Earth Science Informatics
https://doi.org/10.1007/s12145-024-01385-5

RESEARCH

A new approach to dividing the tectonic setting of igneous rocks: machine learning and GeoTectAI software

Ming Lei[1] · Wenyan Cai[1] · Xiao Liu[1] · Chao Zhang[1] · Qingyi Cui[1] · Jian Li[1]

Received: 21 March 2024 / Accepted: 18 June 2024

Abstract

For a long time, elucidating the tectonic setting of unknown rock samples has been a focal point for geologists. Traditional methodologies for this purpose have been scrutinized increasingly due to their inherent limitations. In response to these challenges, this paper applies modern machine learning techniques to analyze the geochemical data of igneous rocks and improve understanding of tectonic settings. By employing a variety of machine learning models, including Decision Trees, K-Nearest Neighbors, Support Vector Machines, Random Forests, Extreme Gradient Boosting, and Artificial Neural Networks, and training with 23 features comprising nine major elements (SiO_2, TiO_2, Al_2O_3, CaO, MgO, MnO, Na_2O, K_2O, and P_2O_5) along with 14 trace elements (La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, and Lu), the study successfully distinguished between seven different tectonic settings. Among these models, Random Forest, Extreme Gradient Boosting, and Artificial Neural Networks demonstrated superior classification accuracy and recall rates, with accuracies of 0.85, 0.87, and 0.86, respectively. This validates the effectiveness and potential of machine learning technologies in distinguishing the tectonic settings of igneous rocks through their geochemical elements. To enable geologists and researchers to more accurately understand and predict the origins of igneous rocks without the need to master machine learning knowledge, a user-friendly software, GeoTectAI, has been developed.

Keywords Machine learning · Igneous rocks · Geochemistry · Tectonic settings

Introduction

Igneous rocks, one of the most common rock types in the Earth's crust, are formed in a variety of tectonic settings, with their geochemical characteristics reflecting the environmental conditions during their formation and evolution (Whalen 1985; Whalen et al. 1987; Chappell and White 1992; Wu et al. 2007). However, traditional methods for discerning these conditions often rely on empirical judgment and simple chemical calculations (e.g., Pearce 2008; Pearce et al. 1984). Furthermore, the inherent ambiguity associated with geochemical data poses a formidable challenge in precisely delineating the genuine genesis of these rocks within intricate geological settings. In traditional geochemical research, utilizing Pearce diagrams and discerning the tectonic source regions of rocks have been prevalent methods (Pearce et al. 1984; Butler and Woronow 1986; Verma et al. 2006; Saccani 2015). However, with the accumulation of geochemical data, it has been discovered that common discriminant diagrams do not always apply to new datasets (Verma 2010; Li et al. 2015; Vermeesch 2006a;

Communicated by Hassan Babaie.

✉ Wenyan Cai
562443758@qq.com

Ming Lei
0328leiming@gmail.com

Xiao Liu
liuxiaogis@163.com

Chao Zhang
czhang@sdut.edu.cn

Qingyi Cui
15265520559@163.com

Jian Li
jianli@sdut.edu.cn

[1] School of Resources and Environmental Engineering, Shandong University of Technology, 266 Xincunxi Road, Zibo 255049, China

Published online: 28 June 2024

Armstrong-Altrin and Verma 2005). Igneous rocks formed under different tectonic settings exhibit unique geochemical signatures, primarily reflected in compositional differences (e.g., Whalen et al. 1987; Eby 1990, 1992; Richards and Kerrich 2007). Understanding these chemical compositions is crucial for accurately identifying the tectonic settings, as the concentrations of major elements, trace elements, and isotopic compositions can be obtained through whole-rock geochemical analysis. Hence, distinguishing the tectonic settings of igneous rocks through their geochemical element composition is feasible.

With the rapid development of artificial intelligence technology, machine learning, as an efficient method of data analysis (Bishop 2006; Jordan and Mitchell 2015; Chen et al. 2020), has become increasingly widespread in the application within Earth sciences. For instance, Snow (2006) explored decision trees as novel approaches, diverging from conventional geochemical discriminant methodologies. Han et al. (2019) delved into a range of machine learning techniques to ascertain the tectonic origins of spinels. Furthermore, Nakamura (2023) conducted experiments with machine learning strategies to classify the tectonic settings of basalts. These studies have demonstrated the feasibility of using machine learning for tectonic setting discrimination.

With the development of big data and machine learning, along with establishing shared databases (Lehnert et al. 2000; Gard et al. 2019), big data analysis now has reliable data support and convenient data access. Alongside improving computing power, the big data analysis and machine learning use have become widely applied (Jordan and Mitchell 2015; Ren et al. 2019; Yaqoob et al. 2016). Therefore, turning the judgment of tectonic settings into a multi-classification problem for machine learning and using supervised learning (Kotsiantis et al. 2007; Hastie et al. 2009) to label the data for training sets has been shown by previous research to be a powerful tool for construction discrimination (Petrelli and Perugini 2016; Doucet et al. 2022; Takaew et al. 2024). Despite this, the application of machine learning still demands that researchers have a certain level of computer science and artificial intelligence foundation, which undoubtedly serves as a limiting factor. Hence, there is an urgent need to introduce a new, convenient, and user-friendly solution to lower the barriers to entry.

This study aims to explore a new approach that utilizes multiple machine learning techniques (including Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, Extreme Gradient Boosting, and Artificial Neural Network) to analyze whole-rock geochemical data of igneous rocks and determine their tectonic settings. Through this approach, we hope to understand igneous rocks' origins and evolutionary history more accurately and systematically. After performing feature engineering on the data collected from the Petrological Database of the Ocean Floor (PetDB) and Geochemistry of Rocks of the Oceans and Continents (GEOROC) Database, the data is trained using machine learning models, followed by analysis, explanation, and comparison of the models. To facilitate the use of the model, corresponding user-friendly software was developed. Finally, the potential and issues of using geochemical elements for tectonic setting judgment are outlined (Fig. 1).

Dataset

The igneous rock geochemical dataset discussed in this paper is derived from two major geochemical databases: PetDB (http://www.earthchem.org/petdb) and GEOROC

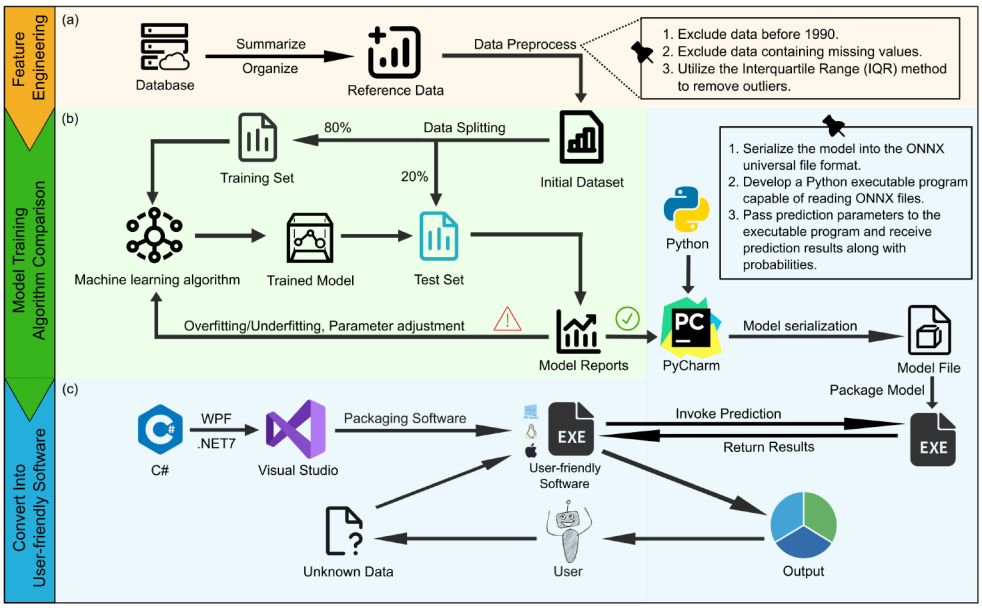

Fig. 1 Workflow for developing machine learning models and software. **a** Data acquisition and preprocessing. **b** Model training and performance analysis. **c** Principles of user-friendly software development

