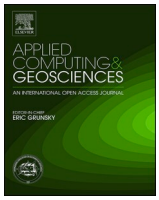
Applied Computing and Geosciences 26 (2025) 100250



Contents lists available at ScienceDirect

Applied Computing and Geosciences

journal homepage: www.sciencedirect.com/journal/applied-computing-and-geosciences



Prediction of rare and anomalous minerals using anomaly detection and machine learning techniques

Abish Sharapatov [a,b], Alisher Saduov [a,*], Nazerke Assirbek [a,b], Madiyar Abdyrov [b], Beibit Zhumabayev [b,c]

[a] Department of Geophysics and Seismology, Satbayev University, Almaty, Kazakhstan
[b] "GeoShar" LLP, Almaty, Kazakhstan
[c] Institute of Ionosphere, Almaty, Kazakhstan

ARTICLE INFO

Keywords:
Anomaly detection
Isolation forest
One-class SVM
KMeans clustering
Mineral composition
Geological processes

ABSTRACT

This study applies machine learning to detect and classify anomalous minerals within a large mineralogical dataset, enhancing geological exploration and resource identification. Using Isolation Forest and One-Class SVM, we identified rare minerals with distinct physical and chemical properties that deviate from common mineral compositions. These anomalies were further grouped using KMeans clustering into three categories, each linked to different geological formation environments: evaporitic, metamorphic, and magmatic processes. The study also evaluates the reliability of these machine learning models using a statistical benchmark and explores the role of deep learning in improving anomaly detection. The findings demonstrate the potential of unsupervised learning to enhance mineral classification, reduce exploration costs, and improve predictive modeling for rare mineral deposits. Future research will refine these methods by integrating Deep Isolation Forest, Autoencoders, and Graph Neural Networks, further strengthening machine learning applications in geosciences.

1. Introduction

The identification and analysis of rare and anomalous minerals play a crucial role in advancing geological research, resource exploration, and industrial applications (Hofstra et al., 2021). Minerals with unique properties provide essential insights into Earth's geological history and contribute to resource mapping, which is invaluable for sustainable exploration and management (Johnson and Emsbo, 2021)These rare minerals often contain economically valuable elements, including rare earth metals and precious compounds that are critical for high-tech industries such as electronics, renewable energy, and aerospace (Golev et al., 2014). Furthermore, understanding the formation and distribution of such minerals enhances geological models and can reveal the conditions under which they formed, offering a deeper perspective on Earth's crustal processes.

Traditionally, geochemical and petrographic methods have been used for detecting mineral anomalies, often relying on factor analysis, multivariate geostatistics, and manual expert interpretation (Luo et al., 2024). However, these approaches have significant limitations in scalability and objectivity, making them less efficient when applied to high-dimensional geochemical datasets. Machine learning (ML) techniques have been increasingly employed to automate the detection and classification of anomalies in mineralogical datasets (Zuo et al., 2020).In this study, Isolation Forest (IF) and One-Class Support Vector Machine (SVM) were employed for anomaly detection due to their robustness in handling high-dimensional and imbalanced data (Zuo et al., 2020). Isolation Forest (an unsupervised machine learning algorithm for anomaly detection that works by recursively splitting the data and isolating outliers based on how quickly they can be separated from the majority of the data) is particularly efficient for large datasets, as it recursively partitions the feature space to isolate outliers. In contrast, One-Class SVM constructs a decision boundary (a machine learning model that learns to distinguish normal samples from anomalies by constructing a decision boundary in a high-dimensional space, classifying points outside this boundary as anomalies) around normal samples, allowing for the detection of subtle deviations. These complementary approaches enhance the accuracy of anomaly detection, capturing both distinct and nuanced patterns within the dataset.

Furthermore, dimensionality reduction (a technique used to simplify complex datasets by reducing the number of features while retaining

* Corresponding author.
E-mail address: a.saduov@satbayev.university (A. Saduov).

https://doi.org/10.1016/j.acags.2025.100250
Received 25 November 2024; Received in revised form 16 March 2025; Accepted 6 May 2025
Available online 10 May 2025
2590-1974/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

important patterns, improving computational efficiency and visualization) using Principal Component Analysis (a dimensionality reduction technique that transforms high-dimensional data into a smaller set of principal components, capturing the most significant variance while preserving essential patterns) (PCA) was applied to reduce dataset complexity while preserving essential geological features. PCA has been widely used in geosciences for feature selection and dimensionality reduction in mineral prospectivity mapping (Luo et al., 2024). Following anomaly detection, KMeans clustering was utilized to group anomalies into distinct clusters, potentially representing unique geological formation environments (Li et al., 2017a,b).

To contextualize the performance and selection of these machine learning algorithms, we present a comparative summary in Table 1. This table outlines the key advantages and applications of each algorithm in geoscientific datasets, emphasizing their suitability for high-dimensional anomaly detection in mineral exploration.

This comparison highlights that each algorithm offers unique strengths, making it essential to choose a method that aligns with the specific dataset characteristics and the nature of anomalies. Isolation Forest and Deep Isolation Forest excel in managing high-dimensional geospatial data, particularly when detecting non-linear relationships, whereas One-Class SVM's ability to identify subtle deviations is beneficial in complex geochemical contexts. Autoencoders, with their feature extraction capabilities, are ideal for large datasets where complex relationships may be present (Tien et al., 2021).

The objectives of this research are twofold: first, to establish a machine learning framework for identifying rare and anomalous minerals based on their physical and chemical attributes; and second, to classify these anomalies into clusters that can provide deeper insight into their geological origins. By applying Isolation Forest and One-Class SVM to detect anomalies and employing KMeans clustering for grouping, this study seeks to bridge traditional geological analysis with modern data-driven approaches. The results contribute to both the academic understanding of rare minerals and practical applications in mineral exploration, setting the stage for future work that incorporates machine learning into geoscience.

2. Materials and methods

2.1. Data description

Data Source: The dataset utilized in this study comprises mineral properties sourced from publicly available databases, specifically Wikipedia, the American Mineralogist Crystal Structure Database (AMCSD), and Kaggle's Comprehensive Database of Minerals. This dataset includes 3112 unique mineral entries, covering an extensive range of chemical compositions, crystal structures, and physical and optical properties. To compile the data, web scraping techniques were employed with Python libraries (BeautifulSoup and requests), allowing structured data extraction directly from online sources. This data, formatted into a CSV, is freely accessible on Kaggle and available for download, ensuring full transparency and reproducibility of the results.

Data Structure: Each mineral entry is represented by 140 columns, encapsulating key physical properties (e.g., crystal structure, Mohs hardness, specific gravity), optical properties (e.g., refractive index), and chemical composition (molar quantities of 87 elements and specific ions like CO_3^{2-}, PO_4^{3-}) as presented in Tables 2 and 3. The inclusion of both general and rare elements supports a detailed examination of mineral properties relevant for machine learning analysis.

Table 1
Summary of key machine learning algorithms in geoscientific applications.

Algorithm	Accuracy (AUC/ Success Rate	Key Advantage	Example Applications
Isolation Forest	80 % (Goldstein and Uchida, 2016)	Efficient for high-dimensional geospatial datasets	Polymetallic deposits, hydrothermal zones
Deep Isolation Forest	>90 % (Saremi et al., 2023)	Superior in handling non-linear data relationships	Hydrothermal copper and polymetallic deposits
One-Class SVM	85 % (AUC) (Goldstein and Uchida, 2016)	High efficiency in detecting geochemical anomalies	Gold prospectivity mapping
Autoencoders	98.36 % (F1) (Ribeiro et al., 2022)	Best for capturing essential features and anomalies	IoT and geosciences, large datasets

2.2. Data preprocessing methods

Effective data preprocessing is a critical step in preparing raw mineral data for machine learning analysis. Given the complexity and high dimensionality of the dataset, proper preprocessing ensures that the data is in an optimal format for further analysis, reducing noise, improving model performance, and enhancing interpretability. The following section outlines the key preprocessing methods applied to the dataset, including *scaling, dimensionality reduction*, and *feature selection*.

2.2.1. Data scaling with StandardScaler

Mineral data often consists of features with varying units and magnitudes. For instance, *Mohs hardness* is measured on a scale from 1 to 10, while *specific gravity and refractive index* are represented by continuous values in a different range. In this scenario, machine learning algorithms may assign higher importance to features with larger numerical ranges, potentially distorting the results. Therefore, to ensure that all features contribute equally to the analysis, *data scaling* was applied.

The *StandardScaler* method was used for this purpose. StandardScaler transforms the data by standardizing it to have a mean of *0* and a standard deviation of *1*, effectively placing all features on the same scale (Lindsay et al., 2019). This approach is particularly beneficial for algorithms such as *Principal Component Analysis (PCA)* and *KMeans clustering*, which are sensitive to the magnitude of the input features.

The scaling procedure can be expressed mathematically as follows:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where.

z is the scaled value;
x is the original feature value;
μ is the mean of the feature;
σ is the standard deviation of the feature.

2.2.2. Dimensionality reduction with Principal Component Analysis (PCA)

In a dataset with numerous features, such as the 87 chemical elements and various physical properties of minerals, dimensionality reduction becomes crucial for simplifying the analysis while retaining the most important information. In this study, Principal Component Analysis (PCA) was applied to reduce the dataset's complexity while preserving its key variability (Liu et al., 2021).The aim of PCA is to identify new, uncorrelated variables known as *principal components*, which capture the maximum variance present in the dataset. This technique is particularly useful when analyzing high-dimensional data, as it allows for the reduction of redundant or less significant features, thereby improving computational efficiency and model performance.

The PCA algorithm works by finding the directions (principal components) along which the variance in the data is maximized. These directions are orthogonal to each other, ensuring that each principal component represents a distinct source of variation in the data. The method relies on the *eigenvalue decomposition* of the covariance matrix of the standardized data.

Given a dataset $\mathbf{X}$ with n observations and p features, PCA seeks to find a linear transformation that projects $\mathbf{X}$ onto a set of orthogonal axes

Table 2
Sample of the minerals dataset.

Name	Crystal Structure	Mohs Hardness	Diaphaneity	Specific Gravity	Optical	Refractive Index	Molar Mass	Molar Volume	Calculated Density
Abenakiite-(Ce)	5.0	4.50	0.0	3.240	3.0	1.580	817.34	0.123	5.50
Abernathyite	4.0	2.75	3.0	3.446	3.0	1.592	435.07	0.056	6.44
Abhurite	5.0	2.00	3.0	4.420	3.0	2.085	921.09	0.122	6.23
Actinolite	2.0	5.50	2.0	1.050	4.0	1.634	861.19	0.112	6.39

Table 3
Key features and codes for mineral dataset.

