The discrimination of tectonic setting Using trace elements in zircons： A machine learning approach

Luyuan Wang
Shandong University of Technology
Chao Zhang (czhang@sdut.edu.cn)
Shandong University of Technology
Rui Geng
Shandong Gold Group Co. LTD
Yuqi Li
Shandong University of Technology
Jijie Song
Shandong University of Technology
Bin Wang
Shandong Provincial No.6 Exploration Institute of Geology and Mineral Resources
Fanghua Cui
Shandong University of Technology

Research Article

Keywords: machine learning, eXtreme Gradient Boosting, rare earth elements of magmatic zircons, tectonic setting

Posted Date: January 9th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2408345/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Earth Science Informatics on November 15th, 2023. See the published version at https://doi.org/10.1007/s12145-023-01142-0.

Abstract

Zircon is the most important accessory mineral in geological research, and they record information on isotopes and trace elements which is of great significance in earth science research. Trace elements in Zircons can be used for analyzing the genesis of zircons, calculating the magma temperature and oxygen fugacity, and tracing the magma source. Due to the limitation of visual dimensions, the information on the zircons is mainly shown with the method of low dimensional diagrams in the present studies, so the high dimensional relationships during trace elements of the zircons are difficult to be discovered. However, with the development of machine learning, mining the high dimensional relationships during the trace elements of the zircons becomes possible. In this paper, four supervised learning algorithms including Random Forest, Support Vector Machine, Decision Tree, and eXtreme Gradient Boosting have been implemented to analyze trace elements of 3907 magmatic zircons from the GEOROC database, and a precise 13-dimensional data classifier model has been established in order to distinguish the tectonic settings of the rift, ocean island, and convergent margin. Based on the results of accuracy, precision, recall, and F1-score, the machine learning approach of eXtreme Gradient Boosting is best in the paper and the results of Accuracy, Precision, Recall, and F1-score are 0.948, 0.941, 0.922, 0.930, respectively. In summary, eXtreme Gradient Boosting in the paper could provide a high-dimensional discriminative approach to distinguish the tectonic settings.

1 Introduction

Zircon is a common accessory mineral in the intermedium-felsic igneous rocks (Zhong *et al.* 2020), they are featured by physical-chemical durability and resistance to alteration. So, the geochemical trace elements of the zircons could effectively record and reveal the process of geological evolutions including crustal assimilation, magma mixing, crustal cycling, and metallogenetic process (Kemp et al. 2007; Szilas et al. 2013; Van Kranendonk and Kirkland 2013; Grimes et al. 2015; Roberts and Spencer 2015; Buret et al. 2016; Lu *et al.* 2016; Spencer et al. 2017; Gao and Santosh 2020; Palin et al. 2020; Xing et al. 2020). With the development of LA-ICP-MS technology in the zircon situ analysis, researchers have obtained a large number of zircons from different tectonic settings, and analyze U-Pb geochronology and the distribution of trace elements in the zircons (Belousova et al. 2002; Grimes et al. 2007; Carley et al. 2014; Kirkland et al. 2015; Grimes et al. 2015; Zou et al. 2021). The trace elements of the zircons provide a sensitive monitor for reflecting its parental magma composition (Barth et al. 2013; Carley et al. 2014; Belousova et al. 2015) and are also used for distinguishing zircon genetic types (Zhong et al. 2018), tracing magmatic source (Bell 2017; Drabon *et al.* 2021), and calculating magmatic temperature (Siégel et al. 2018) and oxygen fugacity (Loader et al. 2017; Zou et al. 2019).

Under thermodynamic equilibrium conditions, the content of trace elements in the zircons could reflect the composition of trace elements in the melt. So, the trace elements in the zircons are important ways to identify the forming conditions, evolution process, and the source of the magma with the method of the indicators, discriminant diagrams, and partitioning coefficients of trace elements in the zircons. The indicators such as U/Yb, Hf, Nb/Yb, Sc/Yb, and Lu/Hf are composed of the content or ratios of a

certain/several elements in the zircons (Grime *et al.* 2007, 2015; Guo et al. 2017) while the discriminant diagrams are developed on the pairs of the indicators. For example, the U/Yb value of the continental zircons is significantly higher than that of the ocean crust zircons, while the contents of Hf and Y in the continental zircons are less than those of the ocean crust Zircons. The indicators mentioned above have been used for discriminating oceanic zircons from continental zircons (Fig. 1a). Moreover, the contents of Sc, Ti, Th, and Nb are combined to establish a discrimination diagram for the tectonic setting (Fig. 1b).

Due to the limitation of the visual dimensions, the studies based on the indicators or diagrams can only simultaneously show the relationships of no more than three elements of content/ratios in the zircons. However, there are more than 50 kinds of trace elements contained in the zircons from the GEOROC database (Zou et al. 2021), all of the elements in the zircons are accord with the principle of charge conservation, lattice strain model, and isomorphic substitution mechanism, which makes cooperative or competitive substitution relationships between any two trace elements of the zircons exist in the forming process of the zircons. However, low dimensional indicators and diagrams of trace elements in the zircons neither fully demonstrate this cooperative or competitive substitution relationship, nor do they effectively reflect the high dimensional relationships of zircon trace elements (Zou et al. 2021).

The machine learning approach could solve the problems of visual dimensional limitations. High-dimensional ways based on the machine learning approach make it possible to mine hidden internal relationships between different elements in the zircons (Zhu et al. 2022).To mine the relationships between zircon rare earth trace elements for establishing tectonic setting classification, we adopt four machine learning algorithms including Random Forest (RF), Support Vector Machine (SVM), Decision Tree (Tree), and eXtreme Gradient Boosting (XGBoost) to analyze rare earth element data of 3907 zircons collected from the convergent margin, oceanic island and rift volcanic settings in the GEOROC database. In addition, sixteen classifier models of reference datasets have also been established based on the different machine learning approaches, classification methods, and normalization processing methods. Compared with the results of accuracy, precision, recall, and F1-score, a tectonic setting discrimination model is finally proposed in the approach of XGBoost.

2 Resource Of Reference Datasets

The reference datasets in this paper are from the REEs of 13475 zircons collected from the convergent margin, ocean island, and rift volcanic tectonic settings in the GEOROC database.

To ensure the reliability and accuracy of the results, we select magmatic zircons and remove the incomplete data of REEs. Three thousand eight hundred and seven data have been selected as the data training set composed of 1203 data from rift volcanic zircons, 299 data from ocean island zircons data, and 2405 data from convergent margin zircons data (Supplementary Table 1). The parameter of the data training set including extremum, average, median, lower, and upper quartiles have been shown in Table 1, while the concentrations of REEs are displayed in Fig. 2.

The REEs contents of zircons from rift volcanic setting are as follows: The value of La is 0.001×10^{-6}~1023.750×10^{-6}, the mean of La is 7.297×10^{-6}; the value of Ce is 0.853×10^{-6}~1965.619×10^{-6}, the mean of Ce is 68.276×10^{-6}; the value of Nd is 0.118×10^{-6}~ 1035.604×10^{-6}, mean of Nd is 13.245×10^{-6}; the value of Sm is 0.210×10^{-6}~ 241.765×10^{-6}, mean of Sm is 10.921×10^{-6}; the value of Eu is 0.005×10^{-6} ~ 45.821×10 – 6, mean is of Eu 1.104×10^{-6}; the value of Gd is 0.857×10^{-6}~ 575.783×10^{-6}, mean of Gd is 60.978×10^{-6}; the value of Tb is 0.270×10^{-6}~ 470.323×10^{-6}, mean of Tb is 255.290×10^{-6}; the value of Dy is 2.338×10^{-6}~ 2206.460×10^{-6}, mean of Dy is 168.997×10^{-6}; the value of Ho is 0.779×10^{-6}~1355.710×10^{-6}, mean of Ho is 158.875×10^{-6}; the value of Er is 2.670×10^{-6}~ 2026.280×10^{-6}, mean of Er is 4387.402×10^{-6}; the value of Tm is 0.500×10^{-6}~ 457.897×10^{-6}, mean of Tm is 95.337×10^{-6}; the value of Yb is 4.064×10^{-6}~ 14867.000×10^{-6}, mean of Yb is 839.110×10^{-6}; the value of Lu is 0.630×10^{-6}~501.149×10^{-6}, mean of Lu is 139.169×10^{-6}.

The REEs contents of zircons from oceanic island setting are as follows: the value of La is 0.004×10^{-6} ~ 1.408×10^{-6}, the mean of La is 0.123×10^{-6}; the value of Ce is 3.298×10^{-6}~ 886.120×10^{-6}, mean of Ce is 86.671×10^{-6}; the value of Nd is 0.313×10^{-6}~ 49.948×10^{-6}, mean of Nd is 6.755×10^{-6}; the value of Sm is 0.941×10^{-6}~ 93.725×10^{-6}, mean of Sm is 15.909×10^{-6}; the value of Eu is 0.309×10^{-6} ~ 20.647×10^{-6}, mean of Eu is 3.354×10^{-6}; the value of Gd is 9.434×10^{-6}~732.299×10^{-6}, mean of Gd is 137.438×10^{-6}; the value of Tb is 3.790×10^{-6}~244.006×10^{-6}, mean of Tb is 46.075×10^{-6}; the value of Dy is 45.601×10^{-6}~ 2518.060×10^{-6}, mean of Dy is 483.443×10^{-6}; the value of Ho is 18.349×10^{-6}~ 875.210×10^{-6}, mean of Ho is 173.005×10^{-6}; the value of Er is 81.122×10^{-6}~ 3299.880×10^{-6}, mean of Er is 688.004×10^{-6}; the value of Tm is 16.796×10^{-6}~589.976×10^{-6}, mean of Tm is 129.433×10^{-6}; the value of Yb is 137.342×10^{-6}~ 4182.950×10^{-6}, mean of Yb is 954.788×10^{-6}; the value of Lu is 24.602×10^{-6}~610.280×10^{-6}, mean of Lu is 149.543×10^{-6}.

