A practical approach for tectonic discrimination of basalts using geochemical data through machine learning

Mengqi Gao[1], Zhaochong Zhang[2,3], Xiaohui Ji[1], Hengxu Li[3], Zhiguo Cheng[3], M. Santosh[3,4]

1. *School of Information Engineering, China University of Geosciences, Beijing, 100083, China.*
2. *Frontiers Science Center for Deep-time Digital Earth, China University of Geosciences (Beijing), Beijing 100083, China.*
3. *State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Beijing 100083, China.*
4. *Department of Earth Sciences, University of Adelaide, Adelaide, SA, Australia.*

Corresponding author: Zhaochong Zhang(zczhang@cugb.edu.cn);

Xiaohui Ji(xhji@cugb.edu.cn);

Key Points:

- XGBoost demonstrates the best performance in discriminating basalts into seven tectonic settings.
- Two schemes for classification of basalt hold significant practical applications.
- Sr, Ba, Ta, FeOt and Nb are the top five elements with the highest average SHAP values in tectonic discrimination.

ABSTRACT

Identifying the tectonic setting of formation of rocks is an essential component in the field of geosciences. The conventional approach is to employ standard tectonic discrimination diagrams based on elemental correlations and ratios, which sometimes are plagued with uncertainties and limitations. The application of machine learning algorithms based on big data can effectively overcome these problems. In this study, three machine learning algorithms, namely Support Vector Machine, Random Forest, and XGBoost, were employed to classify the various types of basalts from diverse settings such as intraplate basalts, island arc basalts, ocean island basalts, mid-ocean ridge basalts, back-arc basin basalts, oceanic flood basalts, and continental flood basalts into seven tectonic environments. For the altered basalts and fresh basalt, we use 22 relatively immobile elements (TiO_2, P_2O_5, Nb, Ta, Zr, Hf, Y, La, Ce, Pr, Nd, Sm, Eu, Gd, Ho, Er, Yb, Lu, Dy, Tb, Cr, Ni) and 35 major plus trace elements to build discrimination models for seven types of tectonic settings of basalt, respectively. The results indicate that XGBoost demonstrates the best performance in discriminating basalts into seven tectonic settings, achieving an accuracy of 85% and 89% respectively. Compared to previous models, our new method presented in this study is expected to have better practical applications.

PLAIN LANGUAGE SUMMARY

Many works have tried to use compositions of young basalts to correlate the geochemical signatures with their specific tectonic settings. These have led to the development of ‘tectonomagmatic discrimination diagrams’. However, the compositions of basalts are dependent upon their source and mineralogy, the depth, degree and mechanism of partial melting and the various fractionation and contamination processes that they went through en route to the surface. Thus, these discrimination diagrams have many uncertainties and limitations. Machine learning algorithms excel at uncovering latent information within extensive datasets and have demonstrated significant advantages and performance in geochemical research. In this study, three machine learning algorithms were employed to discriminate seven tectonic environments based on global big geochemical data of basalt. Considering practicality and accuracy, we use two schemes to build discrimination models. For fresh basalt samples, a combination of major and trace elements is utilized to enhance the model accuracy (89%). In contrast, for altered basalts, we use another model that is based on 22 relatively immobile elements, although the accuracy is slightly lower (85%). The discriminative analysis of basaltic geological tectonic environments based on machine learning holds significant practical application value.

1 Introduction

Basalt, as a derivative of the mantle, is an important proxy for studying mantle composition and evolution, crustal recycling, and interactions among multiple layers of the Earth. Since the introduction of trace element discrimination diagrams in the 1970s (Pearce & Cann, 1973; Pearce & Norry, 1979; Wood, 1980), these diagrams have been widely used to discern the tectonic settings of basalt formation. However, with the accumulation of data, the substantial overlapping regions from different tectonic settings have been revealed (Li et al., 2015). Consequently, these diagrams often yield ambiguous or conflicting results. This ambiguity might be attributed to various factors influencing basalt composition, including mineral composition and components in the source region, depth, degree, and form of partial melting, as well as processes involving fractional crystallization during magmatic evolution as well as crustal assimilation and mixing, leading to uncertainties and challenges in discerning tectonic environments based on geochemical data of basalts (Li et al., 2015). In recent years, the rapid advancement and breakthrough innovations in Earth Science Big Data and artificial intelligence technologies have brought forth new opportunities and challenges for resolving this challenge. Compared to traditional research approaches, machine learning methods have the advantage of performing more comprehensive and in-depth data analysis, enabling to investigate the intrinsic connections and patterns among scattered data points in multidimensional spaces. Currently, extensive geochemical data, including elemental and isotopic composition of basalt, are extracted from relevant databases such as GEOROC and PetDB for addressing the problem of discriminating tectonic environments of basalt using machine learning methods. For example, Petrelli and Perugini (Petrelli & Perugini, 2016) gathered data from GEOROC and PetDB databases, comprising a total of 3095 basalt samples from eight different tectonic backgrounds: continental arc, island arc, intraoceanic arc, back-arc basin, continental flood, midocean ridge, oceanic plateau, and ocean island. This dataset included 24 elements (8 major and 16 trace elements), along with Sr, Nd, and Pb isotope data. They established a classification model for discerning basaltic tectonic backgrounds based on the Support Vector Machine (SVM) method. The model achieved an average accuracy of 0.93, with even the most challenging to differentiate back-arc basin basalt reaching an accuracy of 0.65. Notably, ocean island basalt exhibited an exceptional accuracy of 0.99. Subsequently, Ueki et al. (Ueki et al., 2018), based on the same elements and isotope data from 2074 samples, employed SVM, Random Forest (RF), and Sparse Multinomial Regression algorithms to build machine learning classification models. The outcomes of these models were similar to those reported by Petrelli and Perugini (Petrelli & Perugini, 2016). However, despite the high accuracy of their classification methods, they encountered two significant challenges in practical applications as follows. 1) Basalt samples used to discriminate ancient tectonic environments are generally relatively 'older' and have often undergone post-magmatic weathering or alteration processes, resulting in the migration of mobile elements (such as K, Na, Rb, Sr, Ba, Mg, Ca) and changes in isotopic compositions (e.g., Rb-Sr isotope and Pb isotope) that play crucial roles in their classification. 2) In many analyses, there is often a lack of isotope data, particularly Pb isotopes. Even when such isotope data are available, the limited quantity of samples analyzed undermines their statistical significance.

In order to address the issues faced in practical applications mentioned above, this study omitted isotope data to ensure a sufficient number of samples. Geological data of basalt from various tectonic environments were extracted from GEOROC and PetDB databases, and after data cleansing, 14150 valid samples were retained. Three different methods—SVM, RF, and XGBoost—were employed to establish classification models for seven types of basaltic tectonic environments (Intraplate, continental and oceanic arc, back-arc basin, continental flood, midocean ridge, oceanic plateau, and ocean island). Considering potential alterations in samples, 22 relatively immobile elements (TiO_2, P_2O_5, Nb, Ta, Zr, Hf, Y, La, Ce, Pr, Nd, Sm, Eu, Gd, Ho, Er, Yb, Lu, Dy, Tb, Cr, Ni) were selected to build seven classification models for basalt. These models achieved an overall accuracy of approximately 85%. For fresh samples, 35 major and trace elements were chosen to build basalt classification models, resulting in an impressive overall accuracy of 89%. Hence, compared to previous models, the basalt discrimination model presented in this study is expected to have better practical applications.

2 Data Descriptions and Pre-Processing

The data used in this study are drawn from two public geochemical databases, GEOROC and PETDB. The dataset comprises 68,327 records of basalt samples from seven geological tectonic environments: Intraplate Basalts (IPB), Island Arc Basalts (IAB), Ocean Island Basalts (OIB), Mid-Ocean Ridge Basalts (MORB), Back Arc Basin Basalts (BABB), Ocean Floor Basalts (OFB), and Continental Flood Basalts (CFB).

To improve the quality of the data, preprocessing operations were conducted on the raw data with the following main steps: (1) Data Integration: the consolidation and merging of two databases with different formats were performed, encompassing 37 fields, including major elements, trace elements, latitude, and longitude. (2) Transforming Fe_2O_3 and FeO into FeOt content (Chen et al., 2022). (3) Removing samples with fewer than 20 non-null values. (4) Removing duplicate samples. (5) Impute missing values using the K-nearest neighbors (K=5) interpolation method. Specifically, for each sample with missing values, calculate its distance to other known values in the dataset, select the K nearest known values, and then use the weighted average of these nearest neighbor values as the estimate for the missing values (Troyanskaya et al., 2001). (6) Selecting samples where the total content of major elements (SiO_2, TiO_2, Al_2O_3, FeOt, CaO, MgO, MnO, K_2O, Na_2O, and

P_2O_5) falls within the range of 97.5% to 102.5% (Nakamura, 2023). (7) Rescaling the total content of major elements (SiO_2, TiO_2, Al_2O_3, FeOt, CaO, MgO, MnO, K_2O, Na_2O, and P_2O_5) to 100% anhydrous basis (Ueki et al., 2018). (8) Samples with SiO_2 content between 45% and 52% were selected to ensure they are basaltic compositions. (9) Removing outliers: conducting outlier analysis and processing using boxplots (Liu & Shi, 2022). For detailed methods, please refer to Appendix B. (10) Normalize the data to the range [0,1] (Zhang et al., 2023),The formula is as below (2-1):

$$x_{normalized} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{2-1}$$

After data cleaning and preprocessing, we obtained 14,150 basalt samples (4,582 intraplate basalts, 3,957 island arc basalts, 1,767 ocean island basalts, 687 mid-ocean ridge basalts, 621 back-arc basin basalts, 304 ocean floor basalts, and 2,232 continental flood basalts). The dataset is divided into a training set (75% of the data) and a testing set (25% of the data). The distribution of the basalt samples after preprocessing the global data is illustrated in Figure 1.

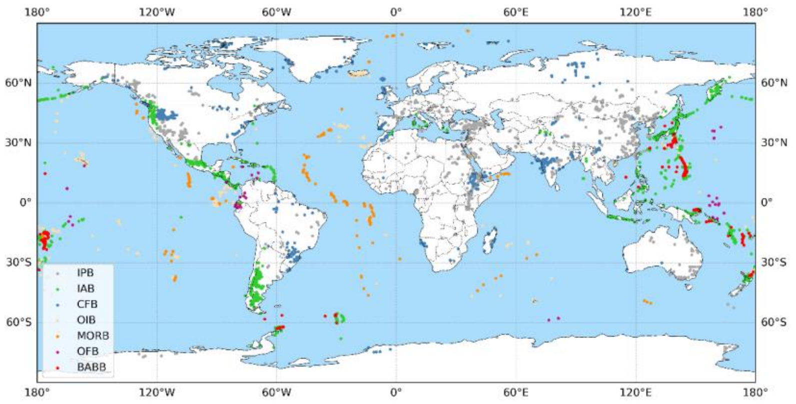


Figure 1. Global Distribution Map of Basalt

3 Basalt Classification Based on Machine Learning

Utilizing machine learning to study basalt data, the latent information and patterns are uncovered, and the learned knowledge is applied to predict outcomes for new and unknown basalt data. Model performance is evaluated using metrics such as accuracy, F1 score, and confusion matrix. Additionally, interpretability analysis is conducted on the model to understand the process of predictions or decisions, enhancing trust and acceptance of the model's predictive process.

3.1 Classification Model

In most of the literature on tectonic environment discrimination, the SVM method is commonly used (Liu & Shi, 2022; Ueki et al., 2018). However, considering the development of machine learning algorithms, ensemble algorithms based on tree models demonstrate better performance in certain scenarios (Chen et al., 2022; Zhang et al., 2023). Combining the results of a preliminary trial comparing various popular machine learning algorithms, this study adopts SVM, RF based on Bagging ensemble, and XGBoost based on Boosting ensemble for classification.