Feature	Description	Code/Scale
Crystal Structure	Classification of the crystal system	1 = Triclinic, 2 = Monoclinic, 3 = Orthorhombic, 4 = Tetragonal, 5 = Hexagonal, 6 = Trigonal, 7 = Cubic, 8 = Amorphous
Mohs Hardness	Hardness on Mohs scale (1–10)	1 to 10
Specific Gravity	Density ratio relative to water	Numeric (e.g., 2.65 for Quartz)
Optical Properties	Behavior of light passing through the mineral	1 = Anisotropic, 2 = Isotropic, 3 = Uniaxial, 4 = Biaxial
Refractive Index	The refractive index of the mineral	Numeric
Oxygen (O)	Number of oxygen atoms per formula unit	Numeric (e.g., 2 for Quartz)
Silicon (Si)	Number of silicon atoms per formula unit	Numeric (e.g., 1 for Quartz)
Iron (Fe)	Number of iron atoms per formula unit	Numeric
Molar Mass	The molecular weight of the mineral	Numeric (e.g., 60.08 for Quartz)
Molar Volume	Volume occupied by 1 mol of the mineral	Numeric
Carbonate (CO_3^{2-})	Presence of carbonate ion	1 if present, 0 if absent
Phosphate (PO_4^{3-})	Presence of phosphate ion	1 if present, 0 if absent

(principal components). This is achieved by calculating the *covariance matrix* of the data, denoted by $\mathbf{\Sigma}$, which is defined as:

$$\sum = \frac{1}{n-1} X^T X \tag{2}$$

The eigenvectors of the covariance matrix correspond to the directions of the principal components, while the eigenvalues represent the variance captured by each principal component. For each eigenvalue $\boldsymbol{\lambda}$, there is an associated eigenvector $\mathbf{v}$, which satisfies the following equation:

$$\sum \vartheta = \lambda \vartheta \tag{3}$$

Here, $\boldsymbol{\lambda}$ represents the variance explained by the principal component corresponding to eigenvector $\mathbf{v}$. The principal components are ranked according to the eigenvalues, with the first principal component explaining the largest amount of variance in the dataset, the second principal component explaining the second-largest amount, and so on.

The data is then projected onto the k principal components that explain the most variance, where k is typically chosen to retain a predefined amount of variance (e.g., 95 %). The projection of the data $\mathbf{X}$ onto the new principal component space is given by:

$$Z = XW \tag{4}$$

Where W is the matrix of eigenvectors corresponding to the top k principal components, and Z represents the transformed dataset in the lower-dimensional space.

In this study, the number of principal components was selected based on a scree plot analysis (Fig. 1), which illustrated the variance captured

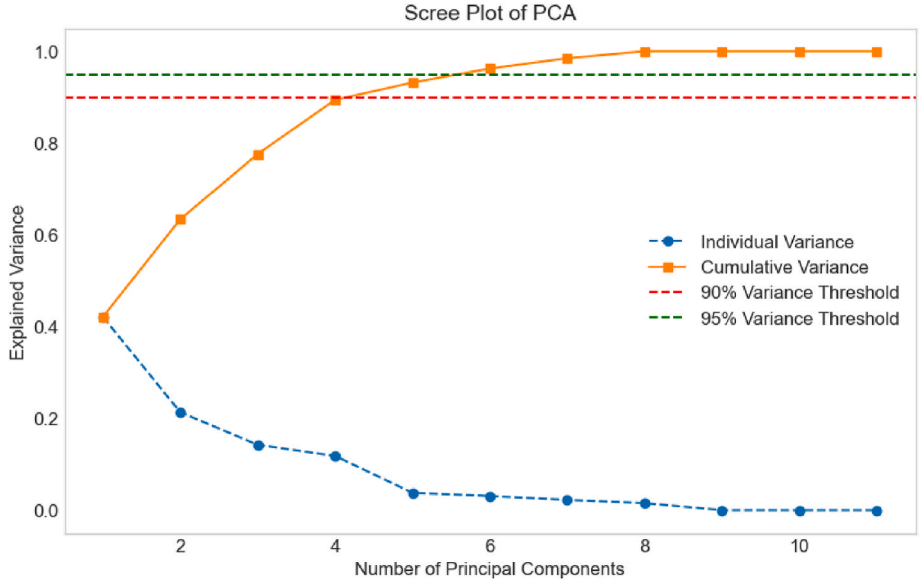


Fig. 1. Scree Plot of PCA: This scree plot illustrates the variance explained by each principal component. The blue dashed line represents the individual variance captured by each component, while the orange solid line shows the cumulative variance. The red and green dashed lines indicate the 90 % and 95 % variance thresholds, respectively. Based on this plot, the first five components account for 90 % of the total variance, while six components capture 95 %. Therefore, for machine learning tasks, we selected five to six principal components, whereas for visualization, only the first two components were used. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

by each component. The first five components accounted for 90 % of the total variance, while six components captured 95 %. Consequently, for machine learning tasks such as anomaly detection, we retained five to six principal components. However, for visualization purposes, only the first two principal components were used to generate two-dimensional PCA plots. This approach allowed for meaningful graphical representation while ensuring that machine learning models had sufficient feature representation.

PCA is not a feature selection method; rather, it reduces dimensionality by creating new uncorrelated principal components that capture the variance present in the data. Unlike feature selection, which retains original variables, PCA transforms them into a new coordinate system where each principal component is a weighted sum of the original features. This distinction is important, as PCA was not used to remove any features but rather to project the data into a more interpretable lower-dimensional space for both visualization and modeling purposes.

Moreover, PCA is particularly valuable in geoscientific data where the relationships between chemical and physical properties can be highly correlated (Su et al., 2023). By transforming the original features into uncorrelated principal components, PCA ensures that these correlations do not distort the results of the subsequent machine learning algorithms. For instance, features such as *specific gravity* and *molar mass* are likely to exhibit high correlation, which can be addressed effectively through PCA.

The application of PCA in this study involved the following steps.

1. *Standardization:* All features were first standardized to ensure that they had a mean of zero and a standard deviation of one, as PCA is sensitive to the scale of the input features;
2. *Covariance Matrix Calculation:* The covariance matrix was computed to capture the relationships between the features;
3. *Eigenvalue Decomposition***:** The eigenvalues and eigenvectors of the covariance matrix were calculated, with the eigenvectors representing the principal components;
4. *Selection of Principal Components***:** The top k principal components were selected based on their associated eigenvalues, ensuring that the selected components explain a significant portion of the variance in the dataset;
5. *Projection:* The original data was projected onto the lower-dimensional space defined by the selected principal components.

By retaining only the most important principal components, this study preserved the essential patterns and structures in the data while reducing its dimensionality. This approach not only made the data more manageable for machine learning algorithms but also allowed for more meaningful insights into the underlying relationships between the physical and chemical properties of minerals.

2.2.3. Feature selection for analysis

Feature selection plays a pivotal role in ensuring the quality and accuracy of machine learning models, especially when dealing with high-dimensional datasets such as the one used in this study (Zhang et al., 2022).With *140 features*, representing both physical properties and chemical compositions of 3112 minerals, it was crucial to focus on the most relevant attributes to reduce noise, prevent overfitting, and improve the model's interpretability. The selection process was driven by both domain knowledge and statistical techniques, ensuring that the chosen features would maximize the efficiency of the subsequent machine learning tasks.

Initially, the dataset comprised various *physical properties* of minerals, such as *Mohs hardness, specific gravity, refractive index,* and *crystal structure*, alongside the *chemical composition* representing the molar quantities of 87 elements and several ionic species. The goal of feature selection was to identify and retain the most significant physical and chemical attributes that contribute to differentiating between rare and common minerals and enable the detection of anomalies.

In selecting features for this study, the primary criteria included relevance to the research objective - detecting rare or anomalous minerals - while maintaining a balance between the physical and chemical properties of the minerals. *Physical properties*, such as *Mohs hardness* and *specific gravity*, are fundamental in describing the physical behavior of minerals and their interaction with the environment. These properties are widely used in mineral classification and are critical in understanding the mineral's potential industrial application and geological significance. Similarly, *refractive index* and *crystal structure* were included due to their importance in determining optical and structural characteristics of minerals, which can provide clues about their formation and environmental conditions.

On the chemical side, the elemental composition of each mineral provides the foundation for understanding its chemical behavior and classification. Chemical elements were selected based on their geochemical significance, such as *silicon, oxygen, iron,* and *calcium,* which are dominant in many silicate and oxide minerals. These elements were chosen for their geological relevance and their ability to distinguish between different mineral groups. Additionally, elements associated with rare or economically important minerals, such as *neodymium* and *thorium*, were retained to ensure that the feature set could support the identification of both common and rare mineral species.

It is important to note that PCA was not used for feature selection in this study. Instead, the selection of relevant features was based on statistical criteria, including variance thresholding and correlation analysis. Although PCA provided insights into feature importance, the actual feature selection process focused on retaining the most informative original variables without altering their structure.

Statistically, feature selection was further refined through the application of *Principal Component Analysis (PCA)*. PCA provided insights into which features contributed the most variance in the dataset, allowing for a reduction in dimensionality without significant loss of information. The eigenvalues associated with each principal component indicated the proportion of variance explained, helping to determine the most critical features. This step ensured that redundant or highly correlated features, such as molar mass and specific gravity, were either combined or removed, leaving only those attributes that were most informative for the machine learning algorithms.

The final set of features selected for the analysis included a balanced mix of physical properties and chemical compositions, ensuring that the models could capture both structural and compositional variations in the dataset. By focusing on features that contribute the most to distinguishing between minerals, the feature selection process helped reduce the computational complexity of the analysis while maintaining the ability to identify rare and anomalous minerals effectively.

Feature selection is a fundamental step in machine learning applications, particularly when working with high-dimensional data. Its primary goal is to enhance model efficiency, reduce redundancy, and maintain interpretability without compromising predictive power (Li et al., 2017a,b,b). In this study, feature selection was performed using a combination of domain expertise and statistical techniques to ensure that the most relevant physical and chemical attributes were retained while avoiding excessive multicollinearity and noise.

For physical properties, key mineralogical attributes such as Mohs hardness, specific gravity, refractive index, and crystal structure were selected due to their established role in mineral classification and differentiation. Chemical composition features were chosen based on geological significance, ensuring that elements contributing to key geochemical processes were retained. Elements with negligible variance (i.e., occurring in fewer than 1 % of minerals) were removed to prevent sparsity issues.

To further refine the feature selection process, correlation analysis was performed to evaluate redundancy among physical features. Principal Component Analysis (PCA) was used only as an exploratory tool to assess feature importance but was not utilized as a feature selection

method. This process ensured that the final set of attributes preserved the interpretive power of the dataset while maintaining computational efficiency.