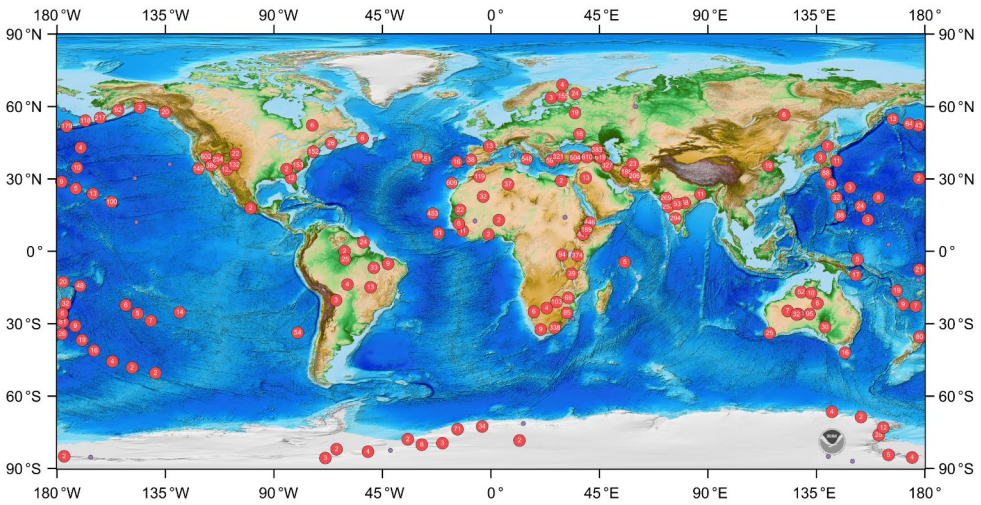

Fig. 2 The sample locations selected for this study. The numeric value of the point in the figure represents how many samples are in the vicinity of that location. The range of colors on the map indicates elevation, spanning from blue to brown. The deeper the color, the higher the elevation of the region

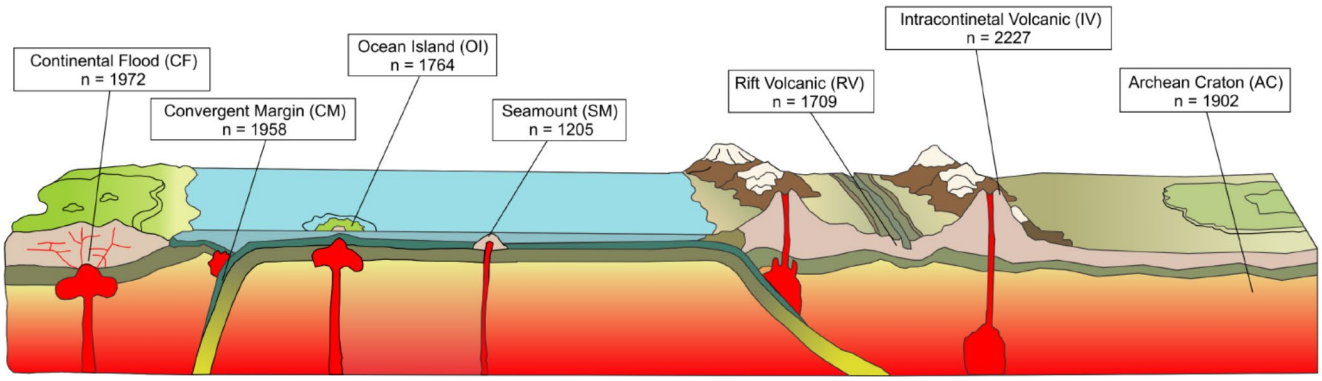

Fig. 3 Schematic diagram of the tectonic setting (after Takaew et al. 2024)

(http://georoc.mpch-mainz.gwdg.de/georoc/). We selected igneous rock samples that have complete major element data (SiO_2, TiO_2, Al_2O_3, CaO, MgO, MnO, Na_2O, K_2O, P_2O_5) and specified rare earth trace elements (La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu), totaling 23 data attributes. The locations of the selected samples are shown in Fig. 2.

The dataset we have collected and compiled comprises a total of 78,867 entries. In pursuit of a more precise analysis of the relationship between tectonics and geochemistry, aiming to minimize factors of interference and uncertainty, we have excluded entries containing null values and those dated prior to 1990. Consequently, the dataset for training the machine learning model consists of 12,737 entries. Within this refined dataset, we have determined seven categories of tectonic settings based on the known tectonic settings within the database. These categories are Archean Cratons (AC) ($n = 1{,}902$), Continental Flood (CF) ($n = 1{,}972$), Convergent Margins (CM) ($n = 1{,}958$), Intracontinental Volcanics (IV) ($n = 2{,}227$), Ocean Islands (OI) ($n = 1{,}764$), Rift Volcanics (RV) ($n = 1{,}709$), and Seamounts (SM) ($n = 1{,}205$). These classifications and their respective data counts are detailed in Fig. 3.

Methods

Feature engineering

It is imperative to filter the existing dataset to uncover the hidden relationships within petrogeochemistry and construct a classifier capable of distinguishing tectonic settings (Jo 2019). Initially, we eliminated entries of magmatic rocks where geochemical elements were either absent or below the detection limit, as these instances could adversely affect the discrimination process. Some machine learning algorithms are particularly sensitive to outliers, which may arise from measurement errors, operational mistakes, or geological factors. To eliminate outliers caused by non-geological errors as much as possible, and to enhance the model's generalizability, we employ the Interquartile Range (IQR) method to exclude outliers (Barbato et al. 2011).

IQR calculation excludes outliers

In machine learning, the IQR is a statistical method used to identify and eliminate outliers within the data. The dataset is initially arranged in ascending order, and the first quartile (Q_1) and the third quartile (Q_3) are calculated. The IQR is the difference between Q_3 and Q_1 (i.e., IQR = Q_3 - Q_1), measuring the data's dispersion. The lower bound is defined as $Q_1 - 1.5 \times$ IQR, and the upper bound is $Q_3 + 1.5 \times$ IQR. These bounds establish the "normal" range for the data. Values falling below the lower bound or above the upper bound are considered outliers and are excluded (Fig. 4).

After preprocessing outliers (Fig. 5), the total number of data entries participating in subsequent training is 9,039. Among these, AC has 1,060, CF has 1,572, CM has 1,689, IV has 1,975, OI has 944, RV has 943, and SM has 856. The dataset is then divided into a training set and a test set at a ratio of 80% and 20%, resulting in 7,231 entries for the training set and 1,808 for the test set.

Feature normalization

Feature normalization is essential because geochemical data come in various ranges and units, necessitating their scaling to a common standard before utilization. Normalization is achieved by scaling or transforming features to a common range, ensuring each feature contributes equally when fed into the classifier (Singh and Singh 2020). Scaling methods resize data from different ranges to a predetermined range, thus preserving the original distribution of the data. Prior to further processing, data must be normalized using the Z-score normalization method:

$$Z = (X - \mu)/\theta$$

where Z is the Z-score result, X is the original data point, μ is the mean, and θ is the standard deviation.

After normalization, the new dataset has a mean of 0 and a standard deviation of 1. As shown in Fig. 6a, the normalized data now approximately follows a normal distribution. Normalization eliminates the influence of magnitudes on the discrimination results, which will be more beneficial for subsequent machine learning training.

Classifiers

In this study, six advanced machine learning algorithms have been adopted to perform complex multi-classification tasks (Fig. 7). The specific methods include Decision Tree (DT), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN). These algorithms are all based on the principles of supervised learning, utilizing seven types of tectonic settings as the labels for features and geochemical elements as the training variables within the target dataset.

Decision tree

Decision trees (DT) (Fig. 7a) are widely utilized algorithms in the field of machine learning (Quinlan 1986; Hastie et al. 2009) and apply both classification and regression problems. This algorithm constructs a tree-shaped model by learning the features within the dataset. In this model, each internal node represents a test on a particular feature, each branch corresponds to one of the possible outcomes of this test, and each leaf node represents a predicted category or numerical value. When employing a DT for prediction, the process begins at the root node and progressively traverses downwards based on the outcomes of the feature tests until reaching a leaf node, thereby deriving the final prediction decision (Suthaharan 2016a). Within the geological sciences, the utilization of DT has been broadly acknowledged and implemented, highlighted by research contributions like Petrelli and Perugini (2016) and Takaew et al. (2024). Specifically, Vermeesch (2006b) has harnessed the power of DT for the tectonic discrimination of basalts. Han et al. (2019)

