Journal of Geomine (2025), 3(3), 188-200
DOI: 10.22077/jgm.2026.10554.1065

Comparative Evaluation of Dimensionality Reduction Methods for Lithological Unit Discrimination Based on Multi-Element Geochemical Data

Soheil Zaremotlagh[1], Asieh Ghanbarpour[2]

Recived: 2025 Nov. 23, Revised: 2026 Jan. 04, Accepted: 2026 Feb. 10, Published: 2026 Feb. 25

ABSTRACT

Identifying and discriminating lithological units from multi-element geochemical data remain fundamental challenges in petrology and mineral exploration. Geochemical datasets are typically high-dimensional, strongly correlated, and exhibit nonlinear relationships arising from petrogenetic processes such as fractional crystallization, magma mixing, and hydrothermal alteration. These complexities necessitate advanced analytical techniques for effective data interpretation. This study presents a systematic comparison of three dimensionality reduction (DR) approaches- Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), and autoencoders (AE)- to improve the discrimination of igneous rock units. The dataset comprises 517 samples representing nine igneous rock types (granite, granodiorite, diorite, quartz diorite, gabbro, dacite, andesite, basalt, and basaltic andesite) analyzed for 22 geochemical components (10 major oxides and 12 trace elements). Following centered log-ratio (CLR) transformation and subsequent MinMax or Z-score normalization, the data were processed using the three dimensionality reduction methods and clustered with the K-means algorithm. Performance was evaluated using multiple internal and external validation indices. The combination of Z-score normalization with UMAP in four-dimensional space yielded superior clustering performance across most metrics, producing compact and well-separated clusters. In contrast, linear PCA and autoencoder methods were less effective in capturing the intrinsic structure of the data. Geochemical validation using Ce/La ratios and Harker diagrams confirmed that clusters generated by the Zscore-UMAP-Kmeans workflow correspond to coherent, petrogenetically meaningful groups. The method effectively discriminates mafic and felsic end-members, while intermediate compositions display greater variability reflecting magmatic differentiation processes. Sensitivity analysis over 30 independent runs demonstrated the stability and reproducibility of the UMAP-based approach. These findings highlight the robustness and interpretability of UMAP for revealing geochemical patterns in complex datasets, with direct implications for mineral exploration targeting and lithological mapping.

KEYWORDS

Geochemical discrimination, igneous rocks, machine learning, mineral exploration, nonlinear dimensionality reduction, unsupervised clustering, UMAP

I. INTRODUCTION

The analysis of multielement geochemical data is essential in geology and mineral exploration for identifying mineralized zones, distinguishing lithological units, and interpreting petrogenetic processes. These datasets—typically comprising major oxides, trace elements, and isotopic ratios—are high-dimensional and exhibit complex nonlinear structures resulting from processes such as fractional crystallization, magma mixing, and hydrothermal alteration (Bévan, Boulvais et al., 2025). Direct analysis is challenged by the "curse of dimensionality," which reduces sample separability and increases computational costs (Grunsky and Caritat, 2020; Jia, Sun et al. 2022).

Dimensionality reduction techniques address this challenge by eliminating redundancy while preserving the essential structure of the data, thereby enabling more effective analysis and visualization. Principal Component Analysis (PCA), a linear technique, has long been employed in geochemical studies to identify compositional trends, reduce noise, and differentiate elemental sources (Zhao and Chen 2021; Acosta-Góngora, Potter et al. 2022; Su, Yu et al. 2023). However, its assumption of linearity is a significant limitation, as many geological processes are inherently nonlinear. Consequently, nonlinear methods such as Uniform Manifold Approximation and Projection (UMAP) and deep learning-based autoencoders have gained prominence for their ability to capture complex

[1]Department of Mining Engineering, Faculty of Engineering, University of Sistan and Baluchestan, Zahedan, Iran, [2]Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Sistan and Baluchestan, Zahedan, Iran
✉ S. Zaremotlagh: s_zare@eng.usb.ac.ir

geochemical relationships (McInnes, Healy et al. 2018). Recent advances have shifted rock classification and rock interpretation from traditional bi- or tri-variate diagrams toward multivariate statistics and machine learning approaches (Zaremotlagh and Hezarkhani 2016, Zaremotlagh and Hezarkhani, 2017; Doucet, Tetley et al. 2022; Badawy, Dmitriev et al. 2023; Cao, Zhang et al. 2023; Abo Khashaba, El-Shibiny et al. 2024; Canbaz and Karaman 2024; Chen, Chen et al. 2024).

PCA remains one of the most widely used techniques for dimensionality reduction and identifying fundamental geochemical trends (Zhao and Chen, 2021; Acosta-Góngora, Potter et al., 2022; Su, Yu et al., 2023). For instance, in the study by Sadeghi et al. (2024), integrating PCA with K-Means clustering enabled the identification of lithium mineralization-related anomalies in multielement soil datasets from the Västernorrland region of Sweden. However, the limitations of linear methods in capturing complex relationships have spurred interest in nonlinear dimensionality reduction techniques. Modern methods like UMAP and t-SNE, which effectively visualize nonlinear patterns and compositional continuity, have been successfully applied in tasks such as lithology classification from drill data (Hansen and Aarset, 2024; Zhang, Liu et al., 2024) and tracing magmatic sources from isotopic data of oceanic basalts (Stracke, Willig et al., 2022).

PCA remains one of the most widely used techniques for dimensionality reduction and identifying fundamental geochemical trends (Zhao and Chen, 2021; Acosta-Góngora, Potter et al., 2022; Su, Yu et al., 2023). For example, in the study by Sadeghi et al. (2024), integrating PCA with K-Means clustering enabled the identification of lithium mineralization-related anomalies in multielement soil datasets from the Västernorrland region of Sweden. However, the limitations of linear methods in capturing complex relationships have spurred interest in nonlinear dimensionality reduction techniques. Modern methods like UMAP and t-SNE, which effectively visualize nonlinear patterns and compositional continuity, have been successfully applied in tasks such as lithology classification from drill data (Hansen and Aarset, 2024; Zhang, Liu et al., 2024) and tracing magmatic sources from isotopic data of oceanic basalts (Stracke, Willig et al., 2022).

This work contributes by (1) providing a quantitative, multi-criteria comparison of dimensionality reduction methods for geochemical data; (2) examining the combined effects of normalization and dimensionality reduction on lithological discrimination; and (3) utilizing a subset of the geochemical dataset compiled by Du Bray, John et al. (2006), which originally consists of volcanic rocks from the western Cascades. This subset includes 517 igneous rock samples from nine lithological units and 22 geochemical components. This approach enables reproducible and comparable evaluations and offers practical insights for selecting optimal data-processing workflows in exploration geochemistry.