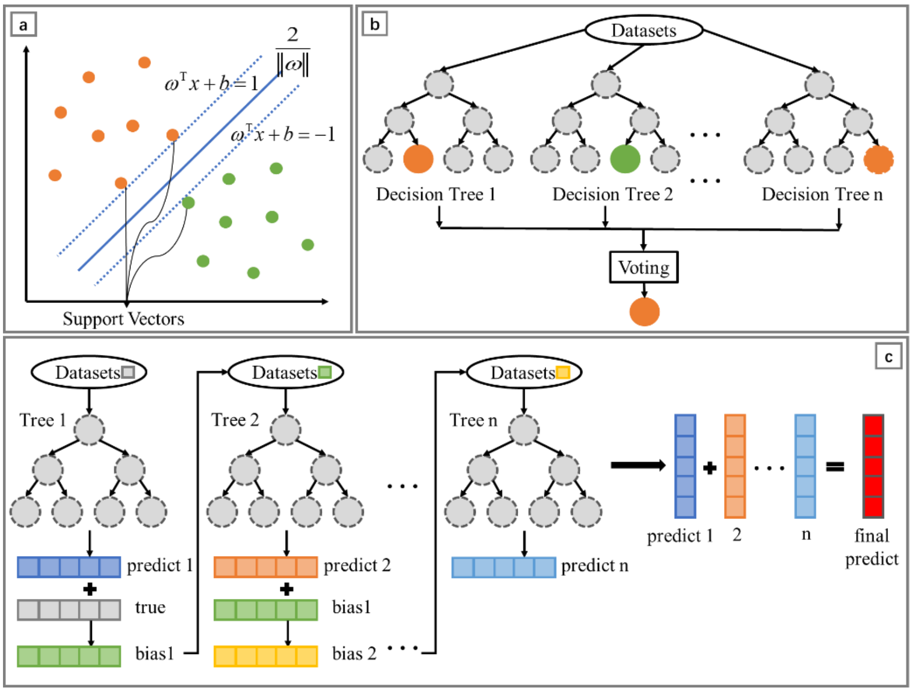


Figure 2. (a) SVM Model, (b) RF Model, (c) XGBoost Model.

Support Vector Machine (SVM), as shown in Figure 2a, belongs to supervised learning. It separates different categories of basalt samples by finding a decision boundary (or hyperplane) and maximizing the distance from the boundary to the nearest basalt samples (support vectors) (Cortes & Vapnik, 1995). To facilitate the linear separation of basalt samples, the data are mapped to a higher-dimensional space. This allows SVM to construct a hyperplane in the high-dimensional space, even when the data are not linearly separable in the original space, effectively separating different categories of basalts.

RF, as shown in Figure 2b, enhances overall performance by ensemble learning with multiple decision trees (Breiman, 2001). Decision trees, resembling binary tree structures, iteratively select the best features as nodes based on entropy calculations, partitioning the dataset into different branches until reaching leaf nodes, which represent the final predicted results. After hyperparameter tuning, the final configuration for constructing the RF includes 300 decision trees with a maximum depth of 30. In each iteration, RF randomly selects all samples with replacement and five random features from the entire basalt training set to train each decision tree. The results from the 300 decision trees are aggregated through a majority voting mechanism, with the most voted result determining the final predicted category for basalts.

The XGBoost, as shown in Figure 2c, achieves the final results by ensemble learning with 500 decision trees (after hyperparameter tuning). During training, it employs a forward distribution algorithm for greedy learning. In each iteration, a decision tree is learned to fit the residual between the predicted values of the previous tree and the actual values. This process continues until the model converges. The final prediction of the entire model is the sum of predictions from all sub-models, with the most significant one determining the corresponding category (Chen & Guestrin, 2016).

3.2 Model Tuning

This study employs grid search algorithm and five-fold cross-validation for hyperparameter optimization of the models, aiming to determine the optimal parameter combinations for SVM, RF, and XGBoost. The grid search algorithm specifies possible values for each hyperparameter and systematically tries all possible combinations. Five-fold cross-validation randomly divides the original dataset into five equally sized subsets, using four subsets sequentially as training sets and the remaining one as a validation set. This process generates five different training and validation sets, and the average of the performance metrics from these five evaluations serves as the model's performance indicator. The combination that performs the best in both grid search and five-fold cross-validation represents the optimal parameters for the model (Pedregosa et al., 2011; Zhao et al., 2019).

3.3 Model Evaluation

The classification models are evaluated using accuracy, F1 score, and confusion matrix. Accuracy represents the proportion of correctly predicted basalt samples out of the total predicted ones, as shown in Formula (3-1). However, in the case of imbalanced samples, accuracy may be influenced by the majority class and may not accurately reflect the model's performance. For instance, with only 304 OFB samples and 621 BABB samples compared to 4582 IPB samples, the number of different basalt samples is highly uneven. Therefore, F1 score is introduced for evaluation, as depicted in Formula (3-2). Precision, denoted as Precision, represents the proportion of correctly classified samples among those predicted as positive class, as shown in Formula (3-3). Recall, denoted as Recall, indicates the probability of correctly predicting positive samples out of the actual positive samples, as shown in

Formula (3-4). In Formulae (3-3) and (3-4), TP is the true positive, representing the number of samples correctly predicted as belonging to a certain category, for example, the number of IPB basalt correctly predicted as IPB. TN is the true negative, representing the number of samples correctly predicted as not belonging to a certain category, for example, the number of non-IPB basalt correctly predicted as not IPB. FP is the false positive, representing the number of samples incorrectly predicted as belonging to a certain category, for example, non-IPB basalt incorrectly predicted as IPB. FN is the false negative, representing the number of samples incorrectly predicted as not belonging to a certain category, for example, IPB basalt incorrectly predicted as not IPB. Ideally, high precision and recall are desired, but in reality, there is a trade-off between the two. The F1 score, as the harmonic mean of precision and recall, provides a comprehensive assessment, considering both metrics effectively for model evaluation.

$$\text{Accuracy}=\frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{3-1}$$

$$\text{F1 score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{3-2}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{3-3}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3-4}$$

The Confusion Matrix, also known as an error matrix, provides a detailed classification of the model's prediction results. It allows for a visual representation of the prediction performance for each class. In the Confusion Matrix, the vertical axis represents the true classes, while the horizontal axis represents the predicted classes. The values in each grid indicate the proportion of samples predicted as the corresponding class in the test set.

3.4 Element Importance Analysis

To gain a better understanding of the model's decision-making process, SHAP (SHapley Additive exPlanations) is employed to provide precise and consistent estimates of the contribution of each feature to the model's classification (Lundberg & Lee, 2017). For each predicted basalt sample, the model generates a prediction value, and the SHAP value represents the numerical allocation of each element in that sample. SHAP not only reflects the impact of element importance in each basalt sample but also indicates the positive or negative influence of these impacts.

4 Results

Rocks may have undergone changes in their structure, texture, and composition due to post-formation geological processes. Generally, the longer a rock has been in existence, the more significant the impact of weathering and alteration, leading to the migration of some mobile elements. Thus, in many cases, the bulk-rock composition may not accurately represent the original composition of the rock. Because of this, our present study explores two scenarios: if basalts have undergone a certain degree of alteration, discriminant analysis is conducted using relatively immobile elements. Conversely, for fresh samples, more elements (including mobile elements) can be included in the analysis. With an increased variety of chemical elements involved in the analysis, the differences between different categories of basalts become more pronounced, resulting in better classification performance by machine learning models. For these two different classification scenarios, SVM, RF, and XGBoost are all employed for analysis.

4.1 Tectonic environment classification based on immobile elements

The 22 relatively immobile elements (TiO_2, P_2O_5, Nb, Ta, Zr, Hf, Y, La, Ce, Pr, Nd, Sm, Eu, Gd, Ho, Er, Yb, Lu, Dy, Tb, Cr, Ni) are employed to classify seven types of basalts.

4.1.1 Classification Results

Table 1. Results of the classification based on immobile elements

	SVM		RF		XGBoost	
	Accuracy	F1_score	Accuracy	F1_score	Accuracy	F1_score
IPB	83.25%	84.80%	85.60%	86.66%	**87.00%**	**88.15%**
IAB	87.68%	82.16%	88.59%	84.12%	**90.51%**	**86.74%**
OIB	76.24%	77.56%	75.34%	80.05%	**80.32%**	**82.56%**
MORB	**83.72%**	78.26%	84.30%	82.15%	**83.72%**	**83.24%**
BABB	41.03%	52.46%	51.28%	57.55%	**60.26%**	**64.16%**
OFB	76.32%	76.82%	**85.53%**	**89.04%**	81.58%	85.52%
CFB	77.60%	79.82%	**84.23%**	83.78%	**84.23%**	**84.84%**

Overall	80.73%	75.98%	83.36%	80.48%	**85.25%**	**82.17%**

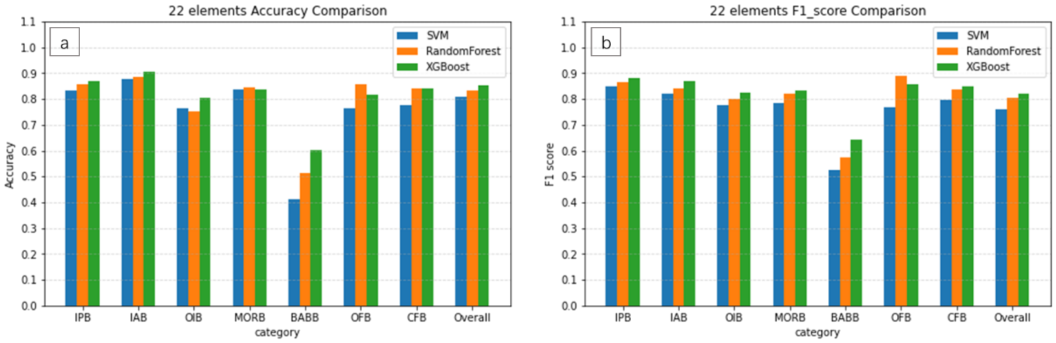


Figure 3. (a) Accuracy and (b) F1 score for classification based on immobile elements

Table 1 and Figure 3 present the accuracy and F1 score of SVM, RF, and XGBoost models for classification based on 22 immobile elements. XGBoost achieves the highest accuracy and F1 score in identifying IPB, IAB, OIB, MORB, BABB, and CFB basalt. MORB has the same accuracy in SVM as XGBoost but lower F1 score. CFB has the same accuracy in RF as XGBoost but lower F1 score. OFB attains the highest accuracy and F1 score in RF. Overall, when using 22 immobile elements for the identification of basalts from seven tectonic settings, XGBoost exhibits the best accuracy and F1 score.

