Machine Learning Investigation of Clinopyroxene Compositions to Evaluate and Predict Mantle Metasomatism Worldwide

Ben Qin[1,2]*, Fang Huang[3], Shichun Huang[4], Andre Python[5], Yunfeng Chen[1,2], J ZhangZhou[1,2]*

1. School of Earth Sciences, Zhejiang University, Hangzhou 310027, China.
2. Key Laboratory of Geoscience Big Data and Deep Resource of Zhejiang Province, Hangzhou 310027, China.
3. CSIRO Mineral Resources, Kensington, WA 6151, Australia.
4. Department of Geoscience, University of Nevada, Las Vegas, 4505 S. Maryland Pkwy, Las Vegas 89154, USA.
5. Center for Data Science, Zhejiang University, Hangzhou, China.

Corresponding author: Ben Qin (charlesbenq@zju.edu.cn); J ZhangZhou (zhangzhou333@zju.edu.cn)

Key points:

- Mantle clinopyroxene major and trace element data compiled and evaluated with machine learning models
- Accuracy comparisons between low- and high-dimensional dataspaces reveal the most important features for classification
- Machine learning models identify clusters of mantle metasomatism worldwide

Plain Language Summary

Clinopyroxene is a major mineral in Earth's upper mantle. Previous studies have attempted to discriminate between reactions modifying the mantle by plotting clinopyroxene major and trace element compositions in two-dimensional (2-D) diagrams. However, these 2-

D methods show poor accuracy when applied to global datasets. Therefore, we suggest a machine learning approach to evaluate clinopyroxene compositional data in higher dimensions. Our results demonstrate that machine learning can significantly improve the accuracy of clinopyroxene compositional predictions over classical methods utilizing elemental ratios. Furthermore, the application of our algorithm to a global clinopyroxene dataset suggests that mantle metasomatism is globally widespread.

Abstract

Clinopyroxene major and trace element compositions document their physicochemical evolution and have been widely used to detect mantle metasomatism. Classical methods typically rely on one or several elemental ratios such as Ca/Al, Mg/Fe, La/Yb, and Ti/Eu to determine whether rocks or minerals have been metasomatized. These methods have proven useful at specific sites, but not globally. In this study, we used machine learning methods to classify the chemical compositions of clinopyroxenes from mantle xenoliths and examine their relationship with mantle metasomatism. We compiled major element data from 8,713 clinopyroxene samples (21,605 analyses) and trace element data from 1,235 clinopyroxene samples (2,967 analyses). Samples were labeled "positive" if clearly affected by patent metasomatism based on petrographic evidence, "negative" if clearly unaffected by metasomatism, or were left unlabeled if neither case applied. We then trained an XGBoost machine learning model, which achieved higher accuracy than traditional methods using a limited number of elemental ratios. Our results identify numerous locations with high mean probabilities of mantle metasomatism and show variability in the probability distributions observed across locations worldwide. These results indicate that metasomatism may be globally widespread, but the probability of metasomatism is not correlated with geophysical

parameters such as crustal thickness, lithospheric thickness, or mantle *S*-wave velocity. Hence, the spatial distribution of metasomatism appears mainly driven by unobserved factors.

1. Introduction

Metasomatism modifies the mineralogy and composition of pre-existing rocks through reaction with melt/fluid at high temperature. This important process produces geochemical and isotopic heterogeneities within Earth's mantle (Aiuppa et al., 2021; Roden & Murthy, 1985; Wang et al., 2022; Zhang et al., 2009), which in turn affect chemical differentiation in the mantle, craton stability, and the physical properties of the lithosphere (Araújo et al., 2009; Dawson, 1984; Liu et al., 2021; Lloyd & Bailey, 1975; Menzies & Murthy, 1980; O'Reilly & Griffin, 2013; Pearson et al., 2021; Peng et al., 2021; Rudnick et al., 1993). Therefore, evaluating mantle metasomatism at the global scale is essential to understanding mantle heterogeneity.

The composition of the upper mantle has been investigated through various approaches. The chemical compositions of basalts and petrological models of peridotite-basalt melting relationships have provided crucial information about the mantle (McDonough & Sun, 1995; Ringwood, 1962). Mantle xenoliths brought to the surface as inclusions in basalts and peridotite massifs directly sample the upper mantle and thus provide direct insights into upper mantle processes and compositions. Knowledge of the petrography and geochemistry of these peridotite samples is therefore likely to provide important information, particularly regarding the nature of upper-mantle partial melting, fractional crystallization, and metasomatism processes (Frey & Green, 1974; Frey & Prinz, 1978).

The occurrence of metasomatism can be revealed by the petrography and geochemistry of mantle xenoliths. Patent metasomatism is identified straightforwardly based on the presence of secondary minerals in xenoliths. In comparison, cryptic metasomatism and stealth

metasomatism are not so easy to identify because the former does not produce new phases (Dawson, 1984) whereas the latter only produces minerals that are indistinguishable from common mantle minerals (e.g., clinopyroxene or garnet; Griffin et al., 2009). Therefore, it is often extremely difficult to identify mantle metasomatism based on petrographic observations alone, and researchers instead turn to the geochemistry of xenolithic minerals. Here, we focus on clinopyroxene compositions for three reasons: 1) clinopyroxene is abundant in the upper mantle, 2) it is a substantial reservoir for various minor and trace elements, and 3) its composition is sensitive to mantle metasomatism. Many studies have proposed using the chemical composition of clinopyroxene to identify metasomatism, including the ratios Ca/Al (Rudnick et al., 1993; Wang et al., 2010), Mg/(Mg + Fe) (Mg#) (Yaxley & Green, 1998), La/Yb (Coltorti et al., 1999; Zong & Liu, 2018), and Ti/Eu (Coltorti et al., 1999; Zong & Liu, 2018). These elemental ratios are based on petrological models describing reactions between silicate minerals and metasomatic agents (e.g., carbonate or silicate melts), which enrich clinopyroxene in incompatible elements (Dalou et al., 2009; Green et al., 1992; Klemme et al., 1995; Rudnick et al., 1993; Sweeney et al., 1995).