II. DATASET DESCRIPTION AND DATA PREPROCESSING

A. Study Dataset

To evaluate and compare the performance of dimensionality reduction methods, we utilized a well-established geochemical dataset of igneous rocks (du Bray, John et al., 2006), widely recognized in petrological and geochemical research. This dataset compiles chemical compositions of diverse rock samples from over one hundred scientific sources. Its geological diversity, high analytical quality, and broad coverage of igneous rock types make it a reliable benchmark for testing data-analysis methodologies. The dataset includes samples from nine igneous lithological units, spanning felsic rocks (granite, granodiorite, and dacite), intermediate rocks (diorite and andesite), and mafic varieties (gabbro, basalt, and basaltic andesite). Each sample was analyzed for a suite of 22 geochemical components, comprising 10 major oxides and 12 trace elements. This compositional diversity reflects a wide range of petrogenetic processes, from fractional crystallization of mafic magmas to the evolution of felsic magmas, thereby enabling a comprehensive assessment of the ability of various methods to discriminate among lithological units with distinct chemical characteristics.

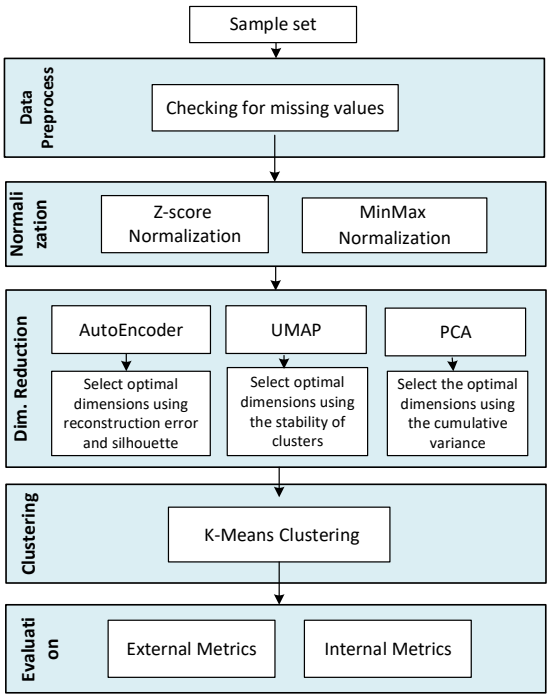

Fig. 1. Research methodology framework for discriminating rock units based on geochemical data

To ensure data quality and the reliability of the analytical results, a screening and quality-control procedure was conducted. During this stage, samples from lithological units with insufficient representation for statistically robust analysis, samples exhibiting strong hydrothermal alteration that could obscure primary magmatic signatures, and samples with uncertain lithological classification were excluded. This screening ensured that each lithological unit used in the modeling process contained a sufficient number of samples with well-defined and reliable chemical compositions. After preprocessing, a total of 517 igneous rock samples with complete chemical analyses for 22 elements and verified lithological labels were selected for further analysis. Fig. 2 presents histograms of the 22 major and trace elements in the igneous rock geochemical dataset. As shown, the measured concentrations display substantial variability across lithological units. Several components, such as SiO_2, Al_2O_3, and FeO*, exhibit broad, multimodal distributions, reflecting fundamental compositional differences among felsic, intermediate, and mafic rocks. In contrast, elements such as TiO_2, MnO, and P_2O_5 show narrower and more concentrated distributions, indicative of the relatively stable geochemical behavior of these components during magmatic processes.

The multimodal nature of several elemental distributions particularly SiO_2, MgO, and CaO indicates the presence of multiple compositional populations within the dataset. Such variability may arise from diverse geological processes, including fractional crystallization, magma mixing, and phase separation during the evolution of igneous rocks. In addition, the long-tailed distributions observed for certain trace elements such as Ba, Sr, and Zr reflect localized enrichment or depletion in some samples, which is likely associated with variations in magmatic source characteristics or late-stage crystallization processes. The heterogeneous and multimodal chemical patterns observed across the dataset provide a suitable context for evaluating the performance of dimensionality reduction and clustering methods. These methods must be capable of preserving intrinsic data structures while effectively capturing distinctions among compositional groups and lithological classifications.

B. Data Preprocessing

The aim of the preprocessing step is to prepare the geochemical dataset for analysis using dimensionality reduction techniques and clustering algorithms. The data consist of chemical compositions obtained through X-ray fluorescence (XRF) analysis. A fundamental challenge in analyzing such datasets is the wide variation in the magnitude of elemental concentrations. For example, SiO_2 typically ranges between 40–80 wt%, whereas TiO_2 generally ranges from 0.1 to 3 wt%. These differences in scale can cause elements with larger absolute values to exert a disproportionate influence on distance calculations in machine learning algorithms, ultimately affecting clustering outcomes. Therefore, prior to dimensionality reduction and clustering, it is essential to normalize the data so that all elements contribute equally to the analysis and the true structure of geochemical relationships is preserved.

Data preprocessing in this study was conducted in two sequential steps. The first step addressed the compositional nature of the geochemical data by applying a Centered Log-Ratio (CLR) transformation. This transformation is commonly used to manage the closed-sum constraint inherent in compositional data and to reduce the risk of spurious correlations in subsequent analyses. The second step involved data normalization to prevent disproportionate influence from features with varying scales. Two widely used normalization techniques were applied: MinMax normalization, which linearly scales data to the interval between 0 and 1 while preserving the original distribution pattern, and Z-score normalization, which transforms data to a standard distribution with a mean of zero and a variance of one, making it appropriate for approximately normally distributed features. These two approaches were selected due to their frequent application in geochemical studies and their distinct impacts on machine learning algorithms. By examining both normalization strategies, this study aims to assess how preprocessing choices influence the performance of dimensionality reduction methods in discriminating lithological units.

C. Dimensionality Reduction

In this study, three distinct approaches for dimensionality reduction of geochemical data were evaluated and compared: Principal Component Analysis (PCA), representing classical linear methods; Uniform Manifold Approximation and Projection (UMAP), representing nonlinear topology-based techniques; and autoencoders, representing deep learning–based models. These methods differ fundamentally in their underlying philosophies and computational mechanisms, collectively encompassing a broad spectrum of dimensionality reduction strategies. Their primary objective is to project high-dimensional, complex datasets into a lower-dimensional space while preserving essential structures and meaningful patterns. The following sections briefly outline the theoretical foundations and operational mechanisms of each method to provide a clearer understanding of how they process geochemical data.