This carefully curated feature set formed the basis for the anomaly detection and clustering tasks that followed, providing a robust framework for analyzing the unique and often subtle differences between common and rare minerals. Through this approach, the study aimed to maximize the model's interpretive power while ensuring that the selected features were scientifically meaningful and relevant to both geoscientific and industrial applications.

2.2.4. Statistical definition of anomalies

To establish a quantitative baseline for anomaly detection, we applied a statistical threshold using standard deviation (σ). A sample was classified as anomalous if its value exceeded two standard deviations (2σ) from the mean in at least one of the three key mineralogical properties.

- Specific Gravity,
- Refractive Index,
- Molar Mass.

This threshold is a widely accepted statistical method for identifying outliers in high-dimensional geoscientific datasets (Barnett and Lewis, 1994; Leys et al., 2013). This statistical reference was used to assess the precision and recall of the machine learning models, particularly Isolation Forest and One-Class SVM, in detecting these anomalies.

2.2.5. Model evaluation metrics

To assess the performance of the anomaly detection models, we employed several widely used classification metrics: Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUROC). These metrics are particularly relevant for imbalanced datasets, such as mineralogical anomaly detection, where the number of anomalous samples is significantly smaller than the total dataset (Chawla et al., 2004; Saito and Rehmsmeier, 2015).

- Precision measures the proportion of correctly identified anomalies among all detected anomalies. Higher precision indicates fewer false positives.
- Recall (also known as sensitivity) quantifies the proportion of actual anomalies that were successfully identified. Higher recall implies fewer false negatives.
- F1-score is the harmonic mean of precision and recall, providing a balanced measure of detection performance.
- AUROC evaluates how well the model differentiates between normal and anomalous samples across various decision thresholds, making it particularly useful for unsupervised anomaly detection (Fawcett, 2006).

Given the absence of labeled ground truth anomalies in our dataset, we established a statistical benchmark using the 2σ threshold, which identifies minerals with at least one extreme deviation in Specific Gravity, Refractive Index, or Molar Mass (Barnett and Lewis, 1994; Leys et al., 2013). This threshold serves as an independent reference to compare machine learning predictions, ensuring a more rigorous evaluation of model performance.

2.3. Machine learning algorithms

2.3.1. Mathematical foundation of anomaly detection algorithms

In this study, *Isolation Forest* and *One-Class* Support *Vector Machine (SVM)* were selected as the primary algorithms for anomaly detection due to their proven effectiveness in dealing with high-dimensional data and their robustness in identifying rare and anomalous samples. Both of these algorithms are designed specifically for anomaly detection tasks and have demonstrated superior performance in various applications, particularly when dealing with imbalanced datasets, where anomalies (outliers) are far less frequent than normal observations (Liu et al., 2008; Schölkopf et al., 2001).

Isolation Forest is an ensemble-based anomaly detection algorithm that isolates outliers using random partitioning. It operates under the assumption that anomalies are more easily isolated than normal instances, meaning they require fewer splits in a binary tree (Liu et al., 2008).

The anomaly score $S(x,n)$ is computed as follows (Liu et al., 2008):

$$S(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{5}$$

where.

$E(h(x))$ is the expected path length of x in the isolation trees,
$c(n)$ is the average path length of a balanced binary tree of size n.

Shorter paths correspond to higher anomaly scores, allowing the method to efficiently separate outliers from normal data points. Studies have demonstrated the effectiveness of Isolation Forest in geochemical anomaly detection and large-scale datasets (Xu et al., 2022).

One-Class SVM is a kernel-based method that models normal data and classifies deviating points as anomalies. Instead of recursively partitioning data, it constructs a hyperplane that encloses the normal observations, treating anything outside as an outlier (Schölkopf et al., 2001).

The optimization problem solved by One-Class SVM is:

$$\min_{\omega,\xi,\rho} \frac{1}{2}\|\omega\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \tag{6}$$

subject to:

$$(\omega \cdot \phi(x_i)) \geq \rho - \xi_i, \xi_i \geq 0 \tag{7}$$

where:

ω is the hyperplane (a decision boundary in a high-dimensional space that separates different classes in machine learning models, such as One-Class SVM, to distinguish normal and anomalous data points) weight vector,
ρ is the offset,
$\phi(x_i)$ maps input data into a high-dimensional space,
ν controls the fraction of expected outliers,
ξ_i are slack variables allowing soft margins.

2.3.2. Justification for algorithm selection

The decision to use Isolation Forest and One-Class SVM was based on their complementary strengths in anomaly detection (Aggarwal, 2017).

- Isolation Forest efficiently detects globally distinct anomalies, making it scalable to large datasets.
- One-Class SVM is sensitive to smaller, more nuanced deviations, capturing anomalies that are close to the normal class but exhibit slight irregularities.

Several alternative machine learning methods were considered but

Table 4
Sample of the minerals dataset.

Method	Reason for Exclusion
KMeans	Assumes clusters are spherical; cannot detect anomalies effectively (Jain, 2010).
DBSCAN	Sensitive to distance thresholds; struggles in high-dimensional spaces (Ester et al., 1996).
Gaussian Mixture Model (GMM)	Assumes Gaussian-distributed clusters, which is unrealistic for mineralogical datasets (Reynolds, 2009).

ultimately found to be less effective (Kriegel et al., 2011; Aggarwal, 2017) (Table 4):

By combining Isolation Forest's efficiency in detecting clear outliers and One-Class SVM's ability to detect subtle deviations, our approach ensures a robust and interpretable anomaly detection framework for rare mineral identification.

2.3.3. KMeans for clustering

After detecting anomalies using *Isolation Forest* and *One-Class SVM*, the next logical step was to analyze whether there were discernible subgroups or patterns among the identified anomalous minerals. For this purpose, *KMeans* clustering was employed to group the anomalies into distinct clusters. KMeans is a widely used algorithm for partitioning data into clusters based on feature similarities, making it well-suited for analyzing the characteristics of anomalies post-detection (Ahmed et al., 2020).

Although *KMeans* may not be ideal for direct anomaly detection in high-dimensional spaces, its application in clustering already-identified anomalies proves to be both effective and efficient (Pérez-Ortega et al., 2018). To ensure that *KMeans* could optimally group these anomalies, the dimensionality of the data was first reduced through *PCA*, focusing on the most significant features and making the data more manageable for clustering (Chouhan, 2021).

The optimal number of clusters was determined using the *Elbow Method*, which involves plotting the sum of squared distances (inertia) within clusters across various values of k (the number of clusters) and identifying the point where the decrease in inertia begins to level off. This "elbow" point, which in our case was found to be at $k = 3$, suggested that three clusters would provide a good balance between detail and interpretability (Vandna et al., 2020). Consequently, KMeans was applied with three clusters to analyze the structure within the subset of anomalies.

Mathematically, *KMeans* works by minimizing the *within-cluster variance* through iterative steps.

1. Randomly initializing k centroids (in this case, k = 3);
2. Assigning each data point to the nearest centroid based on the *Euclidean distance*;
3. Recomputing the centroids based on the new cluster assignments;
4. Repeating the assignment and centroid update steps until convergence

The formula for the assignment step, which assigns each data point x_i to the cluster C_j with the nearest centroid μ_j, is:

$$C_j = \arg\, min_j ||x_i - \mu_j||^2 \tag{8}$$

By clustering the anomalies, we were able to gain deeper insights into potential subgroups, revealing shared physical or chemical characteristics among the minerals within each cluster. For instance, some clusters may contain minerals with similar crystal structures or elemental compositions, providing valuable information for further geological or industrial investigations (Lesouple et al., 2021). This use of KMeans for clustering, rather than anomaly detection, highlights its utility in identifying patterns and similarities within the unique subset of anomalous minerals.

3. Results

3.1. PCA results

To explore the distribution of mineral data and reduce the dimensionality of the dataset for visual interpretation, *Principal Component Analysis (PCA)* was applied, identifying that five to six principal components are necessary to retain 90–95 % of the total variance (Fig. 1). However, for visualization purposes, a two-dimensional representation was used to facilitate interpretability. This approach allows for an intuitive exploration of the dataset while ensuring that machine learning models operate in a more comprehensive feature space.

The *PCA plot* Fig. 2 provides a two-dimensional visualization of the dataset, with each point representing an individual mineral sample. This visualization allows us to observe general patterns, clusters, and potential outliers within the dataset. However, it is important to note that the anomaly detection models were trained using five principal components, ensuring that the analysis retained the majority of the dataset's variance while avoiding excessive dimensionality reduction.

3.1.1. Interpretation of the PCA plot

In Fig. 2, the *x-axis* represents the first principal component, while the *y-axis* represents the second principal component. Together, these two principal components explain 63.39 % of the total variance, which is sufficient for initial visualization but does not encompass the full complexity of the dataset.The majority of minerals are concentrated in dense clusters within the lower-left quadrant of the plot, suggesting similarity in their physical and chemical properties.

Several points are notably distant from the main clusters, indicating that these minerals may exhibit unique characteristics or compositions. Although PCA itself does not directly identify anomalies, these outliers can be visually inspected as potential candidates for further anomaly detection. For this reason, a higher-dimensional PCA transformation (with five components explaining 90 % of variance) was used for machine learning models to improve anomaly detection accuracy.

This dimensionality reduction enables a preliminary analysis of the dataset, setting the foundation for further exploration with dedicated anomaly detection algorithms like *Isolation Forest* and *One-Class SVM*, which will rigorously classify these outliers.

3.2. Anomaly detection results

In this section, we present the results of anomaly detection using two machine learning algorithms: Isolation Forest and One-Class SVM. Both algorithms were employed to identify mineral samples that exhibit distinct physical and chemical properties relative to the overall dataset. The outcomes of each algorithm are discussed below, along with visual representations of detected anomalies in the PCA space.

3.2.1. Anomalies detected by Isolation Forest and One-Class SVM

The *Isolation Forest* algorithm identified a total of *156 anomalous mineral samples*. In contrast, the *One-Class SVM* algorithm detected a larger set of anomalies, marking *663 samples* as potentially anomalous. This difference in the number of identified anomalies reflects the unique characteristics of each algorithm in handling high-dimensional data and detecting outliers based on different underlying assumptions.

3.2.2. Visualization of anomalies in PCA space

To gain further insights into the distribution of anomalies detected by each algorithm, we projected the detected anomalies into PCA space Fig. 3. These visualizations highlight how anomalies identified by *Isolation Forest* and *One-Class* SVM are positioned relative to the normal mineral samples. For visualization purposes, the first two principal components (PC1 and PC2) were used, explaining a total of 63.39 % of the variance, ensuring a clear representation of detected anomalies in two-dimensional space.