Many major and trace element analyses have been reported for clinopyroxenes in xenolith samples worldwide (Figure 1; see section 2 for details on data selection). This highlights the challenge of directly estimating the scale and extent of mantle metasomatism, which requires the development of effective analytical tools capable of exploiting information from indirect indicators of major/trace element compositions. Although elemental ratios have been successfully used to identify metasomatism at specific sites, they have proven inaccurate when applied globally (Figure 2; see section 4 for more details), suggesting that the spatial distribution of metasomatism worldwide is driven by processes that cannot be captured by elemental ratios alone. A chemical model quantifying mantle metasomatism at the global scale has yet to be developed.

Whereas elemental ratios do not seem to capture the main drivers of the spatial distribution of metasomatism worldwide, relevant information may be obtained from large-volume, high-dimensional geochemical data. In recent years, machine learning (ML) approaches have been applied to mineralogy, petrology, and geochemistry datasets to provide new insights and identify trends and patterns that would otherwise be unobservable (e.g., Petrelli & Perugini, 2016; Petrelli et al., 2020; Thomson et al., 2021; Ueki et al., 2018; Valetich et al., 2021; Wang et al., 2021; Zhao et al., 2019), demonstrating their potential to quantify mantle metasomatism worldwide.

Here, we compiled a global dataset of major and trace element compositions of clinopyroxenes from mantle xenoliths and trained a supervised ML algorithm (XGBoost; Chen & Guestrin, 2016) to classify metasomatism in high-dimensional space. We also trained unsupervised machine learning models to ensure that the labeled training and testing dataset and unlabeled application dataset had similar distributions. Finally, we applied our trained ML model to predict the probability of the occurrence of metasomatism at the global scale.

2. Data compilation and labeling

We downloaded compositional data for clinopyroxenes from mantle xenoliths from 972 locations worldwide (Figure 1) from the GEOROC database (http://georoc.mpch-mainz.gwdg.de/georoc/; accessed 14 July 2020) (Sarbas, 2008). Each location includes multiple samples and analyses (Figure S1). To exclude unreliable samples, we used only clinopyroxenes with 40–60 wt.% SiO_2, <40 wt.% MgO, <30 wt.% FeO^T (the superscript 'T' indicates total iron), <26 wt.% CaO, and oxide totals of 98.5–101.5 wt.%. Elements missing from >60% of the entire dataset were not considered.

After this initial filtering, our "Parent" dataset contained 21,605 observations (rows) corresponding to clinopyroxene major element analyses (SiO_2, TiO_2, Al_2O_3, Cr_2O_3, FeO^T, CaO,

MgO, MnO, and Na_2O), and 2,967 rows of trace element analyses (including Sc, Ti, V, Cr, Ni, Rb, Sr, Y, Zr, Nb, Ba, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Hf, Ta, Pb, Th, and U) (Figure S2). In general, the proportion of missing values varies among elements, with major element data being rather complete, e.g., 5% missing analyses for Cr_2O_3 (1,024 of 21,605), and a higher proportion of missing data for trace elements, e.g., 32% missing analyses for Ti (961 of 2,967) (Figure S2).

The Parent dataset was further divided into a labeled training and testing subset ("Labeled dataset") and an unlabeled application subset ("Application dataset"). The Labeled dataset was used to classify clinopyroxene as being affected or unaffected by metasomatism. We classified clinopyroxenes in the Labeled dataset as being affected ("positive") or unaffected by metasomatism ("negative") based on the petrographic descriptions provided in the original literature. A sample was labeled "positive" if its petrographic description contains evidence of metasomatic phases, including silicate glass, calcite, hornblende, phlogopite, and apatite. In contrast, a sample was labeled "negative" if it shows a monotonous increase in chondrite-normalized (McDonough & Sun, 1995) light rare earth element (LREE) concentrations (i.e., La, Ce, Pr, Nd, Sm, and Gd). Based on these criteria, 1,650 major and 539 trace element analyses were labeled "positive" and 439 and 333 respective analyses were labeled "negative". In the Parent dataset, most examples (>70%; i.e., 19,516 major and 2,095 trace analyses) did not contain petrographic descriptions of metasomatic minerals in the original literature; this unlabeled Application dataset was used to test the unsupervised ML algorithm (see Section 3.2).

3. Methods

We employed a three-step modeling process (Figure 3). First, supervised learning models were trained based on the Labeled dataset. Second, unsupervised learning models were

implemented on the Parent dataset to verify that the overall data distributions of the Labeled and Application datasets were similar. Third, the optimal classification model obtained in step one was applied to the unlabeled clinopyroxene compositions in the Application dataset. In this study, all models were implemented using the scikit-learn Python package (Pedregosa et al., 2011).

3.1 Training the supervised learning models

In the first step, we trained supervised ML models to classify clinopyroxenes as a binary variable (1 if affected by metasomatism, 0 otherwise). We tested several ML algorithms, including Random Forest (Breiman, 2001) and Support Vector Machines (Boser et al., 1992), and eventually chose XGBoost (Chen & Guestrin, 2016) due to its flexibility, predictive performance, computational efficiency, and interpretability. Importantly, the Random Forest and Support Vector Machines algorithms are not designed to handle missing values (Boser et al., 1992; Breiman 2001), which are frequent in our dataset. In contrast, XGBoost can accommodate sparse feature formats and can automatically identify the best imputation value for missing values based on reduction on training loss (Chen & Guestrin, 2016). Furthermore, in addition to its high predictive capability and computational efficiency, the tree structure of XGBoost facilitates interpretation of the results (Azodi et al., 2020), which is important for identifying features associated with the occurrence of metasomatism. We directly used the elemental data without any preliminary transformation as the input into the XGBoost classification algorithm.