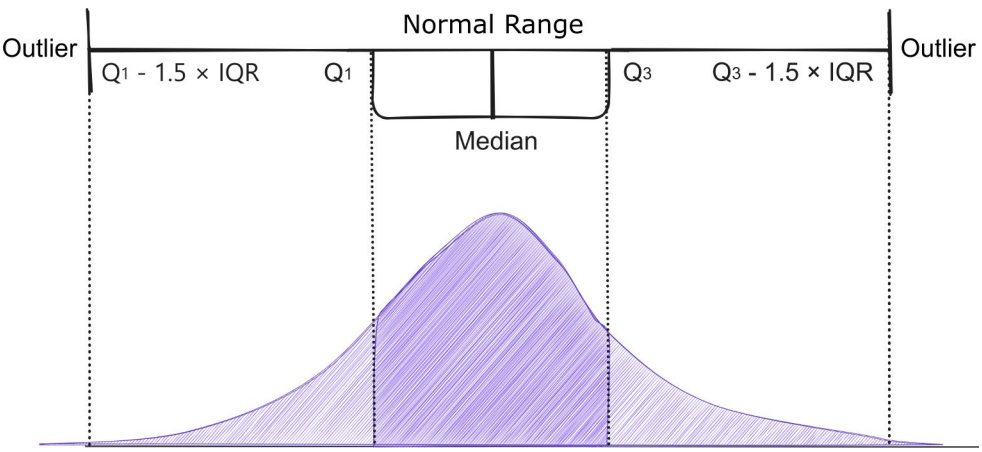

Fig. 4 The range of outliers and normal values in the IQR algorithm

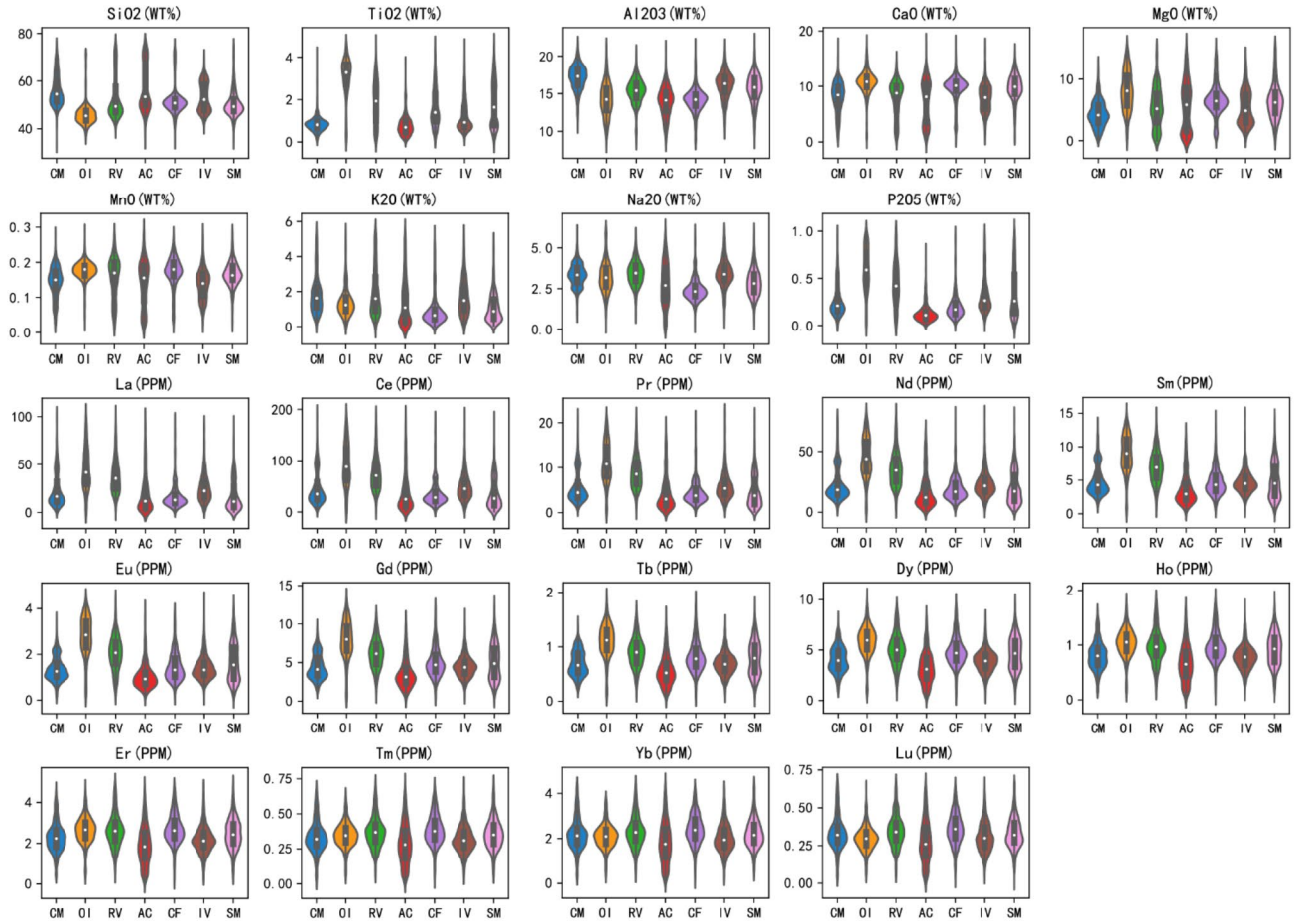

Fig. 5 After removing outliers, the feature distribution of the dataset used for training the model. The vertical axis shows the range of attribute values, while the horizontal axis represents different tectonic settings

have adeptly applied decision trees to distinguish tectonic settings through the analysis of spinel.

K-nearest neighbors

K-Nearest Neighbors (KNN) (Fig. 7b) is a straightforward yet effective machine learning algorithm primarily used for classification tasks (Cover and Hart 1967). It is grounded in the core principle that similar data points are likely to belong to the same category. In KNN, "K" represents the number of neighboring data points the algorithm considers while making a prediction. When classifying a new data point, the algorithm identifies the "K" closest data points in the training dataset to this new point. Subsequently, it classifies the new point into the most frequent category among these neighboring points (Mucherino et al. 2009). In geology, KNN algorithms has been applied to many applications. Notably, Potratz et al. (2021) leveraged KNN algorithms for classifying lithofacies, demonstrating their utility in categorizing geological formations based on their physical characteristics. Bicego et al. (2023) used KNN to decipher volcanic activity signals, showcasing the algorithm's adeptness in interpreting complex geological phenomena. Additionally, Zhang et al. (2019) utilized KNN to study the diorite under different tectonic settings.

Support vector machine

Support Vector Machine (SVM) (Fig. 7c) is an exceedingly effective model within the domain of supervised learning, particularly adept at handling classification problems involving high-dimensional data (Cortes and Vapnik 1995). The primary goal of SVM is to identify an optimal hyperplane that can segregate the data points of different categories with the largest margin possible. In two-dimensional space, this hyperplane manifests as a line that separates categories, while in higher-dimensional spaces, it forms a multidimensional surface that partitions different categories. A notable feature of SVM is its emphasis on maximizing the margin distance between data points and the decision boundary.