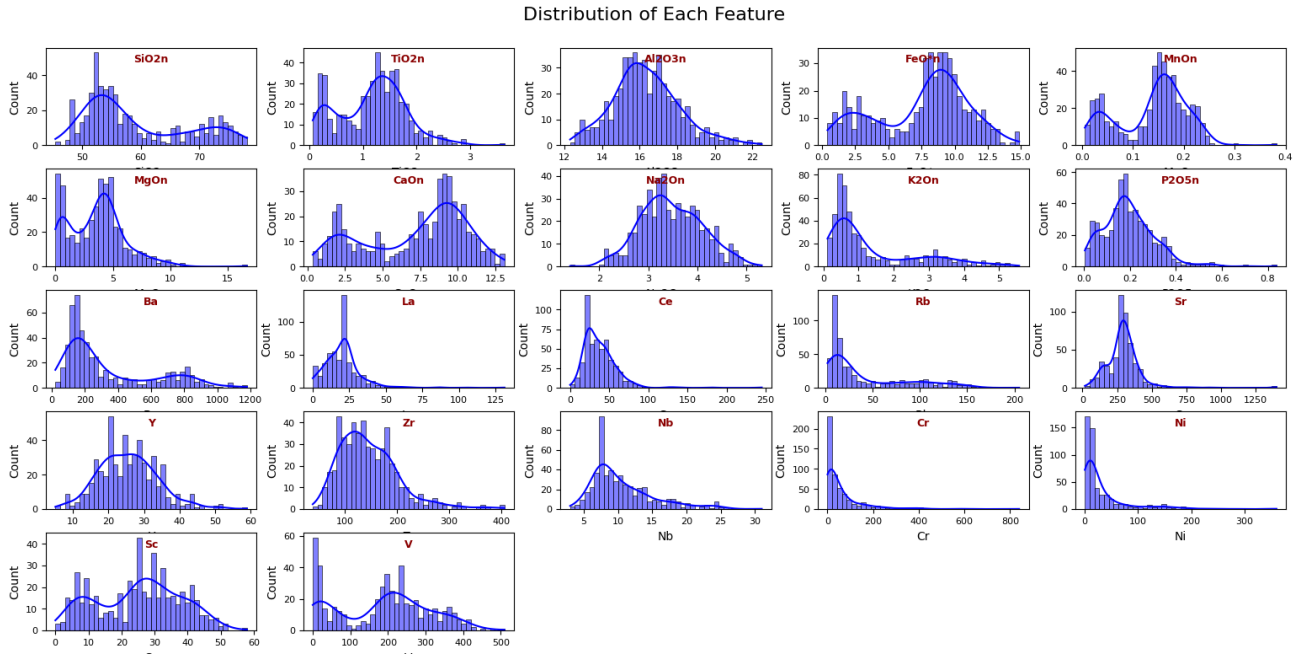

Fig. 2. Frequency distribution of 22 oxides and geochemical elements in rock samples

1) Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is one of the most widely used linear dimensionality reduction techniques, designed to simplify multidimensional datasets while preserving as much of the original information as possible. In geochemical studies, PCA has been extensively applied to identify major compositional trends, reduce noise, and distinguish elemental sources (Xue, Lee et al., 2011; Acosta-Góngora, Potter et al., 2022; Su, Yu et al., 2023; Abu Salem, El Fallah et al., 2024; Abbasi, Yang et al., 2025; Haj, Merheb et al., 2025).

PCA transforms the original dataset into a set of new, orthogonal (uncorrelated) features called principal components. These components represent directions in the feature space that capture the maximum variance in the data. The first component explains the greatest variance, the second component explains the highest remaining variance, and so on. By retaining only the first few components that account for the majority of the total variance, a compact yet effective representation of the dataset can be achieved, while the remaining components are discarded.

The computational procedure of PCA is as follows. Let X denote the standardized data matrix of size (n×m), where n is the number of samples and m is the number of features. The covariance matrix is first computed as:

$$C = \frac{1}{n-1} X^T X \quad (1)$$

The eigenvalues (λ) and eigenvectors of the covariance matrix are computed by solving the characteristic equation $|\lambda I - C| = 0$. The eigenvectors, which define the directions of the principal components, are sorted in descending order based on their corresponding eigenvalues and arranged into the matrix W. Finally, the transformed dataset in the principal component space is obtained using $Z = W \times X$, where Z represents the new reduced-dimensional space.

2) Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection (UMAP) is a modern and highly effective technique for nonlinear dimensionality reduction that has gained significant attention in recent years for analyzing complex, high-dimensional datasets. Unlike linear methods such as PCA, which assume that data lie in a flat Euclidean space, UMAP is based on the assumption that high-dimensional data often lies on a lower-dimensional geometric structure (manifold) within the higher-dimensional space. The method builds upon mathematical foundations in algebraic topology and Riemannian geometry, aiming to map the manifold structure into a low-dimensional space while preserving its topological properties both locally (relationships among nearby samples) and globally (relationships among larger-scale groups or clusters). This characteristic makes UMAP particularly suitable for geochemical datasets, which frequently exhibit complex nonlinear structures resulting from diverse geologic and magmatic processes (Ghojogh, Ghodsi et al., 2021; Zhang, Liu et al., 2024).

UMAP operates in two main stages. The first stage involves constructing a neighborhood graph in the input space to model the local structure of the data. To build this graph, each sample is connected to its k nearest neighbors using a chosen distance metric, commonly the Euclidean distance. For any pair of samples, i and j, the probability of connectivity is computed using a Gaussian kernel:

$$p_{ij} = e^{-\left(\frac{d(x_i,x_j)-\rho_i}{\sigma_i}\right)} \quad (2)$$

where $d(x_i, x_j)$ denotes the Euclidean distance between samples i and j; ρ_i represents the distance to the nearest neighbor, included to prevent excessive point crowding in high-density regions; and σ_i is a local normalization parameter that ensures uniformity of the probability distribution within the neighborhood of each sample. These parameters are calibrated so that the local probability distribution around each point remains balanced and is not disproportionately influenced by variations in sampling density.

The local neighborhood graphs are modeled as fuzzy graphs, where edge weights represent the probability of co-occurrence of two samples within the same local neighborhood. Summarization of these fuzzy graph is performed using:

$$P_{ij} = p_{ij} + p_{ji} - p_{ij}.p_{ji} \quad (3)$$

This formulation ensures that if either sample identifies the other as a neighbor, the corresponding edge is retained in the graph.