4.1.2 Confusion Matrix

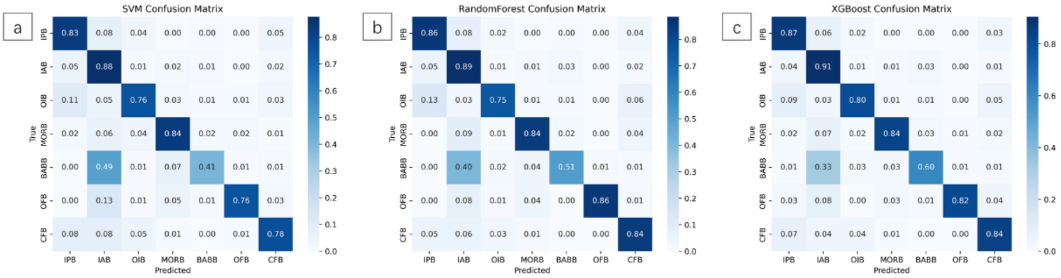


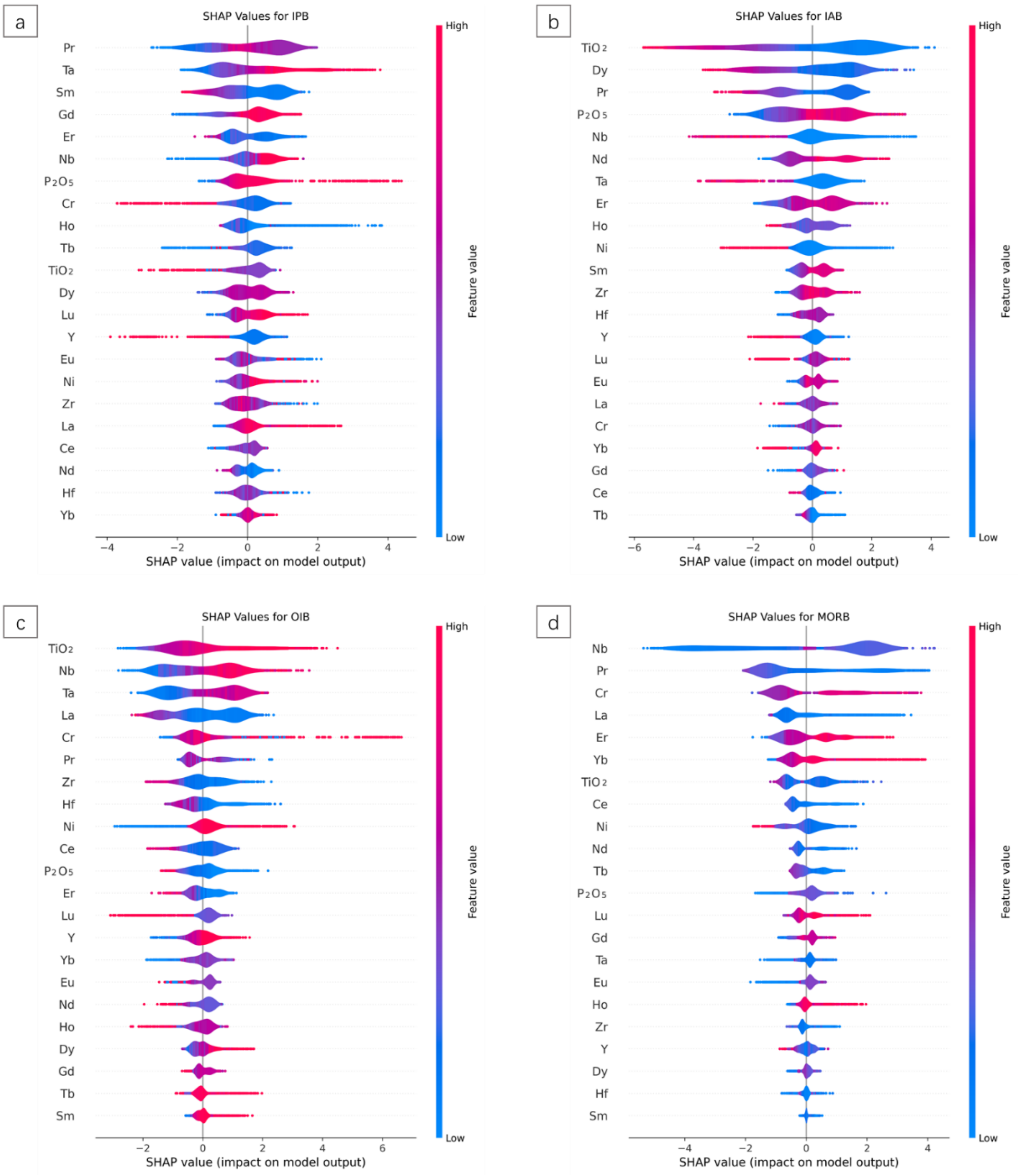
Figure 4. Confusion matrices for the three classification models (a-SVM; b-RF; c-XGBoost) based on immobile elements

In Figure 4c, when using XGBoost for classification based on immobile elements, the model achieves an accuracy of 87% in identifying IPB, with 6% misclassified as IAB. The accuracy for IAB recognition is 91%, the highest among the seven basalt types, with 4% misclassified as IPB. OIB recognition accuracy is 80%, with 9% and 5% misclassified as IPB and CFB, respectively. MORB recognition accuracy is 84%, with 7% misclassified as IAB. BABB recognition accuracy is 60%, with 33% misclassified as IAB. OFB recognition accuracy is 82%, with 8% misclassified as IAB and 4% as CFB. CFB recognition accuracy is 84%, with 7% misclassified as IPB, 4% as IPB, and 4% as OIB.

4.1.3 Element Importance Analysis

In Figure 5a-5g, each row represents an element, and the x-axis represents the SHAP value. Each point represents a basalt sample, where the color of the point indicates the content of that element in the sample. A deeper red color signifies a higher content, while a deeper blue color indicates a lower content. The larger the colored area, the more samples there are. The higher the position of an element in the figure, the higher the corresponding SHAP value, indicating its greater importance in determining the classification of this type of basalt (Lundberg et al., 2018).

For the IPB tectonic environment (Figure 5a), the top five elements with the highest SHAP values are Pr, Ta, Sm, Gd, and Er, indicating their significant impact when predicting the IPB tectonic environment. Yb has the smallest corresponding SHAP value, indicating its minimal influence on predicting the IPB tectonic environment. Specifically, moderate Pr content, higher Ta and Gd content, and lower Sm and Er content make the basalt category more likely to be predicted as IPB.



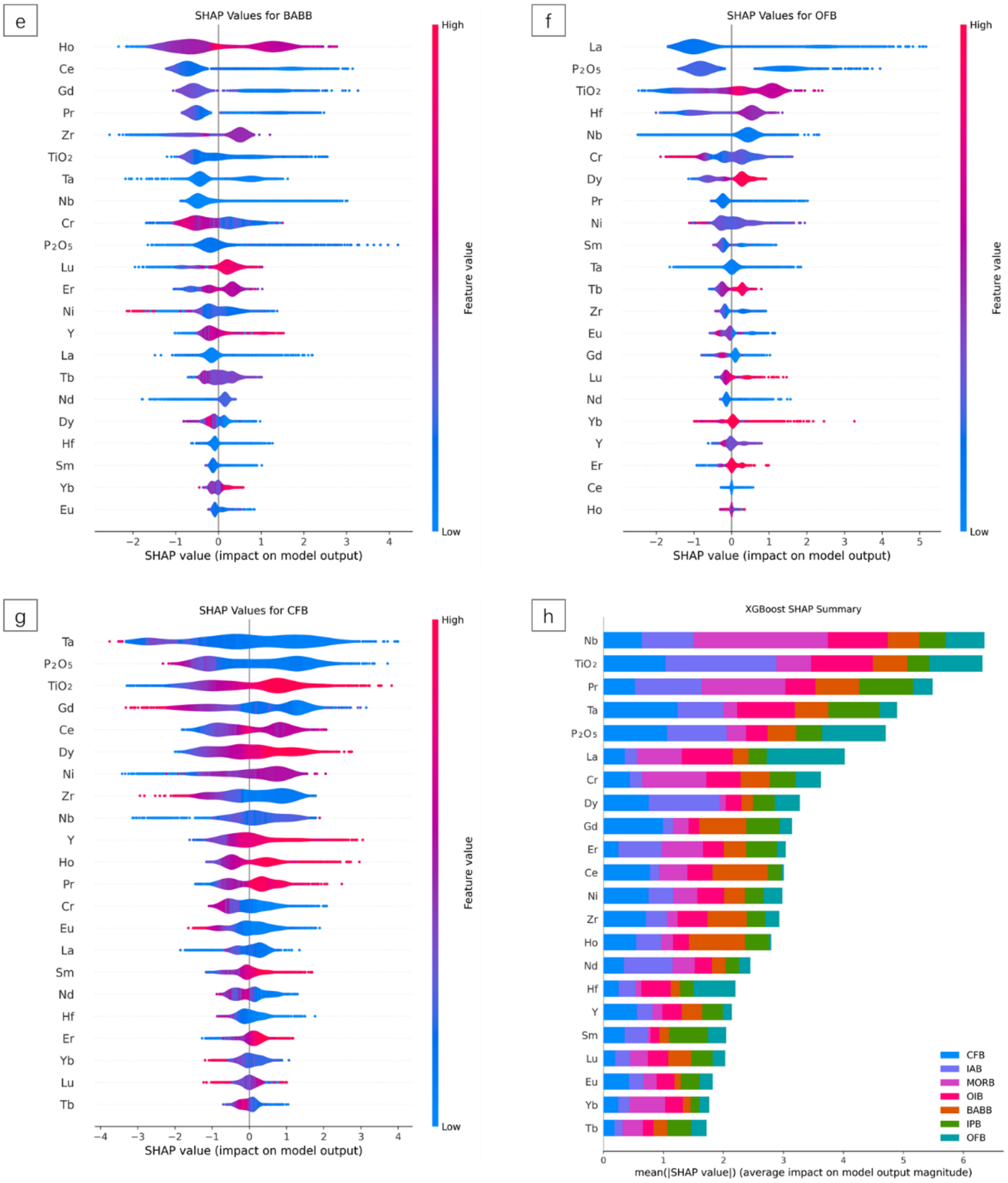


Figure 5. Impact of features for the 22 immobile elements in seven tectonic settings of basalt (a-IPB; b-IAB; c-OIB; d-MORB; e-BABB; f-OFB; g-CFB; h-mean).

For the IAB tectonic environment (Figure 5b), the top five elements with the highest SHAP values are TiO_2, Dy, Pr, P_2O_5, and Nb, showing their major influence when predicting the IAB tectonic environment. Tb has the smallest corresponding SHAP value, indicating its minimal impact on predicting the IAB tectonic environment. In detail, lower TiO_2, Dy, Pr, Nb content, and higher P_2O_5 content make the basalt category more likely to be predicted as IAB.

For the OIB tectonic environment (Figure 5c), the top five elements with the highest SHAP values are TiO_2, Nb, Ta, La, and Cr, indicating their significant impact when predicting the OIB tectonic environment. Sm has the smallest corresponding SHAP value, indicating its minimal influence on predicting the OIB tectonic environment. Specifically, higher TiO_2, Nb, Ta, Cr content, and lower La content make the basalt category more likely to be predicted as OIB.

For the MORB tectonic environment (Figure 5d), the top five elements with the highest SHAP values are Nb, Pr, Cr, La, and Er, showing their major influence when predicting the MORB tectonic environment. Sm has the smallest corresponding SHAP value, indicating its minimal impact on predicting the MORB tectonic environment. Lower Nb, Pr, La content, and higher Cr, Er content make the basalt category more likely to be predicted as MORB.

For the BABB tectonic environment (Figure 5e), the top five elements with the highest SHAP values are Ho, Ce, Gd, Pr, and Zr, indicating their significant impact when predicting the BABB tectonic environment. Eu has the smallest corresponding SHAP value, indicating its minimal influence on predicting the BABB tectonic environment. Specifically, lower Ce, Gd, Pr content, and higher Ho, Zr content make the basalt category more likely to be predicted as BABB.

For the OFB tectonic environment (Figure 5f), the top five elements with the highest SHAP values are La, P_2O_5, TiO_2, Hf, and Nb, indicating their significant impact when predicting the OFB tectonic environment. Ho has the smallest corresponding SHAP value, indicating its minimal influence on predicting the OFB tectonic environment. Specifically, lower La, P_2O_5, Nb content, and higher TiO_2, Hf content make the basalt category more likely to be predicted as OFB.

