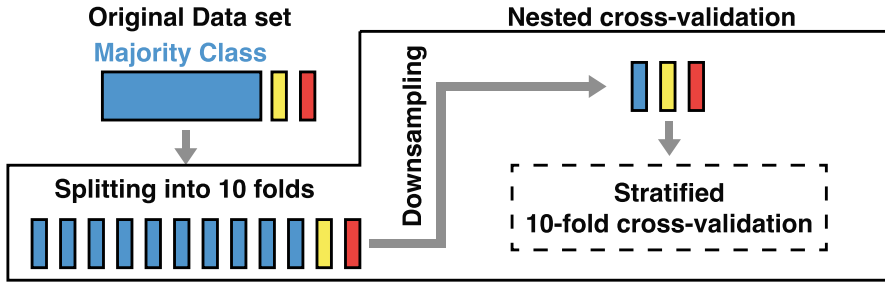
Fig. 3 Schematic overview of nested cross-validation in this study. First, the majority class is divided into 10 folds. Different models are trained and tested with stratified 10-fold cross-validation using the downsampled data set. This process is repeated for each downsampled data set

$$X_{\text{macro}} = \frac{\Sigma_{i=1}^{K} X_i}{K}, \tag{2}$$

where K denotes the number of classes. The macro-average values for each metric were computed by 10-fold cross-validation to avoid overfitting and to make the model more generalizable to the new data (Fig. 3).

3.4 Data Preprocessing

Data preprocessing is a vital step in transforming raw data into a suitable format for a machine learning model. The following steps were performed before model training. Data screening was conducted after selecting the variables used for training, and centered log-ratio (CLR) transformation was further applied to the dataset.

Noisy data caused by the contamination of mineral inclusions were first removed from the dataset. Some data obtained by LA-ICP-MS were affected by REE-bearing minerals such as monazite and apatite. Data with La concentrations of $> 10\,\text{ppm}$ were removed. A few parts per million of La has been used as an empirical criterion (e.g., Sawaki et al. 2022b), and the 10-ppm criterion approximately corresponds to contamination of 0.01 wt% monazite inclusion. Second, outliers that exceeded the interquartile range (IQR) ± 1.5 times the IQR were excluded. Because each class of data was expected to have its own specific data distribution, screening using IQR was performed for each class.

The screened data were converted using CLR transformation. The trace element composition was part of the compositional data with a constant sum (100 wt% in total). Most geochemical data correspond to compositional data (Doucet et al. 2022). The CLR transformation allows statistical analyses to be applied to compositional data (Aitchison 1982). Given a data point x, the ith-centered log-transformed variable z_i is calculated as

$$z_i = \ln(x_i) - \ln(g(x)) \tag{3}$$

$$g(x) = (\Pi_i^N x_i)^{1/N}, \tag{4}$$

where $g(x)$ is the geometric mean for data x, and N is the total number of variables.

The preprocessed data were used for the CART and RF models. In addition, principal component analysis (PCA) was performed on the SVM models as a solution for multicollinearity. Multicollinearity is a problem in which explanatory variables are highly correlated, leading to overfitting or incorrect results. SVM is more susceptible to multicollinearity than RF, and new uncorrelated variables created by PCA were used for the training data.

4 Results

Here, we present a quantitative assessment of classifiability using different input–output combinations. First, we tested the classification of igneous, metamorphic, and hydrothermal origins in a larger framework (16 elements: REE, Th, and U). Secondly, the subdivided categories of igneous rocks were tested (20 elements: Y, Nb, REE, Hf, Th, and U). In addition, the S–I–A–M classification of the granite was tested (21 elements: P, Sc, Y, Nb, REE, Hf, Th, and U). The selection of elements was determined to retain those elements that were geochemically important for classification while still ensuring that the maximum amount of data was utilized in the data analysis.

The first to fourth principal components (PC1–PC4) were used for the prediction models of SVM because these four variables explained the greatest variability in any combination of variables for the classification tasks ($> 80\%$). The results of PCA are summarized in Supplementary Information S1. The optimized hyperparameters for the individual classification models are described in Supplementary Information S1.

4.1 Classification 1: Igneous Versus Metamorphic Versus Hydrothermal

A total of 4,316 data points were employed in the analysis for igneous, 78 for hydrothermal, and 274 for metamorphic after data screening. The igneous data, which is the majority class in this dataset, were randomly divided into 10 splits to balance the imbalanced dataset (Fig. 3). The macro-averaged indices of the prediction accuracy were calculated over stratified 10-fold cross-validations, and the calculated values for each downsampling split were further averaged (Table 2). The prediction accuracy for each class is presented in Supplementary Information S1. The RF and SVM classifications yielded high averaged indices of over 80%, whereas the recall and F1 scores of CART were less accurate. The lower values of recall and F1 score were attributed to the misclassification of hydrothermal and metamorphic origins (Fig. 4a). The values in the diagonal elements of the normalized confusion matrix correspond to the recall values of each class. The variable importance for CART and RF showed a comparable tendency, with U being identified as the most significant variable (Table 3). The selection of La as an important element is a common feature of both algorithms. Heavy REE such as Er and Tm were also selected as relatively important variables for CART, whereas they were not selected for RF. The number of unselected variables was reduced in the case of RF due to the utilization of an ensemble learning approach.

Table 2 Macro-averaged indices of prediction accuracy for each classification model

	Classification 1			Classification 2			Classification 3		
	Precision (%)	Recall (%)	F1 score (%)	Precision (%)	Recall (%)	F1 score (%)	Precision (%)	Recall (%)	F1 score (%)
CART (training)	82	75	79	50	45	47	87	76	81
CART (test)	82	73	77	33	40	36	75	74	74
RF (training)	93	88	90	85	84	84	96	93	94
RF (test)	93	88	90	85	84	85	95	94	95
SVM (training)	89	86	87	87	82	85	98	95	97
SVM (test)	88	84	86	88	81	84	98	96	97

CART — Actual **RF** — **SVM**

(a) Prediction rows: Igneous, Hydrothermal, Metamorphic; columns: I, H, M

CART	I	H	M
Igneous	0.93	0.26	0.30
Hydrothermal	0.01	0.59	0.01
Metamorphic	0.06	0.15	0.69