In the second stage, the low-dimensional embedding is obtained by minimizing the cross-entropy divergence between the fuzzy graph constructed in the high-dimensional space and its corresponding graph in the low-dimensional space. The loss function is defined as:

$$CE = \sum_{(i,j)} \left[P_{ij} log\left(\frac{P_{ij}}{Q_{ij}}\right) + (1 - P_{ij}) log\left(\frac{1 - P_{ij}}{1 - Q_{ij}}\right) \right] \quad (4)$$

where the similarity between points in the low dimensional space is modeled as:

$$Q_{ij} = \frac{1}{1 + a\left(\|y_i - y_j\|\right)^{2b}} \quad (5)$$

Parameters a and b are automatically derived during the UMAP optimization process by fitting the similarity decay curve to the user-specified min_dist value. This procedure ensures that local distances in the low-dimensional manifold remain consistent with the neighborhood structure defined in the high-dimensional space during cross-entropy minimization (Equation 4). The resulting loss function enables the algorithm to preserve the proximity of neighboring samples while simultaneously maintaining adequate separation between distinct clusters (Allaoui, Kherfi et al., 2020).

One of the key advantages of UMAP is its strong ability to preserve both local structure (relationships among neighboring samples, essential for identifying fine-scale subgroups) and global structure (relationships among larger sample groups, important for understanding overall geochemical distribution). This dual preservation enables UMAP to capture the complex nonlinear patterns inherent in geochemical datasets more effectively than linear methods, resulting in significant improvements in clustering performance and lithological discrimination.

In this study, the UMAP hyperparameters were selected based on established methodological guidelines and the statistical characteristics of the dataset. The parameter n_neighbors was set to 50 (approximately 10% of the total samples) to achieve a balanced preservation of both local and global geochemical structures, which is generally appropriate for medium-sized datasets. The default min_dist value (0.1) and the Euclidean distance metric were retained, as they produced stable, well-separated clusters suitable for lithological discrimination without causing excessive fragmentation. To ensure reproducibility, random_state was fixed at 42 during the main comparative experiments, while multiple random seeds were tested in the stability assessment (Section 4.4) to verify the consistency of the UMAP embeddings across repeated runs.

3) Autoencoder Network

The autoencoder is a class of artificial neural networks designed to learn compact and nonlinear representations of high-dimensional data. Autoencoders operate in an unsupervised manner and attempt to reconstruct the input data at the output layer. In this process, the input data are first projected into a reduced latent space and subsequently reconstructed from this compressed representation (Wang, Yao et al. 2016). Formally, let x∈R^d denote an input vector with d original features. A standard autoencoder consists of three major components. The first component is the encoder which transforms the input vector x into a compressed representation h, referred to as the latent code. This transformation is typically implemented using fully connected layers combined with nonlinear activation functions (e.g., ReLU or tanh). The dimensionality of the latent layer h is considerably smaller than that of the input x (e.g., reduced from 22 to 7 dimensions in this study). This dimensional constraint forces the network to extract the most fundamental and discriminative features of the data while suppressing noise and redundancy. The latent, or bottleneck, layer constitutes the core of the autoencoder and contains the compressed, distilled, and dimension-reduced representation of the input. This layer captures the dominant factors of variability within the dataset and effectively serves as the low-dimensional feature space used for subsequent analyses, such as clustering. The third component is the decoder, which defines a mapping $x'=g(h)$, which reconstructs the input data from the latent representation h. The objective is to produce x' that closely approximates the original input x.

Typically, the decoder mirrors the encoder architecture such that if the encoder compresses the data, the decoder expands it back to the original dimensionality.

The learning performance of the autoencoder is evaluated using a loss function, such as mean squared error (MSE), which quantifies the difference between the original input and the reconstructed output. This loss function is minimized during training using optimization algorithms such as Adam or stochastic gradient descent (SGD), enabling the network to learn the most effective low-dimensional representation of the data (Karthick, 2024).

$$MSE(x, x') = \frac{1}{n}\sum(x_i - x_i')^2 \quad (6)$$

The autoencoder in this study was employed as an unsupervised dimensionality reduction method. Since its primary objective was feature extraction rather than prediction, it was trained on the entire dataset without a train–test split to preserve the integrity of the global geochemical structure. The architecture, tailored to the dataset's dimensions (517 samples with 22 geochemical components each), consisted of an input layer with 22 neurons, an encoder with progressively smaller hidden layers, a bottleneck layer containing 7 neurons, and a symmetric decoder, as illustrated in Fig. 3. The autoencoder was trained for up to 300 epochs using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. Mean squared error (MSE) was used as the reconstruction loss, with early stopping (patience = 30) applied to prevent overfitting.

It is important to note that, given the relatively small number of samples, using a deep or highly complex network increases the risk of overfitting—a condition in which the model memorizes specific details of the training set rather than learning generalizable patterns. To prevent this issue, a deliberately simple and compact architecture was chosen to maintain an appropriate balance between model capacity and generalization performance.

III. CLUSTERING AND EVALUATION

After applying dimensionality reduction techniques, the resulting embedded vectors—representing a low-dimensional yet semantically meaningful geochemical feature space—were clustered using the K-Means algorithm. K-Means is one of the most widely used unsupervised clustering methods and operates by minimizing the Euclidean distance between data points and their respective cluster centroids. Although it assumes approximately isotropic cluster geometries and can be sensitive to initialization, it was deliberately chosen as a consistent and widely recognized benchmark to ensure a fair and interpretable comparison of clustering outcomes across different dimensionality reduction (DR) methods. By maintaining a fixed clustering approach, any observed variations in performance can be attributed to the DR techniques themselves rather than to differences in clustering behavior.

Given that the dataset used in this study comprises nine igneous rock units, the number of clusters in K-Means was set to k = 9. This configuration ensured direct comparability between the clustering results and the reference lithological classes, allowing for a rigorous assessment of how effectively each DR method preserves geochemical structure and discriminates rock types. It also enabled the use of external evaluation metrics, providing a quantitative and objective basis for comparing the performance of different dimensionality reduction approaches.