XGBoost is based on a gradient-boosting decision tree method (Friedman, 2001) and has been recently applied in a wide range of applications aiming to predict complex spatial phenomena at the global scale (e.g., Cook-Patton et al., 2020; Python et al., 2021; Zheng et al., 2021). XGBoost uses a gradient-descent algorithm to minimize the loss when adding new

models. In practice, it continuously adds trees to fit the residuals of the previous prediction, and the predictions are computed as the sum of the effects of all trees. For a dataset with n observations, label element $\boldsymbol{y_i} \in R$ with $i = \{1,\ldots,n\}$, and m features composed of feature element $\boldsymbol{x_i} \in R^m$ with $j = \{1,\ldots,m\}$, the predictions $\hat{y}_i$ are obtained by summing the scores obtained in the corresponding leaves, which is expressed as (Chen & Guestrin, 2016):

$$\hat{y}_i = \phi(\boldsymbol{x_i}) = \sum_{k=1}^{K} f_k(\boldsymbol{x_i}), \text{with } f_k \in \mathrm{F} \ , \qquad \text{(Eq. 1)}$$

where $\mathrm{F} = \{f(\boldsymbol{x}) = w_{q(\boldsymbol{x})}\}(q: R^m \to T, w \in R^T)$ is the space of the regression trees, additive function tree $k = \{1,\ldots,K\}$, and each f_k corresponds to an independent tree structure q and leaf weight w. Here, q represents the structure of each tree that maps an observation to a corresponding leaf, with T the total number of leaves in the tree. $w_{q(\boldsymbol{x})}$ represents the set of scores computed in all leaf nodes in a tree. We used XGBoost within a classification framework since the label $\boldsymbol{y_i}$ is binary (1 if affected by metasomatism, 0 otherwise). For each observation i, the output of the classification $\hat{y}_i$ represents the probability that metasomatism is present. In this classification framework, $\hat{y}_i$ is calibrated as a probability by taking only values between 0 and 1. To compute the elements of the confusion matrix, we dichotomize $\hat{y}_i$ as equal to 1 if $\hat{y}_i \geq 0.5$ and 0 otherwise.

To minimize bias and variance in the predictive scores, we performed a ten-fold cross-validation procedure (Kohavi, 1995) by randomly splitting the Labeled dataset into training (70%) and testing subsets (30%). Therefore, overfitting and randomness can be mitigated by cross-validation and the splitting of the training and testing subsets utilized to evaluate the classifier performance. We applied Grid Search Cross-Validation (from the scikit-learn package), which aims to find an optimal hyperparameter combination (eta, gamma, max depth,

and alpha) through an iterative grid-search process. The procedure trains 9,000 candidate models and selects the model with the best predictive performance via ten-fold cross-validation.

To evaluate the performance of XGBoost models, several classification metrics can be defined based on the confusion matrix (Stehman, 1997), a specific table layout that visualizes model performance. Each row of the confusion matrix represents the instances in an actual class, whereas each column represents the instances in a predicted class. We use Accuracy and the F1 score (Dice, 1945; Sørensen, 1948) described below:

Accuracy is the ratio of the total number of correct "positive" and "negative" predictions to the total number of known "positive" and "negative" cases:

$$\text{Accuracy} = \frac{\text{Correct positive + negative predictions}}{\text{Known positive + negative cases}}. \quad \text{(Eq. 2)}$$

The F1 score is the harmonic mean of Precision and Recall:

$$\text{F1 score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}}, \quad \text{(Eq. 3)}$$

where Precision is a measure of accuracy provided that a specific class (here, "positive") has been predicted:

$$\text{Precision} = \frac{\text{Correct positive predictions}}{\text{All positive predictions}}. \quad \text{(Eq. 4)}$$

Recall is a measure of the ability of a model to select instances of a certain class (again, "positive" here) from a dataset:

$$\text{Recall} = \frac{\text{Correct positive predictions}}{\text{Known positive cases}}. \quad \text{(Eq. 5)}$$

Although Accuracy is a common and direct way to evaluate and improve models, the F1 score can give a better measure of the incorrectly classified cases than Accuracy. The F1 score is more suitable when the classes are imbalanced. Given the class imbalance observed in our dataset, we favor the F1 score to compare the predictive performance of our models.

To test the impacts of feature selection on the ML results, we trained the XGBoost model on different major and trace element subsets selected from the Labeled dataset (Table 1). For major elements, we compared the relative feature importance when using two traditional element pairs (CaO and Al_2O_3; MgO and FeO^T), all four of those elements (i.e., both pairs combined), and all nine major elements. For trace elements, we also used two traditional element pairs (Eu and Ti; La and Yb) and both pairs combined, but also considered 13 elements after dimensional reduction by trace element correlation, 25 non-fluid-mobile trace elements (i.e., excluding Rb, Sr, Ba, Pb, and U), and all 30 considered trace elements.

3.2 Unsupervised machine learning models

We used unsupervised machine learning models to assess the degree of similarity between the data distributions of the Labeled and Application datasets. The input data (all major elements) used in the unsupervised machine learning models was centered log-ratio transformed to prevent data closure (Aitchison, 1982).

We applied *k*-means clustering to the Parent dataset to measure the similarity of the Labeled and Application datasets. *k*-means clustering is an unsupervised machine learning algorithm that classifies a given dataset into *k* clusters. It defines *k* (an *a priori* fixed number) centroids, or mean points, one for each cluster S_i that minimizes a norm of the kind:

$$V = \sum_{i=1}^{K} \sum_{X_j \in S_i} (X_j - \bar{X}_t), \quad \text{(Eq. 6)}$$

where $\bar{X}_t$ are the mean points of all $X_j \in S_i$, and V is the objective function. In other words, *k*-means clustering can divide a dataset into several clusters. As a result, data in the same cluster have similar information that is different from data in other clusters. For example, after clustering of the Parent dataset, if the Labeled dataset is distributed in all clusters, then the Labeled dataset contains the same data distribution as that the Application dataset. Therefore,

we can utilize this result to predict the Application dataset. To visualize the *k*-means clustering result, we used principal component analysis (PCA) (Smith, 2002) to reduce the dimensionality to 2 dimensions.