RF	I	H	M
Igneous	0.97	0.07	0.07
Hydrothermal	0	0.73	0.01
Metamorphic	0.03	0.20	0.92

SVM	I	H	M
Igneous	0.95	0.04	0.12
Hydrothermal	0	0.76	0.03
Metamorphic	0.05	0.20	0.85

(b) Prediction rows: Acidic, Intermediate, Basic, Kimberlite, Carbonatite, Ne-syenite; columns: A, I, B, K, C, N

CART	A	I	B	K	C	N
Acidic	0	0	0	0	0	0
Intermediate	0.90	0.92	0.46	0.01	0.31	1
Basic	0.10	0.08	0.52	0.01	0.10	0
Kimberlite	0	0	0.02	0.98	0.59	0
Carbonatite	0	0	0	0	0	0
Ne-syenite	0	0	0	0	0	0

RF	A	I	B	K	C	N
Acidic	0.64	0.10	0.09	0	0	0
Intermediate	0.28	0.85	0.19	0	0.01	0.13
Basic	0.08	0.03	0.71	0	0	0
Kimberlite	0	0	0.01	1	0	0
Carbonatite	0	0	0	0	0.99	0
Ne-syenite	0	0.02	0	0	0	0.87

SVM	A	I	B	K	C	N
Acidic	0.96	0.46	0.09	0.01	0	0.07
Intermediate	0.03	0.48	0.19	0	0	0.20
Basic	0.01	0.04	0.71	0	0	0
Kimberlite	0	0	0.01	0.99	0.04	0
Carbonatite	0	0	0	0	0.96	0
Ne-syenite	0	0.02	0	0	0	0.73

(c) Prediction rows: S-type, I-type, A-type, M-type; columns: S, I, A, M

CART	S	I	A	M
S-type	0.43	0	0	0.01
I-type	0.12	0.77	0	0.11
A-type	0	0	0.89	0.02
M-type	0.45	0.23	0.11	0.87

RF	S	I	A	M
S-type	0.95	0.01	0	0
I-type	0.05	0.93	0	0.11
A-type	0	0	1	0
M-type	0	007	0	0.89

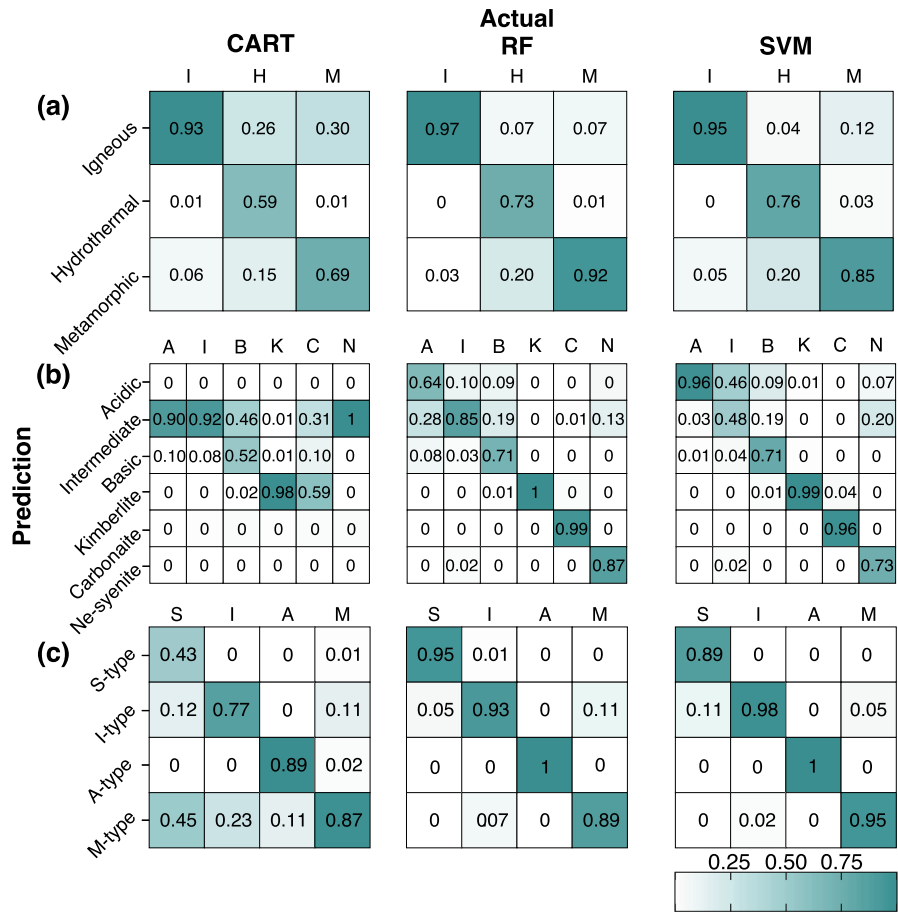| SVM	S	I	A	M
S-type	0.89	0	0	0
I-type	0.11	0.98	0	0.05
A-type	0	0	1	0
M-type	0	0.02	0	0.95

0.25 0.50 0.75

Fig. 4 Normalized confusion matrices for classification between **a** igneous, hydrothermal, and metamorphic origins, **b** acidic, intermediate, basic, kimberlite, carbonatite, and nepheline syenite, and **c** S-, I-, A-, M-type granites. The rows and columns of the matrix represent the predicted and actual classes, respectively. To calculate the average over 10-fold cross-validation for test datasets, the values of each element were normalized assuming the sample size for each class was 1. Diagonal components correspond to the recall of each class

4.2 Classification 2: Source Rock Types of Igneous Zircons

The number of screened data points was 1,936 for acidic, 270 for intermediate, 227 for basic, 166 for kimberlite, 47 for carbonatite, and 30 for nepheline syenite. The acidic class was downsampled in this classification task. The macro-averaged indices demonstrated that the RF and SVM models improved classification performance remarkably (Table 2). Classification of basic, carbonatite, and nepheline syenite was not feasible using CART (Fig. 4b). Although it is still difficult to distinguish between the acidic and intermediate classes, the classification of other classes was improved for the RF and SVM models. The CART model tended to show high variable importance values

Table 3 Averaged variable importance for classification over 10-fold cross-validation