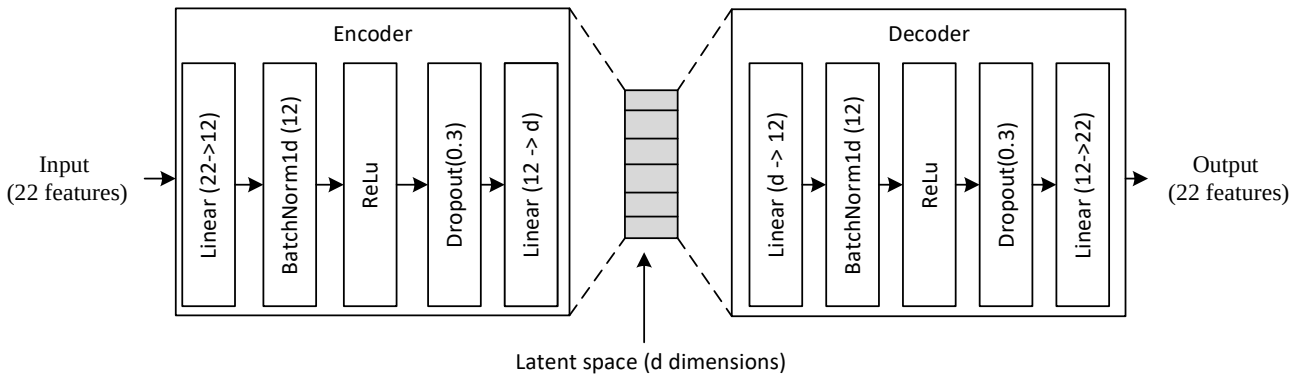

Fig. 3. Autoencoder network architecture with a compact latent layer for dimensionality reduction of multielement geochemical data

A. Selection of the Optimal Dimensionality for Embeddings

Selecting an appropriate number of dimensions for projecting data into a lower-dimensional space is a critical step in the dimensionality reduction workflow, as it directly impacts clustering quality. Choosing too few dimensions may lead to the loss of essential information, whereas selecting too many dimensions may retain noise and reduce clustering performance. For each dimensionality reduction method, a dedicated, method-specific criterion was employed to determine the optimal number of output dimensions. For the PCA method, the cumulative explained variance ratio was used as the primary criterion. This metric indicates the proportion of total information preserved by the selected principal components. Fig. 4 illustrates the cumulative explained variance as a function of the number of principal components. As shown, selecting the first four principal components preserves 83.5% of the total variance in the dataset. This threshold is widely accepted in geochemical studies as sufficient for retaining the essential compositional structure of the data. Therefore, in all PCA-based experiments, the dimensionality of the reduced feature space was set to d=4. This choice provides a suitable balance between information preservation and computational efficiency. Accordingly, PCA reduced the 22-dimensional geochemical dataset to four principal components (PCs). PC1 accounts for 51.9% of the variance and exhibits strong positive loadings with K_2O, Rb, and Na_2O, while showing negative correlations with Sc, V, and TiO_2. This component predominantly represents the alkali-enrichment trend associated with magmatic differentiation. PC2 (14.4% variance) is characterized by positive loadings of incompatible elements (Zr, Nb, Y) and negative loadings of CaO and Al_2O_3, potentially reflecting accessory mineral fractionation or crustal assimilation processes. PC3 (10.1% variance) shows positive associations with compatible elements (Ni, Cr, Sr) and negative loadings with Y and MnO, possibly indicating mafic mineral accumulation. PC4 (7.0% variance) correlates positively with MgO and negatively with Al_2O_3, Ni, and Cr, suggesting variations in ultramafic affinity.

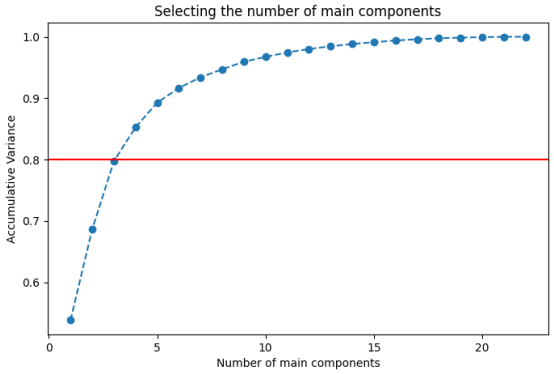

Fig. 4. Cumulative explained variance as a function of the number of PCA components

Fig. 5, which presents the pairwise scatterplots of PCs, demonstrates poor separation of lithological units. The overlap, particularly severe in the PC2-PC4 plane (-0.372), confirms that while PCA captures 83.5% of the variance, its linear nature fails to resolve the nonlinear geochemical relationships necessary to discriminate among the nine rock types.

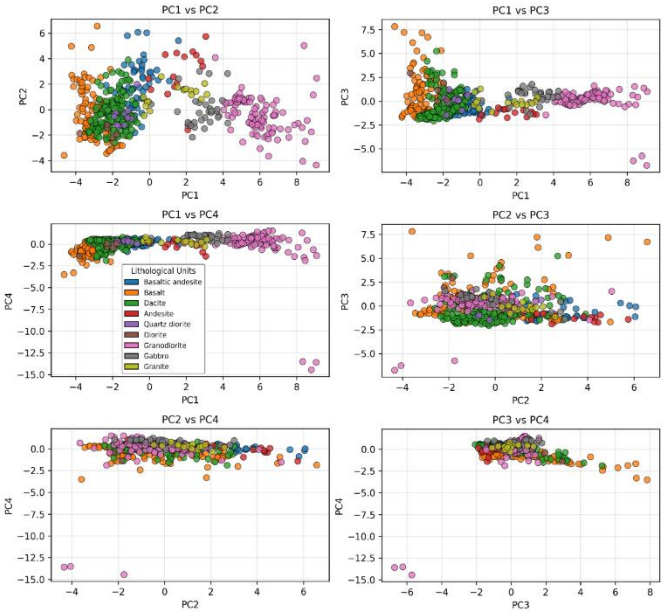

Fig. 5. PCA projection of geochemical data, colored by reference lithological classes

Unlike PCA, UMAP does not have an explicit variance-based metric for selecting the optimal dimensionality. To determine the ideal number of dimensions for the UMAP embedding, a cluster stability evaluation approach was employed. Cluster stability was assessed using the Adjusted Rand Index (ARI), calculated over multiple independent runs of the algorithm across different embedding dimensions. This method allowed for quantifying the consistency and reproducibility of clustering results across stochastic UMAP embeddings. The variation in cluster stability as a function of dimensionality is illustrated in Fig. 6.

As shown in Fig. 6, the ARI steadily increases with dimensionality, indicating improved preservation of the geochemical structure within the reduced space. However, after the fourth dimension, the rate of improvement decreases substantially. This pattern suggests that adding dimensions beyond four does not result in a meaningful enhancement of clustering stability. Accordingly, based on the elbow criterion, the optimal embedding dimensionality for UMAP was set to four.