3.3 Application of trained models to unlabeled data

Finally, we applied the best-performing supervised training model (see section 3.1) to the Application dataset. For an unlabeled clinopyroxene analysis, the XGBoost model classifies it as "positive" (metasomatized) or "negative" (unmetasomatized) based on its chemical composition. For each xenolith sampling site *l*, we defined the mean probability of occurrence of metasomatism ($\bar{p}_l$) by averaging the predictive probabilities obtained in all *n* analyses at that site as:

$$\bar{p}_l = \frac{\sum_{i=1}^{n} p_{l,i}}{n}. \quad \text{(Eq. 7)}$$

Based on the predicted probabilities of metasomatism computed for each sampling location, we mapped the mean predicted probability of metasomatism at the global scale within 1° × 1° bins, for a total of 599 bins that represent our study area.

4. Results and Discussion

4.1 Limitations of traditional models

Previous studies have proposed that metasomatism be identified based on elemental ratios such as CaO/Al_2O_3, Mg#, Ti/Eu, or $(La/Yb)_N$ (Brey et al., 2008; Coltorti et al., 1999; Klemme et al., 1995; Rudnick et al., 1993; Yaxley & Green, 1998; Zong & Liu, 2018). Although these parameters effectively characterize metasomatism at specific sampling locations, they perform poorly when used to predict metasomatism at the global scale (Figure 2).

Classification results using elemental ratios are inconsistent with the petrographically confirmed metasomatized natures of the samples. For example, among the 2,089 samples labeled according to the presence of metasomatic minerals, the classification of metasomatized and unmetasomatized xenoliths using clinopyroxene CaO/Al_2O_3 has an accuracy rate of 47% and 72%, respectively (Figure 2a). Other elemental ratios similarly show low accuracy in predicting metasomatism (17%, 15%, and 53% for Mg#, Ti/Eu, and $(La/Yb)_N$, respectively), and higher accuracy in predicting unmetasomatized observations (69%, 100%, and 100%, respectively; Figure 2b–d). Therefore, it is complicated to predict the occurrence of metasomatism. Classic bi-variate plots are suitable for predicting unmetasomatized samples because the broad data range of those samples overlaps that of metasomatized samples: unmetasomatized samples are typically taken as having Ca/Al < 5, Mg# < 92, Ti/Eu > 1500, or $(La/Yb)_N < 3$, but these ranges also describe numerous metasomatized samples. Our results show that globally, the accuracy of traditional bi-variate plots is relatively poor at only 43–76.5%.

Identifications of metasomatism using different traditional element ratios are also inconsistent (Figure S3). To demonstrate this, we labeled data with $CaO/Al_2O_3 > 5$ as "positive" and <5 as "negative" (Figure S3a). However, plotting these data points, labeled by their CaO/Al_2O_3 ratios, on the MgO vs. FeO^T diagram (and taking Mg# > 92 as also indicating "positive") exhibits a mean accuracy of 67% (Figure S3b). Similarly for trace elements, applying labels based on Ti/Eu values to $(La/Yb)_N$ shows 75% accuracy (Figure S3c, d). Therefore, the traditional elemental ratios CaO/Al_2O_3, MgO/FeO^T, Ti/Eu, and $(La/Yb)_N$ cannot effectively and accurately classify metasomatism across different sampling sites worldwide.

4.2 Classification results and geochemical explanation

Figure 4a presents the confusion matrix (see section 3.1) of the classification results based on the XGBoost model trained using data for all nine major elements from the analyses in the Labeled dataset. Based on the confusion matrix, we obtained a F1 score and accuracy of 0.968 and 0.949, respectively. Consistently, the results based on all 30 trace elements produce an F1 score and accuracy of 0.957 and 0.947, respectively (Figure 4c). These high F1 scores and accuracies suggest that both major and trace element models may outperform traditional ion-pair classification methods.

Based on the results of the XGBoost algorithm, we calculated the relative importance of each feature to the metasomatism classification. As shown in Figure 4b, Na_2O, FeO^T, MnO, and CaO have the highest relative importance scores among the major elements, indicating that they are important for discriminating whether a sample has been metasomatized. Indeed, the presence of melt affects clinopyroxene compositions, producing clinopyroxene with lower Mg# and higher Na/Ca (Yaxley & Green, 1998). Furthermore, Mn's redistribution among garnet, orthopyroxene, clinopyroxene, and olivine is affected by metasomatism (Achterbergh et al., 2001; Norman, 1998). As shown in Figure 4d, Ho, Ce, U, Sr, Yb, and Ba are the most important trace elements for classifying metasomatism, consistent with previous studies evidencing that LREEs preferentially enter the mineral phase compared to HREEs during interactions between peridotite and melts enriched in incompatible elements (Green et al., 1992; Klemme et al., 1995; Rudnick et al., 1993; Sweeney et al., 1995). We note that most elements have positive but small (<17%) feature importance values, suggesting that they may not play a major role in classification of metasomatism, and that metasomatism cannot be effectively identified by using those elements alone.

4.3 Feature correlation and selection

The Parent dataset includes 21,605 observations (rows) corresponding to 9 major elements and 2,967 observations corresponding to 30 trace elements, and we calculated Pearson's correlation coefficients (ρ) between major (Figure 5) and trace elements (Figure S4), where $\rho = 1$ (-1) indicates a perfect positive (negative) correlation, and $\rho = 0$ indicates no correlation. Several major element pairs are highly correlated (e.g., $\rho = -0.81$ for SiO_2 and TiO_2) or moderately correlated (e.g., $\rho = -0.61$ for MgO and TiO_2), but most are poorly correlated ($|\rho| < 0.40$; Figure 5). In comparison, less than a quarter of all trace element pairs are highly correlated ($|\rho| > 0.75$; Figure S4).

Our PCA results (Figure S5) for the Parent dataset show that only 64% and 56% of the variance in the major and trace element data, respectively, can be explained by two dimensions (Figure S5). Therefore, the correlation matrix (Figures 5 and S4) and PCA results further evidence that most elemental ratios provide distinct information and may independently contribute to identifying metasomatism.