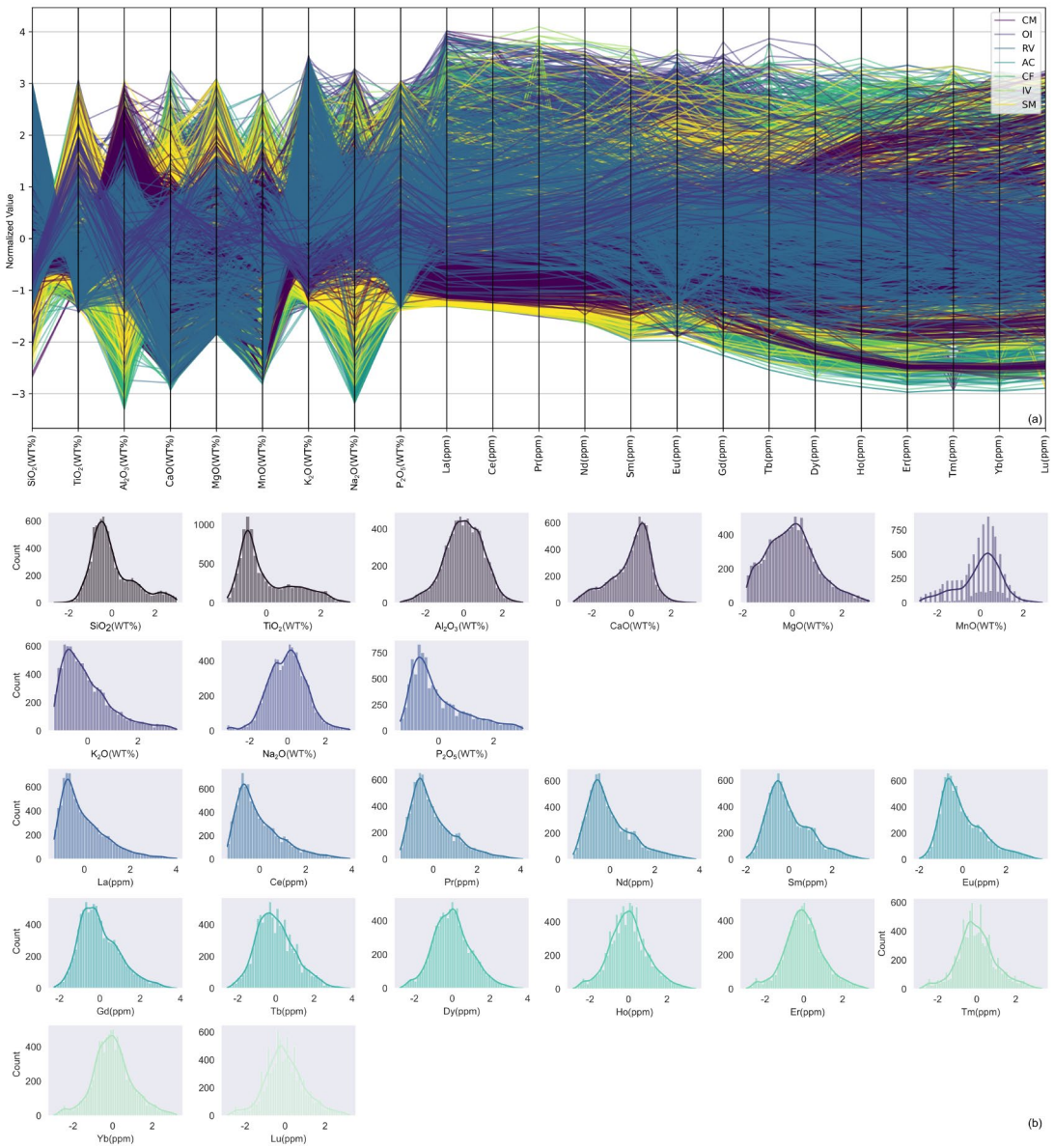

Fig. 6 **a** Parallel coordinates plot. Each line represents a data entry, with different colored lines indicating different tectonic settings. **b** Normalized bar chart of different attribute data

Moreover, by employing kernel function techniques, SVM can process data that is not linearly separable in its original space by mapping the data to a higher-dimensional space, thereby achieving separability in this new space (Patle and Chouhan 2013; Smola and Schölkopf 2004; Suthaharan 2016b).

Random forest

Random Forest (RF) (Fig. 7d) is an efficient ensemble learning algorithm that enhances the accuracy and robustness of predictions by integrating the outcomes of multiple decision trees (Ho 1995; Breiman 2001). Each decision tree is trained independently in this approach, employing a "bagging" (bootstrap aggregating) strategy, which involves randomly selecting samples from the original dataset for training. The RF also introduces randomness selecting features for splitting nodes, enriching the model's diversity and effectively reducing the likelihood of overfitting. This algorithm performs well in classification and regression tasks, particularly in analyzing high-dimensional data. It exhibits strong robustness against noise in the training data and provides more stable and reliable predictions than a single decision tree. RF has been applied in various fields as a model

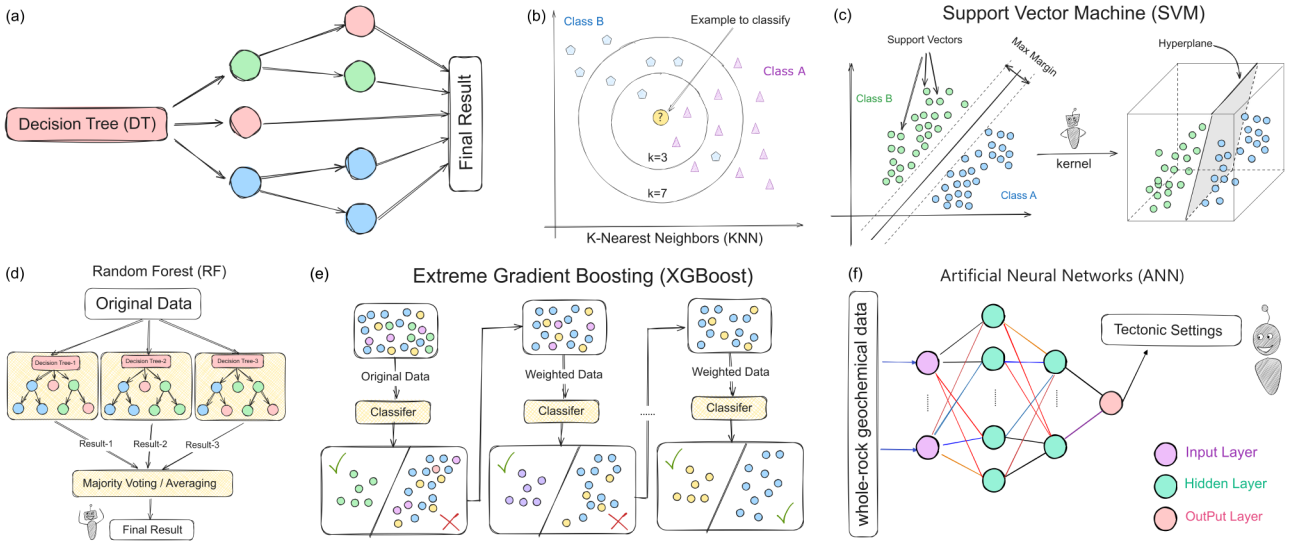

Fig. 7 The corresponding principles of different machine learning algorithms

with commendable comprehensive performance. Ueki et al. (2018) applied RF to geochemically discriminate and characterize magmatic tectonic settings, demonstrating the technique's effectiveness in geological categorization. Similarly, Nakamura (2023) utilized RF for the practical discrimination of tectonic settings of basaltic rocks, further evidencing the method's widespread applicability in geosciences.

Extreme gradient boosting

Extreme Gradient Boosting (XGBoost) (Fig. 7e) is a highly efficient ensemble learning algorithm widely used in classification and regression tasks (Chen and Guestrin 2016). It is based on Gradient Boosted Decision Trees (GBDT) principles and continuously improves model performance by sequentially adding new trees. A distinctive feature of XGBoost is its construction of each tree using a gradient boosting method, where adding each new tree is aimed at reducing the prediction error left by the previous tree. This algorithm excels in handling large datasets and offers a variety of parameters for adjustment to optimize model performance. XGBoost often outperforms traditional gradient boosting methods regarding computational speed and prediction accuracy. It also has capabilities for handling missing data and regularization, which help mitigate the risk of overfitting and enhance the model's generalization ability. XGBoost has found broad applications across multiple domains. Specifically, Wang et al. (2023) applied XGBoost for the determination of tectonic settings through the analysis of trace elements in zircon. Saha et al. (2021) utilized machine learning to discriminate the igneous rocks' tectonic setting based on biotite's major element chemistry, showcasing the algorithm's precision in geological classification.

Artificial neural network

Artificial Neural Networks (ANN) (Fig. 7f) are computational models that simulate the working mechanism of human brain neurons and play a significant role in the applications of machine learning and artificial intelligence (Yegnanarayana 2009). ANNs consist of numerous interconnected nodes (i.e., neurons) distributed across several layers, typically including an input layer, multiple hidden layers, and an output layer. Each neuron is responsible for receiving input signals, processing these signals through a specific activation function, and then transmitting the results to neurons in the next layer. ANNs learn by adjusting the weights of the connections between neurons in the network, a process generally reliant on the backpropagation algorithm and gradient descent method (Zurada 1992). It is noteworthy that Ge et al. (2021) utilized convolutional neural networks (CNN) to identify tectonic settings, successfully recognizing 12 types and achieving commendable results.

Results

This study used precision, recall, and the F1 score as evaluation metrics for machine learning models (Fig. 8a). Precision refers to the proportion of true positive samples among those predicted as positive by the model. Recall reflects the proportion of true positive samples correctly predicted by the represents the harmonic mean between precision and recall providing a comprehensive measure for assessing model performance.