In Fig. 3a, the anomalies detected by Isolation Forest are marked in red, with most of the anomalies clustering around the edges of the main mineral cluster. This distribution reflects Isolation Forest's tendency to isolate outliers that are distanced from dense clusters. In Fig. 3b, the anomalies identified by One-Class SVM are marked inred, showing a denser spread of anomalous points, which indicates the algorithm's sensitivity to subtle deviations within clusters. The distinct distributions of anomalies by each method underscore the complementary nature of these algorithms in identifying outliers in complex, high-dimensional

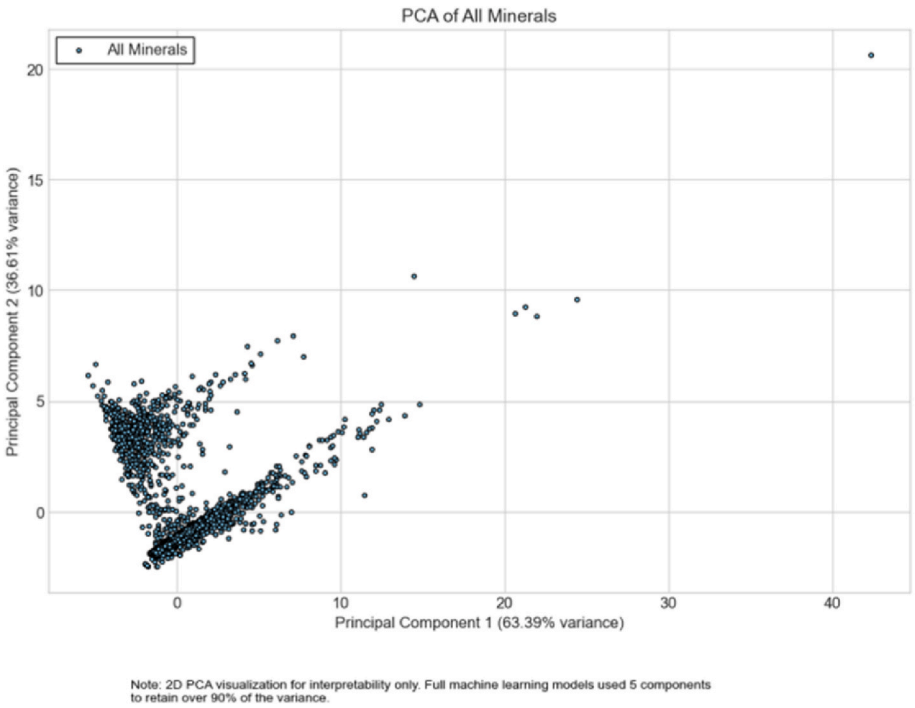


Fig. 2. PCA of All Minerals: This two-dimensional PCA plot illustrates the distribution of mineral samples based on their principal components. The x-axis represents the first principal component explaining 63.39 % of the total variance, and the y-axis represents the second, explaining 36.61 %. Dense clusters in the lower-left quadrant suggest similarity in mineral properties, while points positioned farther from the primary group may indicate minerals with unique physical or chemical characteristics, warranting further anomaly analysis. For anomaly detection and clustering, five principal components were used, capturing over 90 % of the variance, ensuring a more comprehensive representation of mineralogical diversity.

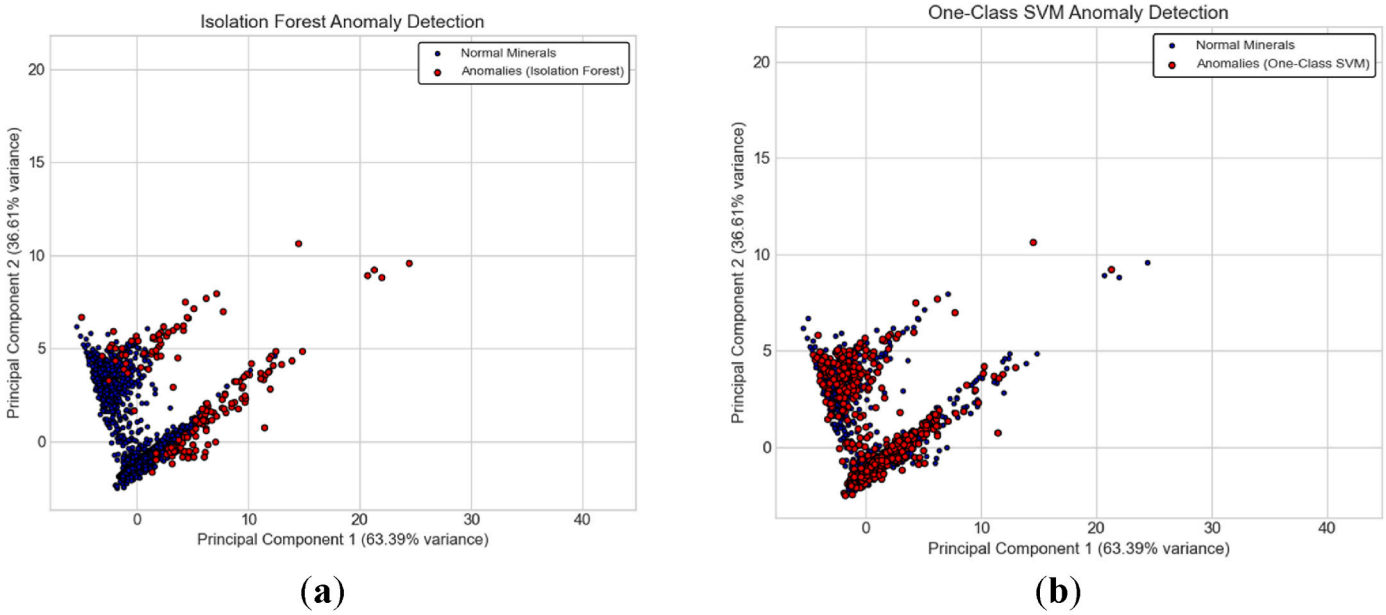


Fig. 3. (**a**) Isolation Forest Anomaly Detection: This figure shows the distribution of normal minerals (blue) and anomalies (red) detected by the Isolation Forest algorithm in PCA space. The x-axis represents the first principal component (PC1, explaining 63.39 % variance), while the y-axis represents the second principal component (PC2, explaining 36.61 % variance). Anomalies tend to cluster on the periphery of the main data distribution; (**b**) One-Class SVM Anomaly Detection: This figure displays the distribution of normal minerals (blue) and anomalies (green) identified by the One-Class SVM algorithm in PCA space. The x-axis represents PC1 (63.39 % variance), and the y-axis represents PC2 (36.61 % variance). Anomalies are more widely dispersed, reflecting One-Class SVM's sensitivity to variations within clusters. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

datasets (Ribeiro et al., 2022).

It is important to highlight that while only two principal components are displayed in Fig. 3 for visualization purposes, the anomaly detection process was conducted using five principal components. These five components capture over 90 % of the total variance, ensuring that the models retain the majority of the data's structure while reducing dimensionality. This approach balances interpretability with analytical rigor, allowing for effective anomaly detection while providing a

meaningful graphical representation of the data.

It is important to note that Fig. 3 includes both normal and anomalous minerals, which explains the higher number of data points in comparison to subsequent clustering visualizations. In contrast, clustering (Section 3.3) was performed exclusively on the subset of minerals identified as anomalies by either Isolation Forest or One-Class SVM.

Also, this visualization is based on only two principal components (PC1 and PC2) for interpretability, capturing 63.39 % of the variance. However, for anomaly detection and clustering, five principal components were used, explaining over 90 % of the total variance to ensure accurate representation of data structure.

3.2.3. Statistical anomaly detection

To provide a quantitative baseline for anomaly detection, we applied a 2σ threshold to identify statistically significant outliers in the dataset. This analysis revealed that 564 mineral samples exceeded the 2σ threshold in at least one of the key mineralogical properties (Specific Gravity, Refractive Index, Molar Mass).

A common characteristic among these statistically identified anomalies is that they often exhibit either significantly higher or lower density, molar mass, or refractive index compared to the general mineral population. Minerals falling into this category frequently contain heavy rare-earth elements (e.g., Europium, Gadolinium) or exhibit unusual crystallographic properties, suggesting formation under extreme geological conditions.

This statistical result serves as a reference for evaluating the performance of Isolation Forest and One-Class SVM in detecting anomalies. Comparing this with the machine learning results, we observe that.

- Isolation Forest identified 156 anomalies, capturing a subset of the statistical anomalies.
- One-Class SVM detected 663 anomalies, encompassing nearly all statistical anomalies but also potentially including false positives.

Interestingly, while 82 % of the anomalies detected by Isolation Forest fall within the 564 statistically defined anomalies, One-Class SVM includes an additional set of minerals that slightly deviate from the general mineralogical trends but do not exceed the strict 2σ threshold. This suggests that One-Class SVM is more sensitive to subtle variations in mineral composition, whereas Isolation Forest isolates the most extreme deviations.

These findings suggest that the true number of anomalous minerals likely lies between 564 and 663, with One-Class SVM being more inclusive and Isolation Forest more conservative.

3.2.4. Model performance evaluation

To quantitatively assess the performance of the anomaly detection models, we compared their outputs to the 2σ statistical threshold, treating the 564 statistically identified anomalies as a reference. Table 5 presents the precision, recall, F1-score, and AUROC values for each model.

Isolation Forest exhibited high precision (1.000), meaning that nearly all detected anomalies were statistically confirmed. However, its low recall (0.277) indicates that many true anomalies were missed. This aligns with its tendency to be more conservative, identifying only the most distinct outliers.

Conversely, One-Class SVM demonstrated high recall (1.000), capturing all statistical anomalies. However, its lower precision (0.851) suggests a higher rate of false positives. The significantly higher AUROC (0.981) for One-Class SVM implies superior ability in distinguishing anomalies from normal samples across various thresholds (Saito and Rehmsmeier, 2015).

Table 5
Model performance evaluation.

Model	Precision	Recall	F1-score	AUROC
Isolation Forest	1.000	0.277	0.433	0.638
One-Class SVM	0.851	1.000	0.919	0.981

These results highlight a trade-off: Isolation Forest provides high confidence but conservative anomaly detection, while One-Class SVM maximizes recall at the cost of some false positives. This suggests that an ensemble approach, combining both models, could yield an optimal balance of precision and recall.

3.2.5. Correlation analysis of features

To validate the feature selection process, a correlation heatmap was generated to assess relationships among the selected physical properties. The heatmap on Fig. 4 confirms strong correlations (>0.7) among Mohs hardness, specific gravity, refractive index, and crystal structure, supporting their inclusion in the dataset. Additionally, molar mass and molar volume exhibit a correlation of 0.75, indicating some level of redundancy. However, given that molar mass represents atomic-scale composition while molar volume reflects macroscopic spatial characteristics, both were retained due to their distinct interpretational significance (Kuhn and Johnson, 2019).