The REEs contents of zircons from convergent margin setting are as follows: the value of La is 0.00028×10^{-6}~20500.000×10^{-6}, the mean of La is 122.222×10^{-6}; the value of Ce is 0.003×10^{-6}~ 20300.000×10^{-6}, mean of Ce is 35.776×10^{-6}; the value of Nd is 0.010×10^{-6}~ 13200.000×10^{-6}, mean of Nd is 11.953×10^{-6}; the value of Sm is 0.172×10^{-6}~ 2810.000×10^{-6}, mean of Sm is 8.868×10^{-6}; the value of Eu is 0.021×10^{-6}~ 105.000×10^{-6}, mean of Eu is 1.832×10^{-6}; the value of Gd is 0.090×10^{-6}~ 2670×10^{-6}, mean of Gd is 40.157×10^{-6}; the value of Tb is 0.530×10^{-6}~ 368.000×10^{-6}, mean of Tb is 14.887×10^{-6}; the value of Dy is 2.320×10^{-6}~ 2070.000×10^{-6}, mean of Dy is 159.761×10^{-6}; the value of Ho is 0.750×10^{-6}~ 1770.000×10^{-6}, mean of Ho is 70.644×10^{-6}; the value of Er is 2.090×10^{-6}~ 2507.860×10^{-6}, mean of Er is 283.136×10^{-6}; the value of Tm is 0.300×10^{-6}~ 2900.000×10^{-6}, mean of Tm is 76.728×10^{-6}; the value of Yb is 1.970×10^{-6}~5081.000×10^{-6}, mean of Yb is 570.656×10^{-6}; the value of Lu is 0.260×10^{-6}~5050.000×10^{-6}, mean of Lu is 140.458×10^{-6}.

‘Spider boxplot’ diagrams for REEs of the zircons in our reference datasets are displayed in Fig. 2. The diagrams represent the statistical distribution of the REEs compositions. It is essential to note that REE

patterns are not entirely discrete or unique, commonly occupying the same model space (Fig. 2d). ‘Spider boxplot’ diagrams can help define the trace element patterns of the zircons and understand the magmatic source and petrogenic processes (Doucet *et al.* 2022). It is also apparent from the Fig. 2 that due to the non-unique distributions of the REEs from each group, distributions of REEs analysis alone cannot definitively classify the tectonic settings of zircons. So, it is necessary to set a flexible approach to solving this limitation.

Table 1
The Extremum, average, median, lower and upper quartile of the training data set

	minimum	lower of quartile	median	upper of quartile	maximum	average
rift volcanic zircons						
La	0.001158	0.017553	0.044496	0.230205	1023.75	7.297253
Ce	0.853	27.45371	46.0755	82	1965.619	68.27596
Nd	0.118125	1.070127	1.770866	6.539723	1035.604	13.2446
Sm	0.2095	3.522988	5.844828	12.27144	241.7648	10.92116
Eu	0.005171	0.111182	0.198122	0.919696	45.82159	1.103603
Gd	0.85675	30.49054	54.70371	80.06089	575.7831	60.97843
Tb	0.269738	17.92728	30.2	97.64207	470.3228	55.29041
Dy	2.338216	23.03766	97	274.8118	2206.461	168.997
Ho	0.779125	65.23952	118.9524	244.172	1355.713	158.875
Er	2.670375	225.8575	441.8409	593.8212	2069.284	438.4024
Tm	0.49975	51	97.68733	127.9862	457.8972	95.33649
Yb	4.06375	460.5995	852.7231	1093.292	14867	839.1095
Lu	0.63	72.92607	137.5466	192.6749	501.1486	139.1694
ocean island zircons						
La	0.004242	0.025872	0.049832	0.129865	1.408443	0.123115
Ce	3.297587	16.07236	32.55352	85.1212	886.12	86.67124
Nd	0.313343	2.102593	3.685601	6.66614	49.94807	6.755016
Sm	0.941297	5.14742	8.656158	17.75065	93.72489	15.90914
Eu	0.309276	1.24386	2.366299	3.965356	20.64747	3.354946
Gd	9.434231	49.56764	81.23011	160.9371	732.2992	137.4384
Tb	3.7903	16.81107	27.96943	54.98023	244.0065	46.07475
Dy	45.60136	195.4219	317.767	573.9016	2518.06	483.4431
Ho	18.34848	72.14706	115.25	209.5005	875.2101	173.0054
Er	81.12233	308.7028	471.8786	845.7849	3299.879	688.0041

	minimum	lower of quartile	median	upper of quartile	maximum	average
Tm	16.796	60.12776	89.59938	159.5752	589.976	129.4326
Yb	137.3415	460.574	671.4237	1174.403	4182.954	954.7877
Lu	24.60187	76.07618	111.3371	183.3614	610.28	149.5427
convergent margin zircons						
La	0.000275	-0.1325	0.031	0.11	20500	122.2223
Ce	0.00302	-17.531	14.5	26.58382	20300	35.77596
Nd	0.01	-4.02223	1.667134	4	13200	11.95301
Sm	0.172113	-6.57238	3.67	7.71492	2810	8.867805
Eu	0.021435	-1.075	0.9	1.55	105	1.832488
Gd	0.0898	-32.6	23	44.9	2670	40.15706
Tb	0.53	-10.8262	8.62	16	368	14.88719
Dy	2.32	-115.088	102.2716	187	2070	159.7605
Ho	0.75	-44.4356	42.06825	75.84042	1770	70.64399
Er	2.09	-179.75	197	341	2507.859	283.1357
Tm	0.3	-36.5	45.26746	76	2900	76.72822
Yb	1.97	-291.485	430	683.6087	5081	570.6558
Lu	0.26	-56.7354	91.87293	144.157	5050	140.4575

3 A Machine Learning Approach

Machine learning presents an ideal framework to perform multivariate analysis, as it is particularly suited to handle and evaluate large volumes of high-dimensional data (Doucet *et al.* 2022). Over the past decade, a number of studies tested the use of a machine learning approach (Ueki et al. 2018; Guo et al. 2021). In this paper, we use machine learning algorithms to set up a classification for distinguishing the tectonic setting based on the REEs of the zircons. The workflow is followed as Fig. 3.

3.1 Data preprocessing

REEs are featured by similar ionic radii and stable + 3-valent ions in nature (Yang *et al.* 2000) and have similar physical and chemical properties. Nevertheless, the lanthanide contraction phenomenon in REEs indicates there is a negative linear relationship between atomic number and the ionic radius. This

phenomenon makes different REEs show different geochemical behaviors, so different rocks or minerals show different distribution characteristics of the REEs. Especially, accessory minerals do cause a great influence on the distribution patterns of the REEs. For example, the zircons and garnets cause the depletion of HREEs; the titanites and apatites cause the depletion of MREES, and the monazites and allanites cause the depletion of LREEs. Besides, the amphibole is compatible with REEs and shows the highest partition coefficient value between Dy and Er (Yang 2000). Because Pm elements in REE do not exist in nature and Pr elements are missing in oceanic island data which will result in model overfitting, in this paper, the REEs without Pm and Pr in the zircons are selected as the eigenvectors of the training data set, and the tectonic setting is used as the judgment label. All the REEs are labeled as the first control group (CG-1:$\sum$REEs), and two control groups are labeled according to the atomic number as LREEs (CG-2: La-Eu without Pr) and HREEs (CG-3: Gd-Lu). In addition, Zhong et al. (2019) and Loader et al. (2017) infer that the contents of La to Pr in the zircons are commonly below the limit of detection and susceptible to the contamination of mineral inclusions, so we define REEs without La, Ce, Pr as the fourth control group (CG-4).

Due to the oddo-Harskin effect, the abundance of REEs with odd atomic numbers is less than that of REEs with even atomic numbers. To eliminate the oddo-Harskin effect, the data of standard rare earth elements are usually used to normalize the REEs data of rocks or minerals (Haskin et al. 1968; Wakita *et al.* 1971; Masua *et al.* 1973; Nakatuura 1974; Evensen et al. 1978; Boynton 1984; Taylor 1985; Mcdough and Sun 1989). Because of the different normalized data, there are differences in the distribution patterns of rare earth elements (Yang 2000). In this paper, we use non-normalized REEs of the zircons from four control groups as the feature vector of the training data set and set up sixteen classifier models with the approaches of four machine learning algorithms. According to the results, an optimal classifier model is finally obtained. Considering the REEs distribution pattern diagrams are usually normalized based on the data of chondrite, primitive mantle, enriched mid-ocean ridge basalt, mid-ocean ridge basalt, and oceanic island basalt, the training data set are normalized by the REEs of chondrites, primitive mantle, enriched mid-ocean ridge basalt, mid-ocean ridge basalt, and oceanic island basalt. Then the normalized data set are used as feature vector to explore the effects of different normalized or non-normalized data in the optimal classification model.