	Classification 1		Classification 2		Classification 3	
	CART	RF	CART	RF	CART	RF
P	–	–	–	–	31.1	54.7
Sc	–	–	–	–	72.9	85.8
Y	–	–	54.7	53.6	72.2	77.2
Nb	–	–	0	56.7	72.8	96.4
La	62.9	41.3	0	19.8	0	8.9
Ce	27.3	38.2	2.6	49.2	21.0	61.4
Pr	0.8	11.1	0	38.0	0	2.3
Nd	0	0.3	0	3.4	0	0.8
Sm	0.1	15.5	0	1.3	0	0.3
Eu	41.5	44.8	0	46.9	15.0	40.8
Gd	0.7	24.7	27.9	17.2	0	7.4
Tb	0.4	10.0	30.7	35.0	0	18.5
Dy	0.4	6.4	6.5	15.4	4.5	26.3
Ho	9.6	10.2	12.9	32.7	40.2	49.2
Er	61.5	16.5	97.0	60.5	75.4	95.2
Tm	58.9	16.7	98.8	85.0	38.8	60.2
Yb	28.5	12.6	98.8	95.2	0	31.4
Lu	9.7	8.8	99.7	90.6	0	18.9
Hf	–	–	40.0	74.8	0	18.0
Th	0	2.1	38.7	76.6	14.5	31.2
U	99.7	100.0	32.3	73.7	18.9	41.8

for certain elements, such as heavy REE (Table 3). In contrast, the RF model selected Y, Nb, Ce, Eu, Hf, Th, and U in addition to heavy REE as important variables.

4.3 Classification 3: Granitic Rocks

The number of screened data points used in the analysis was 30 for S-type, 134 for I-type, 45 for A-type, and 124 for M-type granites. The skewness of the dataset was relatively small; therefore, only stratified 10-fold cross-validation was employed without downsampling. The classification models using RF and SVM were more precise than that using CART (Table 2). In particular, it significantly improved the precision and recall of S-type granite classification (Fig. 4c). The CART and RF models exhibit similar variable importance tendencies. Sc, Nb, and Er are the most influential elements in the classification, followed by P, Eu, and Ce (Table 3).
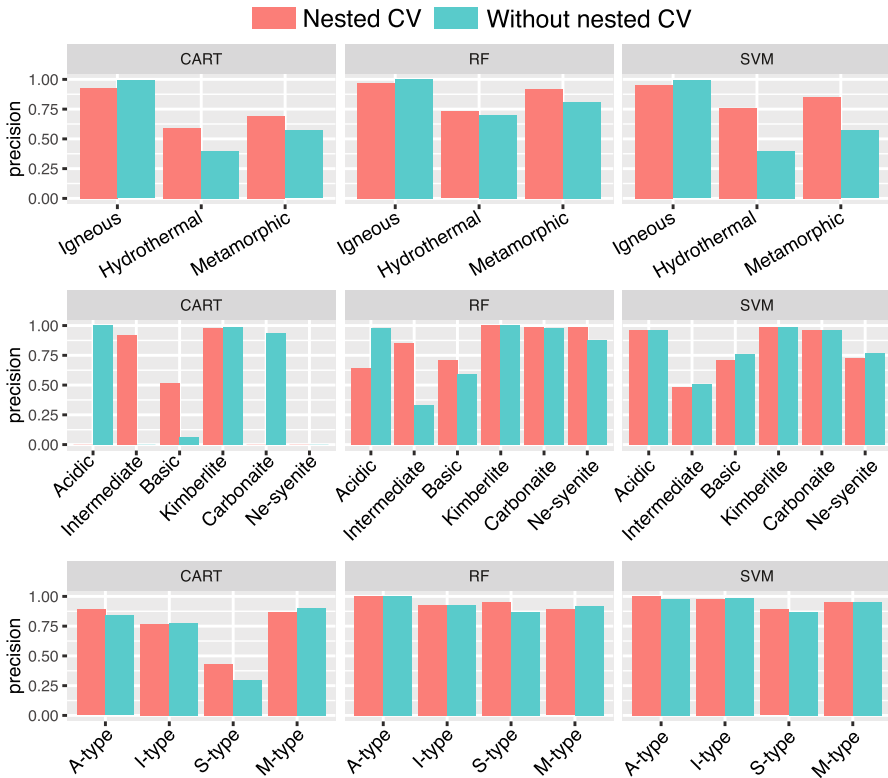
Fig. 5 Effects of nested cross-validations for the CART, RF, and SVM models. The red and blue bars represent the averaged accuracy with and without nested cross-validation, respectively

5 Discussion

5.1 Prediction Performance

Down-sampling and stratified cross-validation were combined to address the issue of an imbalanced dataset (Fig. 3). The precision of hydrothermal and metamorphic classes for SVM and the basic class for RF were improved (Fig. 5); therefore, the nested cross-validation contributes to better classification.

The CART model is too simple to handle the new dataset. This study found that it was difficult to classify alkaline and non-alkaline rocks using the CART algorithm (Figs. 4, 5; Table 2), which was valid in a previous study (Belousova et al. 2002). The RF and SVM models solved this underfitting issue (Fig. 4; Table 2). Furthermore, the small differences between the results of the training and test datasets suggest that overfitting was suppressed (Table 2). Thus, the obtained results indicate the reliability and effectiveness of the classification models for predicting the origins of zircons.
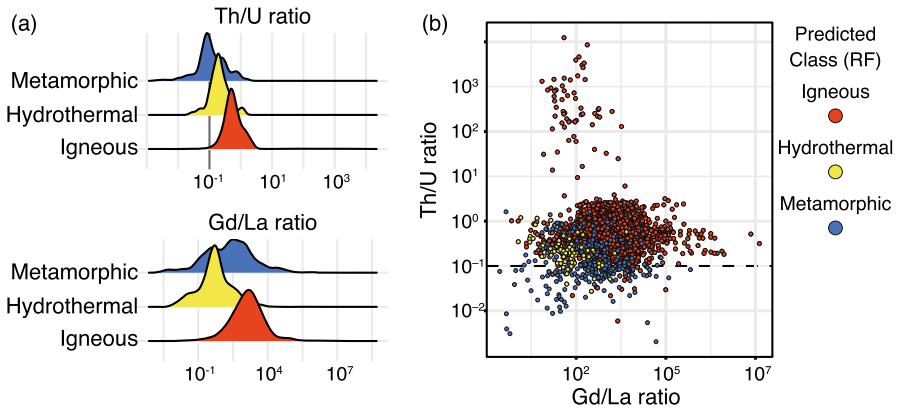
Fig. 6 Variations in Th/U and Gd/Lu ratios and visualization of predicted results in Th/U and Gd/Lu space. Igneous zircon shows different data distributions from other types despite some degree of overlapping. **a** Probability density distributions of Th/U and Gd/Lu ratios. **b** Predicted rock types by random forest classification in Th/U and Gd/Lu space