For chemical composition, a full correlation matrix was initially generated but was excluded from visualization due to excessive dimensionality. Instead, the selection of chemical elements focused on geochemically significant components, ensuring that rare elements contributing to anomaly detection remained in the dataset while minimizing redundancy.

3.3. Results of KMeans clustering

3.3.1. Clustering anomalous minerals

The KMeans clustering was applied exclusively to the samples that were previously identified as anomalies. These anomalies were selected based on the results of the Isolation Forest and One-Class SVM methods. The union of anomalies detected by both methods was used as the input for clustering. The application of PCA helped to reduce dimensionality and improve cluster separation.

Only the samples classified as anomalous by at least one of the anomaly detection methods were included in clustering (Fig. 6). Normal mineral samples were excluded from clustering, as the goal was to analyze patterns within the detected anomalies.

Using the *Elbow Method* Fig. 5, we determined that $k = 3$ provided the most efficient partitioning, minimizing within-cluster variance. Post-PCA dimensionality reduction, KMeans effectively grouped the anomalies into three distinct clusters, as illustrated in Fig. 6. This clustering helps reveal inherent patterns and shared characteristics among the anomalies, offering deeper insights into potential geological formation processes (Talusani et al., 2008).

3.3.2. Characteristics of each cluster

The characteristics of each cluster reflect distinct geological and mineralogical contexts, providing insight into the environments and processes contributing to their formation.

1. *Cluster 1*: This cluster predominantly includes minerals with specialized crystal structures and unusual compositions, often associated with highly alkaline or evaporitic environments. Notable minerals like Natron, Hydrohalite, and Lanarkite suggest formation in hypersaline lakes or evaporite deposits, where high concentrations of specific ions and extreme pH levels facilitate the crystallization of such unique compositions. The presence of sulfate and halide minerals points to arid and evaporative conditions, where water evaporation precipitates these rare mineral forms;
2. *Cluster 2*: Minerals in this cluster are silicate-rich and commonly linked to metamorphic or metasomatic environments. Minerals such as Afghanite, Charoite, and Stilpnomelane are typically formed under high-pressure and moderate-temperature conditions, often in

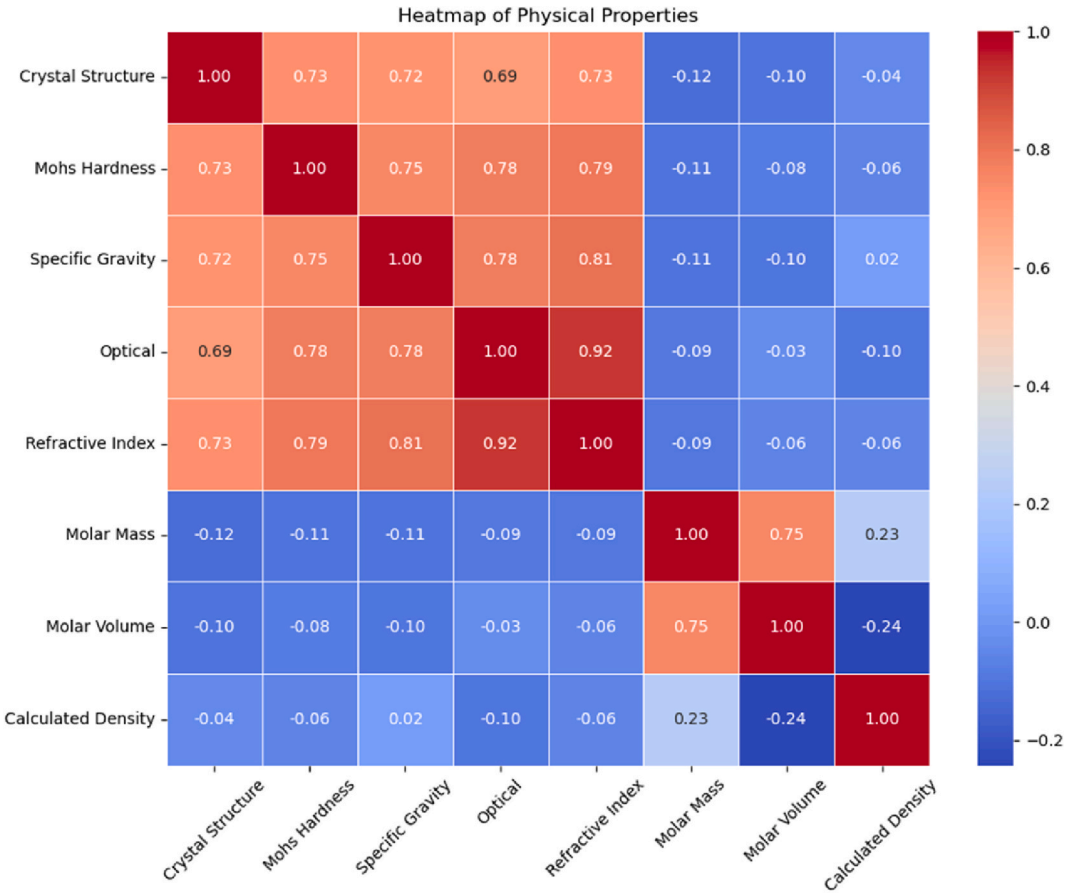


Fig. 4. Heatmap of physical properties.

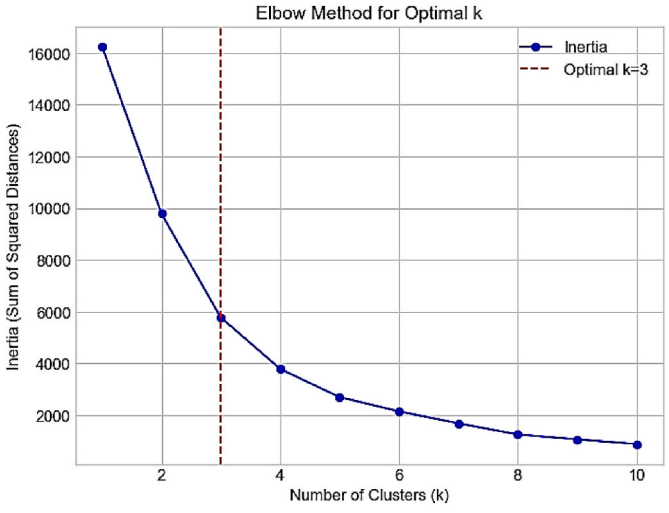


Fig. 5. Elbow Method for Optimal k: This figure illustrates the Elbow Method applied to determine the optimal number of clusters for KMeans clustering. The blue line represents the inertia (sum of squared distances) for each tested number of clusters, and the red dashed line marks the optimal k = 3, where the curve shows a notable "elbow," indicating the most efficient clustering solution. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

regions affected by tectonic activity or hydrothermal alteration. This cluster may represent minerals that undergo recrystallization due to thermal and mechanical stresses, reflecting crustal processes where silicate structures stabilize under varying pressure and temperature. The predominance of aluminosilicate phases in this cluster further suggests deep crustal origins or metamorphic transformations within subduction zones;

3. Cluster 3: Minerals in this cluster are derived from both igneous and sedimentary environments. The presence of minerals such as *Beryl*, *Zircon*, and *Corundum* indicates a primary magmatic origin, suggesting crystallization from silicate melts under high-temperature conditions. Other minerals in this cluster, like *Cerussite* and *Bromellite*, imply secondary formation processes, including weathering and oxidation in the Earth's surface layers. This diversity suggests that Cluster 3 captures minerals that have undergone multiple geological processes, reflecting both primary crystallization from magmas and secondary alteration in sedimentary or oxidizing environments.

The identified clusters not only reflect distinct geological environments but also provide valuable insights for mineral exploration and resource assessment. Their practical applications include guiding mineral prospecting, refining geological models, and improving the efficiency of mineral extraction processes.

Cluster 1, which predominantly includes minerals such as Natron and Hydrohalite, is significant for identifying evaporitic environments where high ion concentrations lead to crystallization. These settings are often associated with lithium-rich brine deposits, crucial for battery production and renewable energy applications. The clustering of these minerals highlights regions that may be promising for future lithium exploration, particularly in arid or semi-arid regions where such deposits are known to form.

Cluster 2 consists of silicate-rich minerals commonly linked to metamorphic and hydrothermal environments. This cluster is of particular importance for predicting ore-bearing zones, as metasomatic

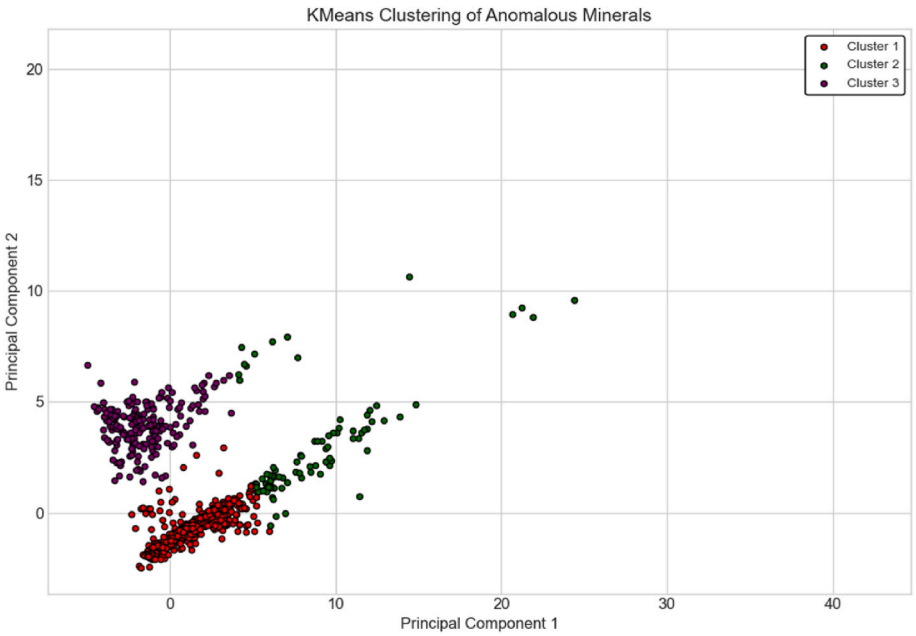


Fig. 6. KMeans Clustering of Anomalous Minerals After PCA Dimensionality Reduction: This figure presents the clustering results of anomalous minerals after applying PCA-based dimensionality reduction. The x-axis and y-axis represent the first and second principal components, respectively. These two components capture 63.39 % of the total variance, making the visualization informative while acknowledging that higher-dimensional features were used in model training. The clustering was performed using KMeans (k = 3), as determined by the Elbow Method. Each cluster highlights a distinct subgroup of anomalous minerals, reflecting potentially different geological formation processes.

and hydrothermal processes are key drivers of mineralization for gold, copper, and rare earth elements. The identification of this cluster may assist in prioritizing regions with high exploration potential, particularly within subduction zones or tectonically active regions.