3.2 Classifier models set up

Four machine learning approaches are composed of four supervised learning algorithms in this paper including Random Forest (RF), Support Vector Machine (SVM), Decision Tree (Tree), and eXtreme Gradient Boosting (XGBoost). The reference datasets are divided into training data and test data according to the ratio of 9:1. RF is an ensemble learning method based on decision tree classifiers (Breiman 2001). Traditional decision trees are focused on an optimal attribute among the n attributes (Zhou 2016). Based on each decision tree node, an optimal parameter is selected from the k subsets after a subset including k attributes is set up. Meanwhile, RF is characterized by easy implementation, low computational overhead, good classification effect, high stability, and fast operation speed, but it is easy to overfit when it solves problems with large noise (Zhu et al. 2022). The number of trees constructed in

the RF classifier model in this paper is 25 (25 weak classifiers), and the randomness parameter value (random_state) is 42. The SVM makes the linearly inseparable training samples in the low-dimensional space be mapped into the high-dimensional feature space, and the optimal classification hyperplane in the high-dimensional feature space is determined. The principle of SVM is to convert the multi-classification problem into a binary classification problem. Unfortunately, the dimension of the feature space may be very high, even infinitely multidimensional, it could be impossible to calculate the inner product of the function in the SVM. But the kernel function makes it possible to calculate the inner product of the function in the SVM. The kernel function of the SVM classifier model in this paper is a Gaussian function. The Tree contains a root node, several internal nodes, and several leaf nodes. The root node containing all sample collections is aimed to classify with the process from the internal nodes of the attribute test to the leaf nodes. Tree has the advantage of being easy to understand, interpret and visualize. The maximum depth (max_depth) of the Tree classifier model in this paper is 3, the random parameter (random_state) is 42, and the numbers of leaf nodes including min_samples_leaf and min_samples_split are 10 respectively. The XGBoost (Formula 1) is an enhanced algorithm of the Gradient Boosting Decision Tree. Its core principle is to generate a new weak classifier by generating a new tree to fit the residuals from the first *n* weak classifiers, and all weak classifiers are combined and become part of the final strong classifier. The advantage of XGBoost is that when predicting the *t*-th weak classifier value, Taylor expansion is performed on the predicting value f_t(xi) (Formula 2) and the second-order expansion item is retained, which makes the prediction accuracy of each layer higher. So it makes the overall convergence of the model faster, and it is more advantageous to obtain the dependencies between complex data. The maximum depth of the XGBoost classifier model in this paper is 26, the learning rate (eta) is 0.1, and the minimum value of loss reduction (gamma) required for leaf nodes to branch is 0.

$$L\left(f_t\right)=\sum_{i=1}^{n} L(y_i, \widehat{y}_i^{t-1}+ft\left(x_i\right))+\Omega(f_t)+C \text{ (Formula 1)}$$

$$f\left(x+\Delta x\right) \approx f\left(x\right)+f^{\prime}\left(x\right) \Delta x+\frac{1}{2} f^{\prime\prime}\left(x\right) \Delta x \text{ (Formula 2)}$$

(1)$L\left(f_t\right)$ is the loss function of each weak classifier

(2)y_i is the true value of layer *t*

(3)$\widehat{y}_i^{t-1}$ is the predicted value of the *t-1*th layer

(4)$ft\left(x_i\right)$ is the predicted value of the *t*-th layer

(5)$\Omega\left(f_t\right)$ is a regular term

3.3 Model evaluation

Model evaluation can not only evaluate the effectiveness of classification results but also determine the most effective classifier algorithm. As evaluation indicators for the classifier models, the accuracy rate (Accuracy), precision rate (Precision), recall rate (Recall), and F1 score are obtained on the test data.

Accuracy is the ratio of the numbers of correctly predicted samples to the total numbers of samples. Precision is the ratio of the numbers of actual positive samples predicted as positive samples to the numbers of all predicted positive samples in the result (Formula 3). Recall is the ratio of the number of actual positive samples predicted as positive samples to all the actual positive samples (Formula 4). F1 score with the range of 0–1 is the average value of Precision and Recall (Formula 5) and is an efficient way to evaluate the quality of the model in an imbalanced dataset.

Precision = TP/(TP + FP) (Formula 3)

Recall = TP/(TP + TN) (Formula 4)

F1 score = 2*1/(1/Precison + 1/Recall) (Formula 5)

Where TP = number of true positives, FP = number of false positives, TN = number of true negative

4 Results

4.1 Results of sixteen classifier models

In this paper, training data has been trained independently for 50 times. The statistical results including the average and variances of different learning models have been shown in Table 2 and Fig. 4.

In the RF classifier model, the averages of Accuracy in control groups are 0.857, 0.890, 0.941, and 0.917, respectively. The averages of Precision in the control groups are 0.803, 0.878, 0.940, and 0.916, respectively. The averages of Recall in the control groups are 0.702, 0.844, 0.907, and 0.899, respectively. The averages of F1 score in the control groups are 0.733, 0.854, 0.922, and 0.900, respectively.

In the SVM classifier model, the averages of Accuracy in control groups are 0.806, 0.623, 0.619, and 0.623, respectively. The averages of Precision in the control groups are 0.816, 0.454, 0.466, and 0.505, respectively. The averages of Recall in the control groups are 0.573, 0.336, 0.337, and 0.339, respectively. The averages of F1 score in the control groups are 0.596, 0.261, 0.262, and 0.266, respectively.

In the Tree classifier model, the averages of Accuracy in control groups are 0.666, 0.729, 0.708, and 0.723, respectively. The averages of Precision in the control groups are 0.429, 0.733, 0.746, and 0.727, respectively. The averages of Recall in the control groups are 0.412, 0.498, 0.459, and 0.490, respectively. The averages of F1 score in the control groups are 0.394, 0.522, 0.463, and 0.511, respectively.

In the XGBoost classifier model, the averages of Accuracy in control groups are 0.860, 0.897, 0.948, and 0.92, respectively. The averages of Precision in the control groups are 0.801, 0.878, 0.941, and 0.913, respectively. The averages of Recall in the control groups are 0.730, 0.858, 0.922, and 0.892, respectively. The averages of F1 score in the control groups are 0.755, 0.866, 0.930, and 0.901, respectively.

Table 2 Average and variance of Accuracy, Precision, Recall, and F1 scores of RF, SVM, Tree and XGBoost

RF	Accuracy	σ	Precision	σ	Recall	σ	F1 score	σ
CG-1	0.857	0.00027	0.803	0.01654	0.702	0.06753	0.733	0.04278
CG-2	0.89	0.00031	0.878	0.00345	0.844	0.00907	0.854	0.0039
CG-3	0.941	0.00017	0.94	0.00099	0.907	0.00432	0.922	0.00158
CG-4	0.917	0.00019	0.916	0.00153	0.889	0.00513	0.9	0.00164
SVM								
CG-1	0.806	0.00042	0.816	0.02351	0.573	0.12824	0.596	0.09393
CG-2	0.623	0.00068	0.454	0.1697	0.336	0.22035	0.261	0.12789
CG-3	0.619	0.00055	0.466	0.17027	0.337	0.21961	0.262	0.12577
CG-4	0.623	0.00033	0.505	0.00033	0.339	0.21841	0.266	0.12544
Tree								
CG-1	0.666	0.00039	0.429	0.09441	0.412	0.15463	0.394	0.10387
CG-2	0.729	0.00035	0.733	0.05095	0.498	0.12823	0.522	0.06297
CG-3	0.708	0.00046	0.746	0.08215	0.459	0.15494	0.463	0.08638
CG-4	0.723	0.00034	0.727	0.03328	0.49	0.12986	0.511	0.06775
XGBoost								
CG-1	0.86	0.00032	0.801	0.01271	0.73	0.04999	0.755	0.03176
CG-2	0.897	0.00018	0.878	0.00345	0.858	0.00757	0.866	0.00757
CG-3	0.948	0.00007	0.941	0.00107	0.922	0.00287	0.93	0.00132
CG-4	0.92	0.00015	0.913	0.00116	0.892	0.00417	0.901	0.00149

4.2 the results of the normalized $\sum$REE data with different methods

Based on the results in this paper, the optimal XGBoost classifier model (discussion in sections 5.1 and 5.2) is used to discuss the differences of the normalized $\sum$REE data with different methods, the $\sum$REE data as feature vectors are normalized by the data of chondrites, primitive mantle, enriched mid-ocean ridge basalts, mid-ocean ridge basalts, and oceanic island basalts, respectively. The experimental results have been displayed in Fig. 5 and Table 3.