5.2 Distinguishing Igneous Zircon from Metamorphic and Hydrothermal Zircons

Distinguishing igneous origins from hydrothermal and metamorphic origins is fundamental for interpreting the U–Pb ages of zircons. The results for the RF and SVM models indicate that igneous zircons can be identified with approximately 90% accuracy using rare earth elements, Th, and U. Uranium is the most important explanatory variable for this classification based on the variable importance (Table 3), which is consistent with previous geochemical studies.

The Th/U ratio has been found to be an effective indicator for distinguishing between igneous and metamorphic origins (Rubatto 2017). It is commonly observed that igneous zircon tends to have a Th/U ratio above 0.1, whereas metamorphic zircon tends to have a Th/U ratio below 0.1, although the systematics of Th and U in zircon are not always simple because of the relationship between zircon growth and the breakdown of coexisting minerals, such as monazite (Yakymchuk et al. 2018). The dataset used in this study included zircon data from high-grade metamorphic rocks with Th/U ratios exceeding 0.1 (e.g., Yang et al. 2021). Nevertheless, the distributions of the igneous and metamorphic zircons were different (Fig. 6a).

Lanthanum, which possesses highly variable importance values, demonstrates distinct distribution patterns across various rock types. The igneous zircon exhibits a higher Gd/La ratio than the other zircon types (Fig. 6a). The lower Gd/La ratios of hydrothermal and metamorphic zircons can be attributed to the enrichment in light REE due to metamictization (Hoskin and Schaltegger 2003). These discrepancies also contribute to the high accuracy of the igneous zircon classification (Fig. 6b). However, the large overlap between metamorphic and hydrothermal zircons could lead to misclassification of these origins.

The geochemical differences between igneous zircons and the other types are also represented in the PCA space used for SVM (Fig. 7). Based on the principal component loadings, each component can be geochemically interpreted to reflect the following
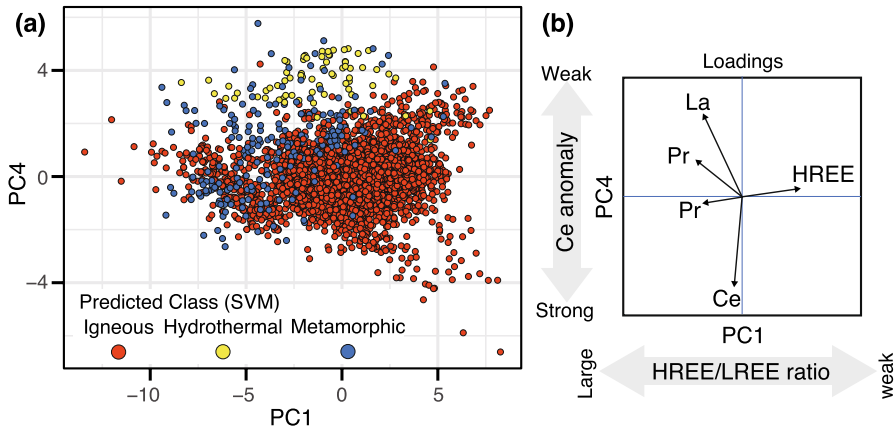
Fig. 7 Data distribution in principal component (PC) space. **a** PC scores of zircon data with predicted rock type label by SVM classification model in PC1 and PC4 space. **b** Loadings of PC1 and PC4. PC1 and PC4 represent the ratio of heavy rare earth elements (HREE) to light rare earth elements (LREE) and the degree of Ce anomaly, respectively

factors: light REE/heavy REE ratio (PC1), the relative abundance of U–Th and middle REE (PC2), Eu anomaly (PC3), and Ce anomaly (PC4). Distinctions in the rock types were most apparent in the PC1–PC4 space (Fig. 7a), and the other components are summarized in Supplementary Information S1. The metamorphic zircons were plotted within the region of the igneous zircons; in contrast, the hydrothermal zircons tended to have higher PC4 scores.

This observation is attributed to the classification of hydrothermal and igneous origins with high accuracy by SVM (Fig. 4a). The PC4 loading indicated that this component reflected the degree of decoupling of Ce from La and Pr, known as the Ce anomaly (Fig. 7b), consistent with the small Ce anomalies in naturally observed cases (e.g., Toscano et al. 2014; Yuan et al. 2018).

Overall, igneous zircons can be distinguished from metamorphic and hydrothermal zircons, whereas the classification of hydrothermal and metamorphic zircons remains imprecise. Cathodoluminescence (CL) imaging is a reliable method for the practical classification of detrital zircon (Corfu et al. 2003; Zheng et al. 2022). The geochemical classification can serve as a complementary approach to independently provide information on the origin of detrital zircons, particularly when classification using CL images is inconclusive.