In general, XGBoost provides better classification results when it is trained on more elements (Table 1). For example, the respective F1 scores and accuracies of models trained on major element data from the Labeled dataset improved from 0.891–0.899 and 0.821–0.833 when only two elements were used to 0.941 and 0.910 for four elements and 0.968 and 0.949 for all nine major elements. For models trained on trace element data from the Labeled dataset, the respective F1 scores and accuracies improved from 0.818–0.933 and 0.771–0.916 for two elements to 0.945 and 0.931 for four elements and 0.960 and 0.950 for 13 elements, but do not improve markedly when using 25 (0.954 and 0.943) or 30 elements (0.957 and 0.947). Our results show that XGBoost performs optimally when trained on 13 features (elements) and does not improve when more features are used. Despite that each feature shows a relatively low variable importance value, the best predictive performance is achieved when most features are

included. Given the data and within the limitations of the models, our results suggest that each feature may contribute to partially explain metasomatism.

4.4 Evaluating ML model performance and applicability

The ten-fold cross-validation procedure we performed on the Labeled dataset before it was randomly split into training (70%) and testing subsets (30%) resulted in a F1 score of 0.871 with standard deviation (s.d.) = 0.073 for the major element data and 0.918 (s.d. = 0.122) for the trace element data. These results demonstrate that the Labeled dataset is relatively balanced.

In Table 1, the mean F1 score of the best model as determined by Grid Search Cross-Validation on the major element training data was 0.950 (MajorI-9) and that for the trace element data was 0.973 (TraceI-25). The XGBoost models can then be further evaluated by applying these best models to the testing set (30% of the Labeled data). The best major and trace element models achieved accuracies of 0.949 (MajorI-9) and 0.950 (TraceI-13), respectively, when applied to the testing set (Table 1).

Unsupervised learning is useful for discerning patterns from the characteristics of the data itself (Figure 6). In our *k*-means unsupervised model trained on the Parent dataset, the highest silhouette coefficient (a measure of how similar an object is to its cluster compared to other clustered, with high values indicating objects are well-matched to their clusters and poorly matched to neighboring clusters; Rousseeuw, 1987) corresponds to two major element clusters and two trace element clusters. The Labeled dataset is distributed across all clusters in which the Application dataset is distributed, indicating that both the Application and Labeled datasets have similar distributions. These results indicate that the model trained on the Labeled dataset can be confidently applied to the Application dataset.

4.5 Probability of mantle metasomatism at the global scale

When applied to the Application dataset, our model predictions are presented as the probability of metasomatism, which ranges from 0 to 1 by definition. The global map of the predicted mean probability of metasomatism identifies locations with high probabilities of metasomatism (Figure 7). Here, we computed the mean for each location because multiple analytical points were available at each location.

The map highlights variations in the distribution of the probability of metasomatism. Figure 8 shows the predicted probability distributions at four localities. These locations were chosen because they cover four continents and because a sufficient number of samples (>100) were available at each to accurately estimate metasomatism. The results suggest bimodal distributions at Hannuoba (North China Craton) and Zealandia (South Pacific Ocean), and unimodal distributions with high probabilities of metasomatism at Pulpwood Harbour (South Canadian Shield) and Finsch (Kaapvaal Craton). The variability observed in the results may indicate that mantle metasomatism occurs widely but heterogeneously. Therefore, the probability of metasomatism at the global scale is generally high, and melt heterogeneity may reduce the likelihood of metasomatism in some locations. Alternatively, the machine learning algorithms work well for moderately metasomatized samples from the classic stable cratons (e.g., South Canadian Shield, Kaapvaal Craton), yet misclassify extensively metasomatized samples affected by multiple metasomatic agents that first fertilized and later depleted mineral chemical compositions (Zhang, 2009).

In addition, we also compared the probability of metasomatism to xenolith rock type. The results show that no correlation exists between the probability of metasomatism and rock type, including clinopyroxenite, dunite, harzburgite, lherzolite, peridotite, pyroxenite, and wehrlite (Figure S6). Indeed, it has been suggested that metasomatism may occur in various tectonic settings (Aiuppa et al., 2021; Dawson, 1984; Liu et al., 2021; Menzies & Murthy, 1980; Roden & Murthy, 1985; Wang et al., 2022). In particular, carbon and water lower the melting

temperatures of peridotites, and carbonated and hydrous silicate melts have been suggested as effective metasomatic agents (Hirschmann, 2000; Dasgupta and Hirschmann, 2006; Sarafian et al., 2017; Sun and Dasgupta, 2019; Thomson et al., 2016). However, considering only mantle xenoliths might present a bias because they are preferentially affected by melt when brought to the surface by eruptions, but cannot represent the average mantle composition (Artemieva, 2009).

To assess whether the probability of metasomatism is related to local lithospheric structures, we compared our results with geophysical observations of crustal thickness, lithospheric thickness, and *S*-wave velocity (at 50–200 km depth in 25-km depth intervals) and parameterized the globe into a 1° × 1° grid (Figure S7). Within each cell, we averaged the probabilities of metasomatism for each location and compared them with geophysical observations (Figures S7, S8). We also used unsupervised machine learning to search for clustering of metasomatism probabilities and geophysical parameters, but did not observe any correlations (Figure S8). We identified three possible reasons for this. First, mantle metasomatism may not necessarily be related to specific tectonic settings. Second, metasomatism produces only secondary effects on geophysical parameters such as seismic wave velocity, and a full separation of compositional from thermal factors is required to identify potential metasomatic modifications to the lithospheric mantle. Third, our dataset does not provide information on the depth or age distributions of the clinopyroxenes, making it difficult to relate the predicted metasomatic probabilities to the lithospheric mantle at a specific spatiotemporal location. Therefore, further efforts are required to reconcile the effects of metasomatism on both the chemical and physical properties of the lithospheric mantle at the global scale.

5. Conclusions

We developed a model to predict whether xenolithic clinopyroxenes have been metasomatized by applying a multidimensional approach using the XGBoost machine learning algorithm. Our model can predict whether a given sample has been metasomatized with better accuracy (95%) than traditional approaches using elemental ratios (43–77%). Our results indicate that models trained on clinopyroxene compositions, including all major and at least 13 trace elements, achieve the best prediction accuracy compared to traditional methods using only two or four elements. Furthermore, *k*-means clustering showed that the Application and Labeled datasets had similar data distributions, indicating that the models trained on the Labeled data can confidently be used to predict whether clinopyroxenes experienced metasomatism based on unlabeled data. Finally, our results show that many locations are likely to have undergone metasomatism and that metasomatism is heterogeneously distributed worldwide.