Table 3
Average and variance of accuracy, precision, recall, and F1 scores of normalization data and raw data. CL- the data normalized after chondrites; PM- the data normalized after primitive mantle; MORB- the data normalized after mid-ocean ridge basalt; EMORB-the data normalized after enriched mid-ocean ridge basalt; OIB- the data normalized after oceanic island basalt

	Accuracy	Precision	Recall	F1 score
CL	0.949	0.943	0.923	0.932
PM	0.950	0.938	0.925	0.930
MORB	0.950	0.940	0.923	0.931
EMORB	0.949	0.940	0.917	0.927
OIB	0.946	0.931	0.918	0.923
Raw Data	0.948	0.941	0.922	0.930

5 Discussion

5.1 Comparison of classifier models

The optimal classifier model is proposed based on the Accuracy, Precision, Recall, and F1 score. The different CGs have different results in different classifier models of RF, SVM, Tree, and XGBoost. The average accuracy rates of 0.941 and 0.948 for $\sum$REE (CG-3) are higher than the ones of other CGs in RF and XGBoost, indicating the classifier models using $\sum$REE (CG-3) as the feature vector are the optimal classifier models in the RF and XGBoost, while their variances are 0.00017 and 0.00007, respectively. In the SVM, the classifier model using LREE (CG-1) as the feature vector has the highest quality with the highest average accuracy rate of 0.806 and lowest variance of 0.00042, while the optimal classifier model in the RF is the one using HREE (CG-2) as the feature vector based on the average accuracy rate of 0.929 and the variance of 0.0003 in Tree.

The average accuracy rate and the variance indicate the models in RF and XGBoost are significantly more effective than the ones in SVM and Tree (Fig. 7). The reasons may be that the final prediction results of RF and XGBoost are achieved through decision integration which has three advantages as follows (Dietterrich, 2000): (1) Due to the large hypothesis space of the classification tasks, the classification results of multiple hypotheses in the training data set are the same. Although a single hypothesis could cause the results of poor generalization performance, combining multiple learners could reduce the risk of poor generalization performance; (2) A single learner is easy to fall into the local minimum value problem in the calculation process, resulting in poor generalization performance. By running multiple decisions at the same time, the possibility of falling into the local minimum value may be reduced; (3) The hypothesis adopted by a single strategy may not be in the hypothesis space of the current learning

algorithm, which makes a single learner invalid. By combining multiple learners, the hypothesis space becomes bigger, making the prediction results closer to the actual value.

Although the average accuracy in RF and XGBoost classifier model is similar, the average accuracy of RF is lower than that of the XGBoost, while the variance of RF is larger than that of the XGBoost. In addition, the precision and recall of $\sum$REE classifier method in RF are 0.940 and 0.907, they are also lower than the precision and recall of $\sum$REE classifier method in XGBoost, while the variance of $\sum$REE classifier method in RF is greater than that of $\sum$REE classifier method in XGBoost. In summary, the XGBoost machine learning algorithm is the best algorithm of the four machine learning algorithms, and the classifier model in XGBoost is proposed for distinguishing the tectonic setting based on the $\sum$REE.

5.2 Comparison of the classifier models

The results clearly show that XGBoost is the most effective of the four machine learning algorithm (Fig. 6 and Fig. 7). Furthermore, the $\sum$REE classification model is the optimal classifier model in all classifier models (Fig. 6 and Fig. 7) and has more elements as feature vector. The reason could be that REEs usually appear in groups in any geological process because of similar ionic radii and similar physical and chemical properties (Zhang *et al.* 2012), in addition, REEs with stable geochemical properties remain relatively inactive during the processes of low-grade metamorphism, weathering, and hydrothermal alteration (Michard 1989; Yang 2000). Thereby, the $\sum$REEs can be used as an eigenvector. Based on the discussion mentioned above, there is a positive correlation between the quality of the model prediction results and the numbers of feature vectors. In summary, the $\sum$REE classifier model is the optimal classifier model for discriminating the tectonic setting.

5.3 Comparison of five kinds of normalization

To confirm the effects on the normalized REEs, five groups of factor vectors are obtained based on the different normalization data including CL, PM, MORB, OIB, and EMORB and are used for calculating the Accuracy, Precision, Recall and F1 score of different classifier models (Fig. 8a). The results have been furtherly compared to the ones of raw data (RD). The Accuracy of all data is about 0.95, and all the data except the ones of normalization data after OIB have similar Precision and F1 scores. In addition, the Recalls of normalization data after OIB and EMORB are slightly smaller than the ones of the RD and normalization data after CL, PM, and MORB. Based on the results, the Accuracy, Precision, Recall, and F1 score of all the data are 0.9 and have a tiny difference between the normalization data and the RD. In summary, we believe that whether the REEs in zircons are normalized or not has little effect on the classifier model to distinguish tectonic settings.

6 Conclusion

(1) We select four Machine Learning methods and four control groups, which are used to handle and evaluate large volumes of REEs in zircons. The Accuracy, Precision, Recall, and F1 score of the sixteen classifier models are compared to determine the suitable Machine Learning method and feature vector. The XGBoost classifier using $\sum$REEs as feature vector has the highest average expectation and the

smallest variance in terms of accuracy, precision, recall, and F1 score, which is the most suitable for distinguishing the tectonic setting among the four machine learning approaches in this paper.

(2) Based on the XGBoost classifier model using ∑REEs as feature vector, the Accuracy, Precision, Recall, and F1 score of the raw data with normalization data is compared to determine the influence of normalization data. The result shows that whether the REEs in zircons are normalized or not has little effect on the classifier model to distinguish tectonic settings.

This work demonstrates the feasibility of using big data to predict the tectonic settings of zircons.

Declarations

authorship contribution statement: Luyuan Wang: Conceptualization, Investigation, Methodology, Writing - original draft. **Chao Zhang**: Project administration, Resources, Writing - review & editing. **Rui Geng**: Writing - review & editing. **Yuqi Li**: Data curation. **Jijie Song**: Data curation. Bin Wang: Writing - review & editing. **Fanghua Cui**: Supervision, Data curation

Declaration of competing interest: We declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper entitled “The discrimination of tectonic setting using trace elements in zircons: A machine learning approach”.

Availability of Data and Material: If someone wants to request the reference data and code, wangluyuan9@outlook.com can provide these data.

Funding：This study was co-supported by the NSFC (Grant No.41802238), and the Science Foundation of Shandong Province (ZR2021MD104, ZR2019PD010).

References

1. Barth AP, Wooden JL, Jacobson CE, Economos RC (2013) Detrital zircon as a proxy for tracking magmatic arc systems: the California arc example. Geology 41:223–226. doi:10.1130/G33619.1
2. Bell EA (2017) Petrology: Ancient magma sources revealed. Nature Geoscience,10(6): 397 - 398
3. Belousova EA, Jiménez JMG, Graham I, Griffin WL, O’Reilly SY, Pearson N, Martin L, Craven S, Talavera C (2015) The enigma of crustal zircons in upper-mantle rocks: clues from the Tumut ophiolite, southeast Australia. Geology 43:119–222. doi:10.1130/G36231.1
4. Belousova EA, Griffin WL, O’Reilly SY, Fisher NJ (2002) Igneous zircon: trace element composition as an indicator of source rock type. Contrib Miner Petrol 143:602–622. doi:10.1007/s00410-002-0364-7
5. Boynton WV (1984) Geochemistry of the rare earth elements: meteorite studies, In: Henderson P. (ed.), Rare earth element geochemistry. Elservier, pp.63-114.
6. Breiman L (2001) Random forests. Machine learning. 45(1):5-32

7. Buret Y, von Quadt A, Heinrich C, Selby D, Wälle M, Peytcheva I (2016) From a longlived upper-crustal magma chamber to rapid porphyry copper emplacement: Reading the geochemistry of zircon crystals at Bajo de la Alumbrera (NW Argentina). Earth Planet. Sci. Lett. 450, 120–131.
8. Carley TL, Miller CF, Wooden JL, Padilla AJ, Schmitt AK, Economos RC, Bindeman IN, Jordan BT (2014) Iceland is not a magmatic analog for the Hadean: evidence from the zircon record. Earth Planet Sci Lett 405:85–97. doi:10.1016/j.epsl.2014.08.015
9. Dietterich TG (2000) “Ensemble methods in machine learning.”In Proceedings of the 1st International Workshop on Multiple Classifier Systems(MCS),1-15,Cagliari, Italy.
10. Doucet LS, Gamaleldien H, Li ZX (2022a) Pitfalls in using the geochronological information from the EarthChem Portal for Precambrian time-series analysis. Precambrian Res 369, 106514.
11. Doucet LS, Tetley MG, Li ZX, Liu YB, Gamaleldien H (2022b) Geochemical fingerprinting of continental and oceanic basalts: A machine learning approach. Earth-Science Reviews.104192, 0012-8252. doi:10.1016/j.earscirev.2022.104192.
12. Drabon, Byerly BL, Byerly GR, Wooden JL, Keller CB, Lowa DR (2021) Heterogeneous Hadean crust with ambient mantle affinity recorded in detrital zircons of the Green Sandstone Bed, South Africa. Proceedings of the National Academy of Sciences,118 (8) : e2004370118
13. Evensen NM, Hamilton PJ, O'Nions RK (1978) Rare earth abundances in chondritic meterorites. Gcochsn. Cosmochim. Acta. 42, 1199-1212
14. Gao P, Santosh M (2020) Mesoarchean accretionarymélange and tectonic erosion in the Archean Dharwar Craton, southern India: Plate tectonics in the early Earth. Gondwana Res. 85, 291–305.
15. Guo L, Zhang HF, Harris N, Xu WC, Pan FB (2017) Detrital zircon U-Pb geochronology trace-element and Hf isotope geochemistry of the metasedimentary rocks in the Eastern Himalayan syntaxis: Tectonic and paleogeographic implications. Gondwana Research, 41: 207 - 221
16. Guo, P., Yang, T., Xu, W.L., Chen, B., 2021. Machine learning reveals source compositions of intraplate basaltic rocks. Geochem. Geophys. Geosyst. 22 (9) e2021GC009946
17. Grimes CB, John BE, Kelemen PB, Mazdab F, Wooden JL, Cheadle MJ, Hanghøj K, Schwartz JJ (2007) The trace element chemistry of zircons from oceanic crust: a method for distinguishing detrital zircon provenance. Geology 35:643–646. doi:10.1130/G23603A.1
18. Grimes CB, Wooden, JL, Cheadle MJ, John BE (2015) “Fingerprinting” tectono-magmatic provenance using trace elements in igneous zircon. Contrib Mineral Petrol 170, 46. Doi:10.1007/s00410-015-1199-3
19. Haskin LA, Haskin MA, Frey FA, Wildman TR (1968) Relative and absolute terrestrial abundances of the rare earths, ln: Ahrens LH. (ed), Origin and distribution of the elements, vol.1. Oxford: Pergamon, pp.889 – 911
20. Kemp A, Hawkesworth C, Foster G, Paterson B, Woodhead J, Hergt J, Gray C, Whitehouse M (2007) Magmatic and crustal differentiation history of granitic rocks from Hf-O isotopes in zircon. Science 315, 980–983.