For the CFB tectonic environment (Figure 5g), the top five elements with the highest SHAP values are Ta, P_2O_5, TiO_2, Gd, and Ce, indicating their significant impact when predicting the CFB tectonic environment. Tb has the smallest corresponding SHAP value, indicating its minimal influence on predicting the CFB tectonic environment. Specifically, lower Ta, P_2O_5, Gd content, and higher TiO_2, Ce content make the basalt category more likely to be predicted as CFB.

Figure 5h is an overall stacked bar chart of element importance, sorted according to element importance, indicating the overall importance of different elements when classifying the seven types of basalts. It can be seen that when considering all tectonic environments (Figure 5h), the top five elements with the highest average SHAP values are Nb, TiO_2, Pr, Ta, and P_2O_5, while Tb, Yb, and Eu have the lowest average SHAP values, indicating that Nb, TiO_2, Pr, Ta, and P_2O_5 are the most important elements for classifying the seven types of tectonic environments for basalts.

4.2 Tectonic environment classification based on 35 elements

In this section, we add 13 more elements to improve the overall classification performance of the model by increasing the classification features. All the 35 elements are included such as SiO_2, TiO_2, Al_2O_3, FeOt, CaO, MgO, MnO, K_2O, Na_2O, P_2O_5, Rb, Sr, Ba, Th, U, Nb, Ta, Zr, Hf, Y, La, Ce, Pr, Nd, Sm, Eu, Gd, Ho, Er, Yb, Lu, Dy, Tb, Cr, Ni.

4.2.1 Classification Results

Table 2. Results of the classification based on 35 elements

	SVM		RF		XGBoost	
	Accuracy	F1_score	Accuracy	F1_score	Accuracy	F1_score
IPB	87.61%	89.44%	87.52%	88.53%	**89.70%**	**91.09%**
IAB	90.51%	87.80%	90.20%	86.57%	**91.82%**	**89.03%**
OIB	**87.78%**	**88.69%**	83.26%	87.51%	**87.78%**	**88.69%**
MORB	**94.19%**	88.04%	91.28%	**90.49%**	91.28%	89.97%
BABB	58.97%	65.95%	64.10%	67.80%	**66.67%**	**71.72%**
OFB	80.26%	83.56%	**90.79%**	**95.17%**	85.53%	91.55%
CFB	88.71%	87.92%	88.89%	88.33%	**89.25%**	**88.53%**
Overall	87.51%	84.49%	87.18%	86.34%	**88.95%**	**87.23%**

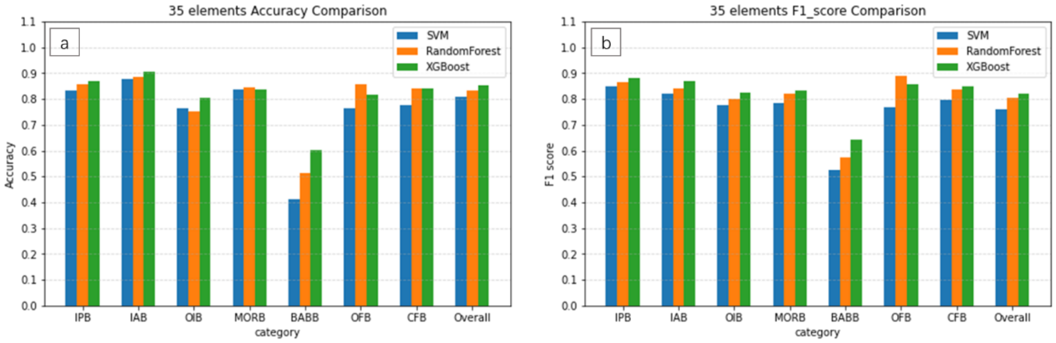


Figure 6. (a) Accuracy and (b) F1 score for classification based on 35 elements

[Table 2](#) and [Figure 6](#) present the accuracy and F1 score of SVM, RF, and XGBoost models in classifying the 35 elements. IPB, IAB, OIB, BABB, and CFB achieved the highest accuracy and F1 score in XGBoost. OIB had the same accuracy and F1 score in SVM and XGBoost. MORB had the highest accuracy in SVM, and RF had the highest F1 score. OFB achieved the highest accuracy and F1 score in RF. Overall, using 35 elements for the identification of seven tectonic settings of basalt, XGBoost exhibited the best accuracy and F1 score.

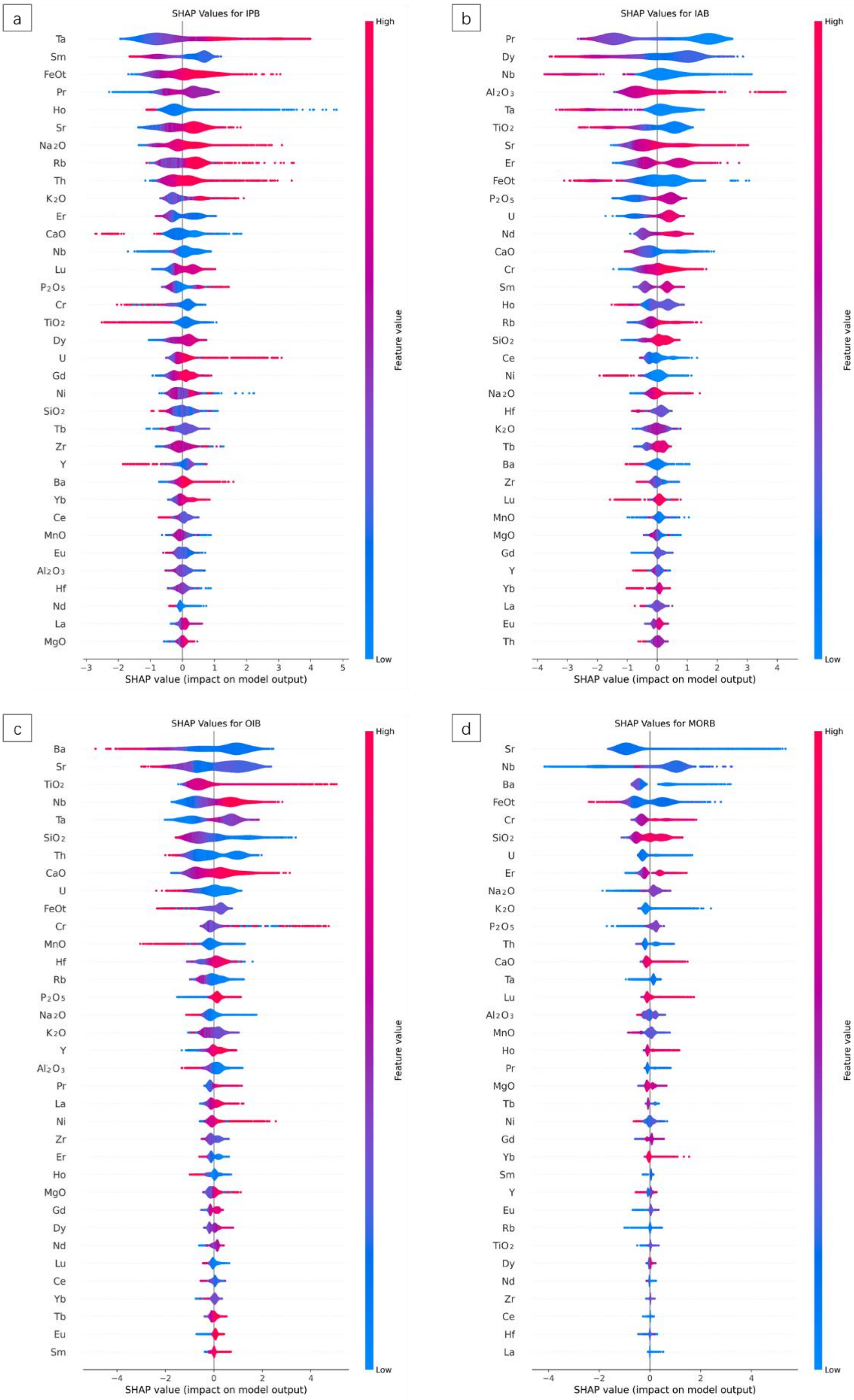
4.2.2 Confusion Matrix

Figure 7. Confusion matrices for the three classification models (a-SVM; b-RF; c-XGBoost) based on 35 elements

In Figure 7c, when the number of classification elements increased to 35 and XGBoost was used for classification, the model achieved an accuracy of 90% in identifying IPB, with 5% misclassification as IAB. The accuracy of identifying IAB reached 92%, the highest among the seven types of basalt, with 3% misclassification as IPB. The accuracy of identifying OIB was 88%, with 5% misclassification as IPB. The accuracy of identifying MORB was 91%, with 5% misclassification as IAB. The accuracy of identifying BABB was 67%, with 29% misclassification as IAB. The accuracy of identifying OFB was 86%, with 5% misclassification as IAB, 4% as IPB, and 4% as CFB. The accuracy of identifying CFB was 89%, with 5% misclassification as IPB. Compared to the classification using immobile elements, the accuracy of IPB increased by 3%, IAB increased by 1%, OIB increased by 8%, MORB increased by 7%, BABB increased by 7%, OFB increased by 4%, and CFB increased by 5%.

4.2.3 Element Importance Analysis

For the IPB tectonic environment (Figure 8a), the top five elements with the highest SHAP values are Ta, Sm, FeOt, Pr, and Ho. This indicates that these elements have the greatest influence when predicting the IPB tectonic environment. MgO has the smallest corresponding SHAP value, indicating the least impact on predicting the IPB tectonic environment. Specifically, higher concentrations of Ta, FeOt, and Pr and lower concentrations of Sm and Ho make it easier to predict basalt types as IPB.


a
SHAP Values for IPB
Ta
Sm
FeOt
Pr
Ho
Sr
Na₂O
Rb
Th
K₂O
Er
CaO
Nb
Lu
P₂O₅
Cr
TiO₂
Dy
U
Gd
Ni
SiO₂
Tb
Zr
Y
Ba
Yb
Ce
MnO
Eu
Al₂O₃
Hf
Nd
La
MgO
SHAP value (impact on model output)
b
SHAP Values for IAB
Pr
Dy
Nb
Al₂O₃
Ta
TiO₂
Sr
Er
FeOt
P₂O₅
U
Nd
CaO
Cr
Sm
Ho
Rb
SiO₂
Ce
Ni
Na₂O
Hf
K₂O
Tb
Ba
Zr
Lu
MnO
MgO
Gd
Y
Yb
La
Eu
Th
SHAP value (impact on model output)
c
SHAP Values for OIB
Ba
Sr
TiO₂
Nb
Ta
SiO₂
Th
CaO
U
FeOt
Cr
MnO
Hf
Rb
P₂O₅
Na₂O
K₂O
Y
Al₂O₃
Pr
La
Ni
Zr
Er
Ho
MgO
Gd
Dy
Nd
Lu
Ce
Yb
Tb
Eu
Sm
SHAP value (impact on model output)
d
SHAP Values for MORB
Sr
Nb
Ba
FeOt
Cr
SiO₂
U
Er
Na₂O
K₂O
P₂O₅
Th
CaO
Ta
Lu
Al₂O₃
MnO
Ho
Pr
MgO
Tb
Ni
Gd
Yb
Sm
Y
Eu
Rb
TiO₂
Dy
Nd
Zr
Ce
Hf
La
SHAP value (impact on model output)
High
Feature value
Low