Data Availability Statement

All data and code used in this study are available at https://doi.org/10.5281/zenodo.6466993.

Acknowledgments

The authors appreciate the assistance of Yifan Zhang, Yunzhu He, and Jinfeng Sun for data compilation. We thank Can He, Shengxin Wang, and Anzhou Li in for their contribution in running the models. The work benefits from discussions with Prof. Hongfu Zhang, Xinmiao Zhao, and Yan Xiao. We thank Robert Dennen for polishing the language of the paper. J ZhangZhou acknowledges support from NSFC grant No. 42072066 and startup funds from Zhejiang University. Andre Python was funded by Zhejiang University under grant No. 2021QN81029 (fundamental research funds for the central universities).

References

Achterbergh, E. V., Griffin, W. L., & Stiefenhofer, J. (2001). Metasomatism in mantle xenoliths from the Letlhakane kimberlites: estimation of element fluxes. *Contributions to Mineralogy and Petrology*, *141*(4), 397-414. https://doi.org/10.1007/s004100000236

Aiuppa, A., Casetta, F., Coltorti, M., Stagno, V., & Tamburello, G. (2021). Carbon concentration increases with depth of melting in Earth's upper mantle. *Nature Geoscience*, *14*(9), 697-703. https://doi.org/10.1038/s41561-021-00797-y

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological), 44(2)*, 139-160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

Araújo, D. P., Griffin, W. L., & O'Reilly, S. Y. (2009). Mantle melts, metasomatism and diamond formation: Insights from melt inclusions in xenoliths from Diavik, Slave Craton. *Lithos*, *112*, 675-682. https://doi.org/10.1016/j.lithos.2008.09.015

Artemieva, I. M. (2009). The continental lithosphere: reconciling thermal, seismic, and petrologic data. *Lithos*, *109*(1-2), 23-46. https://doi.org/10.1016/j.lithos.2008.09.015

Azodi, C. B., Tang, J., & Shiu, S. H. (2020). Opening the Black Box: Interpretable machine learning for geneticists. *Trends in Genetics*, *36*(6), 442-455. https://doi.org/10.1016/j.tig.2020.03.005

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In Haussler, D. (Ed.), *5th Annual ACM Workshop on COLT* (pp. 144–152). Pittsburgh, PA: ACM Press.

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

Brey, G. P., Bulatov, V. K., Girnis, A. V., & Lahaye, Y. (2008). Experimental melting of carbonated peridotite at 6–10 GPa. *Journal of Petrology*, *49*(4), 797-821. https://doi.org/10.1093/petrology/egn002

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Coltorti, M., Bonadiman, C., Hinton, R. W., Siena, F., & Upton, B. G. J. (1999). Carbonatite metasomatism of the oceanic upper mantle: evidence from clinopyroxenes and glasses in ultramafic xenoliths of Grande Comore, Indian Ocean. *Journal of Petrology*, *40*(1), 133-165. https://doi.org/10.1093/petroj/40.1.133

Cook-Patton, S. C., Leavitt, S. M., Gibbs, D., Harris, N. L., Lister, K., Anderson-Teixeira, K. J., et al. (2020). Mapping carbon accumulation potential from global natural forest regrowth. *Nature* 585, 545–550. https://doi.org/10.1038/s41586-020-2686-x

Dalou, C., Koga, K. T., Hammouda, T., & Poitrasson, F. (2009). Trace element partitioning between carbonatitic melts and mantle transition zone minerals: Implications for the source of carbonatites. *Geochimica et Cosmochimica Acta*, *73*(1), 239-255. https://doi.org/10.1016/j.gca.2008.09.02

Dasgupta, R., & Hirschmann, M. M. (2006). Melting in the Earth's deep upper mantle caused by carbon dioxide. *Nature*, *440*(7084), 659-662. https://doi.org/10.1038/nature04612

Dawson, J. B. (1984). Contrasting types of upper-mantle metasomatism? *Developments in Petrology*, *11*(2), 289-294. https://doi.org/10.1016/B978-0-444-42274-3.50030-5

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297-302. https://doi.org/10.2307/1932409

485 Frey, F. A., & Green, D. H. (1974). The mineralogy, geochemistry and origin of Iherzolite
486 inclusions in Victorian basanites. *Geochimica et Cosmochimica Acta*, *38*(7), 1023-1059.
487 https://doi.org/10.1016/0016-7037(74)90003-9
488 Frey, F. A., & Prinz, M. (1978). Ultramafic inclusions from San Carlos, Arizona: petrologic
489 and geochemical data bearing on their petrogenesis. *Earth and Planetary Science*
490 *Letters*, *38*(1), 129-176. https://doi.org/10.1016/0012-821X(78)90130-9
491 Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals*
492 *of Statistics*, *29*(5), 1189-1232. https://www.jstor.org/stable/2699986
493 Green, T. H., Adam, J., & Siel, S. H. (1992). Trace element partitioning between silicate
494 minerals and carbonatite at 25 kbar and application to mantle metasomatism. *Mineralogy*
495 *and Petrology*, *46*(3), 179-184. https://doi.org/10.1007/BF01164645
496 Griffin, W. L., Kobussen, A. F., Babu, E. V. S. S. K., O'Reilly, S. Y., Norris, R., & Sengupta,
497 P. (2009). A translithospheric suture in the vanished 1-Ga lithospheric root of South India:
498 evidence from contrasting lithosphere sections in the Dharwar Craton. *Lithos*, *112*, 1109-
499 1119. https://doi.org/10.1016/j.lithos.2009.05.015
500 Hirschmann, M. M. (2000). Mantle solidus: Experimental constraints and the effects of
501 peridotite composition. *Geochemistry, Geophysics, Geosystems*, *1*(10).
502 https://doi.org/10.1029/2000GC000070
503 Klemme, S., Van der Laan, S. R., Foley, S. F., & Günther, D. (1995). Experimentally
504 determined trace and minor element partitioning between clinopyroxene and carbonatite
505 melt under upper mantle conditions. *Earth and Planetary Science Letters*, *133*(3-4), 439-
506 448. https://doi.org/10.1016/0012-821X(95)00098-W
507 Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model
508 selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial*
509 *Intelligence* (Vol. 2, pp. 1137-1145).