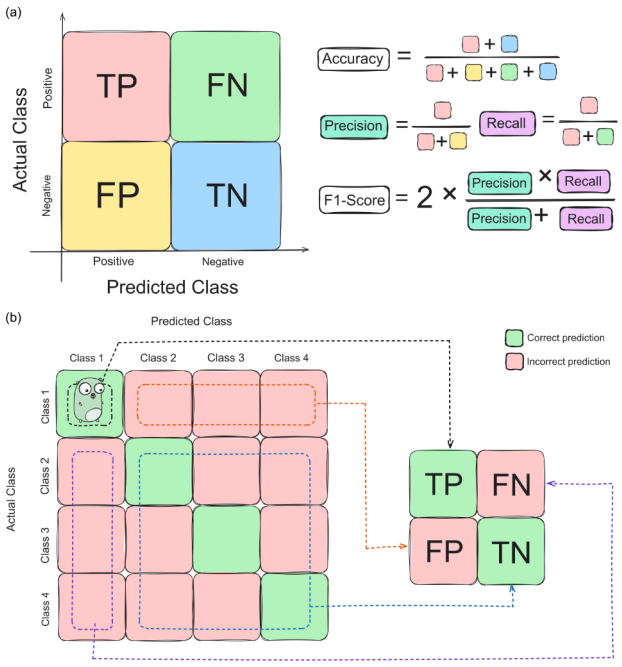

Fig. 8 Explanation of the confusion matrix principle and various evaluation indicators (after Zhong et al. 2023)

The overall classification results of the machine learning models in this study are visualized using confusion matrices (Fig. 8b). This study presents performance data for various machine learning models concerning tectonic settings, as illustrated in Fig. 9 and the attached table. The comprehensive accuracy comparison (Fig. 10a) shows that, all algorithms achieved an accuracy rate above 80% except for DT, which performed poorly. Among them, XGBoost, RF, and ANN exhibited the best performance, achieving accuracy rates of 87%, 85%, and 86%, respectively, with XGBoost demonstrating the highest performance across all algorithms. Regarding F1 scores, XGBoost, RF, and ANN also achieved the best results among all algorithms.

When analyzing the different tectonic settings individually, the DT algorithm shows the worst performance among all tectonic settings (Fig. 10b). XGBoost, ANN, RF, and SVM show good discriminative ability on AC (Fig. 10c), with scores of 0.83, 0.82, 0.81, and 0.81, respectively. In contrast, KNN is slightly worse, with a score of 0.75. On the contrary, the DT algorithm is the least effective scoring 0.68. According to the CF discrimination results (Fig. 10c), XGBoost, RF, ANN, KNN, and SVM had similar results, with scores of 0.91, 0.89, 0.88, 0.87, and 0.86. DT algorithm improved slightly in discriminating CF, reaching a score of 0.78. Various algorithms generally exhibit relatively good improvements in CF discrimination compared to AC discrimination. Similarly, the discrimination results of CM

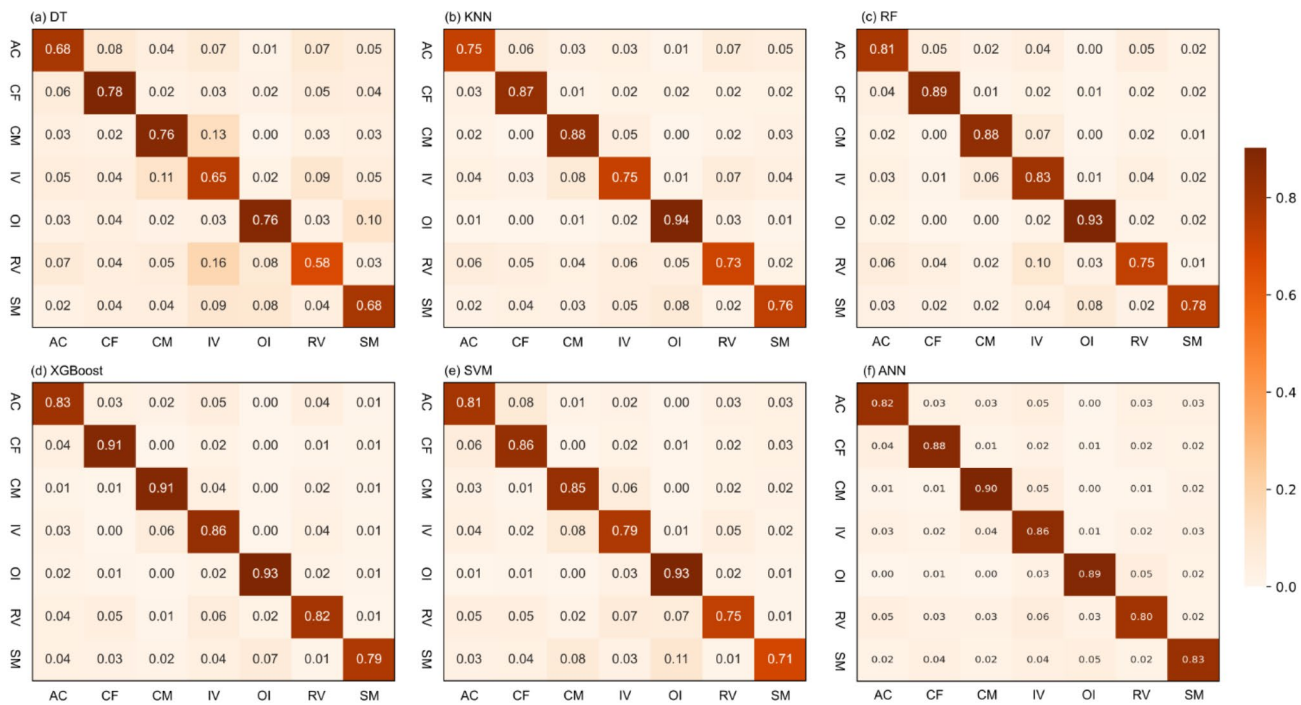

Fig. 9 Confusion matrices of different machine learning algorithms under different classes. Each small square represents the relationship between a specific true category (vertical axis) and the category predicted by the model (horizontal axis)

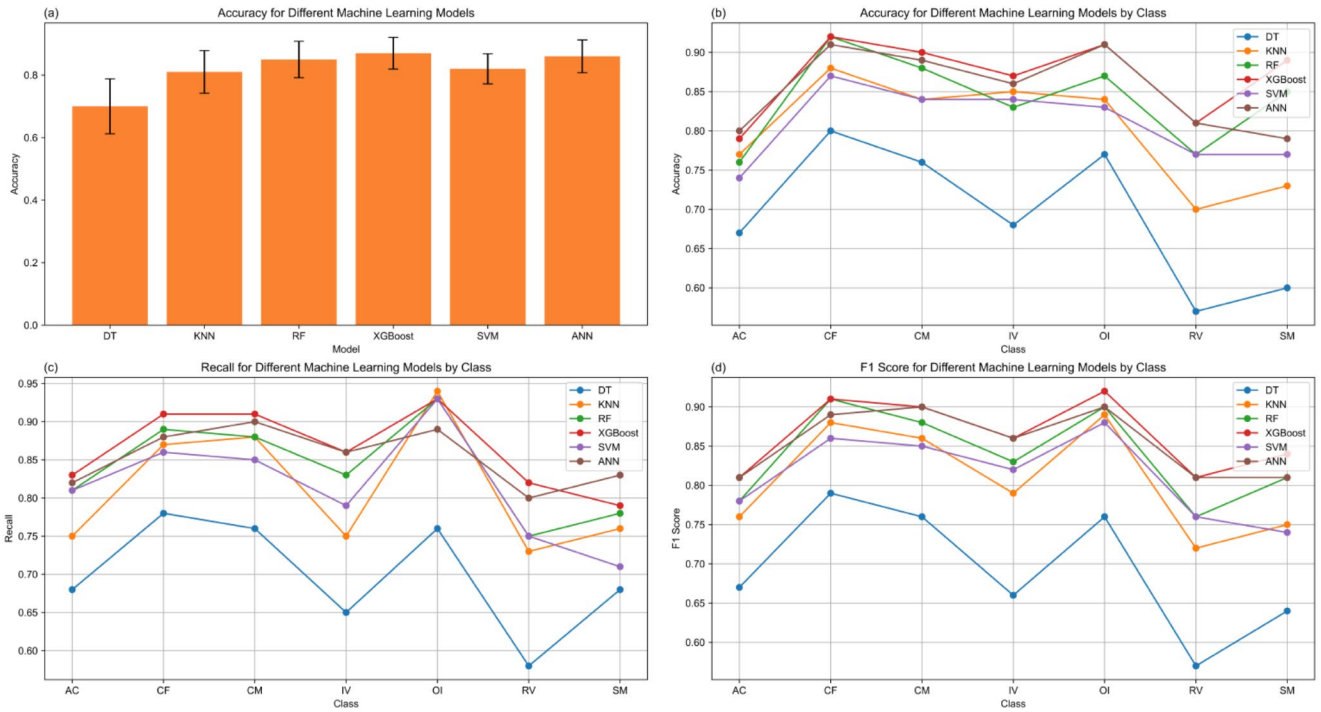

Fig. 10 Performance comparison of algorithms: **a** Accuracy across tectonic settings, **b** F1 scores across tectonic settings, **c** Recall rates across tectonic settings, **d** Overall accuracy of different algorithms

are similar to those of CF. XGBoost, ANN, RF, KNN, and SVM also show similar effectiveness (Fig. 10c), with scores of 0.91, 0.90, 0.88, 0.88, and 0.85, respectively. Compared to CF discrimination, the DT algorithm's score slightly decreases to 0.76.