5.3 Identification of Source Magma Type

A substantial amount of newly added data highlights the challenge of this classification task using the traditional CART (Fig. 4b, Belousova et al. 2002). In contrast, the RF and SVM models allowed us to distinguish zircons crystallized from alkaline (kimberlite, carbonatite, and nepheline-syenite) and non-alkaline magma (acidic, intermediate, and basic) with high precision (Fig. 4b). One interesting finding is that the RF and SVM
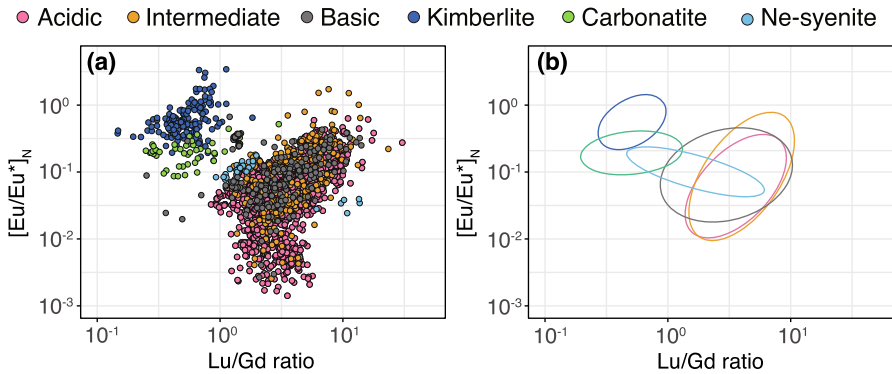
Fig. 8 Variations in elemental concentrations for igneous zircons. **a** Scatter plot of Lu versus Eu and **b** 95% confidence ellipses for each rock type assuming multivariate t distributions

models show promise and potential for the classification of the basic class. At the same time, it is still challenging to distinguish the acidic and intermediate classes (Fig. 4b).

Heavy rare earth elements such as Lu play the most important roles in classification based on variable importance values (Table 3). The selection of these elements for the classification trees was consistent with the results of Belousova et al. (2002). The elemental ratios related to these elements can help visualize their contributions to the discrimination ability (Fig. 8): Lu/Gd ratio and Eu anomaly ($[Eu/Eu^*]_N$: CI-chondrite-normalized Eu and Eu^*, defined by the geometric mean of the CI-chondrite-normalized Sm and Gd).

These elemental ratios, commonly used in geochemistry, are predominantly controlled by the source magma chemistry and the crystallographic characteristics of zircon. The positive correlations between Lu/Gd ratio and Eu anomaly within each source rock type reflect differences in the degree of magma differentiation. The Eu anomaly is generally caused by the decoupling of Eu^{2+} from other REE^{3+} via feldspar fractionation. Therefore, a wide range of magma differentiation results in a correlation between the Lu/Gd and Eu anomaly. Furthermore, zircons from alkaline rocks show weaker Eu anomalies and lower Lu/Gd ratios, reflecting the low oxidation state and low Lu/Gd ratio of alkaline magmas (Möller et al. 1980; Shnyukov et al. 1989). On the basis of this rationale, the established models accurately distinguish zircons from alkaline rocks. However, a significant overlap within non-alkaline magmas makes it difficult to discriminate the differences in SiO_2 content using the available elemental data.

Discriminating the differences in SiO_2 in the PC space is also challenging (Fig. 9). While the alkaline rocks were separated in the space of PC2–PC3–PC4, the non-alkaline rocks overlapped (Fig. 9a). Each principal component was related to the relative abundance of light REE to heavy REE (PC2), the correlation of high-field-strength elements (Nb, Th, and U) and light REE (PC3), and the decoupling of La and Ce (Fig. 9b). SVM analysis with data preprocessing in principal component analysis has improved classification performance; however, distinguishing zircons sourced from acidic and intermediate rocks remains a challenge.
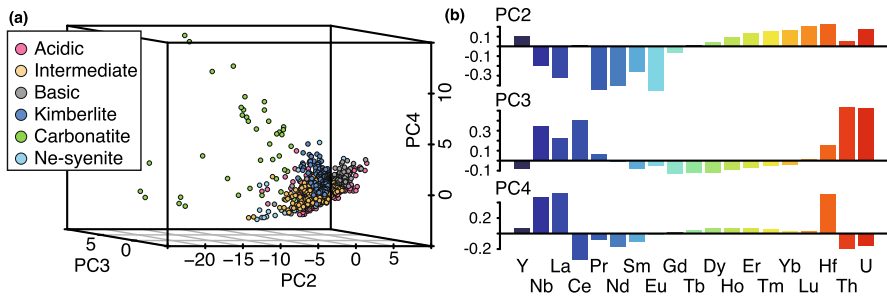
Fig. 9 Results of principal component analysis for igneous zircon data. **a** Three dimensional scatter plots for second to fourth principal components (PC2–PC4) with actual labels of rock type. **b** Loadings for the principle components

5.4 S–I–A–M Granite Classification

The RF and SVM algorithms successfully achieved unified classification models for all types of S–I–A–M granites (Fig. 4). Because Sc, Nb, and heavy REE were found to be important factors for predictive performance in these classification models (Table 3), we focused on identifying the factors controlling their behaviors.

The abundance of P in zircon is key for discriminating between S-type and I-type granites. Burnham and Berry (2017) demonstrated that the molar concentrations of P and (REE + Y) in S-type granite zircon exhibit a 1:1 correlation, whereas such correlation is absent in zircon from I-type granite (Fig. 10a). The total amount of REE + Y mainly depends on that of Y + heavy REE. The strong 1:1 correlation for the S-type granite was interpreted to be the result of the coupled substitution of P and REE for Zr (Burnham and Berry 2017). In contrast, the vacancy-related mechanism for REE substitution (Burnham and Berry 2012; Trail et al. 2011) can be attributed to deviations from the 1:1 straight line for other types. Although this diagram was originally developed to discriminate between S-type and I-type, we found different slopes for the A-type and M-type (Fig. 10a). This may be related to the different proportions of the substitution processes during crystallization.

It was recently revealed that high-field-strength elements (HFSE), such as Nb and Sc, are useful for monitoring the difference in source materials and subsequent crystal fractionation (Sawaki et al. 2022b; Grimes et al. 2015), which is consistent with our observation that Sc and Nb were selected as important variables (Table 3). Figure 10b visualizes the difference in zircons sourced from the S–I–A–M granite using these elements.