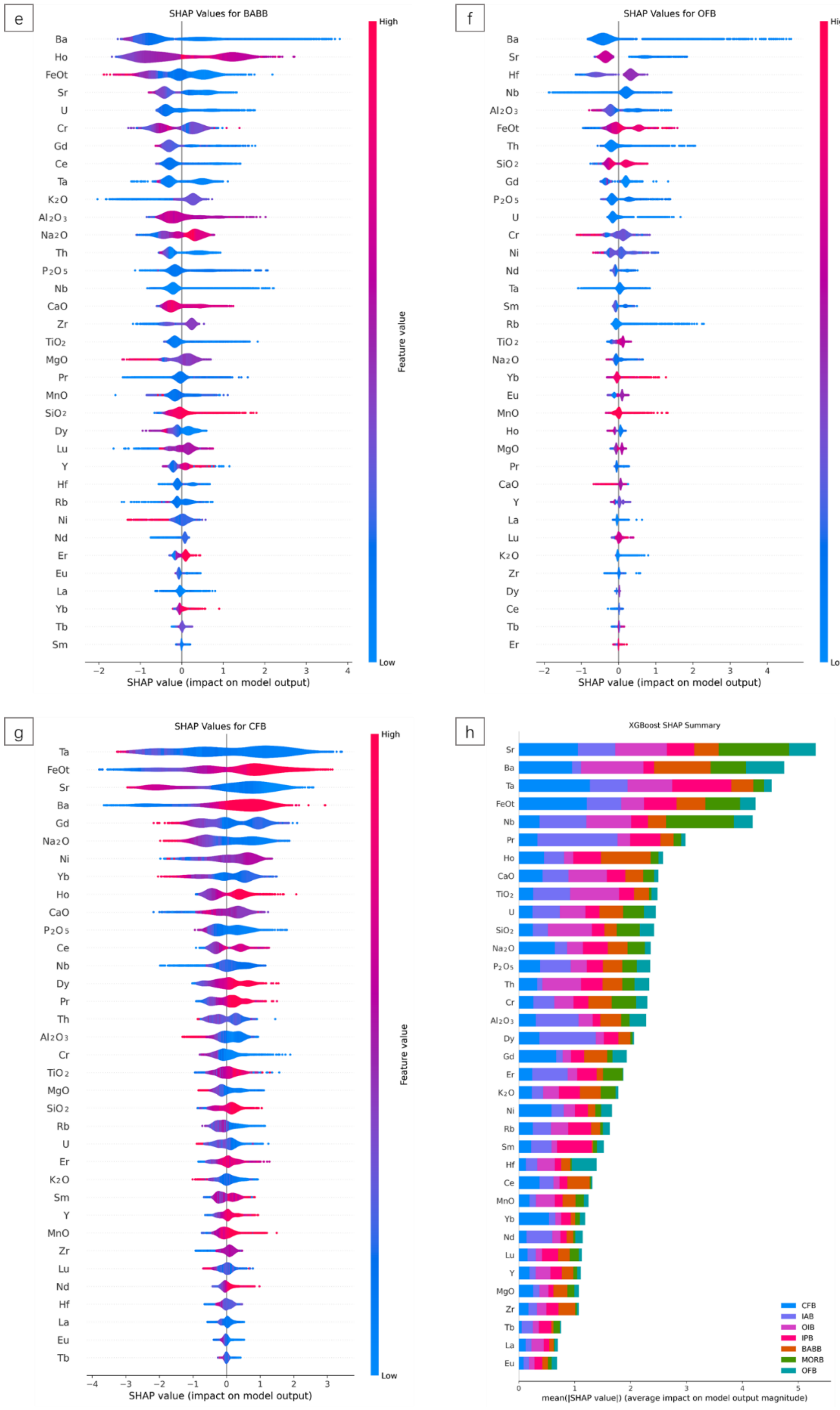


Figure 8. Impact of features for the 35 elements (a-IPB; b-IAB; c-OIB; d-MORB; e-BABB; f-OFB; g-CFB; h-mean).

For the IAB tectonic environment (Figure 8b), the top five elements with the highest SHAP values are Pr, Dy, Nb, Al_2O_3, and Ta. These elements have the greatest impact when predicting the IAB tectonic environment. Th has the smallest corresponding SHAP value, indicating the least impact on predicting the IAB tectonic environment. Specifically, lower concentrations of Pr, Dy, Nb, and Ta and higher concentrations of Al_2O_3 make it easier to predict basalt types as IAB.

For the OIB tectonic environment (Figure 8c), the top five elements with the highest SHAP values are Ba, Sr, TiO_2, Nb, and Ta. These elements have the greatest impact when predicting the OIB tectonic environment. Sm has the smallest corresponding SHAP value, indicating the least impact on predicting the OIB tectonic environment. Specifically, lower concentrations of Ba and Sr and higher concentrations of TiO_2, Nb, and Ta make it easier to predict basalt types as OIB.

For the MORB tectonic environment (Figure 8d), the top five elements with the highest SHAP values are Sr, Nb, Ba, FeOt, and Cr. These elements have the greatest impact when predicting the MORB tectonic environment. Specifically, lower concentrations of Sr, Nb, Ba, and FeOt and higher concentrations of Cr make it easier to predict basalt types as MORB.

For the BABB tectonic environment (Figure 8e), the top five elements with the highest SHAP values are Ba, Ho, FeOt, Sr, and U. These elements have the greatest impact when predicting the BABB tectonic environment. Sm has the smallest corresponding SHAP value, indicating the least impact on predicting the BABB tectonic environment. Specifically, lower concentrations of Ba, FeOt, Sr, and U and higher concentrations of Ho make it easier to predict basalt types as BABB.

For the OFB tectonic environment (Figure 8f), the top five elements with the highest SHAP values are Ba, Sr, Hf, Nb, and Al_2O_3. These elements have the greatest impact when predicting the OFB tectonic environment. Er has the smallest corresponding SHAP value, indicating the least impact on predicting the OFB tectonic environment. Specifically, lower concentrations of Ba, Sr, Nb, Al_2O_3, and higher concentrations of Hf make it easier to predict basalt types as OFB.

For the CFB tectonic environment (Figure 8g), the top five elements with the highest SHAP values are Ta, FeOt, Sr, Ba, and Gd. These elements have the greatest impact when predicting the CFB tectonic environment. Tb has the smallest corresponding SHAP value, indicating the least impact on predicting the CFB tectonic environment. Specifically, lower concentrations of Ta, Sr, and Gd and higher concentrations of FeOt and Ba make it easier to predict basalt types as CFB.

Considering all tectonic settings (Figure 8h), the top five elements with the highest average SHAP values are Sr, Ba, Ta, FeOt, and Nb. The lowest average SHAP values are for Eu, La, and Tb. This indicates that Sr, Ba, Ta, FeOt, and Nb are the most important elements for classifying basalt types in the seven tectonic environments.

5 Discussion

5.1 The impact of imbalanced data sets

The training set has a significant imbalance in the number of samples for each class, and an imbalanced dataset can lead classification models to focus more on the majority class, resulting in biased classification results and reduced model performance. Therefore, Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) is employed to increase the quantity of IAB, OIB, MORB, BABB, OFB, and CFB samples, aiming to balance the number of samples for each class. In specific terms, SMOTE assumes that points in the feature space with proximity in features are also similar. It involves randomly selecting a sample point from the minority class, identifying its K-nearest neighbors, and then randomly choosing one neighbor. The difference between this chosen neighbor and the current sample point is calculated. To ensure diversity, this difference is multiplied by a random threshold within the [0,1] range. The obtained result represents the newly added sample point. This process is repeated until the sample size of each category reaches the target sample size. SMOTE processing is applied only to the training set in this study, with K set to 5.

The oversampled quantities for each basalt category after SMOTE processing are presented in Table 3. The classification results for basalt are presented in Tables 4 and 5, while the confusion matrices are illustrated in Figures A1 and A2.

Table 3. The quantity of basalt in the training set before and after SMOTE processing

	IPB	IAB	OIB	MORB	BABB	OFB	CFB
No SMOTE	3436	2967	1325	515	465	228	1674
SMOTE	3436	3436	3436	3436	3436	3436	3436

Table 4. Accuracy before and after SMOTE - immobile elements

	SVM		RF		XGBoost	
	No SMOTE	SMOTE	No SMOTE	SMOTE	No SMOTE	SMOTE
IPB	**83.25%**	78.10%	**85.60%**	82.11%	**87.00%**	84.55%
IAB	**87.68%**	79.39%	**88.59%**	85.25%	**90.51%**	87.17%
OIB	76.24%	**82.81%**	75.34%	**80.54%**	80.32%	**83.94%**
MORB	83.72%	**91.28%**	84.30%	**88.95%**	83.72%	**86.63%**
BABB	41.03%	**80.77%**	51.28%	**76.92%**	60.26%	**73.08%**
OFB	76.32%	**85.53%**	85.53%	**88.16%**	81.58%	**85.53%**
CFB	77.60%	**81.90%**	**84.23%**	83.69%	84.23%	**85.48%**
Overall	**80.73%**	80.56%	**83.36%**	83.28%	**85.25%**	84.97%

Table 5. F1 score before and after SMOTE - immobile elements

	SVM		RF		XGBoost	
	No SMOTE	SMOTE	No SMOTE	SMOTE	No SMOTE	SMOTE

IPB	**84.80%**	83.80%	**86.66%**	85.82%	**88.15%**	87.61%
IAB	82.16%	**82.17%**	84.12%	**85.17%**	**86.74%**	86.73%
OIB	77.56%	**79.14%**	80.05%	**81.09%**	82.56%	**82.81%**
MORB	78.26%	**80.10%**	82.15%	**85.24%**	**83.24%**	83.01%
BABB	52.46%	**61.92%**	57.55%	**65.22%**	64.16%	**67.06%**
OFB	**76.82%**	75.58%	**89.04%**	83.23%	85.52%	**85.53%**
CFB	79.82%	**80.53%**	**83.78%**	82.00%	**84.84%**	84.35%
Overall	75.98%	**77.61%**	80.48%	**81.11%**	82.17%	**82.44%**

Table 6. Accuracy before and after SMOTE – 35 elements

	SVM		RF		XGBoost	
	No SMOTE	SMOTE	No SMOTE	SMOTE	No SMOTE	SMOTE
IPB	**87.61%**	85.60%	**87.52%**	84.90%	**89.70%**	89.01%
IAB	**90.51%**	85.56%	**90.20%**	87.98%	**91.82%**	91.31%
OIB	87.78%	**89.14%**	83.26%	**87.78%**	87.78%	**89.82%**
MORB	**94.19%**	93.60%	91.28%	**93.60%**	91.28%	**93.60%**
BABB	58.97%	**90.38%**	64.10%	**84.62%**	66.67%	**73.72%**
OFB	80.26%	**89.47%**	90.79%	**93.42%**	85.53%	**89.47%**
CFB	88.71%	**89.43%**	88.89%	**90.50%**	89.25%	**90.14%**
Overall	**87.51%**	87.32%	87.18%	**87.60%**	88.95%	**89.49%**

Table 7. F1 score before and after SMOTE - 35 elements

	SVM		RF		XGBoost	

	No SMOTE	SMOTE	No SMOTE	SMOTE	No SMOTE	SMOTE
IPB	**89.44%**	88.94%	88.53%	**88.62%**	91.09%	**91.32%**
IAB	**87.80%**	87.18%	86.57%	**87.45%**	89.03%	**89.77%**
OIB	88.69%	**89.04%**	87.51%	**88.48%**	88.69%	**88.72%**
MORB	88.04%	**89.69%**	90.49%	**91.48%**	89.97%	**91.48%**
BABB	65.95%	**73.06%**	67.80%	**74.58%**	71.72%	**73.72%**
OFB	83.56%	**85.00%**	**95.17%**	91.61%	91.55%	**91.89%**
CFB	**87.92%**	87.47%	**88.33%**	87.52%	88.53%	**89.42%**
Overall	84.49%	**85.77%**	86.34%	**87.11%**	87.23%	**88.04%**

As shown in Tables 4-7 and Figures A1-A6(Appendix A), after balancing the number of samples for each class of basalt in the training set using SMOTE, the overall accuracy of the models for classifying the seven types of basalts is quite similar, and the overall F1 scores have improved. Specifically, for each class of basalts, whether based on immobile elements or 35 elements, SVM, RF, and XGBoost all exhibit a decrease in accuracy for IPB and IAB, while an increase is observed for OIB, MORB, BABB, OFB, and CFB. For SVM and RF, the accuracy for IPB and IAB decreases by approximately 2% to 5%, and for XGBoost, the decrease is around 1% to 2%. On the other hand, OIB, MORB, OFB, and CFB show an improvement of 2% to 10%, with BABB showing the most significant increase, ranging from 7% to 40%. Overall, although the accuracy for certain classes of basalts decreases, the affected classes are few, and the decrease is small. Considering the improved accuracy for other classes and the overall enhancement in F1 scores, the loss is justified. Therefore, utilizing SMOTE to balance the training samples proves to be an effective method for enhancing model performance.