Cluster 3 includes minerals such as Beryl, Zircon, and Corundum, which are typically associated with high-temperature magmatic systems. These minerals serve as indicators of primary igneous deposits, particularly pegmatites and kimberlites, which are primary sources of gemstones, strategic metals, and industrial minerals. Additionally, the presence of secondary minerals like Cerussite and Bromellite suggests that this cluster captures zones of supergene enrichment, where weathering processes concentrate valuable metals, improving the economic viability of mineral deposits.

4. Discussion

This study provides a detailed exploration into the detection and clustering of anomalous minerals using machine learning methods, revealing significant geological implications. Isolation Forest and One-Class SVM, applied to high-dimensional mineralogical data, enabled the identification of rare minerals characterized by unique physical and chemical properties. Among these, Specific Gravity and Refractive Index emerged as particularly distinguishing features. Minerals identified as anomalies often exhibited higher-than-average values in these areas, potentially indicating high atomic weight elements or complex bonding structures, which, in turn, suggest formation under specific, often extreme, geological conditions. The presence of elements like Europium or other rare compounds among these anomalies aligns with known processes in high-pressure, high-temperature environments, where unique bonding configurations can arise due to extreme geological constraints (Liu et al., 2023).

The clustering analysis further facilitated a refined geological interpretation by dividing the anomalous minerals into three distinct groups, each reflecting unique formation environments and geological processes. *Cluster 1,* containing minerals with high occurrences of halides and sulfates, indicates formation in highly saline or evaporitic settings. This environment, marked by high ion concentration and alkalinity, often leads to the precipitation of minerals such as Natron and Hydrohalite, common in evaporite deposits where water evaporation concentrates ions in the remaining fluid. These conditions are typically associated with arid or semi-arid climates, where rapid evaporation drives mineral precipitation, hinting at ancient geological settings akin to modern saline basins. The mineralogical characteristics in this cluster support interpretations of restricted marine or lacustrine environments that produce mineral deposits through prolonged evaporation cycles.

Cluster 2 primarily encompasses silicate minerals rich in elements like aluminum and magnesium, which are indicative of high-pressure and moderate-temperature metamorphic settings. Minerals such as Afghanite and Charoite in this cluster are typically formed in regions undergoing intense tectonic activity or hydrothermal alteration. This suggests the potential for these minerals to serve as indicators of tectonically active regions where pressure-induced recrystallization dominates. Their formation in tectonic or metamorphic zones underscores the dynamic geological processes that shape mineral characteristics through recrystallization under varying conditions of pressure and temperature. Cluster 2's mineral compositions suggest geological environments associated with subduction zones or orogenic belts, where silicate structures are subjected to significant tectonic forces that drive mineral transformation and formation.

Cluster 3 presents a mix of igneous and secondary minerals, including Beryl, Zircon, and Corundum, which are typical of primary magmatic environments. These minerals form under high-temperature conditions within silicate melts, often found in igneous intrusions and associated with magmatic differentiation processes. Other minerals in this cluster, such as Cerussite and Bromellite, suggest secondary formation through surface weathering and oxidation. This combination of primary and secondary minerals implies that Cluster 3 captures a spectrum of minerals with origins ranging from deep magmatic processes to surface alterations driven by weathering. The presence of both magmatic and altered minerals in this cluster points to geological environments where igneous activity is followed by subsequent weathering, highlighting the influence of both internal and surface processes in mineral formation.

The clustering results provide direct applications for mineral exploration, resource classification, and economic evaluation of mineral deposits. By distinguishing mineral assemblages associated with different geological settings, this study presents a data-driven framework for identifying prospective mineral deposits. Cluster 1 highlights potential sites for evaporitic lithium and rare earth element deposits, which are increasingly in demand due to their role in renewable energy technologies. Cluster 2, with its association with metasomatic and hydrothermal mineralization, suggests potential zones of economic interest for base and precious metals, particularly in subduction-related geological settings. Cluster 3, which includes both primary igneous minerals and secondary weathering products, highlights mineral deposits formed by magmatic differentiation and later enrichment through oxidation and leaching.

In a broader geoscientific context, these clustering results can be utilized in mineral deposit classification, improving geological mapping by refining the understanding of mineralogical zonation in different geological settings. The ability to distinguish between mineral assemblages through machine learning enhances the efficiency of exploration campaigns, reducing the need for extensive field sampling by providing preliminary insights into likely mineralization environments. Furthermore, these results may be of interest for the mining industry, particularly in optimizing resource extraction by understanding the mineralogical composition of deposits before processing.

The practical implications of this study extend beyond theoretical anomaly detection, providing tangible benefits for mineral exploration and resource assessment. One of the key applications is the early-stage identification of rare or economically valuable mineral deposits. By distinguishing mineralogical anomalies associated with specific geological processes, our approach enables geologists to prioritize exploration targets, reducing the need for extensive field sampling and minimizing exploration costs. This is particularly relevant for the discovery of lithium, rare earth elements, and strategic metals, where machine learning-driven anomaly detection can highlight previously overlooked mineralization zones (Golev et al., 2014; Hofstra et al., 2021).

Additionally, anomaly detection models can be integrated into geophysical and geochemical survey pipelines to enhance deposit characterization. For example, the ability to identify subtle geochemical variations linked to hydrothermal alteration or magmatic differentiation can improve predictive modeling of ore-forming environments. This approach aligns with existing exploration workflows but enhances them with automated, data-driven insights that refine drill targeting and geological mapping.

Another critical advantage is the applicability of this methodology to mine planning and resource classification. Once an anomaly is detected and validated through geological sampling, its clustering characteristics can inform decisions about extraction feasibility and processing methods. For instance, minerals in Cluster 1, associated with evaporitic settings, may indicate the presence of lithium-bearing brines, whereas minerals in Cluster 3, linked to igneous processes, could highlight primary magmatic deposits such as pegmatites or carbonatites. Such insights allow mining companies to optimize resource allocation and processing strategies.

Finally, this study contributes to the growing trend of AI-driven mineral exploration, where machine learning models serve as decision-support tools rather than replacements for geological expertise. Future research should explore integrating additional geophysical datasets—such as aeromagnetic, electromagnetic, and hyperspectral data—into machine learning frameworks, further refining mineral prospectivity mapping (Sharapatov and Shayakhmet, 2017; Li et al., 2017a, b).

By integrating machine learning-based anomaly detection with clustering, this study demonstrates a scalable approach for analyzing high-dimensional geochemical datasets. Future applications of this methodology could include predictive modeling of undiscovered mineral deposits, aiding in the prioritization of exploration targets, and refining mineral classification techniques through the integration of additional geochemical and geophysical data sources.

While traditional unsupervised learning models, such as Isolation Forest and One-Class SVM, effectively detect anomalies, they may have limitations in capturing complex non-linear feature interactions in mineralogical datasets. Recent advancements in deep learning offer promising solutions to these challenges. For instance, Deep Isolation Forest (Saremi et al., 2023) extends the standard Isolation Forest by incorporating deep neural networks, allowing for more robust anomaly detection in high-dimensional spaces. Additionally, Autoencoders have demonstrated the ability to reconstruct normal geochemical distributions, flagging deviations as potential anomalies (Xu et al., 2022). Another approach, Graph Neural Networks (GNNs), provides a framework to incorporate spatial and geochemical relationships between minerals, which can improve anomaly detection by leveraging geological context. These deep learning techniques present an opportunity for further refinement of anomaly classification and exploration strategies, bridging the gap between traditional methods and more advanced AI-driven models.

Comparing Isolation Forest and One-Class SVM revealed that the two algorithms capture different aspects of anomaly detection. Isolation Forest identified 156 anomalies, focusing on isolating data points with clear separation from the majority, aligning well with minerals that exhibit strongly distinctive properties. One-Class SVM, detecting 663 anomalies, demonstrated greater sensitivity to subtle deviations within clusters, making it more attuned to anomalies that slightly differ from the broader dataset but retain notable variance. For instance, certain minerals were only detected as anomalies by One-Class SVM, suggesting it excels in identifying smaller, more nuanced deviations. This complementary nature of the algorithms allows for a comprehensive approach to anomaly detection in geological data, as each method prioritizes different aspects of data variability, thereby providing a fuller spectrum of rare and anomalous minerals within the dataset.

To further evaluate the reliability of these anomaly detection methods, we introduced a statistical benchmark using a 2σ threshold, which identified 564 anomalies within the dataset. This statistical reference provides an independent assessment of the machine learning models' performance.

Comparing these findings, we observe that Isolation Forest (156 anomalies) correctly detects only 27.7 % of the statistical anomalies, suggesting it is too conservative. Meanwhile, One-Class SVM (663 anomalies) captures 117.5 % of the statistical anomalies, indicating a broader detection range but a higher likelihood of false positives. These results suggest that the true number of anomalous minerals likely falls between 564 and 663, making One-Class SVM slightly more inclusive but still more aligned with statistical expectations than Isolation Forest.

These performance differences have direct implications for mineral exploration. The high precision of Isolation Forest suggests that it is well-suited for applications where minimizing false positives is critical, such as the identification of rare and economically significant minerals that warrant detailed follow-up analysis. However, its low recall indicates that it may overlook some potentially valuable mineral anomalies. In contrast, One-Class SVM's high recall allows for a more inclusive identification of anomalies, making it particularly useful for preliminary screening in large-scale mineral exploration. The trade-off between these approaches suggests that an ensemble model, leveraging Isolation Forest's precision with One-Class SVM's broader anomaly detection capability, could optimize mineral prospectivity analysis. Such a hybrid strategy could refine anomaly classification by first applying One-Class SVM to identify a comprehensive set of potential anomalies, followed by Isolation Forest to prioritize the most geochemically distinct samples.

Despite the advantages of the proposed methodology, certain limitations must be acknowledged. One of the primary challenges is the potential bias in the dataset, as mineralogical databases may contain incomplete or unevenly distributed information. The presence of rare

minerals is inherently limited in such datasets, which may lead to overrepresentation of certain geological environments while underrepresenting others. As a result, some anomalies detected by the machine learning models may arise from data sparsity rather than genuine geological significance. Addressing this limitation would require expanding the dataset with additional field-collected mineralogical samples and incorporating more diverse geochemical and petrophysical datasets (Sharapatov and Shayakhmet, 2017; Sharapatov et al., 2023a).