21. Kirkland CL, Smithies RH, Taylor RJM, Evans N, McDonald B (2015) Zircon Th/U ratios in magmatic environs. Lithos 212–215:397–414
22. LA, Kobussen A (2016) Zircon compositions as a pathfinder for porphyry Cu±Mo±Au deposits. Econ. Geol. Spec. Pub. 19, 329–347.
23. Loader MA, Wilkinson JJ, Armstrong RN (2017) The effect of titanite crystallisation on Eu and Ce anomalies in zircon and its implications for the assessment of porphyry cu deposit fertility. Earth Planet Sci Lett 472:107–119
24. Lu YJ, Loucks RR, Fiorentini M, McCuaig TC, Evans NJ, Yang ZM, Hou, ZQ, Kirkland CL, Parra-Avila Roberts NM, Spencer CJ (2015) The zircon archive of continent formation through time. Geological Society. 389. Special Publications, London, pp. 197–225.
25. Nakamura N (1974) Determination of REE, Ba, Fe, Mg, Na, and K in carbonaceous and ordinary chondrites. Geochim. Cosmochimn. Acta, 38, 757-775
26. Masuda A, Nakamura N, Tanaka T (1973) Fine structures of mutually normalised rare-earth patterns of chondrites. Geochim. Cosmochim. Acta. 37, 239-248
27. Michard A (1989) Rare earth element systematics in hydrothermal fluids. Geochim. Cosmochim. Acta, 53, 745 - 750.
28. Palin RM, Santosh M, Cao W, Li SS, Hernández-Uribe D, Parsons A (2020) Secular metamorphic change and the onset of plate tectonics. Earth Sci. Rev. 207, 103172.
29. 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-Science Reviews, 176: 87 - 116
30. Spencer CJ, Cavosie AJ, Raub TD, Rollinson H, Jeon H, Searle MP, Miller JA, McDonald BJ, Evans NJ, 2017. Evidence formelting mud in Earth'smantle fromextreme oxygen isotope signatures in zircon. Geology 45, 975–978.
31. Sun SS, McDonough WF (1989) Chemical and isotopic systematics of oceanic basalts: implications for mantle composition and processes. Geol Soc Lond Spe Publ 42:313–345. doi:10.1144/GSL.SP.1989.042.01.19
32. Szilas K, Hoffmann JE, Scherstén A, Kokfelt TF, Münker C (2013) Archaean andesite petrogenesis: insights from the Grædefjord Supracrustal Belt, southern West Greenland. Precambrian Res. 236, 1–15.
33. Tayloy SR, Mclennan SM (1985) The continental crust: its composition and evolution. Oxford: Blackwell
34. 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.
35. Van Kranendonk MJ, Kirkland CL (2013) Orogenic climax of Earth: the 1.2–1.1 Ga Grenvillian superevent. Geology 41, 735–738.

36. Xing K, Shu Q, Lentz DR, Wang F (2020) Zircon and apatite geochemical constraints on the formation of the Huojihe porphyry Mo deposit in the Lesser Xing'an Range, NE China. Am. Mineral. 105, 382–396.
37. Yang XM (2000) Petrological geochemistry. In: Yang X.Y., Chen S.X., University of Science and Technology of China Press.7-312-01190-X, China
38. Zou XY, Qin KZ, Han XL, Li GM, Evans NJ, Li ZZ, Yang W (2019) Insight into zircon REE oxy-barometers: A lattice strain model perspective. Earth and Planetary Science Letters,506: 87 – 96
39. Zou XY, Jiang JL, Qin KZ, Zhang YF, Yang W, Li XH (2021) Progress in the principle and application of zircon trace element. Acta Petrologica Sinica, 37(4): 985 - 999. doi: 10. 18654 /1000-0569
40. Zhang H.F., 2012. Geochemical. Geological Press.978-7-116-07710-2, China
41. Zhong SH, Feng C, Seltmann R, Li D, Qu H (2018) Canmagmatic zircon be distinguished from hydrothermal zircon by trace element composition? The effect ofmineral inclusions on zircon trace element composition. Lithos 314–315, 646–657.
42. Zhong SH, Seltmann R, Qu HY Song YX (2019) Characterization of the zircon Ce anomaly for estimation of oxidation state of magmas: a revised Ce/Ce* method. Miner Petrol 113, 755–763.doi: 10.1007/s00710-019-00682-y
43. Zhong SH, Li SZ, Seltmann R, Lai ZQ, Zhou J (2022) The influence of fractionation of REE-enriched minerals on the zircon partition coefficients. Geoscience Frontiers, 12(3), 1674-9871. doi: 10.1016/j.gsf.2020.10.002
44. Zhu ZY, Zhou F, Wang Y, Zhou T, Hou ZL, Qiu KF (2022) Machine learning-based approach for zircon classification and genesis determination. Earth Science Frontiers, 29(5): 464-475
45. Zhou ZH (2016) Machine learning. Tsinghua University Press.978-7-302-42328-7, China

Figures

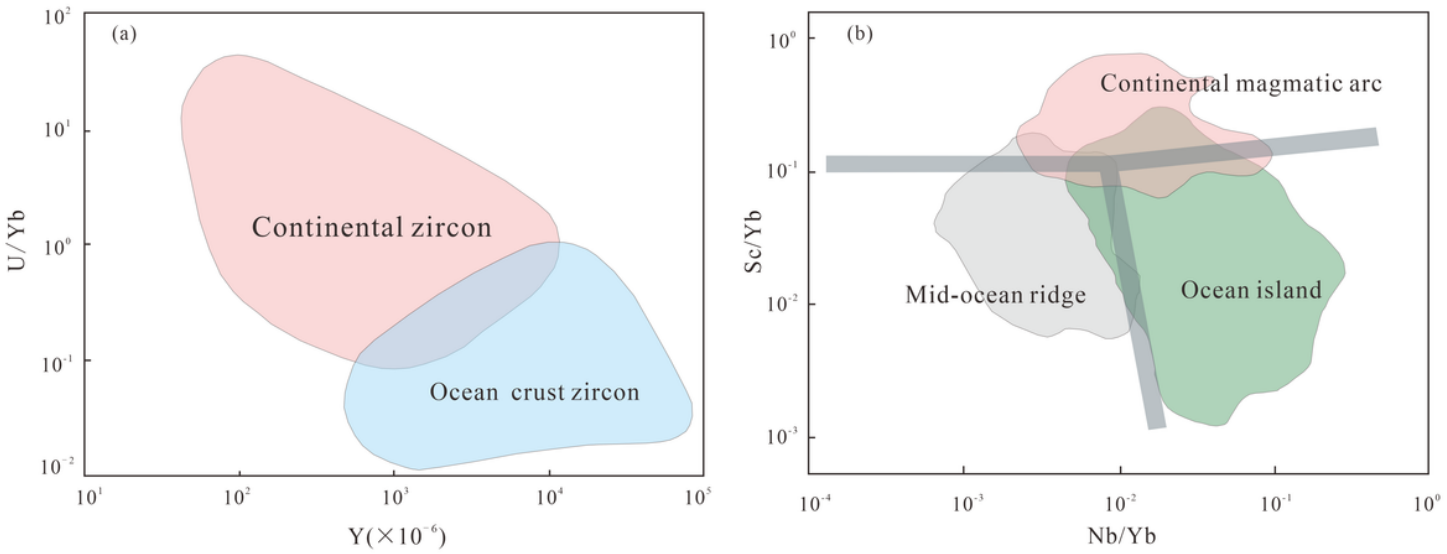

Figure 1

tectonic-magmatic setting discriminant diagram(a) Discriminant diagrams with continental and ocean crust zircon fields;(b) Discriminant diagrams with continental magmatic arc, ocean crust, and mid-oceanic ridge zircon fields (modified after *Grime et al.*2007, 2015)