5.2 Comparison of Machine Learning Methods

In terms of the various algorithms, SVM is suitable for small to medium-sized datasets with low-dimensional features, RF is suitable for medium-sized datasets with high-dimensional features, and XGBoost typically performs well in large datasets and complex problem scenarios. Combining two sets of different classification features, XGBoost performs relatively well in identifying IPB, IAB, OIB, BABB, and CFB (with respective data sizes of 3436, 2967, 1325, 465, 1674). RF performs relatively well in identifying MORB and OFB (with respective data sizes of 515, 228). The experimental results align well with the characteristics of the algorithms.

When the number of classification features increases from 22 to 35, all models show an increase in accuracy, suggesting that at this point, the models have not yet experienced overfitting due to an excessive number of classification features. Therefore, it is inferred that the main factors limiting model accuracy are concentrated in the data itself. For the original real dataset, the sample quantity relationship is IPB > IAB > CFB > OIB > MORB > BABB > OFB. When using two sets of elements and three classification models separately, IAB and MORB perform the best, while BABB performs the worst.

There is no clear pattern among the accuracy rates of other types of basalts, indicating that there is no apparent positive or negative correlation between accuracy and sample quantity.

Table 8. Comparison of Accuracy for Different Test Set Sizes

Elements	Test Set	SVM		RF		XGBoost	
		Accuracy	F1_score	Accuracy	F1_score	Accuracy	F1_score
22	unbalance	81%	78%	83%	81%	85%	82%
	balance	83%	83%	84%	84%	84%	84%
35	unbalance	87%	86%	88%	87%	89%	88%
	balance	87%	87%	89%	89%	88%	88%

In addition, Section 5.1 compared the impact of the quantities of various types of basalt in the training set, considering both unequal and equal quantities. The overall accuracy of the model was found to be similar in both scenarios. To explore the relationship between the test set size and accuracy, while maintaining a balanced quantity of each type of basalt in the training set, 76 samples were randomly selected from each basalt type (the original test set had a minimum of 76 samples for OFB) to create a balanced test set. A comparison of accuracy and F1 score between the imbalanced and balanced test sets is presented in Table 8. Although there are slight variations in individual accuracy and F1 score values, these fluctuations are within a normal range. Thus, it can be concluded that the proportional quantities of different basalt types in the test set do not significantly affect the model's performance.

This study also compared the distribution of testing samples of basalt that were misclassified by SVM, RF, and XGBoost with the distribution of training samples, as shown in Appendix C. The comparative results indicate that the elemental content of misclassified samples exceeds the numerical range learned by the model through training samples, deviating from the potential patterns and rules learned by the model.

5.3 Reason of mis-discrimination of tectonic environments

As evident from the confusion matrices shown in Figures 4 and 7, IPB is frequently misclassified as IAB, possibly due to the influence of crustal contamination on IPB (Hawkesworth & Gallagher, 1993). Additionally, some IPB instances are misclassified as OIB, which may be attributed to their shared intra-plate environment, exhibiting similar mantle sources or partial melting processes (Kovalenko et al., 2007). Misclassification of IAB as BABB may be explained by the fact that both are related to subduction processes, with BABB typically forming after IAB; thus, early-stage BABB often exhibits geochemical characteristics similar to IAB (Ishizuka et al., 2009).The misclassification of OIB as OFB occurs because both are formed in intra-oceanic plate environments, sharing similar mantle source components (Niu et al., 2011). Misclassification of MORB as BABB may be due to their evolutionary relationship, as late-stage BABB tends to evolve toward environments associated with mid-ocean ridges, resulting in similar characteristics of light rare earth element depletion. The misclassification of OFB as IAB may be attributed to the fact that some IAB is an early product of subduction, and this subset of IAB has a lower influence from subduction components, thus sharing similar source components with OFB. The misclassification of CFB as OIB and IPB is also related to their common intra-plate environment, sharing similar source components and partial melting processes (Farmer, 2014).

5.4 The role of elements in tectonic discrimination

Ta, Sm, and FeOt have the most significant impact on distinguishing IPB, with higher Ta and FeOt content resulting in better differentiation of IPB. Although the source regions of IPB are typically heterogeneous and often influenced by crustal contamination, most IPB source regions are enriched in incompatible elements (Kovalenko et al., 2007), such as high field strength elements (HFSE). Therefore, IPB tends to enrich these elements, and the enrichment of FeOt in IPB may be related to the inclusion of eclogite or pyroxenite in the source region (Sobolev et al., 2005) or partial melting of mantle at deep level.

Pr, Dy, and Nb are the three most important elements for distinguishing IAB, with lower concentrations of Pr, Dy, and Nb favoring better differentiation of IAB. The source region of IAB is generally considered to be a depleted mantle source with varying proportions of subducted slab contributions. Therefore, most IAB exhibits depleted rare earth element (REE) signatures

(Labanieh et al., 2012; Stern, 2002). Lower concentrations of Pr and Dy are favorable for distinguishing IAB. Additionally, during the partial melting process that forms IAB in the source region, residual minerals enriched in Nb and Ta may lead to Nb depletion in IAB (Schmidt & Jagoutz, 2017).

Ba, Sr, and TiO_2 display the greatest impact in distinguishing OIB, with lower Ba and Sr concentrations favoring better differentiation, while higher TiO_2 content is advantageous for distinguishing OIB. OIB generally forms in enriched mantle source regions (Hofmann, 1997) and tends to enrich in HFSE, such as Ta, Nb, and Ti. Ba and Sr are elements that are relatively mobile in fluids; hence, volcanic rocks associated with subduction are typically enriched in Ba, Th, and other elements. OIB formation, however, involves minimal fluid involvement, resulting in relatively lower Ba concentrations.

Sr and Nb are most critical in distinguishing MORB, and the lower the content of Sr and Nb, the better the discrimination of MORB. The depletion of Sr in MORB may be related to the early crystallization of certain calcium-rich minerals, such as calcium plagioclase due to low water in melt. Additionally, some MORB samples exhibit characteristics of depleted trace elements, leading to lower Nb and Ta contents compared to other tectonic environments of basalts (Hofmann, 1997).

Ba, Ho, and FeOt are most effective for distinguishing BABB, with lower Ba and FeOt concentrations favoring better differentiation, and higher Ho concentrations being advantageous for differentiation. BABB is characterized by the relative depletion of light rare earth elements (LREE) compared to heavy rare earth elements (HREE); therefore, BABB typically exhibits higher Ho content compared to other tectonic environments (Ishizuka et al., 2009). Ba is a relatively mobile element in fluids, and as a product of island arc evolution, BABB has essentially no fluid involvement in its formation, resulting in relative Ba depletion (Conder et al., 2002). As BABB evolves towards a more calc-alkaline composition during magmatic evolution, it tends to deplete in FeOt.

Ba, Sr, and Hf hold the highest importance for distinguishing OFB, with lower Ba and Sr concentrations favoring better differentiation, and higher Hf concentrations being advantageous for differentiation. The lower Ba and Sr concentrations in OFB may be due to the early crystallization of certain calcium-rich minerals, such as clinopyroxene and calcium plagioclase. Additionally, Ba and Sr are relatively mobile elements in fluids, and since OFB formation involves minimal fluid involvement, the Ba and Sr concentrations are relatively lower.

The content of Ta and FeOt are most important for distinguishing CFB, with lower Ta concentrations leading to better differentiation, and higher FeOt concentrations being advantageous for differentiation. The source region of CFB generally undergoes modification, resulting in heterogeneous source composition. However, most CFB source regions exhibit enrichment in large-ion lithophile elements (LILE) and depletion in HFSE (Farmer, 2014), leading to lower Ta concentrations in CFB compared to other tectonic environments. The higher FeOt concentrations in CFB may be attributed to the participation of pyroxenite or garnet pyroxenite in the partial melting process, resulting in higher FeOt content in the melt (Sobolev et al., 2005).

6 Concluding remarks and future work

When discriminating the tectonic environments of basalt using 22 immobile elements (TiO_2, P_2O_5, Nb, Ta, Zr, Hf, Y, La, Ce, Pr, Nd, Sm, Eu, Gd, Ho, Er, Yb, Lu, Dy, Tb, Cr, Ni), the model with the best classification performance is XGBoost, followed by RF and SVM. XGBoost achieves an overall accuracy of 85%, with the highest accuracy in classifying IAB (91%) and the lowest in classifying BABB (60%).

When discriminating the tectonic environments of basalts using 35 elements (SiO_2, TiO_2, Al_2O_3, FeOt, CaO, MgO, MnO, K_2O, Na_2O, P_2O_5, Rb, Sr, Ba, Th, U, Nb, Ta, Zr, Hf, Y, La, Ce, Pr, Nd, Sm, Eu, Gd, Ho, Er, Yb, Lu, Dy, Tb, Cr, Ni), the model with the best classification performance is XGBoost, with an overall accuracy of 89%, with the highest accuracy in classifying IAB (92%) and the lowest in classifying BABB (67%).

Hence, in practical applications, if the samples have undergone alteration, it is recommended to use immobile elements for discrimination. If the samples have not undergone alteration and are relatively fresh, it is advisable to use major elements along with trace elements for higher classification accuracy.

In the data processing section, due to the limited number of samples, for samples with a relatively small proportion of missing values, this study adopted the K-nearest neighbors (KNN) interpolation method, followed by outlier handling using box plots. Both KNN and box plots are classical algorithms widely applied in numerous studies, known for their versatility and effectiveness. However, with the rapid development of deep learning, more complex algorithms for handling missing values and outliers have been proposed and successfully applied in various cases. In future studies, more advanced methods for handling missing values and outliers to maximize data accuracy and utility are recommended.

Based on the experimental results, the deviation of various element concentrations in the test samples from those in the training samples appears to be a major cause of classification errors. If a large number of misclassified samples are obtained, conducting a detailed analysis of the element concentrations in these error samples would provide more specific insights into the erroneous elements. Correcting such errors could lead to an improvement in experimental accuracy.

As a whole, although machine learning approaches are particularly useful, caution should be made when this is applied to geochemical problems, particularly on the selection of the appropriate machine learning methods. Information scientists and geochemists need to work together for an objective evaluation of data and a multi-disciplinary approach for successful results.

Data Availability Statement

The data used in this study were drawn from two public geochemical databases, GEOROC and PETDB. Figures were made with Matplotlib version 3.5.1 (Caswell et al., 2021; Hunter, 2007), available under the Matplotlib license at https://matplotlib.org/. Part of the software associated with this manuscript for the calculation and storage is licensed under MIT and published on GitHub https://github.com/MinkiGao/TectonicDiscrimination-.

References

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Caswell, T., Droettboom, M., Lee, A., Hunter, J., Firing, E., Stansby, D., et al. (2021). Matplotlib v3.5.1 [Software]. Zenodo. https://doi.org/10.5281/zenodo.5773480

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA.* https://doi.org/10.1145/2939672.2939785

Chen, Z., Wu, Q., Han, S., Zhang, J., Yang, P., & Liu, X. (2022). A study on geological structure prediction based on random forest method. *Artificial Intelligence in Geosciences*, *3*, 226-236.