Furthermore, the unsupervised nature of anomaly detection introduces assumptions regarding what constitutes an anomaly. Isolation Forest, for example, assumes that anomalies exhibit unique feature values and are well-separated from the majority of data points, which may not always be geologically valid. Conversely, One-Class SVM constructs a decision boundary based on the overall distribution of mineral properties, potentially classifying extreme but geochemically valid samples as anomalies (Goldstein and Uchida, 2016). Future research could refine anomaly detection criteria by integrating domain-specific geological knowledge and employing hybrid methods that combine unsupervised learning with semi-supervised techniques, where a limited subset of known anomalies is used to guide model training.

Additionally, the assumption that mineralogical features contribute equally to anomaly classification may not always hold. Some properties, such as specific gravity and refractive index, were found to be more influential, while others, like trace element concentrations, may play a secondary role. Future improvements should explore feature weighting mechanisms, such as SHAP values or domain-specific ranking, to better represent the geological significance of different mineral attributes (Kuhn and Johnson, 2019).

Finally, the clustering method used to categorize anomalies assumes that mineralogical variations can be represented by spherical clusters, which may oversimplify natural geological formations. Given that real-world geological processes produce highly non-spherical, stratified, or overlapping mineralogical distributions, future studies should evaluate alternative clustering approaches, such as density-based or hierarchical clustering (Kriegel et al., 2011).

The impact of feature selection extends beyond computational efficiency, directly influencing the interpretability of the machine learning results. The correlation analysis of physical properties validated the selection of Mohs hardness, specific gravity, refractive index, and crystal structure, confirming their complementary roles in mineral classification. Prior research highlights that these features provide key discriminative properties in geoscience applications, supporting their retention in anomaly detection frameworks (Kuhn and Johnson, 2019).

Despite their moderate correlation (0.75), both molar mass and molar volume were preserved due to their distinct geological significance. While high correlation between features often suggests redundancy, domain knowledge remains essential in determining whether both attributes provide unique interpretational value. The selected chemical features further ensure that geochemical variability is adequately captured without introducing excessive dimensionality or sparsity-related issues.

This structured feature selection approach aligns with best practices in high-dimensional geoscientific data analysis, ensuring that retained features maximize anomaly detection performance while preserving geological interpretability.

The 564 statistically defined anomalies provide an independent benchmark for evaluating these machine learning methods. A detailed examination of these minerals shows that they often share specific characteristics, such as high specific gravity and the presence of rare elements like Europium and Thorium. This suggests that these anomalies may represent unique mineralogical formations, potentially formed in rare or extreme geological conditions.

One-Class SVM appears to capture a broader spectrum of anomalies, including both the 564 statistical anomalies and additional minerals with minor deviations. This aligns with its algorithmic approach, where the decision boundary encloses most of the dataset, treating even slight deviations as potential anomalies. Conversely, Isolation Forest isolates only the most extreme cases, leading to a more conservative anomaly definition.

This distinction is crucial for practical mineral exploration. If a study prioritizes high-confidence anomalies with clear geological significance, Isolation Forest may be the preferred model. However, if a broader, more exploratory approach is required—such as identifying previously unrecognized mineral types—One-Class SVM may be more suitable. A hybrid strategy, incorporating both precision-focused and recall-focused anomaly detection, could further improve results.

Recent studies in machine learning-based mineral anomaly detection have primarily focused on supervised classification approaches, often utilizing labeled training data from well-characterized mineral deposits (Zhao et al., 2021; Xiong et al., 2023). While these methods demonstrate high predictive accuracy, they are inherently limited by the availability of labeled anomalies, making them less effective for detecting previously unknown mineral occurrences. Our study overcomes this limitation by implementing an unsupervised approach, allowing for the identification of novel mineralogical anomalies without relying on predefined classes.

Additionally, prior research has largely emphasized geochemical datasets, whereas our work integrates both geochemical and physical properties to enhance interpretability and robustness (Liu et al., 2022). By applying and comparing two distinct anomaly detection algorithms—Isolation Forest and One-Class SVM—we provide a more comprehensive evaluation of machine learning techniques for mineral prospectivity analysis. This dual-model approach addresses a key gap in previous studies, where the relative performance of different anomaly detection models has not been systematically assessed.

The results of our analysis indicate that One-Class SVM exhibits high recall, capturing a broad spectrum of potential anomalies, while Isolation Forest provides high precision, ensuring that the detected anomalies are geochemically distinct. This trade-off highlights the need for hybrid approaches, combining the strengths of both methods, which could further refine mineral exploration strategies.

The choice of principal components significantly influenced the anomaly detection process. While a two-component PCA transformation was used for visualization, all machine learning models were trained using five principal components, which capture 90 % of the dataset's total variance (Fig. 1). This approach ensured that key geochemical and physical characteristics were preserved while reducing computational complexity.

By increasing the number of principal components from two to five, we achieved a more refined anomaly detection process. Using only two components, key mineralogical variations may have been lost, leading to suboptimal model performance. Conversely, the inclusion of additional components provided a richer representation of mineral properties, improving the ability of both Isolation Forest and One-Class SVM to distinguish subtle anomalies.

Furthermore, the analysis of the explained variance curve (Fig. 1) demonstrated that incorporating more than six components resulted in diminishing returns, as the additional variance captured was minimal. Therefore, we selected five principal components as a balance between model interpretability and performance. This decision aligns with best practices in anomaly detection within high-dimensional geochemical datasets, where preserving a sufficient proportion of variance is essential to avoid loss of relevant information while ensuring efficient processing (Hajihosseinlou et al., 2024).

This insight highlights a potential improvement in anomaly detection strategies. Instead of relying solely on a single model, a hybrid approach combining the strengths of Isolation Forest (precision) and One-Class SVM (recall), refined using the 2σ statistical threshold, could lead to a more robust anomaly detection framework. Future studies could integrate ensemble methods or probabilistic modeling to further refine these classifications and reduce uncertainty.

A key challenge in applying explainable machine learning to anomaly detection lies in the unsupervised nature of the models used.

Unlike classification or regression models, where post hoc explainability techniques such as SHAP provide clear feature attributions, anomaly detection methods do not operate with explicit class labels. Isolation Forest isolates anomalies based on recursive partitioning, while One-Class SVM constructs a decision boundary in high-dimensional space. Consequently, the feature contributions inferred through these models lack direct interpretability in a traditional sense. Furthermore, when applying SHAP to One-Class SVM, excessive memory requirements render it computationally infeasible for large mineralogical datasets. Instead of relying on explainability techniques designed for supervised learning, we focus on geological clustering and feature analysis as the primary means of interpreting anomalies. By grouping detected anomalies into distinct clusters and analyzing their shared mineralogical properties, we establish a domain-specific context for understanding the nature and significance of these anomalous samples. This approach aligns with practical applications in geology, where clustering patterns and mineral associations often hold more interpretative value than model-specific feature attributions.

This research, however, has certain limitations. The reliance on PCA for dimensionality reduction, though effective, may not fully preserve all intricate relationships within the dataset's high-dimensional structure. Data quality remains a crucial factor, and potential inaccuracies or incompleteness could impact the reliability of results. Additionally, KMeans clustering assumes spherical clusters, which may oversimplify the complex structures and interdependencies within mineralogical data, especially in natural settings where geological processes create non-spherical formations. Future research could benefit from advanced anomaly detection techniques such as Deep Isolation Forest or Autoencoders, which offer enhanced sensitivity to high-dimensional and non-linear patterns, thereby potentially capturing additional layers of variability among mineral characteristics. Furthermore, investigating temporal changes in mineralogical data could provide deeper insights into the evolutionary pathways of mineral deposits, thereby contributing to a more comprehensive understanding of rare minerals and their applications in geology and resource management (Sharapatov et al., 2023a).

5. Conclusions

This study successfully applied machine learning techniques to identify and categorize anomalous minerals, revealing distinct physical and chemical properties that set them apart from more common mineral types. By employing Isolation Forest and One-Class SVM, we detected a unique subset of rare minerals, and KMeans clustering further grouped these anomalies into three clusters, each suggesting specific geological formation environments. This approach demonstrates the potential of unsupervised learning in mineral exploration, particularly in detecting previously unrecognized mineralogical patterns. The study highlighted certain influential factors, such as high specific gravity, distinctive crystalline structures, and the presence of rare elements, as key indicators of geological rarity. These features may serve as potential markers for identifying unusual mineral deposits and offer a foundation for future research in mineral exploration.

One of the key findings of this research is the effectiveness of integrating geochemical and physical properties for mineral classification, which improves anomaly detection accuracy compared to purely geochemical approaches used in previous studies (Zhao et al., 2021; Xiong et al., 2023). Furthermore, the combination of Isolation Forest and One-Class SVM provides complementary perspectives on mineral anomaly detection: while Isolation Forest is effective in identifying distinct, well-separated anomalies, One-Class SVM is more sensitive to subtle variations, making it suitable for detecting minerals with rare but geochemically meaningful properties. This insight can guide geologists in selecting appropriate machine learning techniques depending on the specific characteristics of mineral datasets.

To further enhance anomaly detection accuracy, future research should explore deep learning-based techniques, such as Deep Isolation Forest, Autoencoders, and Graph Neural Networks. These models have demonstrated the potential to capture complex relationships within high-dimensional geological datasets, improving classification robustness and reducing false positives. By integrating deep learning approaches, anomaly detection can be refined, leading to more effective mineral exploration and predictive modeling of undiscovered mineral deposits.

Building on these findings, future research should focus on a more comprehensive validation using field-collected mineralogical samples, ensuring that the identified anomalies correspond to actual rare or economically significant minerals. Additionally, exploring alternative machine learning techniques, such as Autoencoders or Deep Isolation Forest, could improve the robustness of anomaly detection in high-dimensional mineralogical data (Liu et al., 2022). Comparative studies with larger and more diverse datasets would further validate the effectiveness of these methods, expanding our understanding of the complex formation processes of rare minerals and contributing to more targeted and efficient mineral exploration efforts.

Another promising direction involves integrating multi-source geophysical data, including petrophysical (laboratory, field), geophysical (aero, surface, borehole), and lithochemical information, into machine learning models to enhance predictive capabilities (Sharapatov and Shayakhmet, 2017; Sharapatov et al., 2023b, Sharapatov et al., 2024). Such an approach would extend the applicability of the proposed methodology to various stages of mineral deposit studies, from early-stage exploration to resource assessment and extraction monitoring. Incorporating deep learning-based geospatial analysis could further refine anomaly detection by leveraging spatial relationships between mineral occurrences, potentially leading to more efficient exploration strategies.