S-type zircons with lower Nb/P and Ce/P ratios could be attributed to the characteristics of parental magmas derived from sediments with high amounts of P relative to HFSE or the high solubility of apatite due to its peraluminous composition (Chappell and White 2001). In contrast, A-type zircons with higher Nb/P and Ce/P ratios mirror the plume-related characteristics of A-type granite magma with high HFSE content (Bonin 2007; Grimes et al. 2015). I- and M-type zircons show a similar intermediate range of Ce/P ratios, likely reflecting a common mantle-derived source material; however, the Nb/P ratio depicts a parallel array (Fig. 10b). Sawaki et al. (2022b)
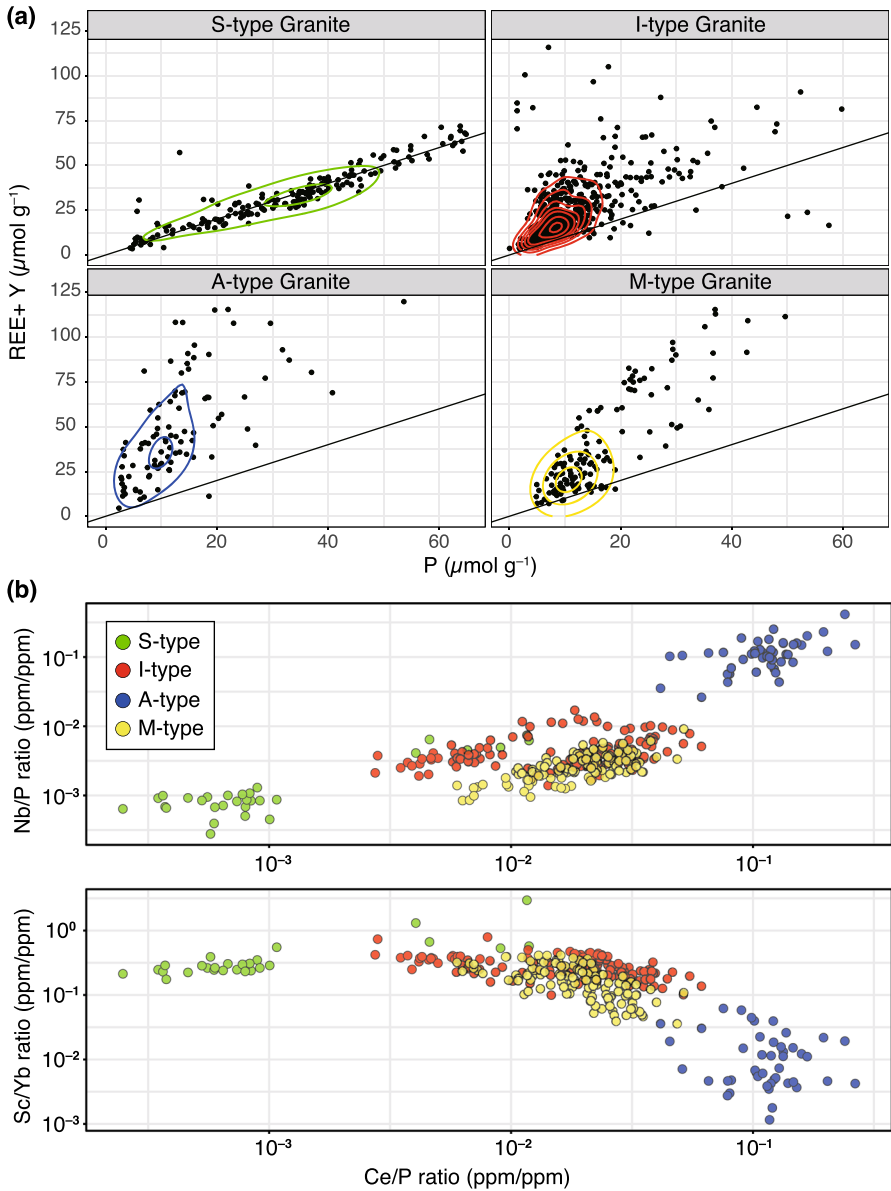
Fig. 10 Geochemical interpretation diagrams for zircon sourced from granitic rocks. **a** Binary plots with probability density contours depicting the relationship between the concentrations of REE + Y and P. Straight lines of slope 1 represent the relationship of these elements expected for REE incorporation via the xenotime substitution. All data before screening are plotted. **b** The fields of zircon composition used for training the classification model. Ce/P and Nb/P ratios were previously used as proxies for the involvement of sediments in magma (Sawaki et al. 2022b, a), and the Sc/Yb ratio was originally used as a proxy for tectono-magmatic provenance related to amphibole fractionation (Grimes et al. 2015)
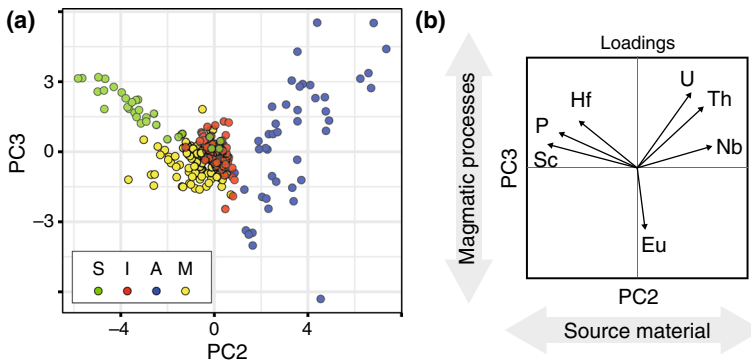
Fig. 11 Data distribution in principal component (PC) space for zircon sourced from S–I–A–M granitic rocks. **a** Scores of PC2 and PC4. **b** Loadings of PC2 and PC3. Representative variables with large values of loading are displayed

reported that the elevated Nb/P ratio of I-type zircon reflects the sediment involvement compared to M-type zircon. The metaluminous composition of the I- and M-type zircons would have resulted in the early crystallization of apatite and subsequent zircon crystallization, unlike zircon crystallization prior to apatite for the peraluminous composition of S-type granite. The Sc/Yb ratio diagram highlights the differences in A-type zircons caused by significant amphibole fractionation (Grimes et al. 2015).

The different features of the granite types were also visualized in the PC space (Fig. 11a). The S- and A-type zircons are plotted in different domains, and the I- and M-type zircons are distributed in different areas with some overlap. The variables with higher loading values were consistent with those with higher variable importance (Fig. 11b). PC2 representing the differences in HFSE and P content best reflects the differences in the source material of the granitic magma. On the other hand, PC3 can be interpreted as magmatic processes rather than the difference in source materials. The loading of PC3 showed that Eu and U were most strongly associated with large variations in S- and A-type zircons (Fig. 11b). The Eu content of felsic magma is generally controlled by fractionation of feldspars (Bea 1996). In contrast, the behavior of U is complicated due to the various timings of U-bearing accessory minerals including zircon, monazite, xenotime, and titanite (Sawaki et al. 2022b, a). Therefore, PC3 may represent a magmatic differentiation trend for each type.