Unlike the principal components in PCA, the individual dimensions in the UMAP embedding do not correspond to specific elemental concentrations or simple geochemical processes. However, the pairwise scatterplots of these dimensions (Fig. 7) reveal substantially improved visual separation of lithological

units compared to PCA (Fig. 5). The samples appear more distinct, with markedly reduced overlap, particularly between compositionally intermediate rock types. This enhanced visual clustering directly reflects UMAP's ability to capture complex, nonlinear relationships inherent in the geochemical data that linear methods like PCA cannot resolve.

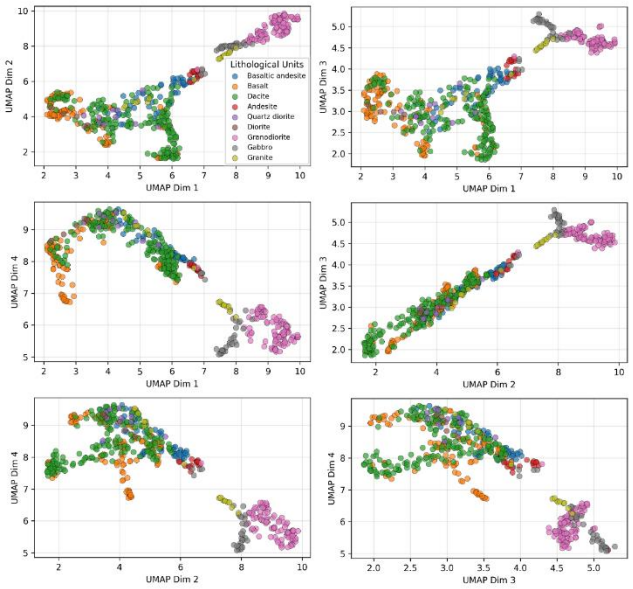

Fig. 7. UMAP Projection of Geochemical Data, Colored by Reference lithological classes

To determine the optimal dimensionality for the autoencoder, a hybrid approach combining two complementary criteria—reconstruction error and the silhouette coefficient—was employed. The reconstruction error measures the model's ability to accurately reproduce the original inputs, while the silhouette coefficient assesses the separability and compactness of the resulting clusters. The variations of these two metrics across different latent dimensions are presented in Fig. 8. As shown, the highest silhouette coefficient occurs at two dimensions (d = 2); however, this configuration is associated with a substantial reconstruction error, indicating a loss of critical geochemical information. Therefore, the optimal dimensionality must strike a balance, ensuring the reconstruction error remains sufficiently low while preserving or enhancing the clustering structure.

It is important to note that, due to the stochastic nature of neural network training, autoencoder results may vary across different runs. To address this variability, the network was trained multiple times, and the outcomes were analyzed statistically. Based on these repeated experiments, a latent dimensionality of seven (d = 7) consistently provided an optimal balance between achieving a relatively high silhouette coefficient and maintaining a low reconstruction error. Therefore, seven dimensions were selected as the optimal reduced dimensionality for the autoencoder, and all reported results in this study are based on this configuration.

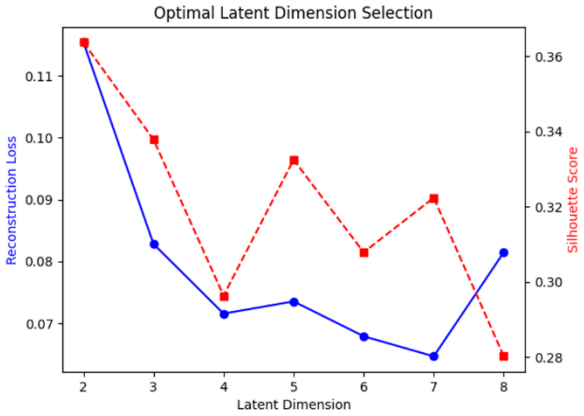

Fig. 8. Variations of reconstruction error and the silhouette coefficient across different latent dimensions in the autoencoder

B. Cluster Evaluation Results

To obtain a comprehensive and multi-perspective assessment of clustering performance resulting from the combination of different normalization and dimensionality reduction techniques, two categories of evaluation metrics were employed. These categories provide complementary insights into clustering quality. The first category comprises internal validation indices, which operate independently of the reference class labels and assess cluster quality solely based on the geometric properties of the data, specifically within-cluster compactness and between-cluster separation. These metrics are particularly useful when ground-truth labels are unavailable or when the goal is to evaluate the intrinsic structure of the data. The second category includes external validation indices, which quantify the degree of agreement between the clustering results and the reference class labels of the samples (in this study, the lithological units determined through petrographic analyses). These metrics directly measure how well the cluster assignments correspond to the underlying geological classification. Using both internal and external metrics provides a more holistic evaluation of clustering performance. A method may exhibit strong internal cohesion yet show weak correspondence with geological reality, or vice versa. Therefore, incorporating both perspectives ensures a balanced and reliable interpretation of the results. Table 1 presents a complete list of the evaluation indices used in this study, along with brief descriptions and their corresponding categories.

The results of applying the K-Means algorithm to dimension-reduced datasets produced by different combinations of normalization and dimensionality reduction methods are presented in Table 2. This table reports clustering performance based on seven internal evaluation metrics, enabling a quantitative comparison of different normalization and dimensionality reduction configurations. Given the multiplicity of evaluation metrics and the possibility of inconsistencies among them, selecting the best method is only feasible by considering the consensus and overall performance

patterns across all metrics. To facilitate this comprehensive comparison, Fig. 9 presents a combined plot of the internal metrics for the different methods. The vertical axis is scaled logarithmically to allow simultaneous visualization of metrics with different ranges. Metrics for which lower values indicate better performance (Davies-Bouldin, Compactness, and Within-Cluster Variance) are shown with dotted lines, while metrics for which higher values are preferred (Silhouette, Calinski-Harabasz, Separation, and Between-Cluster Variance) are shown with solid lines.

Fig. 9 clearly demonstrates that UMAP-based methods consistently outperform PCA and autoencoder approaches in both two- and four-dimensional settings. UMAP, especially when combined with MinMax or Z-score normalization, achieves the highest scores in positive metrics (Silhouette, Calinski–Harabasz, Separation, and Between-Cluster Variance) and the lowest scores in negative metrics (Davies–Bouldin, Compactness, and Within-Cluster Variance). In contrast, PCA and autoencoder exhibit lower performance, with metrics close to or below those of the original high-dimensional data. This suggests that linear or simple neural network-based dimensionality reduction techniques may fail to enhance clustering quality and, in some cases, may obscure important data structures. These findings underscore the necessity of advanced nonlinear dimensionality reduction methods, such as UMAP, for geochemical datasets characterized by complex nonlinear relationships. Since the lithology of the samples is definitively known from petrographic and laboratory analyses, clustering performance was also evaluated using external metrics. Table 3 presents the results based on the Adjusted Rand Index, Normalized Mutual Information, and Purity, with the number of dimensions used for each method reported in the final column.