Looking at the IV's discrimination results, all algorithms show a decrease in their discriminative ability, and the KNN and DT algorithms show a particularly significant decrease (Fig. 10c), reaching scores of 0.75 and 0.65. Compared with the previous tectonic settings discrimination performance, XGBoost, ANN, and RF, although slightly decreased, maintain relatively high discrimination ability, respectively: 0.86, 0.86, and 0.83. On the contrary, in the case of using SVM for IV discrimination, the score fell below 0.79 to 0.80. In contrast, regarding the discrimination performance for OI, all algorithms demonstrated a significant improvement in their discriminative abilities, with XGBoost, KNN, SVM, and RF showing particularly close performance (Fig. 10c), scoring respectively 0.93, 0.94, 0.93, and 0.93. Following closely behind was ANN, with a score of 0.89. Although DT showed improvement in discriminating OI, its score was only 0.76.

The discrimination performance for RV indicates a significant decrease in the discriminative abilities of all algorithms (Fig. 10c). Despite this, XGBoost and ANN maintained relatively high scores, with 0.82 and 0.80. RF, SVM, and KNN scored closely to each other, with scores of 0.75, 0.75, and 0.73, respectively. The DT algorithm exhibited the poorest performance in discriminating RV, with a score of only 0.58. In contrast, regarding the discriminative performance for SM, the performances of various algorithms are mixed (Fig. 10c). ANN exhibits the best performance with a score of 0.83, while the scores of the remaining algorithms all fall below 0.80. Specifically, the XGBoost, RF, KNN, SVM, and DT scores are 0.79, 0.78, 0.76, 0.71, and 0.68, respectively.

Discussion

Different machine learning methods

The DT algorithm performs worst among the six machine learning classifiers. This could be partly due to data imbalance and partly because DT algorithms have inherent weaknesses compared to ensemble learning algorithms. Although the effectiveness of the SVM algorithm in geology-related fields has been proven and documented in previous studies (Melgani and Bruzzone 2004; Ueki et al. 2018;), its performance is moderate when distinguishing among multi-class tectonic settings. The performance of the KNN algorithm shows that it has good fitting and discriminant ability for small sample petrological data consistent with the previous research results (Nakamura 2023). XGBoost, RF and ANN exhibit the most robust overall performance, highlighting the advantages of ensemble learning in handling large-scale,

high-dimensional data. Furthermore, ANNs use multiple layers of neurons to nonlinearly explore relationships in big data, making them uniquely discriminative in certain feature discriminations (Fig. 10c). Therefore, the simple KNN algorithm can perform relatively well with small-sample data, while nonlinear models like RF, XGBoost, and ANN yield better results with large-scale data.

Feature discrimination and selection

XGBoost, RF, and ANN can understand their feature selection and importance in different tectonic settings by calculating SHAP values and plotting SHAP summary diagrams (Lundberg and Lee 2017) (Fig. 11). Features are ranked by the sum of SHAP values across all samples in the test set (except for ANN, which uses 50% of test samples for SHAP value calculation). The vertical order from top to bottom indicates the importance of the feature in the model. Red indicates high feature values, while blue indicates low values. The horizontal axis represents the impact of feature values on the output (i.e., the SHAP value). A SHAP value above 0 indicates a positive impact of the feature, and vice versa.

Overall, RF and XGBoost algorithms prioritize major elements as the most essential features for discrimination in most tectonic settings. However, ANN demonstrates a broader sensitivity to both major and trace elements, indicating their capacity to discern subtle relationships missed by more traditional machine learning approaches. This nuanced understanding is particularly evident in the discrimination of RV and SM features, where trace elements play a more significant role. Agrawal et al. (2008) and Li et al. (2015) have confirmed this perspective, emphasizing the application of trace elements in differentiating tectonic settings. Such insights reveal a clear link between trace elements and tectonic settings, underscoring the advanced analytical capabilities of ANNs. Through their multi-layered structure, these networks uncover complex interrelations that RF and XGBoost might overlook, as illustrated in Fig. 10c, showcasing ANN's superior ability to leverage both major and trace elements for a more comprehensive feature importance analysis.

In evaluating tectonic settings, feature selection is crucial. XGBoost and RF algorithms highlighted SiO_2, TiO_2, Al_2O_3, CaO, and P_2O_5 as the top five features in overall importance. In contrast, although ANN prioritizes SiO_2, TiO_2, Al_2O_3, CaO, and Nd as its overall top features, it relies more on trace elements when analyzing specific tectonic settings. The integrated use of a wider array of geochemical elements may be crucial for surpassing the limitations of traditional charts in the discrimination of tectonic settings.

The potential and problems of machine learning in tectonic setting discrimination

Our findings indicate that under large-scale, high-dimensional geochemical data, the performance of ensemble learning and artificial neural networks surpasses that of conventional machine learning methods. These advanced models offer a powerful means for identifying known tectonic settings or assessing unknown or uncertain ones. The relationship between geochemical elements and tectonic settings can be inferred by examining the feature importance analyses provided by RF, XGBoost, and ANN used in this study. Similarly, machine learning is not limited to classification tasks but can also be applied to clustering (Jain et al. 1999; Lavallin and Downs 2021; Ezugwu et al. 2022), actively learning to identify relevant relationships within large datasets and thereby validating related findings.

The success of machine learning in geochemistry depends on both the volume and integrity of the data. Despite the existence of vast geochemical databases, issues with data accuracy and completeness persist. Thus, selecting machine learning strategies for geochemical analysis should be guided by the specific context and objectives rather than the allure of novel or intricate methods. Incorporating expert knowledge into machine learning applications is crucial for achieving enhanced outcomes in traditional challenges (Hronsky and Kreuzer 2019).

Transform machine learning models into user-friendly software

To enhance the versatility of the application, user-friendly software was created by leveraging three of the best-performing machine learning models. This software is named GeoTectAI. It relies on the .NET 7 framework, making it operable on Linux, MacOS, and Windows platforms (Fig. 1c). Through this software (Fig. 12), users can quickly and conveniently obtain classification results of the whole-rock geochemical element dataset of igneous rocks without programming and AI-related knowledge. Readers can access and download all resources from the following GitHub repository: https://github.com/MaxwellLei/GeoTectAI.

Using the previous research data from the Austral-Cook Islands (Jackson et al. 2020; Takamasa et al. 2009) as an example, this application allows for the input of data from single or batch samples, including information on nine major elements and fourteen trace elements in whole-rock geochemistry. As shown in Fig. 12, the software has loaded whole-rock data samples of igneous rocks from the Austral-Cook Islands study area. You can select different pre-trained machine learning models from the dropdown menu, using the Random Forest model as an example. After clicking