Conder, J. A., Wiens, D. A., & Morris, J. (2002). On the decompression melting structure at volcanic arcs and back‐arc spreading centers. *Geophysical Research Letters*, *29*(15), 17-11-17-14.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Farmer, G. L. (2014). 4.3 - Continental Basaltic Rocks. In H. D. Holland & K. K. Turekian (Eds.), *Treatise on Geochemistry (Second Edition)* (pp. 75-110). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-08-095975-7.00303-X

Hawkesworth, C., & Gallagher, K. (1993). Mantle hotspots, plumes and regional tectonics as causes of intraplate magmatism. *Terra Nova*, *5*(6), 552-559.

Hofmann, A. W. (1997). Mantle geochemistry: the message from oceanic volcanism. *Nature*, *385*(6613), 219-229.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Ishizuka, O., Yuasa, M., Taylor, R. N., & Sakamoto, I. (2009). Two contrasting magmatic types coexist after the cessation of back-arc spreading. *Chemical Geology*, *266*(3-4), 274-296.

Kovalenko, V., Naumov, V., Girnis, A., Dorofeeva, V., & Yarmolyuk, V. (2007). Average compositions of magmas and mantle sources of smid-ocean ridges and intraplate oceanic and continental settings estimated from the data on melt inclusions and quenched glasses of basalts. *Petrology*, *15*, 335-368.

Labanieh, S., Chauvel, C., Germa, A., & Quidelleur, X. (2012). Martinique: a clear case for sediment melting and slab dehydration as a function of distance to the trench. *Journal of Petrology*, *53*(12), 2441-2464.

Li, C., Arndt, N. T., Tang, Q., & Ripley, E. M. (2015). Trace element in discrimination diagrams. *Lithos*, *232*, 76-83.

Liu, B., & Shi, J. (2022). A Machine Learning-Based Approach to Discriminating Basaltic Tectonic Settings. *International Journal of Computational Intelligence and Applications*, *21*(02), 2250012.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Nakamura, K. (2023). A practical approach for discriminating tectonic settings of basaltic rocks using machine learning. *Applied Computing and Geosciences*, *19*, 100132.

Niu, Y., Wilson, M., Humphreys, E. R., & O'Hara, M. J. (2011). The origin of intra-plate ocean island basalts (OIB): the lid effect and its geodynamic implications. *Journal of Petrology*, *52*(7-8), 1443-1468.

Pearce, J. A., & Cann, J. R. (1973). Tectonic setting of basic volcanic rocks determined using trace element analyses. *Earth and planetary science letters*, *19*(2), 290-300.

Pearce, J. A., & Norry, M. J. (1979). Petrogenetic implications of Ti, Zr, Y, and Nb variations in volcanic rocks. *Contributions to mineralogy and petrology*, *69*(1), 33-47.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Petrelli, M., & Perugini, D. (2016). Solving petrological problems through machine learning: the study case of tectonic discrimination using geochemical and isotopic data. *Contributions to mineralogy and petrology*, *171*, 1-15.

Schmidt, M. W., & Jagoutz, O. (2017). The global systematics of primitive arc melts. *Geochemistry, Geophysics, Geosystems*, *18*(8), 2817-2854.

Sobolev, A. V., Hofmann, A. W., Sobolev, S. V., & Nikogosian, I. K. (2005). An olivine-free mantle source of Hawaiian shield basalts. *Nature*, *434*(7033), 590-597.

Stern, R. J. (2002). Subduction zones. *Reviews of geophysics*, *40*(4), 3-1-3-38.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520-525. https://doi.org/10.1093/bioinformatics/17.6.520

Ueki, K., Hino, H., & Kuwatani, T. (2018). Geochemical discrimination and characteristics of magmatic tectonic settings: A machine‐learning‐based approach. *Geochemistry, Geophysics, Geosystems*, *19*(4), 1327-1347.

Wood, D. A. (1980). The application of a ThHfTa diagram to problems of tectonomagmatic classification and to establishing the nature of crustal contamination of basaltic lavas of the British Tertiary Volcanic Province. *Earth and planetary science letters*, *50*(1), 11-30.

Zhang, R., Cheng, Z., Zhang, Z., Chen, Z., Ernst, R., & Santosh, M. (2023). Formation of Tarim Large Igneous Province and

Strengthened Lithosphere Revealed Through Machine Learning. *Journal of Geophysical Research: Solid Earth*, *128*(1), e2022JB025772.

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. *Geophysical Research Letters*, *46*(10), 5234-5242.

Author Contributions:

Conceptualization: Zhaochong Zhang

Methodology: Xiaohui Ji, Mengqi Gao

Investigation: Mengqi Gao

Software: Mengqi Gao

Data curation: Mengqi Gao, Hengxu Li

Visualization: Mengqi Gao

Validation: Zhaochong Zhang, Xiaohui Ji and Hengxu Li

Funding acquisition: Xiaohui Ji, Zhaochong Zhang

Writing – original draft: Mengqi Gao, Zhaochong Zhang, Hengxu Li

Writing – review & editing: Zhaochong Zhang, Zhiguo Cheng, M. Santosh

Appendix

A The impact of imbalanced data sets

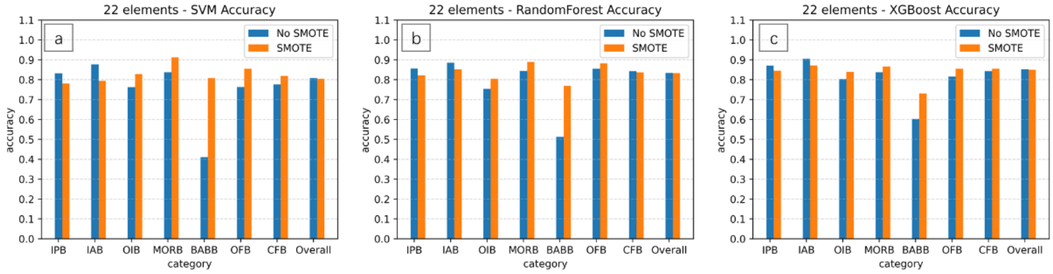


Figure A1. Accuracy before and after SMOTE - immobile elements

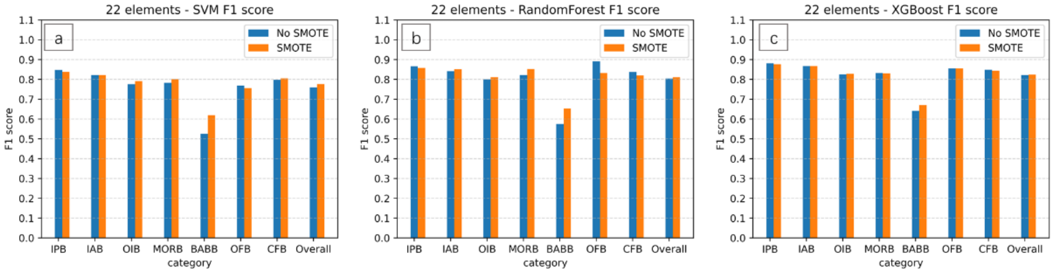


Figure A2. F1 score before and after SMOTE - immobile elements

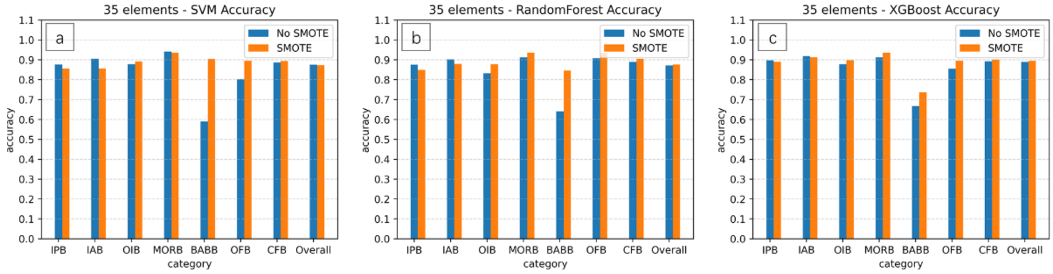


Figure A3. Accuracy before and after SMOTE - 35 elements

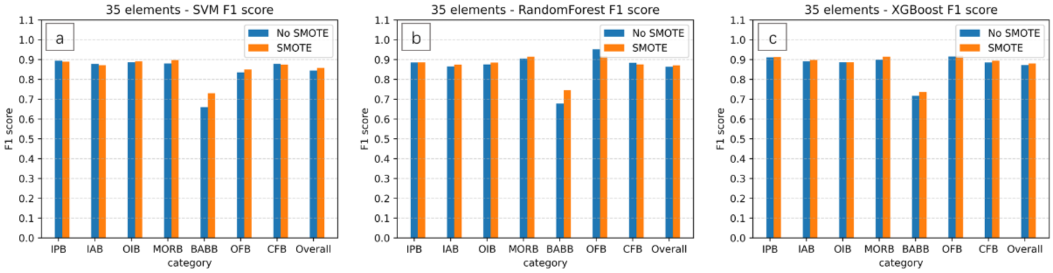


Figure A4. F1 score before and after SMOTE - 35 elements

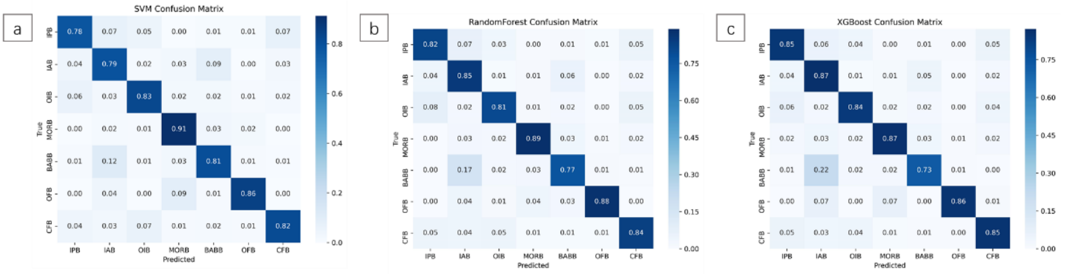


Figure A5. SMOTE-Confusion matrices for the three models based on immobile elements

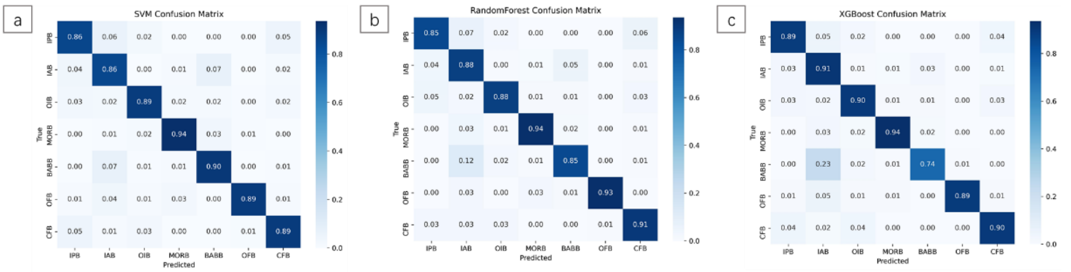


Figure A6. SMOTE-Confusion matrices for the three models based on 35 elements

B Analysis and Processing of Elemental Outliers in Basalt Using Box Plots

operation involves analyzing and addressing outliers using box plots. Specifically, this process begins by calculating the upper and lower quartiles (Q_3 and Q_1 , respectively) of the elemental content in basalt samples. Subsequently, the upper and lower boundaries are determined using the formula provided in Appendix-B1. Any data points beyond these boundaries are considered outliers and are removed directly.

$$\begin{aligned} down\_margin &= Q_1 - 1.5(Q_3 - Q_1) \\ up\_margin &= Q_3 + 1.5(Q_3 - Q_1) \end{aligned} \qquad \text{（Appendix-B1）}$$

The blue part in Figures B1 and B2 represents the main body of the box plot. The yellow line in the middle of the box represents the median of the data. The lower boundary of the box represents the lower quartile Q_1, and the upper boundary of the box represents the upper quartile Q_3. The bottom horizontal line represents the lower bound, and the top horizontal line represents the upper bound. Data points outside the upper bound are considered outliers.