5.5 Limitations and Future Research Directions

The P, Sc, and Nb data discussed in the previous section have typically been used in studies on granites to elucidate the chemical evolution of granitic magmas. However, isobaric and polyatomic ion interferences have been obstacles to the accurate determination of Sc and P by the widely used quadrupole-type ICP-MS. Data for Sc have been limited to SIMS data (e.g., Carley et al. 2011; Grimes et al. 2015), and triple quadrupole ICP-MS, also referred to as ICP-MS/MS, has recently enabled LA-ICP-MS analysis of these elements (e.g., Sawada et al. 2019, 2022a). The development of

a comprehensive dataset is necessary to decipher the differences in the source magma composition and various crystal fractionation trends from zircon trace element chemistry. In particular, P and Sc, which could not be included in all classification analyses due to the limited amount of data, could play critical roles in distinguishing zircons sourced from intermediate and basic rocks.

The distinction between the acidic and intermediate classes remained challenging; however, the classification of the basic rock class yielded relatively accurate predictions (Fig. 4b; Supplementary Information S1). This unexpected result suggests a feasible rough estimate of the degree of magmatic differentiation (SiO_2). The increasing availability of trace element data for zircons in mafic rocks in the future could allow for the verification of reliability, which would be beneficial for detrital zircon studies investigating the geochemical characteristics of the early Earth's crust.

6 Conclusions

This study quantitatively assessed the geochemical classification of zircon source rocks and demonstrated their potential utility for identifying zircon origins in different layers. The recent increase in zircon trace element data has rendered CART-based classification more arduous, owing to the overlapping composition data of individual rock types. However, the random forest (RF) and support vector machine (SVM) models achieved classification with overall F1 scores exceeding 80%. The variables that significantly affected the classification performance were consistent with the elements commonly used in geochemical indices, which also supports the reliability of the classification. In terms of future work, it would be interesting to establish a dataset of diverse rock types, including key elements that are difficult to measure, such as P and Sc, for better classification models.

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

Aitchison J (1982) The statistical analysis of compositional data. J R Stat Soc Ser B Methodol 44(2):139–160

Allègre CJ, Rousseau D (1984) The growth of the continent through geological time studied by Nd isotope analysis of shales. Earth Planet Sci Lett 67(1):19–34

Bea F (1996) Residence of REE, Y, Th and U in granites and crustal protoliths; implications for the chemistry of crustal melts. J Petrol 37(3):521–552

Belgiu M, Drăguţ L (2016) Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogramm Remote Sens 114:24–31

Belousova E, Griffin WL, O'Reilly SY, Fisher N (2002) Igneous zircon: trace element composition as an indicator of source rock type. Contrib Miner Petrol 143(5):602–622

Bonin B (2007) A-type granites and related rocks: evolution of a concept, problems and prospects. Lithos 97(1–2):1–29

Bowring SA, Schmitz MD (2003) High-precision U-Pb zircon geochronology and the stratigraphic record. Rev Mineral Geochem 53(1):305–326

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

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

Carley TL, Miller CF, Wooden JL, Bindeman IN, Barth AP (2011) Zircon from historic eruptions in Iceland: reconstructing storage and evolution of silicic magmas. Mineral Petrol 102:135–161

Cawood PA, Hawkesworth C, Dhuime B (2012) Detrital zircon record and tectonic setting. Geology 40(10):875–878

Chappell BW, White AJ (2001) Two contrasting granite types: 25 years later. Aust J Earth Sci 48(4):489–499

Cherkassky V, Mulier FM (2007) Learning from data: concepts, theory, and methods. Wiley, New York

Corfu F, Hanchar JM, Hoskin PW, Kinny P (2003) Atlas of zircon textures. Rev Mineral Geochem 53(1):469–500

Do KT, Wahl S, Raffler J, Molnos S, Laimighofer M, Adamski J, Suhre K, Strauch K, Peters A, Gieger C, Langenberg C, Stewart ID, Theis FJ, Grallert H, Kastenmüller G, Krumsiek J (2018) Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. Metabolomics 14(10):1–18

Doucet LS, Tetley MG, Li ZX, Liu Y, Gamaleldien H (2022) Geochemical fingerprinting of continental and oceanic basalts: a machine learning approach. Earth Sci Rev 66:104192

Ewing R, Lutze W, Weber WJ (1995) Zircon: a host-phase for the disposal of weapons plutonium. J Mater Res 10:243–246

Gaschnig RM (2019) Benefits of a multiproxy approach to detrital mineral provenance analysis: an example from the Merrimack River, New England, USA. Geochem Geophys Geosyst 20(3):1557–1573

Gehrels G (2014) Detrital zircon U–Pb geochronology applied to tectonics. Annu Rev Earth Planet Sci 42(11):127–149

Grimes CB, John BE, Kelemen P, Mazdab F, Wooden J, Cheadle MJ, Hanghøj K, Schwartz J (2007) Trace element chemistry of zircons from oceanic crust: a method for distinguishing detrital zircon provenance. Geology 35(7):643–646

Grimes C, Wooden J, Cheadle M, John B (2015) "fingerprinting" tectono-magmatic provenance using trace elements in igneous zircon. Contrib Mineral Petrol 170:1–26

Hoskin PW, Ireland TR (2000) Rare earth element chemistry of zircon and its use as a provenance indicator. Geology 28(7):627–630

Hoskin PW, Schaltegger U (2003) The composition of zircon and igneous and metamorphic petrogenesis. Rev Mineral Geochem 53(1):27–62

Iizuka T, Yamaguchi T, Itano K, Hibiya Y, Suzuki K (2017) What Hf isotopes in zircon tell us about crust-mantle evolution. Lithos 274:304–327

Itano K, Ueki K, Iizuka T, Kuwatani T (2020) Geochemical discrimination of monazite source rock based on machine learning techniques and multinomial logistic regression analysis. Geosciences 10(2):63