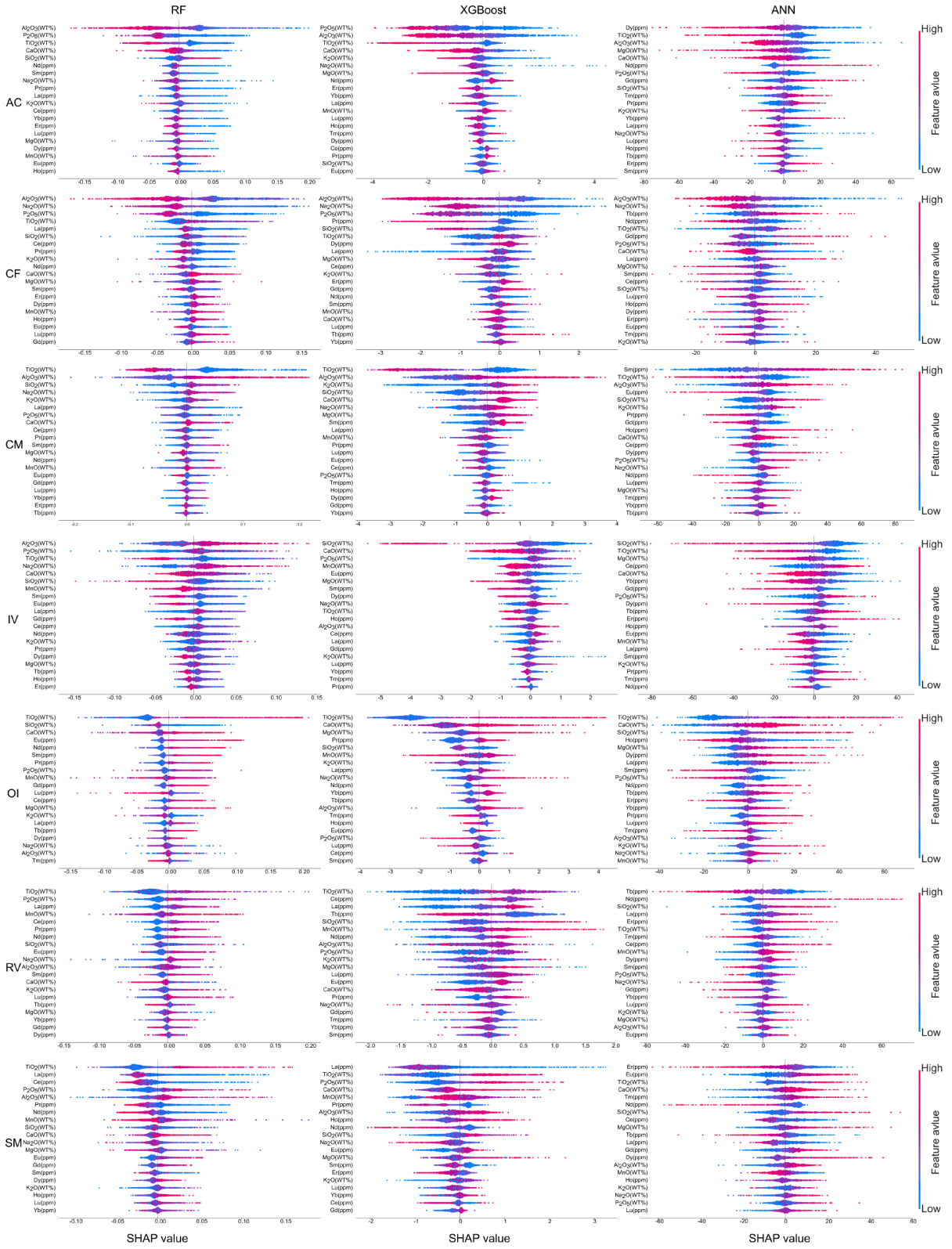

Fig. 11 SHAP summary plots for XGBoost, RF, and ANN across different tectonic settings

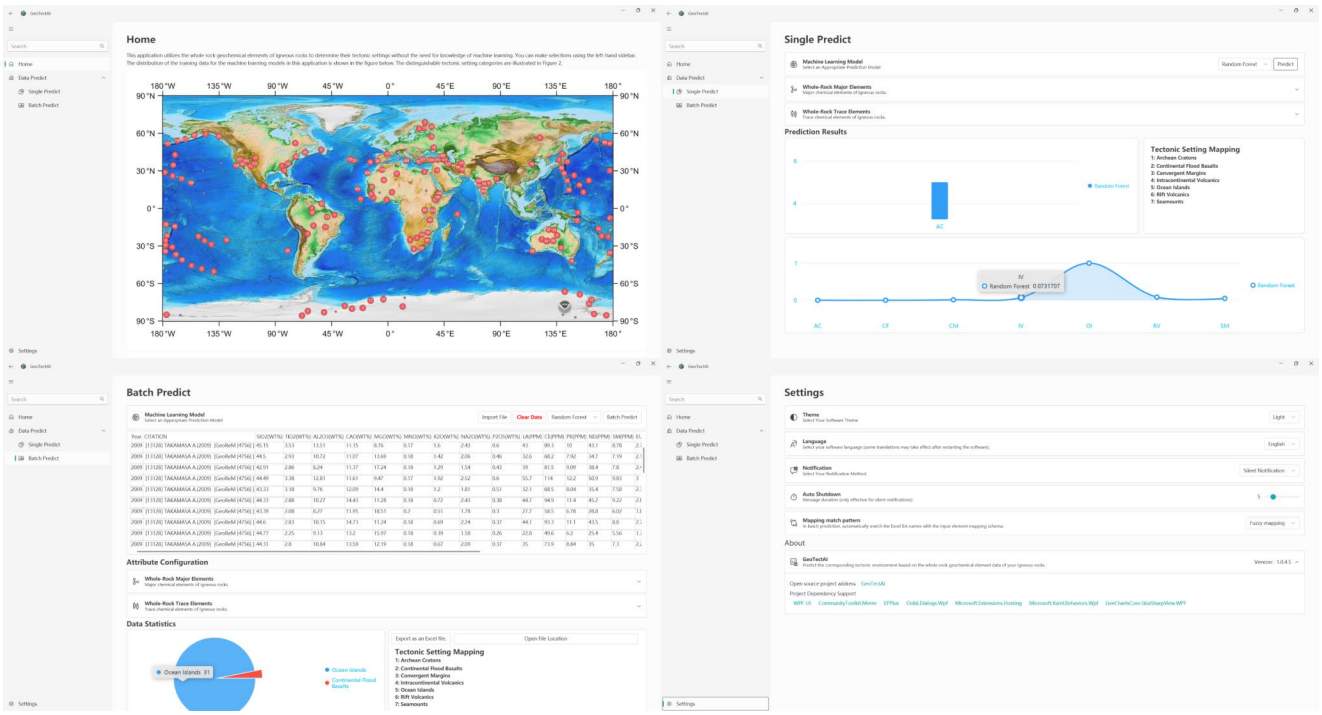

Fig. 12 Schematic diagram of the user-friendly software interface

"Predict," the prediction results will be displayed at the bottom, and the mapping results will be visible on the right side. In this application example, 31 out of 32 samples were identified as OI type. You can also obtain results through single-sample predictions and view the probability values of the predictions. For instance, sample number RTG301-1 has a significantly higher probability of being classified as OI type compared to other probabilities. Therefore, it can be preliminarily determined that these samples belong to the OI tectonic setting.

Conclusions

1. This study employed six distinct machine learning methodologies—DT, KNN, SVM, RF, XGBoost, and ANN—to discern tectonic settings based on whole-rock geochemistry of magmatic rocks from eight different tectonic settings. Among these, RF, XGBoost, and ANN demonstrated superior performance, achieving approximately 87% in average accuracy and recall rates.
2. XGBoost and RF algorithms have identified SiO_2, TiO_2, Al_2O_3, CaO, and P_2O_5 as the most crucial features, whereas ANN emphasizes SiO_2, TiO_2, Al_2O_3, CaO, and Nd, focusing more on trace elements in specific tectonic settings. The integrated use of a more comprehensive array of geochemical elements may be crucial for surpassing the limitations of traditional charts in the discrimination of tectonic settings.
3. Leveraging the three most effective models, a cross-platform, user-friendly visual software was developed to offer a tool for discriminating tectonic settings.
4. Machine learning is an efficient tool in geochemical research, particularly valuable for analyzing high-dimensional and large geochemical datasets. Through the outcomes of machine learning models, it is possible to identify differences and relationships between tectonic settings retrospectively.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12145-024-01385-5.

Acknowledgements The authors thanks PetDB and GEOROC databases for providing data support. We gratefully acknowledge Prof. Hassan A. Babaie and the three anonymous reviewers for their insightful comments, which greatly improved the manuscript.

Author contributions Conceptualization, Wenyan Cai, Ming Lei, Jian Li and Chao Zhang; methodology, Wenyan Cai, Ming Lei and Jian Li; software, Wenyan Cai, Ming Lei and Xiao Liu; validation, Wenyan Cai and Ming Lei; formal analysis, Wenyan Cai, Ming Lei and Jian Li; data curation, Ming Lei and Qingyi Cui; writing—original draft preparation, Ming Lei; writing—review and editing, Wenyan Cai and Jian Li; visualization, Ming Lei; supervision, Wenyan Cai and Jian Li; project administration, Wenyan Cai; funding acquisition, Wenyan Cai.

Funding This work was financially supported by the Shandong Natural Science Foundation (No. ZR2021QD106 and No. ZR2021QD056)

and National Natural Science Foundation of China (42202087 and 42203071).

Data availability Data were downloaded from the GEOROC database (https://georoc.eu/) on 4 November 2023, using the following parameters: Query by Geological Setting = Archean Cratons (incl. Greenstone Belts) , Continental Flood Basalts, Convergent Margins, Intracontinental Volcanics, Ocean Islands, Seamounts, Rift Volcanics；ROCK TYPE = VOLCANIC ROCK and PLUTONIC ROCK.The data were downloaded from the PetDB Database (www.earthchem.org/petdb) on 6, November2023,, using the following parameters: By Tectonic Setting = OCEAN_ISLAND and rock classification = Igneous.