Liu, J., Pearson, D. G., Wang, L. H., Mather, K. A., Kjarsgaard, B. A., Schaeffer, A. J., et al. (2021). Plume-driven recratonization of deep continental lithospheric mantle. *Nature*, *592*(7856), 732-736. https://doi.org/10.1038/s41586-021-03395-5

Lloyd, F. E., & Bailey, D. (1975). Light element metasomatism of the continental mantle: the evidence and the consequences. *Physics and Chemistry of the Earth*, *9*, 389-416. https://doi.org/10.1016/0079-1946(75)90030-0

McDonough, W. F., & Sun, S. S. (1995). The composition of the Earth. *Chemical Geology*, 120(3-4), 223-253. https://doi.org/10.1016/0009-2541(94)00140-4

Menzies, M., & Murthy, V. R. (1980). Nd and Sr isotope geochemistry of hydrous mantle nodules and their host alkali basalts: implications for local heterogeneities in metasomatically veined mantle. *Earth and Planetary Science Letters*, *46*(3), 323-334. https://doi.org/10.1016/0012-821X(80)90048-5

Norman, M. D. (1998). Melting and metasomatism in the continental lithosphere: laser ablation ICPMS analysis of minerals in spinel lherzolites from eastern Australia. *Contributions to Mineralogy and Petrology*, *130*(3), 240-255. https://doi.org/10.1007/s004100050363

O'Reilly, S. Y., & Griffin, W. L. (2013). Mantle metasomatism. In Harlov, D. E., & Austrheim, H. (Eds.), *Metasomatism and the Chemical Transformation of Rock, Lecture Notes in Earth System Sciences* (pp. 471-533). Springer, Berlin, Heidelberg. https://link.springer.com/book/10.1007%2F978-3-642-28394-9

Pearson, D. G., Scott, J. M., Liu, J., Schaeffer, A., Wang, L. H., van Hunen, J., et al. (2021). Deep continental roots and cratons. *Nature*, *596*(7871), 199-210. https://doi.org/10.1038/s41586-021-03600-5

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830. https://arxiv.org/abs/1201.0490v4

Peng, Y., Manthilake, G., & Mookherjee, M. (2021). Electrical conductivity of metasomatized lithology in subcontinental lithosphere. *American Mineralogist*. https://doi.org/10.2138/am-2021-7942

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*(10), 1-15. https://doi.org/10.1007/s00410-016-1292-2

Petrelli, M., Caricchi, L., & Perugini, D. (2020). Machine Learning Thermo-Barometry: Application to Clinopyroxene-Bearing Magmas. *Journal of Geophysical Research: Solid Earth*, *125*(9), e2020JB020130. https://doi.org/10.1029/2020JB020130

Python, A., Bender, A., Nandi, A. K., Hancock, P. A., Arambepola, R., Brandsch, J., & Lucas, T. C. (2021). Predicting non-state terrorism worldwide. *Science Advances*, 7(31), eabg4778. https://doi.org/10.1126/sciadv.abg4778

Ringwood, A. E. (1962). A model for the upper mantle. *Journal of Geophysical Research*, *67*(2), 857-867. https://doi.org/10.1029/JZ067i002p00857

Roden, M. F., & Murthy, V. R. (1985). Mantle metasomatism. *Annual Review of Earth and Planetary Sciences*, *13*(1), 269-296. https://doi.org/10.1146/annurev.ea.13.050185.001413

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Rudnick, R. L., McDonough, W. F., & Chappell, B. W. (1993). Carbonatite metasomatism in the northern Tanzanian mantle: petrographic and geochemical characteristics. *Earth and Planetary Science Letters*, *114*(4), 463-475. https://doi.org/10.1016/0012-821X(93)90076-L

559 Sarafian, E., Gaetani, G. A., Hauri, E. H., & Sarafian, A. R. (2017). Experimental constraints
560 on the damp peridotite solidus and oceanic mantle potential
561 temperature. *Science*, *355*(6328), 942-945.
562 https://www.science.org/doi/10.1126/science.aaj2165
563 Sarbas, B. (2008). The GEOROC database as part of a growing geoinformatics network. In
564 Brady, S. R., Sinha, A. K., Gundersen, L. C. (Eds.), *Geoinformatics 2008—Data to*
565 *Knowledge, Proceedings, Geoinformatics 2008—Data to Knowledge* (Scientific
566 Investigations Report 2008-5172, pp. 42-43). Reston, VA: USGS. https://gfzpublic.gfz-
567 potsdam.de/pubman/item/item_10062
568 Smith, L. I. (2002). *A tutorial on Principal Components Analysis* (Technical report). University
569 of Otago. http://hdl.handle.net/10523/7534
570 Sørensen, T. J. (1948). *A method of establishing groups of equal amplitude in plant sociology*
571 *based on similarity of species content and its application to analyses of the vegetation on*
572 *Danish commons*. Copenhagen: I kommission hos E. Munksgaard.
573 Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy.
574 *Remote Sensing of Environment*, *62*(1), 77-89.
575 https://www.sciencedirect.com/science/article/abs/pii/S0034425797000837
576 Sun, C., & Dasgupta, R. (2019). Slab–mantle interaction, carbon transport, and kimberlite
577 generation in the deep upper mantle. *Earth and Planetary Science Letters*, *506*, 38-52.
578 https://doi.org/10.1016/j.epsl.2018.10.028
579 Sweeney, R. J., Prozesky, V., & Przybylowicz, W. (1995). Selected trace and minor element
580 partitioning between peridotite minerals and carbonatite melts at 18–46 kb
581 pressure. *Geochimica et Cosmochimica Acta*, *59*(18), 3671-3683.
582 https://doi.org/10.1016/0016-7037(95)00270-A