Kampichler C, Wieland R, Calmé S, Weissenberger H, Arriaga-Weiss S (2010) Classification in conservation biology: a comparison of five machine-learning methods. Ecol Inform 5(6):441–450

Manning CD (2009) An introduction to information retrieval. Cambridge University Press, Cambridge

Mathur A, Foody GM (2008) Multiclass and binary SVM classification: implications for training and classification users. IEEE Geosci Remote Sens Lett 5(2):241–245

Möller P, Morteani G, Schley F (1980) Discussion of ree distribution patterns of carbonatites and alkalic rocks. Lithos 13(2):171–179

Noble WS (2006) What is a support vector machine? Nat Biotechnol 24(12):1565–1567

Olden JD, Lawler JJ, Poff NL (2008) Machine learning methods without tears: a primer for ecologists. Q Rev Biol 83(2):171–193

Pal M, Mather PM (2004) Assessment of the effectiveness of support vector machines for hyperspectral data. Future Gener Comput Syst 20(7):1215–1225

Pereira I, Storey CD (2023) Detrital rutile: records of the deep crust, ores and fluids. Lithos 66:107010

Petrelli M, Perugini D (2016) Solving petrological problems through machine learning: the study case of tectonic discrimination using geochemical and isotopic data. Contrib Miner Petrol 171(10):1–15

Rino S, Komiya T, Windley BF, Katayama I, Motoki A, Hirata T (2004) Major episodic increases of continental crustal growth determined from zircon ages of river sands; implications for mantle overturns in the Early Precambrian. Phys Earth Planet Inter 146(1–2):369–394

Rubatto D (2017) Zircon: the metamorphic mineral. Rev Mineral Geochem 83(1):261–295

Sawada H, Isozaki Y, Sakata S, Hirata T, Maruyama S (2018) Secular change in lifetime of granitic crust and the continental growth: a new view from detrital zircon ages of sandstones. Geosci Front 9(4):1099–1115

Sawada H, Isozaki Y, Aoki S, Sakata S, Sawaki Y, Hasegawa R, Nakamura Y (2019) The late Jurassic magmatic protoliths of the Mikabu greenstones in sw Japan: a fragment of an oceanic plateau in the Paleo-Pacific Ocean. J Asian Earth Sci 169:228–236

Sawada H, Niki S, Nagata M, Hirata T (2022a) Zircon U–Pb–Hf isotopic and trace element analyses for oceanic mafic crustal rock of the neoproterozoic-Early Paleozoic Oeyama ophiolite unit and implication for subduction initiation of proto-Japan arc. Minerals 12(1):107

Sawaki Y, Asanuma H, Sakata S, Abe M, Ohno T (2022a) Trace-element composition of zircon in Kofu and Tanzawa granitoids, Japan: quantitative indicator of sediment incorporated in parent magma. Island Arc 31(1):e12455

Sawaki Y, Asanuma H, Sakata S, Abe M, Ohno T (2022b) Zircon trace-element compositions in Miocene granitoids in Japan: discrimination diagrams for zircons in M-, I-, S-, and A-type granites. Island Arc 31(1):e12466

Shnyukov S, Cheburkin A, Andreev A (1989) Geochemistry of wide-spread coexisting accessory minerals and their role in investigation of endogenetic and exogenetic processes. Geol J 2:107–14

Toscano M, Pascual E, Nesbitt R, Almodóvar G, Sáez R, Donaire T (2014) Geochemical discrimination of hydrothermal and igneous zircon in the Iberian Pyrite Belt, Spain. Ore Geol Rev 56:301–311

Trail D, Watson EB, Tailby ND (2011) The oxidation state of hadean magmas and implications for early earth's atmosphere. Nature 480(7375):79–82

Ueki K, Hino H, Kuwatani T (2018) Geochemical discrimination and characteristics of magmatic tectonic settings: a machine-learning-based approach. Geochem Geophys Geosyst 19(4):1327–1347

Valley JW (2003) Oxygen isotopes in zircon. Rev Mineral Geochem 53(1):343–385

Vapnik V (1999) The nature of statistical learning theory. Springer, Berlin

Wang J, Hattori K, Yang Y, Yuan H (2021) Zircon chemistry and oxidation state of magmas for the Duobaoshan–Tongshan ore-bearing intrusions in the Northeastern Central Asian Orogenic Belt, NE China. Minerals 11(5):503

Whalen JB (1985) Geochemistry of an island-arc plutonic suite: the Uasilau–Yau Yau intrusive complex, New Britain, PNG. J Petrol 26(3):603–632

Yakymchuk C, Kirkland CL, Clark C (2018) Th/u ratios in metamorphic zircon. J Metamorph Geol 36(6):715–737

Yang Y, Liang C, Zheng C, Xu X, Zhou J, Zhou X, Cao C (2021) Metamorphic evolution of high-grade granulite-facies rocks of the Mashan Complex, Liumao area, eastern Heilongjiang Province, China: evidence from zircon U–Pb geochronology, geochemistry and phase equilibria modelling. Precambr Res 355:106095

Yuan F, Liu JJ, Carranza EJM, Zhang S, Zhai DG, Liu G, Wang GW, Zhang HY, Sha YZ, Yang SS (2018) Zircon trace element and isotopic (Sr, Nd, Hf, Pb) effects of assimilation-fractional crystallization of pegmatite magma: a case study of the Guangshigou biotite pegmatites from the North Qinling Orogen, central China. Lithos 302:20–36

Zhao Y, Zhang Y, Geng M, Jiang J, Zou X (2019) Involvement of slab-derived fluid in the generation of cenozoic basalts in Northeast China inferred from machine learning. Geophys Res Lett 46(10):5234–5242

Zheng D, Wu S, Ma C, Xiang L, Hou L, Chen A, Hou M (2022) Zircon classification from cathodoluminescence images using deep learning. Geosci Front 13(6):101436

Zhong S, Liu Y, Li S, Bindeman I, Cawood P, Seltmann R, Niu J, Guo G, Liu J (2023) A machine learning method for distinguishing detrital zircon provenance. Contrib Miner Petrol 178(6):35

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

Ziyi Z, Fei Z, Yu W, Tong Z, Zhaoliang H, Kunfeng Q (2022) Machine learning-based approach for zircon classification and genesis determination. Earth Sci Front 29(5):464