Table 1. Evaluation metrics for clustering quality

Metric	Description	Metric Type
Silhouette Score	Measures the difference between the intra-cluster distance and the nearest other cluster; values close to 1 indicate good clustering.	Internal
Davies–Bouldin Index	Computes the ratio of intra-cluster compactness to inter-cluster separation; lower values indicate more distinct clusters.	Internal
Calinski–Harabasz Index	Measures the ratio of between-cluster variance to within-cluster variance; higher values indicate better separation.	Internal
Within-Cluster Compactness	Calculates the average distance of points from the cluster center; lower values indicate denser clusters.	Internal
Between-Cluster Separation	Evaluates the distance between cluster centers; higher values indicate better separation.	Internal
Within-Cluster Variance	Measures the dispersion of points within each cluster; lower values indicate higher homogeneity.	Internal
Between-Cluster Variance	Measures the dispersion of cluster centers; higher values indicate greater diversity.	Internal
Adjusted Rand Index	Measures agreement between two partitions with random chance adjustment; ranges from -1 (complete disagreement) to 1 (perfect agreement).	External
Normalized Mutual Information	Measures the normalized information similarity between two partitions; ranges from 0 (independent) to 1 (identical).	External
Purity	Calculates the percentage of points that match the dominant cluster label; higher values indicate greater cluster homogeneity.	External

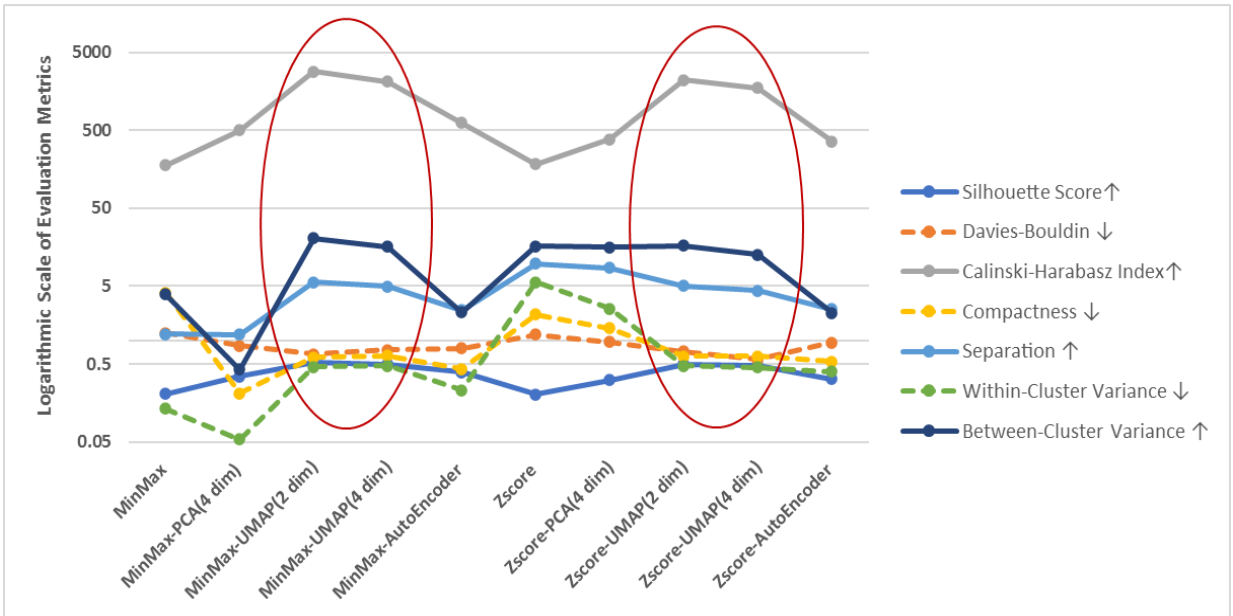

Fig. 9. Comparative evaluation of clustering quality based on internal metrics

Table 2. Evaluation of clustering quality based on internal metrics

Run K-means on:	Silhouette Score↑	Calinski-harabasz index↑	Separation ↑	Between-cluster variance ↑	Within-cluster variance ↓	Compactness ↓	Davies-bouldin ↓
MinMax of raw data (22d)	0.208	179.117	1.221	3.913	0.133	4.119	1.244
MinMax+PCA(4d)	0.346	503.487	1.189	0.428	0.054	0.210	0.855
MinMax+UMAP (2d)	0.521	2,846.01	5.665	20.548	0.458	0.606	0.671
MinMax+UMAP (4d)	0.494	2,125.26	4.923	15.930	0.475	0.634	0.758
MinMax+AE(7d)	0.395	625.613	2.440	2.287	0.232	0.431	0.798
Zscore of raw data (22d)	0.205	185.478	9.713	16.406	5.594	2.149	1.199
Zscore+PCA (4d)	0.310	387.321	8.539	15.780	2.582	1.431	0.966
Zscore+UMAP (2d)	0.498	2,220.69	5.029	16.535	0.472	0.628	0.722
Zscore+UMAP (4d)	0.479	1754.56	4.348	12.604	0.455	0.627	0.575
Zscore+AE (7d)	0.321	360.499	2.516	2.270	0.399	0.539	0.945

Notes: ↑ = higher values indicate better performance; ↓ = lower values indicate better performance

Based on the results presented in Table 3, the clustering method that combines Z-score normalization with UMAP up to four dimensions achieved the highest scores across all three external evaluation metrics, outperforming all other approaches. This method attained an Adjusted Rand Index of 0.332, a Normalized Mutual Information of 0.522, and a Purity of 0.684, indicating the strongest agreement with the reference lithological classification. These findings demonstrate that UMAP not only excels at preserving the geometric structure of clusters (internal metrics) but also provides superior alignment with geological reality (external metrics).

Table 3. Comparative evaluation of clustering quality based on external metrics.