583 Thomson, A. R., Walter, M. J., Kohn, S. C., & Brooker, R. A. (2016). Slab melting as a barrier
584 to deep carbon subduction. Nature, 529(7584), 76-79.
585 https://www.nature.com/articles/nature16174
586 Thomson, A. R., Kohn, S. C., Prabhu, A., & Walter, M. J. (2021). Evaluating the formation
587 pressure of diamond-hosted majoritic garnets: A machine learning majorite
588 barometer. *Journal of Geophysical Research: Solid Earth*, *126*(3), e2020JB020604.
589 https://doi.org/10.1029/2020JB020604
590 Ueki, K., Hino, H., & Kuwatani, T. (2018). Geochemical discrimination and characteristics of
591 magmatic tectonic settings: A machine-learning-based approach. *Geochemistry,*
592 *Geophysics, Geosystems*, *19*(4), 1327-1347. https://doi.org/10.1029/2017GC007401
593 Valetich, M. J., Le Losq, C., Arculus, R. J., Umino, S., & Mavrogenes, J. (2021). Compositions
594 and Classification of Fractionated Boninite Series Melts from the Izu–Bonin–Mariana Arc:
595 A Machine Learning Approach. *Journal of Petrology*, *62*(2), egab013.
596 https://doi.org/10.1093/petrology/egab013
597 Wang, C., Jin, Z., Gao, S., Zhang, J., & Zheng, S. (2010). Eclogite-melt/peridotite reaction:
598 Experimental constrains on the destruction mechanism of the North China Craton. *Science*
599 *China Earth Sciences*, *53*(6), 797-809. https://doi.org/10.1007/s11430-010-3084-2
600 Wang, X., Wang, Z., Cheng, H., Zong, K., Wang, C. Y., Ma, L., et al. (2022). Gold endowment
601 of the metasomatized lithospheric mantle for giant gold deposits: Insights from
602 lamprophyre dykes. *Geochimica et Cosmochimica Acta*, *316*, 21-40.
603 https://doi.org/10.1016/j.gca.2021.10.006
604 Wang, Y., Qiu, K. F., Müller, A., Hou, Z. L., Zhu, Z. H., & Yu, H. C. (2021). Machine Learning
605 Prediction of Quartz Forming-Environments. *Journal of Geophysical Research: Solid*
606 *Earth*, *126*(8), e2021JB021925. https://doi.org/10.1029/2021JB021925

Yaxley, G. M., & Green, D. H. (1998). Reactions between eclogite and peridotite: mantle refertilisation by subduction of oceanic crust. *Schweizerische mineralogische und petrographische Mitteilungen*, *78*(2), 243-255.

Zhang, H. F. (2009). Peridotite-melt interaction: a key point for the destruction of cratonic lithospheric mantle. *Chinese Science Bulletin*, *54*, 3417-3437. https://doi.org/10.1007/s11434-009-0307-z

Zhang, H. F., Goldstein, S. L., Zhou, X. H., Sun, M., & Cai, Y. (2009). Comprehensive refertilization of lithospheric mantle beneath the North China Craton: further Os–Sr–Nd isotopic constraints. *Journal of the Geological Society*, *166*(2), 249-259. https://doi.org/10.1144/0016-76492007-152

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. https://doi.org/10.1029/2019GL082322

Zheng, Z., Zhao, L. & Oleson, K.W. Large model structural uncertainty in global projections of urban heat waves. *Nature Communications* 12, 3736 (2021). https://doi.org/10.1038/s41467-021-24113-9

Zong, K., & Liu, Y. (2018). Carbonate metasomatism in the lithospheric mantle: Implications for cratonic destruction in North China. *Science China Earth Sciences*, *61*(6), 711-729. https://doi.org/10.1007/s11430-017-9185-2

Figure Captions

Figure 1. Locations of sample analyses used in this study. The color of each sampling point represents the number of analyses performed on clinopyroxenes in mantle xenoliths from that location.

Figure 2. The application of elemental ratios proposed in previous studies to attempt to identify metasomatism in the global dataset. Symbols indicate whether each sample was petrographically identified as metasomatized ('positive') or not affected by metasomatism ('negative'). (a) The accuracy (Eq. 2) of CaO versus Al_2O_3 is 59.5%; (b) that of MgO versus FeO^T is 43%; (c) that of Ti versus Eu is 57.5%; and (d) that of La versus Yb is 76.5%.

Figure 3. Operational flow chart of the methods used in this study. Step I: the Labeled dataset was used to train the XGBoost models and evaluate model performance. Step II: the preprocessed Parent dataset was used to train *k*-means clustering models to verify that the data distributions of the Labeled and Application datasets were similar. Step III: the best model was used to predict the probability of metasomatism worldwide within 1° × 1° grid cells based on the Application dataset.

Figure 4. Results of the XGBoost model trained on the Labeled dataset to classify clinopyroxenes as affected or unaffected by metasomatism. (a, c) Confusion matrices of classification results based on major and trace element compositions from the testing subset, respectively. (b, d) The relative feature importances of major and trace elements, respectively.

Figure 5. Heat-map matrix of linear correlations (Pearson coefficients) between major elements concentrations in clinopyroxenes of the Parent dataset (21,605 observations).

Figure 6. Unsupervised learning results illustrating the similarity of the (a) major and (b) trace element data distributions in the Labeled (training, orange diamonds) and Application datasets (gray circles).

Figure 7. Probability map of mantle metasomatism at 972 unique sampling locations. Symbol color indicates the predicted probability of metasomatism from 0 (blue) to 1 (red).

Figure 8. Probability distributions of metasomatism at four selected sampling locations: (a) Hannuoba, North China Craton; (b) Zealandia, South Pacific Ocean; (c) Pulpwood Harbour, South Canadian Shield; and (d) Finsch, Kaapvaal Craton.

657 **Table 1.** Summary of XGBoost model performance.