CRediT authorship contribution statement

Abish Sharapatov: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Alisher Saduov:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Nazerke Assirbek:** Writing – review & editing, Formal analysis, Conceptualization. **Madiyar Abdyrov:** Writing – review & editing, Formal analysis, Conceptualization. **Beibit Zhumabayev:** Writing – review & editing, Visualization.

Code availability section

The code for the project is available under the following Github.

Repository

https://github.com/Kazinage/Mineral-anomaly.

Data availability

The "Comprehensive Database of Minerals" dataset is publicly available in the Kaggle data repository.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used GPT-4 from OpenAI in order to improve the readability and language of some paragraphs in the text. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alisher Saduov reports a relationship with K I Satbayev Kazakh National Research Technical University that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

It is open source data from Kaggle. Link for the data and code available in the article

References

Aggarwal, C.C., 2017. Outlier Analysis. Springer. https://doi.org/10.1007/978-3-319-47578-3.

Ahmed, M., Seraj, R., Islam, S.M.S., 2020. The K-means algorithm: a comprehensive survey and performance evaluation. Electronics 9 (8), 1295. https://doi.org/10.3390/electronics9081295.

Barnett, V., Lewis, T., 1994. Outliers in Statistical Data. John Wiley & Sons. https://doi.org/10.1007/978-1-4615-3182-4.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2004. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953.

Chouhan, H., 2021. Elbow method for optimal customer segmentation using K-means clustering. Int. J. Sci. Res. Eng. Dev. 7 (3), 92. https://ijsred.com/volume7/issue3/IJSRED-V7I3P92.pdf.

Comprehensive Database of Minerals. Available online: https://www.kaggle.com/datasets/vinven7/comprehensive-database-of-minerals (accessed on 22 February 2023).

Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. KDD'96), pp. 226–231.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27 (8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Golev, A., Scott, M., Erskine, P.D., Ali, S.H., Ballantyne, G.R., 2014. Rare earths supply chains: current status, constraints and opportunities. Resour. Policy 41, 52–59. https://doi.org/10.1016/j.resourpol.2014.03.004.

Goldstein, M., Uchida, S., 2016. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLoS One 11 (4), e0152173. https://doi.org/10.1371/journal.pone.0152173.

Hajihosseinlou, M., Abdi, G., Ghodsi, M., Arjmand, R., 2024. Deep learning-based outlier detection for high-dimensional geoscientific data. Earth Sci. Informat. (in press).

Hofstra, A.H., Kreiner, D.C., Emsbo, P., Johnson, C.A., Marsh, E.E., 2021. Rare earth element deposits: an overview of past, present, and future exploration trends. Minerals 11 (7), 712. https://doi.org/10.3390/min11070712.

Jain, A.K., 2010. Data clustering: 50 Years beyond K-means. Pattern Recognit. Lett. 31 (8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.

Johnson, C.A., Emsbo, P., 2021. Trace element geochemistry of rare earth deposits. Econ. Geol. 116 (3), 555–578. https://doi.org/10.5382/econgeo.4823.

Kriegel, H.P., Kröger, P., Sander, J., Zimek, A., 2011. Density-based clustering. WIREs Data Mining Knowledge Discove. 1 (3), 231–240. https://doi.org/10.1002/widm.30.

Kuhn, M., Johnson, K., 2019. Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press. https://doi.org/10.1201/9781315108230.

Lesouple, J., Baudoin, C., Spigai, M., Tourneret, J.-Y., 2021. Generalized Isolation Forest for Anomaly Detection. HAL Archives. Available online: https://hal.science/hal-03382634/document.

Liu, F., Zhou, W., Liu, B., 2021. Flow field description and simplification based on principal component analysis downscaling and clustering algorithms. Front. Earth Sci. 9, 804617. https://doi.org/10.3389/feart.2021.804617.

Liu, Z.-J., Yang, L.-Q., Xie, D., Yang, W., Li, D.-P., Feng, T., Deng, J., 2023. Hydrothermal alteration processes of Xincheng gold deposit Jiaodong Peninsula, China: constraints from composition of hydrothermal Rutile. Minerals 14 (4), 417. https://doi.org/10.3390/min14040417.

Lindsay, J.B., Newman, D.R., Francioni, A., 2019. Scale-optimized surface roughness for topographic analysis. Geosciences 9 (7), 322. https://doi.org/10.3390/geosciences9070322.

Liu, C., Wang, P., Zhou, J., 2022. Integration of geochemical and physical parameters for improved mineral anomaly detection using AI techniques. J. Geochem. Explor. 232, 106903. https://doi.org/10.1016/j.gexplo.2022.106903.

Li, X., Zhang, C., Li, W., Ford, J.P., 2017a. A K-means clustering approach for mapping geological structures using airborne magnetic data: a case study from the Eastern Lachlan orogen, Australia. J. Appl. Geophys. 139, 1–12. https://doi.org/10.1016/j.jappgeo.2017.02.001.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H., 2017b. Feature selection: a data perspective. ACM Comput. Surv. 50 (6), 1–45. https://doi.org/10.1145/3136625.

Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L., 2013. Detecting outliers: do not use standard deviation around the mean, use the median absolute deviation instead. Front. Psychol. 4, 138. https://doi.org/10.3389/fpsyg.2013.00138.

Liu, F.T., Ting, K.M., Zhou, Z.H., 2008. Isolation forest. In: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM), pp. 413–422. https://doi.org/10.1109/ICDM.2008.17.

Luo, Z., Farahbakhsh, E., Müller, R.D., Zuo, R., 2024. Multivariate statistical analysis and bespoke deviation network modeling for geochemical anomaly detection of rare earth elements. Appl. Geochem. 150, 106146. https://doi.org/10.1016/j.apgeochem.2024.106146.

Pérez-Ortega, J., Almanza-Ortega, N.N., Romero, D., 2018. Balancing effort and benefit of K-means clustering algorithms in big data realms. PLoS One 13 (8), e0201874. https://doi.org/10.1371/journal.pone.0201874.

Reynolds, D., 2009. Gaussian mixture models. Encycloped. Biomet. 659–663. https://doi.org/10.1007/978-0-387-73003-5_196.

Ribeiro, D., Matos, L.M., Moreira, G., Pilastri, A., Cortez, P., 2022. Isolation forests and deep Autoencoders for industrial screw tightening anomaly detection. Computers 11 (4), 54. https://doi.org/10.3390/computers11040054.

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10 (3), e0118432. https://doi.org/10.1371/journal.pone.0118432.

Saremi, M., Bagheri, M., Mirzabozorg, S.A.A., Hassan, N.E., Hoseinzade, Z., Maghsoudi, A., Rezania, S., Ranjbar, H., Zoheir, B., Beiranvand Pour, A., 2023. Evaluation of deep isolation forest (DIF) algorithm for mineral prospectivity mapping of polymetallic deposits. Minerals 14 (10), 1015. https://doi.org/10.3390/min14101015.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. Neural Comput. 13 (7), 1443–1471. https://doi.org/10.1162/089976601750264965.

Sharapatov, A., Saduov, A., Assirbek, N., 2023a. Comparative analysis of machine and deep learning algorithms in geology [Сравнительный анализ возможностей алгоритмов машинного и глубокого обучения в геологии]. Min J. Kaz. 223 (11), 14–20. https://doi.org/10.48498/minmag.2023.223.11.002.

Sharapatov, A., Shayakhmet, M., 2017. Physico-geological basis of efficiency of application of aeromagnetic method in oil-gas caspian lowland. News. Nationa. Acad. Sci. Repub. Kaz. Series Geol. Tech. Sci. 3 (423), 95–99. https://www.scopus.com/record/display.uri?eid=2-s2.0-85020483566&origin=resultslist.

Sharapatov, A., Assirbek, N., Saduov, A., Abdyrov, M., Zhumabayev, B., 2024. Consolidated geological and geophysical characteristics of uranium deposit rocks and prospects for their utilization (Shu-Sarysu province, Kazakhstan). News of the National Academy 6 (467), 210–229. https://doi.org/10.32014/2024.2518-170X.471.

Sharapatov, A., Assirbek, N.A., Samanbetov, N.Z., 2023b. Evaluation of the effectiveness of electric logging methods in studying the tightness of casing strings in wells of uranium deposits [Оценка эффективности Методов электрокаротажа при изучении герметичности обсадных колонн в скважинах урановых месторождений]. Bull. Tomsk Polytech. Univer. Geo Asset. Eng. 334 (10), 7–15. https://doi.org/10.18799/24131830/2023/10/4336.

Su, Q., Yu, H., Xu, X., Chen, B., Yang, L., Fu, T., Liu, W., Chen, G., 2023. Using principal component analysis (PCA) combined with multivariate change-point analysis to identify brine layers based on the geochemistry of the core sediment. Water 15 (10), 1926. https://doi.org/10.3390/w15101926.

Tien, C.-W., Huang, T.-Y., Chen, P.-C., Wang, J.-H., 2021. Using Autoencoders for anomaly detection and transfer learning in IoT. Computers 10 (7), 88. https://doi.org/10.3390/computers10070088.

Talusani, R.V.R., Sivell, W.J., Ashley, P.M., 2008. Mineral Chemistry, Petrogenesis, and Tectonic Setting of the Wateranga Layered Intrusion. ResearchGate, Southeast Queensland, Australia. Available online: https://www.researchgate.net/publication/237169318_Mineral_chemistry_petrogenesis_and_tectonic_setting_of_the_Wateranga_layered_intrusion_southeast_Queensland_Australia.

Vandna, K., Yadav, P., Kumar, V., 2020. K-means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula. ResearchGate. Available online: https://www.researchgate.net/publication/346349075.

Xiong, X., Zhang, H., Wang, Y., 2023. Supervised deep learning for mineral prospectivity mapping: advances and limitations. Comput. Geosci. 173, 105503. https://doi.org/10.1016/j.cageo.2023.105503.

Xu, H., Pang, G., Wang, Y., Wang, Y., 2022. Deep isolation forest for anomaly detection. arXiv preprint arXiv:2206.06602. https://arxiv.org/abs/2206.06602.

Zhang, N., Chen, M., Yang, F., Yang, C., Gao, Y., Shang, Y., Peng, D., 2022. Forest height mapping using feature selection and machine learning by integrating multi-source satellite data in Baoding City, North China. Remote Sens. 14 (18), 4434. https://doi.org/10.3390/rs14184434.

Zhao, Y., Li, J., Wang, X., 2021. Machine learning-based anomaly detection in mineral exploration: a review and case study. Ore Geol. Rev. 135, 104235. https://doi.org/10.1016/j.oregeorev.2021.104235.

Zuo, R., Wang, J., Zhang, S., 2020. Recognizing multivariate geochemical anomalies for mineral exploration using machine learning algorithms. Comput. Geosci. 140, 104484. https://doi.org/10.1016/j.cageo.2020.104484.