Declarations

Competing interests The authors declare no competing interests.
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Agrawal S, Guevara M, Verma SP (2008) Tectonic discrimination of basic and ultrabasic volcanic rocks through log-transformed ratios of immobile trace elements. Int Geol Rev 50:1057–1079

Armstrong-Altrin JS, Verma SP (2005) Critical evaluation of six tectonic setting discrimination diagrams using geochemical data of Neogene sediments from known tectonic settings. Sed Geol 177:115–129

Barbato G, Barini EM, Genta G, Levi R (2011) Features and performance of some outlier detection methods. J Applied Statistics 38:2133–2149

Bicego M, Rossetto A, Olivieri M, Londoño-Bonilla JM, Orozco-Alzate M (2023) Advanced KNN approaches for Explainable Seismic-Volcanic Signal classification. Math Geosci 55:59–80

Bishop C (2006) Pattern recognition and machine learning, vol 2. Springer, pp 531–537

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

Butler JC, Woronow A (1986) Discrimination among tectonic settings using trace element abundances of basalts. J Geophys Research: Solid Earth 91:10289–10300

Chappell BW, White AJR (1992) I-and S-type granites in the Lachlan Fold Belt. Earth Environ Sci Trans Royal Soc Edinb 83:1–26

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794

Chen LR, Wang L, Miao JL, Gao H, Zhang Y, Yao Y, Bai M, Mei LS, He J (2020) Review of the application of big data and artificial intelligence in geology. Journal of Physics: Conference Series 1684, 012007

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297

Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13:21–27

Doucet LS, Tetley MG, Li ZX, Liu Y, Gamaleldien H (2022) Geochemical fingerprinting of continental and oceanic basalts: a machine learning approach. Earth Sci Rev 233:104192

Eby GN (1990) The A-type granitoids: a review of their occurrence and chemical characteristics and speculations on their petrogenesis. Lithos 26:115–134

Eby GN (1992) Chemical subdivision of the A-type granitoids: petrogenetic and tectonic implications. Geology 20:641–644

Ezugwu AE, Ikotun AM, Oyelade OO, Abualigah L, Agushaka JO, Eke CI, Akinyelu AA (2022) A comprehensive survey of clustering algorithms: state-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Eng Appl Artif Intell 110:104743

Gard M, Hasterok D, Halpin JA (2019) Global whole-rock geochemical database compilation. Earth Syst Sci Data 11:1553–1566

Ge C, Huo J, Gu HO, Wang FY, Sun H, Li XY, Li WW, Yuan F (2021) Tectonic discrimination and application based on convolution neural network and incomplete big data. J Geochem Explor 220:106662

Han S, Li M, Ren Q (2019) Discriminating among tectonic settings of spinel based on multiple machine learning algorithms. Big Earth Data 3:67–82

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York, pp 1–758

Ho TK (1995) Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, 278–282

Hronsky JMA, Kreuzer OP (2019) Applying spatial prospectivity mapping to exploration targeting: fundamental practical issues and suggested solutions for the future. Ore Geol Rev 107:647–653

Jackson MG, Halldórsson SA, Price A, Kurz MD, Konter JG, Koppers AAP, Day JMD (2020) Contrasting Old and Young Volcanism from Aitutaki, Cook Islands: implications for the origins of the Cook–Austral volcanic chain. J Petrol 61:egaa037

Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv (CSUR) 31:264–323

Jo JM (2019) Effectiveness of normalization pre-processing of big data to the machine learning performance. J Korea Inst Electron Communication Sci 14:547–552

Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. Science 349:255–260

Kotsiantis SB, Zaharakis I, Pintelas P (2007) Supervised machine learning: A review of classification techniques. In Emerging Artificial Intelligence Applications in Computer Engineering, vol. 160, pp. 3–24

Lavallin A, Downs JA (2021) Machine learning in geography–Past, present, and future. Geogr Compass 15:e12563

Lehnert K, Su Y, Langmuir CH, Sarbas B, Nohl U (2000) A global geochemical database structure for rocks. Geochemistry, Geophysics, Geosystems 1

Li C, Arndt NT, Tang Q, Ripley EM (2015) Trace element indiscrimination diagrams. Lithos 232:76–83

Lundberg SM, Lee SI (2017) Consistent feature attribution for tree ensembles. arXiv Preprint arXiv :170606060

Melgani F, Bruzzone L (2004) Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans Geosci Remote Sens 42:1778–1790

Mucherino A, Papajorgji PJ, Pardalos PM (2009) K-nearest neighbor classification. In Data Mining in Agriculture, 83–106

Nakamura K (2023) A practical approach for discriminating tectonic settings of basaltic rocks using machine learning. Appl Comput Geosci 19:100132

Patle A, Chouhan DS (2013) SVM kernel functions for classification. In Proceedings of the 2013 International Conference on Advances in Technology and Engineering (ICATE), 1–9

Pearce JA (2008) Geochemical fingerprinting of oceanic basalts with applications to ophiolite classification and the search for Archean oceanic crust. Lithos 100:14–48

Pearce JA, Lippard SJ, Roberts S (1984) Characteristics and tectonic significance of supra-subduction zone ophiolites. Geol Soc Lond Special Publications 16:77–94

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

Potratz L, Canchumuni SW, Castro JDB, Potratz J, Pacheco MAC (2021) Automatic Lithofacies Classification with t-SNE and K-Nearest Neighbors Algorithm. Anuário Do Instituto De Geociências 44

Quinlan JR (1986) Induction of decision trees. Mach Learn 1:81–106

Ren Q, Li M, Han S, Zhang Y, Zhang Q, Shi J (2019) Basalt tectonic discrimination using combined machine learning approach. Minerals 9:376

Richards JP, Kerrich R (2007) Special paper: Adakite-like rocks: their diverse origins and questionable role in metallogenesis. Econ Geol 102:537–576

Saccani E (2015) A new method of discriminating different types of post-archean ophiolitic basalts and their tectonic significance using Th-Nb and Ce-Dy-Yb systematics. Geosci Front 6:481–501

Saha R, Upadhyay D, Mishra B (2021) Discriminating tectonic setting of igneous rocks using biotite major element chemistry–A machine learning approach. Geochem Geophys Geosyst 22, e2021GC010053.

Singh D, Singh B (2020) Investigating the impact of data normalization on classification performance. Appl Soft Comput 97:105524

Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14:199–222

Snow CA (2006) A reevaluation of tectonic discrimination diagrams and a new probabilistic approach using large geochemical databases: moving beyond binary and ternary plots. J Phys Res 111:B06206

Suthaharan S (2016a) Decision tree learning. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, 237–269

Suthaharan S (2016b) Support vector machine. Machine learning models and algorithms for big data classification: thinking with examples for effective learning, 207–235

Takaew P, Xia JC, Doucet LS (2024) Machine learning and tectonic setting determination: bridging the gap between Earth scientists and data scientists. Geosci Front 15:101726

Takamasa A, Nakai S, Sahoo Y, Hanyu T, Tatsumi Y (2009) W isotope compositions of oceanic islands basalts from French Polynesia and their meaning for core–mantle interaction. Chem Geol 260:37–46

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

Verma SP (2010) Statistical evaluation of bivariate, ternary and discriminant function tectonomagmatic discrimination diagrams. Turkish J Earth Sci 19:185–238

Verma SP, Guevara M, Agrawal S (2006) Discriminating four tectonic settings: five new geochemical diagrams for basic and ultrabasic volcanic rocks based on log—ratio transformation of major-element data. J Earth Syst Sci 115:485–528

Vermeesch P (2006a) Tectonic discrimination diagrams revisited. Geochemistry, Geophysics. Geosystems 7

Vermeesch P (2006b) Tectonic discrimination of basalts with classification trees. Geochim Cosmochim Acta 70:1839–1848

Wang L, Zhang C, Geng R, Li Y, Song J, Wang B, Cui F (2023) The discrimination of tectonic settings using trace elements in magmatic zircons: a machine learning approach. Earth Sci Inf 16:4097–4112

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

Whalen JB, Currie KL, Chappell BW (1987) A-type granites: geochemical characteristics, discrimination and petrogenesis. Contrib Miner Petrol 95:407–419

Wu FY, Li XH, Yang JH, Zheng YF (2007) Discussions on the petrogenesis of granites. Acta Petrologica Sinica 23:1217–1238 (in Chinese with English abstract)

Yaqoob I, Hashem IAT, Gani A, Mokhtar S, Ahmed E, Anuar NB, Vasilakos AV (2016) Big data: from beginning to future. Int J Inf Manag 36:1231–1247

Yegnanarayana B (2009) Artificial neural networks. PHI Learning Pvt. Ltd

Zhang BY, Sun JK, Luo XJ, Jin WJ, Wang L, Du XL, Chen WF, Du J, Zhang Q, Zhu YQ (2019) Data analysis of major and trace element of gabbro clinopyroxene from different tectonic setting. Earth Sci Front 26:33–44

Zhong SH, Liu Y, Li SZ, Bindeman IN, Cawood PA, Seltmann R, Niu JH, Guo GH, Liu JQ (2023) A machine learning method for distinguishing detrital zircon provenance. Contrib Miner Petrol 178:35

Zurada J (1992) Introduction to artificial neural systems. West Publishing Co