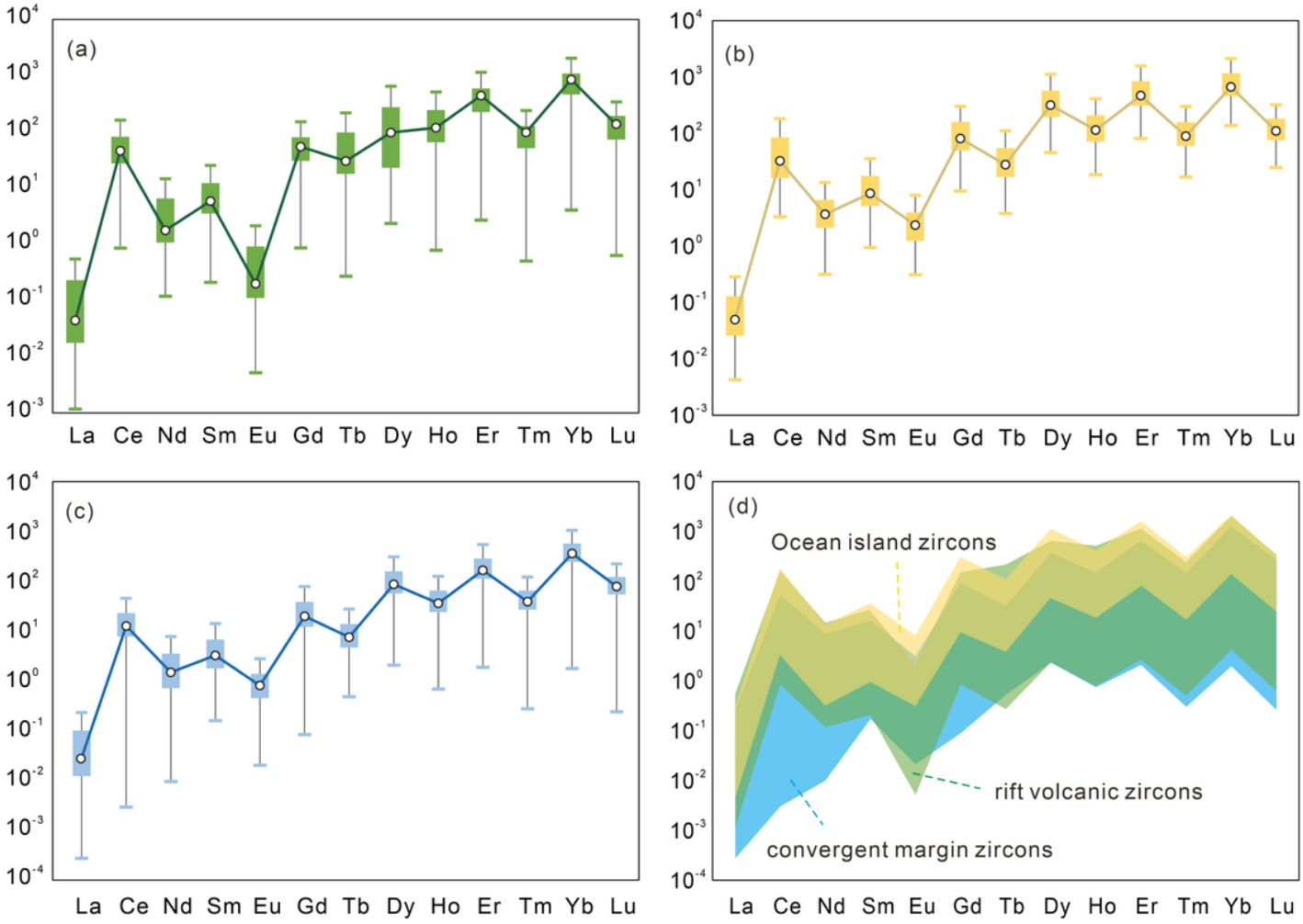

Figure 2

The distribution the REEs patterns of the zircons from the different tectonic settings in our reference datasets: (a) the zircons from the rift volcanic tectonic setting; (b) the zircons from the ocean island tectonic setting;(c) the zircons from the convergence margin tectonic setting; (d) rift volcanic, ocean island, convergence margin overlay.

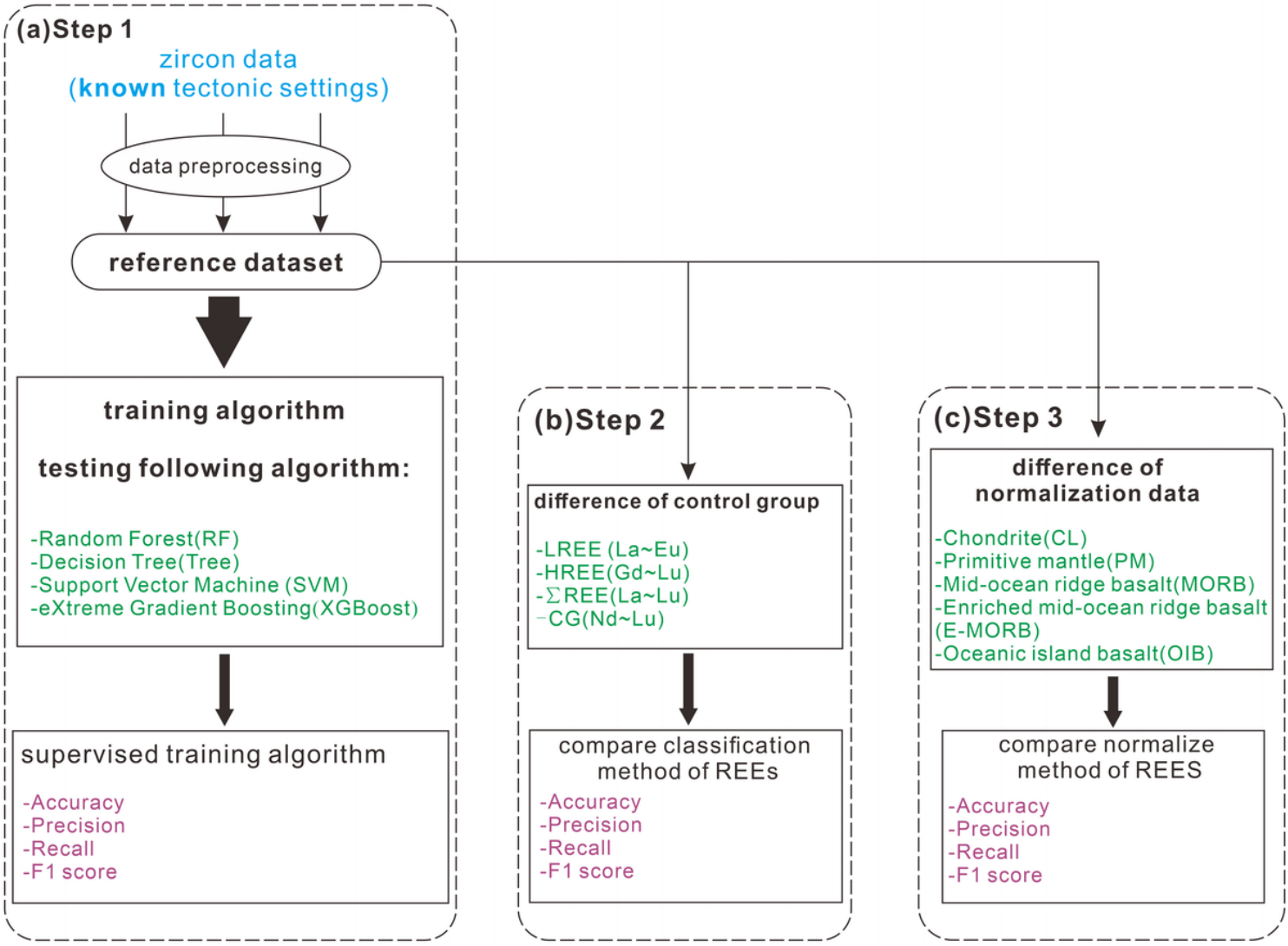

Figure 3

Supervised machine learning workflow block diagram used in this study: (a) Step 1, data preprocessing and determining the suitable algorithm;(b) Step 2, determining the suitable classification methods;(c) Step 3, determining the influence of standardization data