As seen in Figure B1, the SiO_2 element has a relatively large proportion of normal values compared to the other eight elements. This is because, in the data processing process, to ensure that the selected samples are basalt samples, only samples with SiO_2 content in the range of 45% to 52% were chosen, which is equivalent to having already performed outlier processing. Therefore, the SiO_2 content of various types of basalt is relatively concentrated in the box plot. For the remaining eight elements, due to the presence of extreme outliers, the span of element content (vertical axis range) is large, and the region occupied by normal values is small. For example, the Ba content of IAB, some samples are close to 20,000, and some samples even exceed 350,000, while the content of most other samples is within 5,000. Due to the presence of outlier samples, the normal blue part is not displayed completely compared to SiO_2.

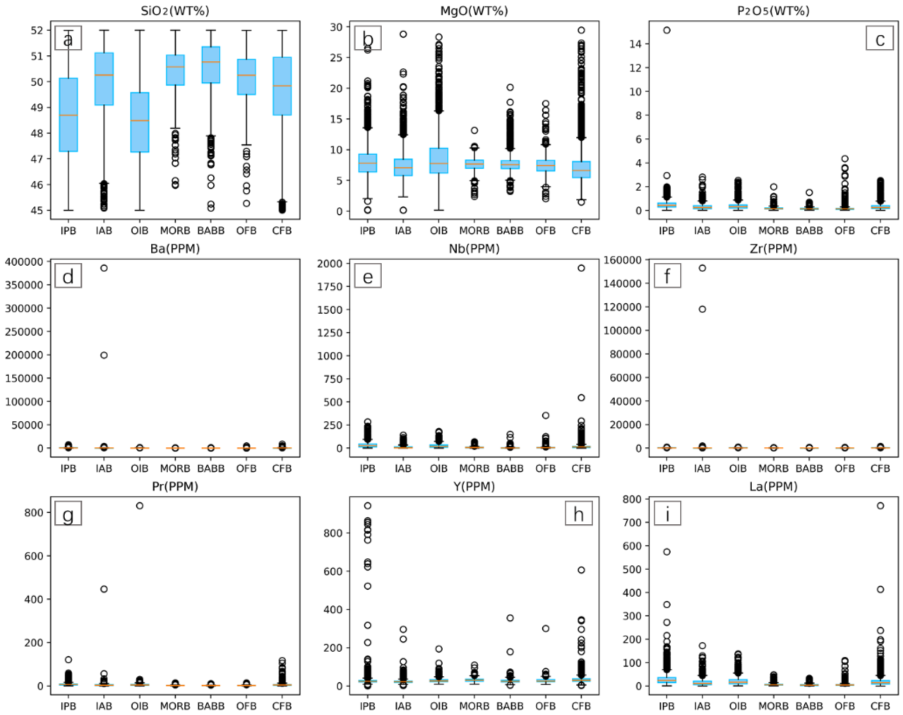


Figure B1. The distribution of element content before removing outliers.

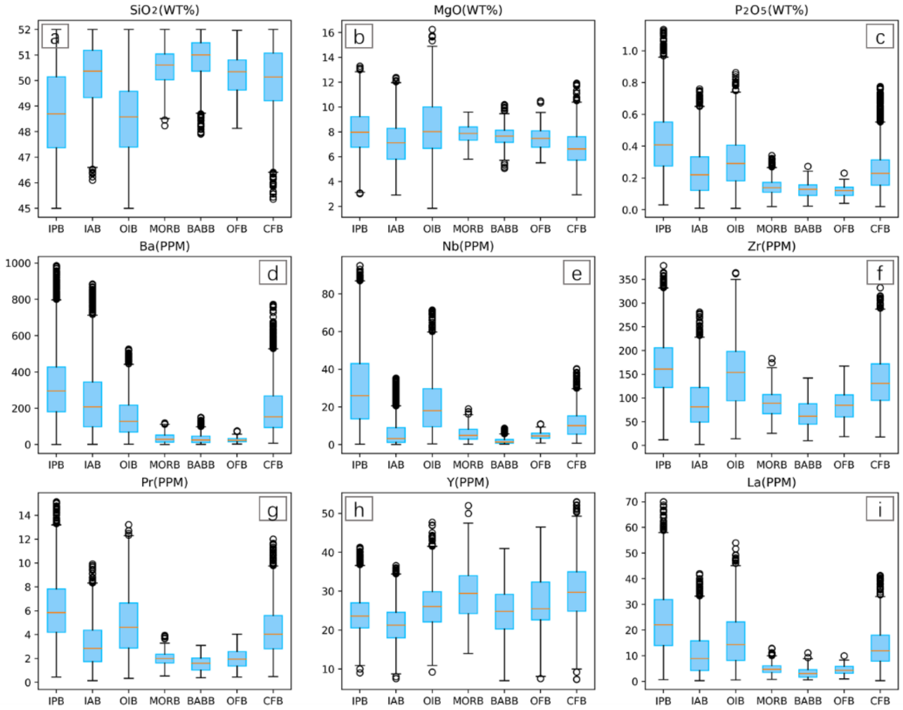


Figure B2. The distribution of element content after removing outliers.

In Figure B2, after removing extreme outliers, the box plot shows a more concentrated range of element content (vertical axis range). Normal data from various types of basalt occupy the main part of the plot, and the data is more centralized. Taking Ba content as an example, after removing outlier samples, the Ba content of the seven types of basalt does not exceed 1000.

C Error Sample Analysis

The distribution comparison between misclassified test samples and training samples for SVM, RF, and XGBoost is shown in Figure C1. Visualizing three elements randomly selected for each tectonic environment. As shown in the figure, misclassified samples in IPB, IAB, OIB, MORB, and OFB exhibit deviations in the values of two or more elements from the distribution of training samples. In BABB, misclassified samples show a similar trend in the distribution of Ce and Gd elements compared to training samples, but Ho element deviates noticeably. The distribution of the three elements in CFB is roughly consistent. Therefore, the reason for misclassification is that the elemental content of the volcanic rocks exceeds the numerical range learned by the model, deviating from the latent patterns and rules the model has learned.

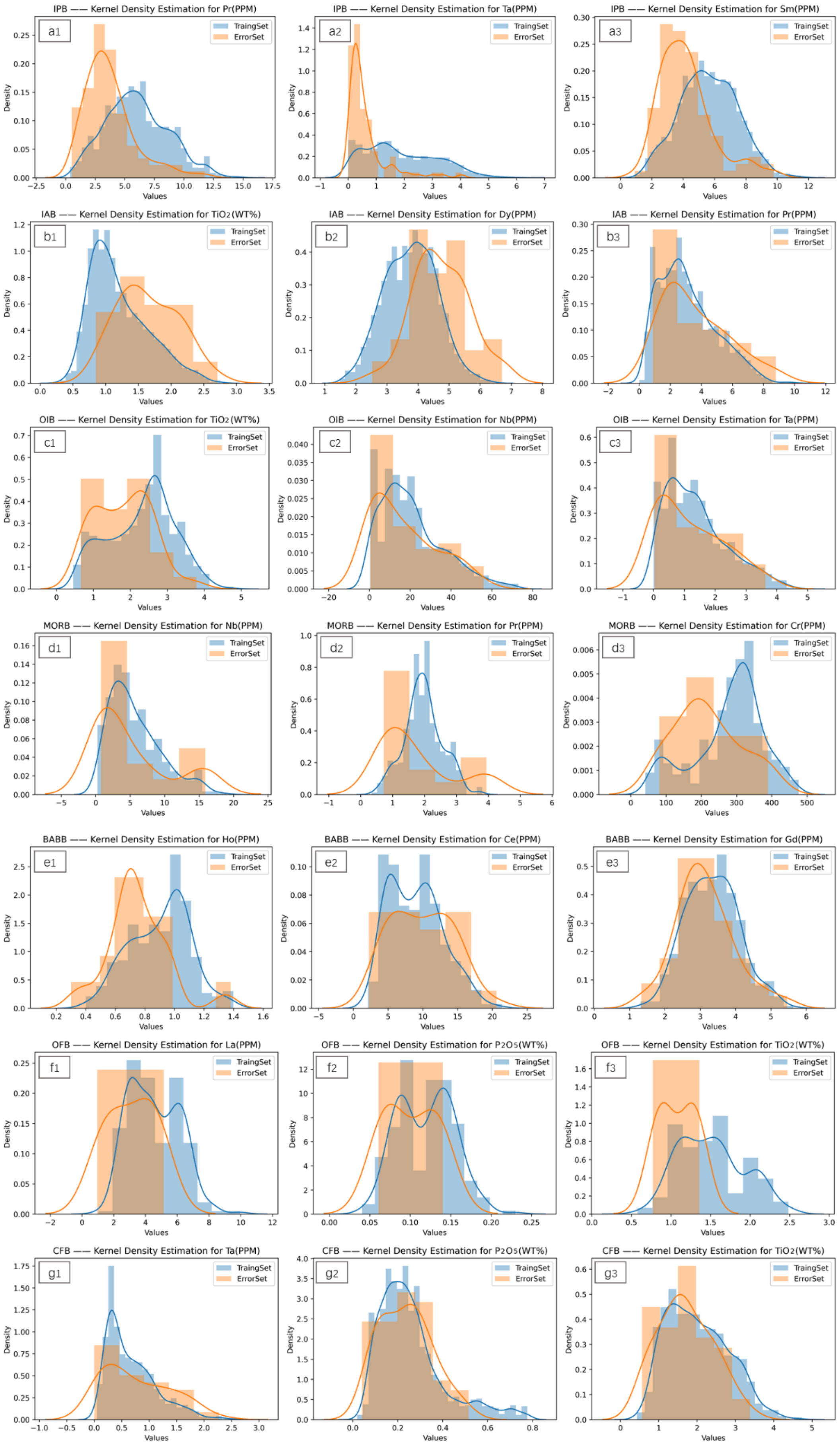

IPB —— Kernel Density Estimation for Pr(PPM)
a1
IPB —— Kernel Density Estimation for Ta(PPM)
a2
IPB —— Kernel Density Estimation for Sm(PPM)
a3
IAB —— Kernel Density Estimation for TiO2(WT%)
b1
IAB —— Kernel Density Estimation for Dy(PPM)
b2
IAB —— Kernel Density Estimation for Pr(PPM)
b3
OIB —— Kernel Density Estimation for TiO2(WT%)
c1
OIB —— Kernel Density Estimation for Nb(PPM)
c2
OIB —— Kernel Density Estimation for Ta(PPM)
c3
MORB —— Kernel Density Estimation for Nb(PPM)
d1
MORB —— Kernel Density Estimation for Pr(PPM)
d2
MORB —— Kernel Density Estimation for Cr(PPM)
d3
BABB —— Kernel Density Estimation for Ho(PPM)
e1
BABB —— Kernel Density Estimation for Ce(PPM)
e2
BABB —— Kernel Density Estimation for Gd(PPM)
e3
OFB —— Kernel Density Estimation for La(PPM)
f1
OFB —— Kernel Density Estimation for P2O5(WT%)
f2
OFB —— Kernel Density Estimation for TiO2(WT%)
f3
CFB —— Kernel Density Estimation for Ta(PPM)
g1
CFB —— Kernel Density Estimation for P2O5(WT%)
g2
CFB —— Kernel Density Estimation for TiO2(WT%)
g3
TraingSet
ErrorSet
Density
Values

Figure C1. Comparison between misclassified test samples and training samples