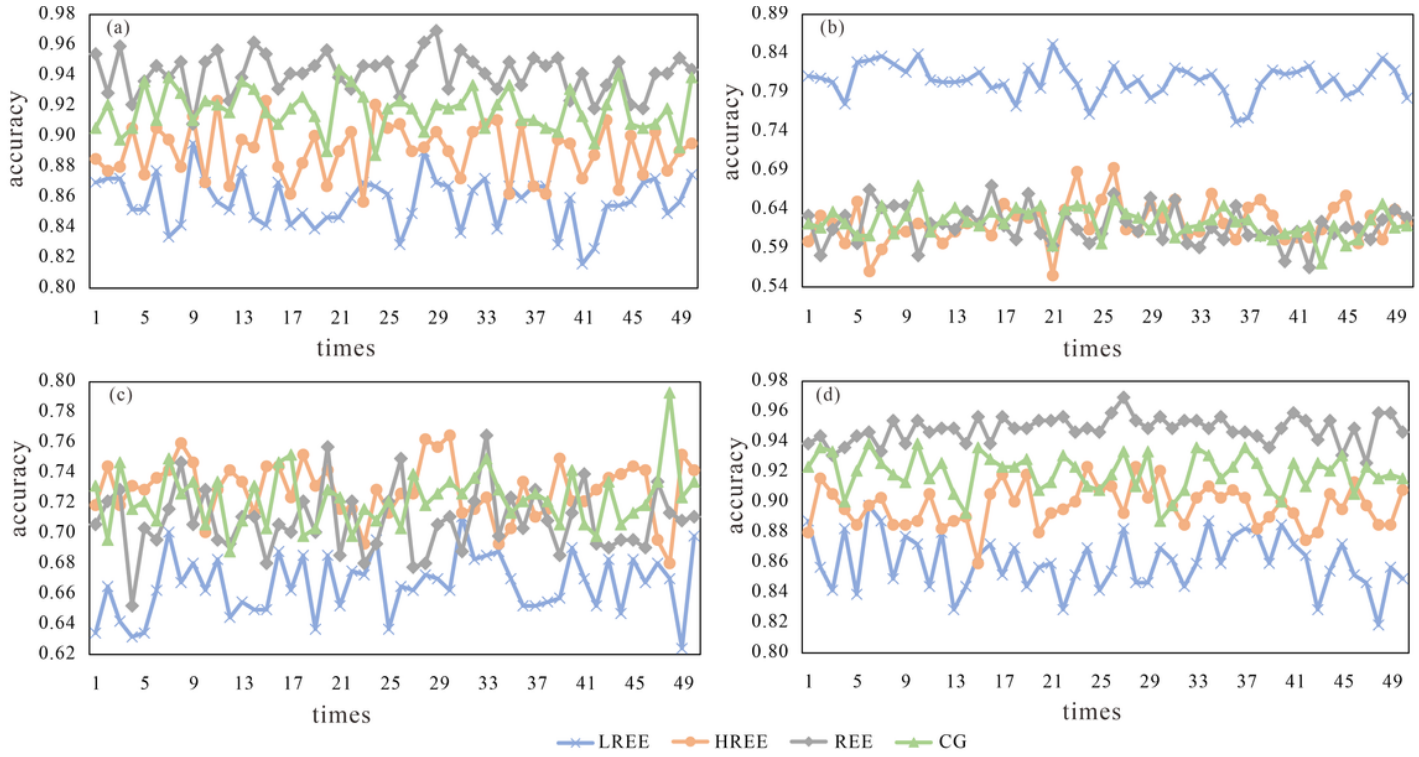

Figure 4

(a) The accuracy score line of the RF classifier model. (b) The accuracy score line of the SVM classifier model. (c) The accuracy score line of the Tree Classifier model. (d) The accuracy score line of the XGBoost classifier

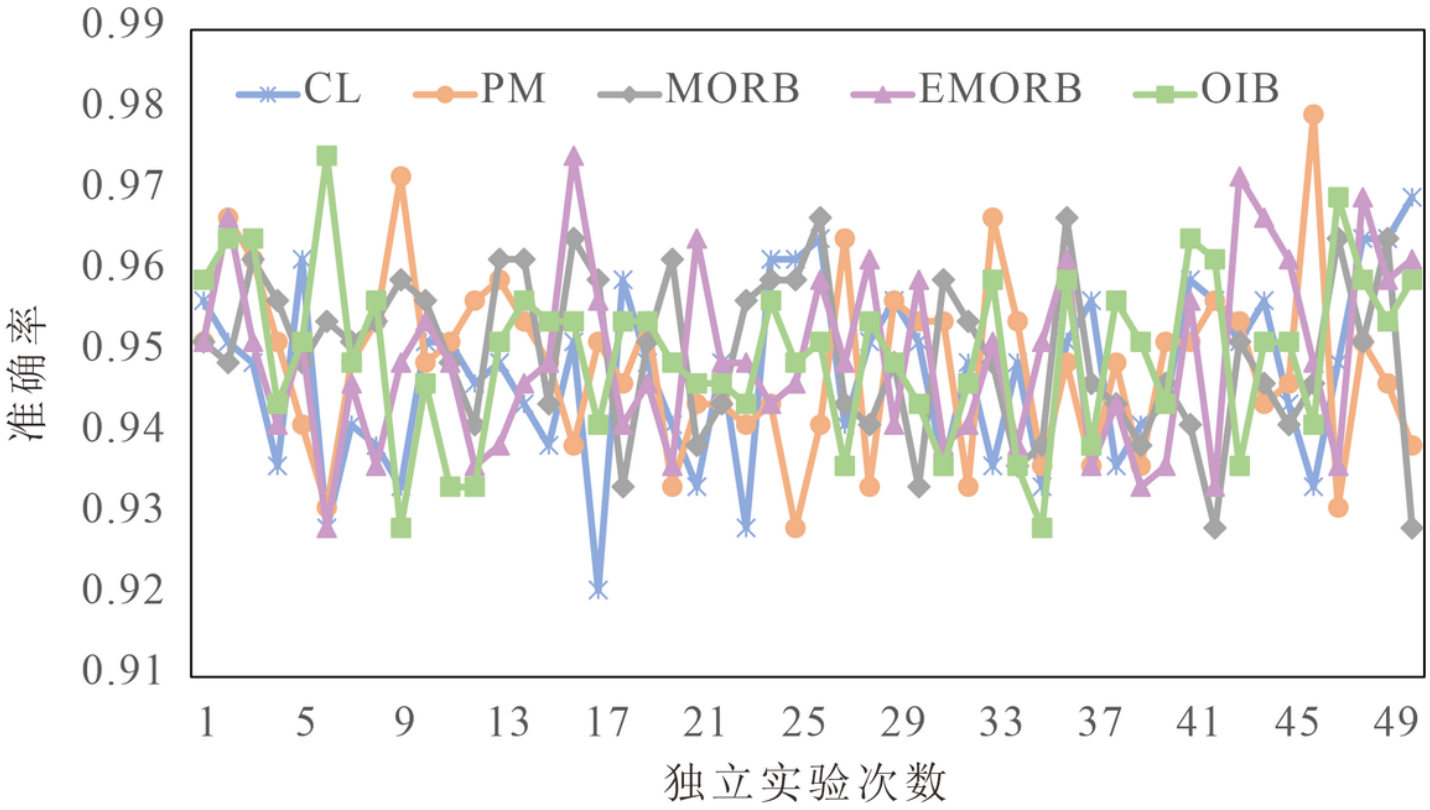

Figure 5

Accuracy scores line of different normalization data

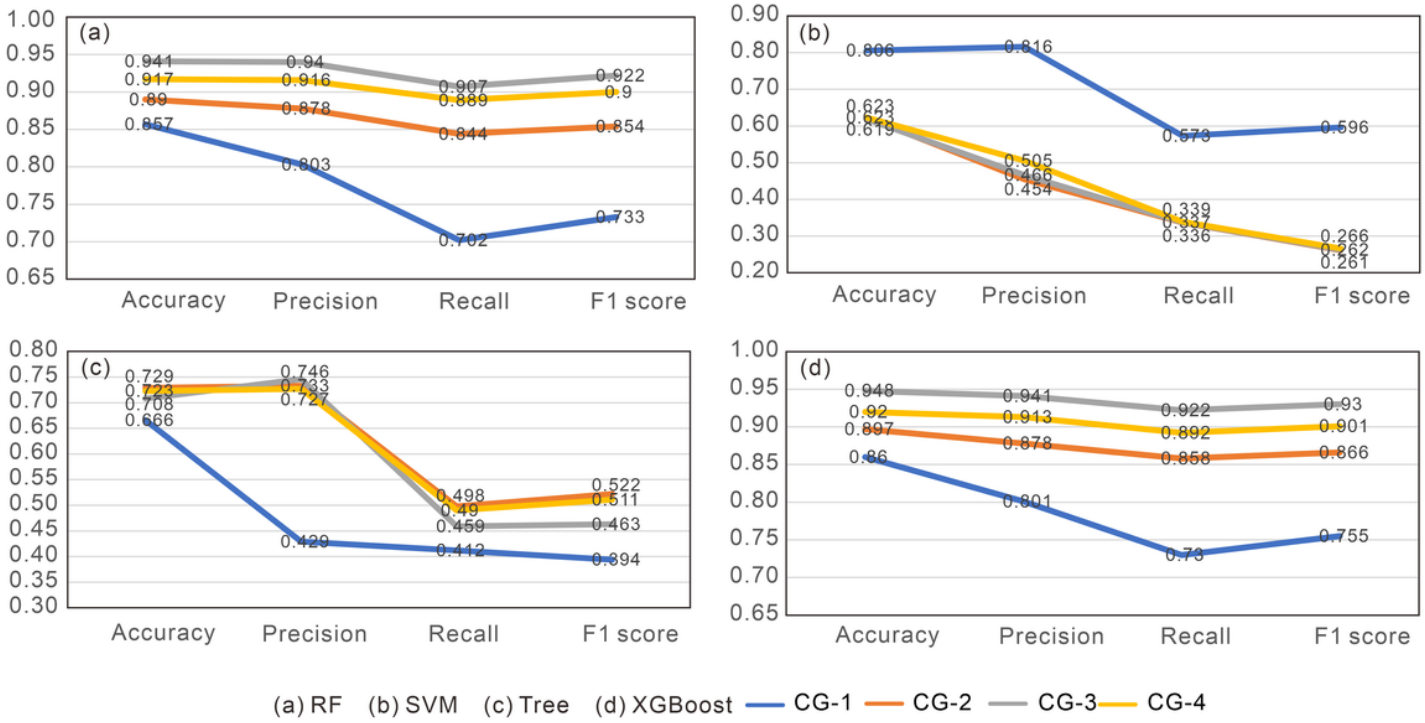

Figure 6

Plot of Accuracy, Precision, Recall, and F1 score for sixteen classifier models.

(a) classifier models with the different control groups in RF. (b) classifier models with the different control groups in SVM. (c) classifier models with the different control groups in Tree. (d) classifier models with the different control groups in XGBoost.

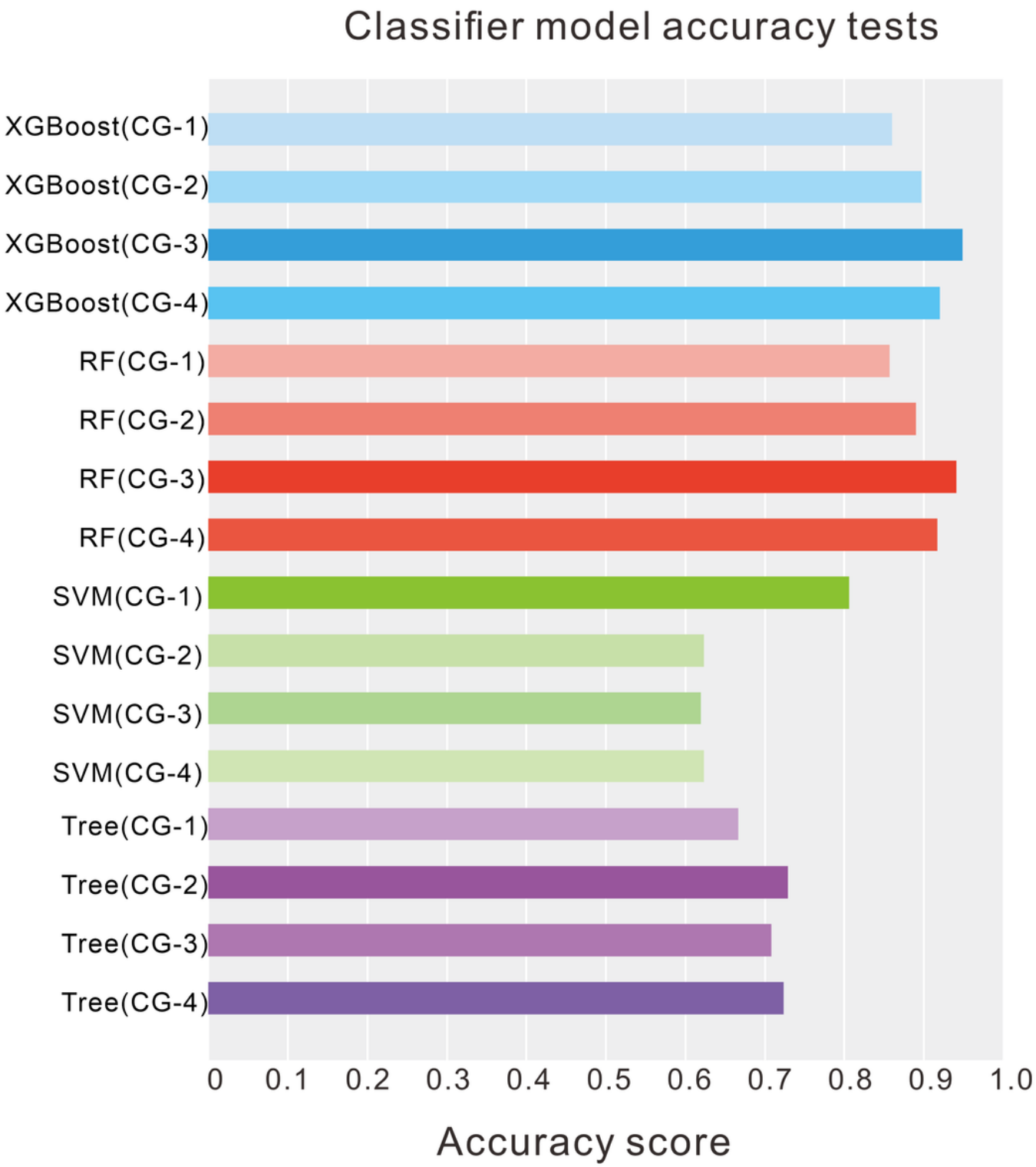

Figure 7

Comparison of different classifier models in this study

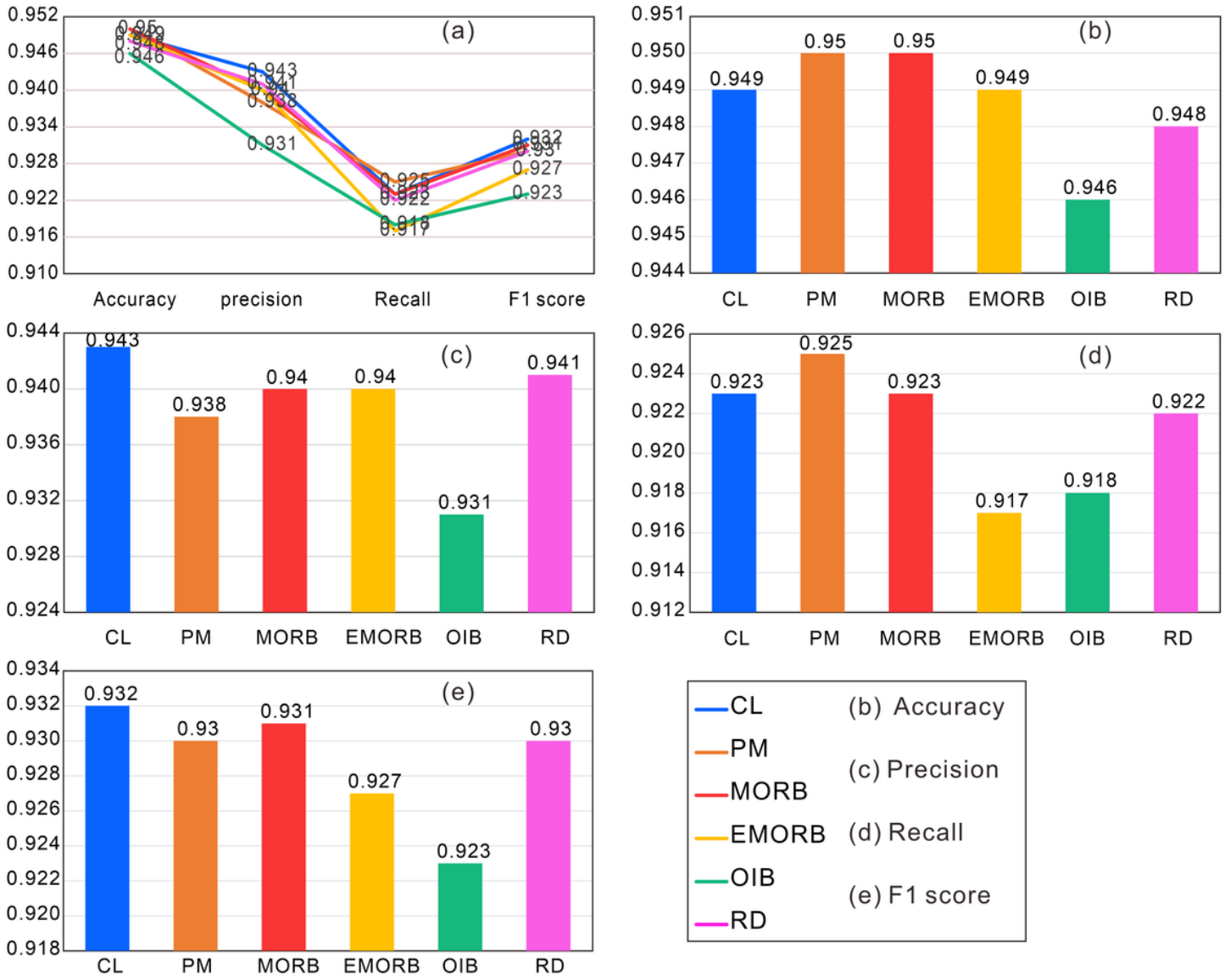

Figure 8

(a) Plot of Accuracy, Precision, Recall, and F1 score for normalization classifier models; (b) comparison of Accuracy between normalization data and raw data in XGBoost classifier model; (c) comparison of Precision between normalization data and raw data in XGBoost classifier model; (d) comparison of Recall between normalization data and raw data in XGBoost classifier model; (e) comparison of F1score between normalization data and raw data in XGBoost classifier model

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarytalble1.xlsx