Run K-means on:	Adjusted rand index	Normalized mutual information	Purity	Number of dimensions
MinMax (22d)	0.308	0.515	0.651	22
MinMax+PCA(4d)	0.296	0.503	0.647	4
MinMax+UMAP (2d)	0.273	0.475	0.651	2
MinMax+UMAP (4d)	0.311	0.500	0.676	4
MinMax+AE(7d)	0.221	0.447	0.585	7
Zscore (22d)	0.277	0.461	0.622	22
Zscore+PCA (4d)	0.255	0.450	0.599	4
Zscore+UMAP (2d)	0.303	0.482	0.667	2
Zscore+UMAP (4d)	0.332	0.522	0.684	4
Zscore+AE (7d)	0.320	0.483	0.663	7

A notable observation in Table 3 is that UMAP combined with MinMax normalization performs worse than when combined with Z-score normalization, highlighting the importance of selecting an appropriate normalization method alongside the dimensionality reduction technique.

C. Geochemical Interpretation of Clusters

To validate the geological significance of the algorithmically derived clusters, their average compositions were examined using standard geochemical diagnostic diagrams.

The Harker diagrams (Fig. 10) display systematic geochemical trends across the clusters identified by the Zscore-UMAP-Kmeans method, highlighting meaningful magmatic differentiation processes.

The negative correlations between SiO_2 and both MgO and CaO observed in most clusters indicate progressive fractional crystallization and the removal of mafic minerals from the melt. Clusters characterized by low SiO_2 and high MgO–CaO contents are primarily associated with basaltic and basaltic–andesitic compositions, representing relatively primitive magmas. Dacite-dominated clusters occupy intermediate positions between mafic and felsic end-members, suggesting magma evolution through fractional crystallization and possible magma mixing. In contrast, the granodiorite-dominated cluster is distinguished by high SiO_2 and very low MgO–CaO values, reflecting an evolved felsic magma. The concurrent increase in Na_2O and K_2O with SiO_2 suggests progressive alkali enrichment during late-stage magmatic differentiation. Overall, the Zscore-UMAP-Kmeans workflow produced stable and geologically coherent clusters that align closely with magmatic differentiation trends observed in Harker diagrams. This consistency demonstrates the method's capability to preserve key geochemical relationships while maintaining interpretability.

Fig. 11 presents normalized trace-element spider diagrams. The left panel shows the geochemical patterns of clusters obtained from the Z-score UMAP-K-means workflow, while the right panel displays the corresponding patterns of the reference lithological classes. The overall similarity between the two panels—particularly in incompatible element enrichment, HFSE depletion, and relative LILE behavior—indicates that the clustering results successfully capture meaningful geochemical signatures consistent with magmatic differentiation from mafic to felsic compositions. The similarity between the spider diagrams of algorithmic clusters and the reference lithological classes, along with comparable Ce/La trends, further supports that the clustering results reflect petrogenetically meaningful geochemical variations.

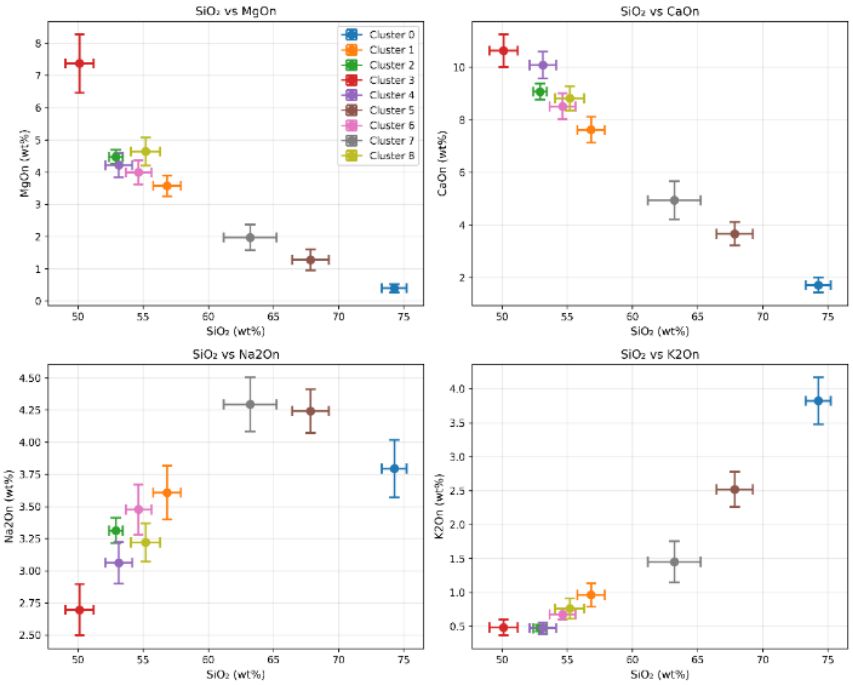

Fig. 10. Harker diagrams showing SiO_2 versus major oxides (MgO, CaO, Na_2O, K_2O) for clusters derived from Zscore-UMAP-Kmeans. Error bars indicate within-cluster variability

Geochemical validation using rare earth element ratios (Ce/La) in Fig. 12 demonstrated a strong correlation between algorithmically derived clusters and reference lithological units for most samples. The basaltic, andesitic, and granodioritic clusters exhibited excellent agreement, whereas moderate discrepancies were noted for some intermediate rock types (dacite and diorite), likely reflecting the continuous nature of magmatic differentiation or minor alteration effects.

The effective discrimination of end-member compositions (mafic versus felsic) by our Z-score-UMAP-K-means pipeline, as supported by coherent Ce/La geochemical systematics, demonstrates its capability for first-order lithological classification. The higher misclassification rate for intermediate compositions aligns with known challenges in geochemical classification, where gradual transitions blur discrete lithological boundaries. These deviations may, in fact, highlight the algorithm's sensitivity to subtle geochemical variations that conventional petrographic classification might overlook.

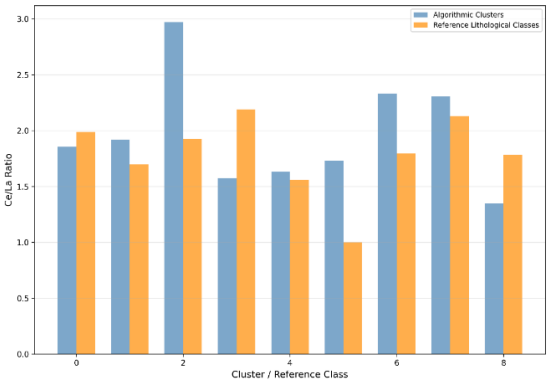

Fig. 12. REE Fractionation, Ce/La Ratio Comparison

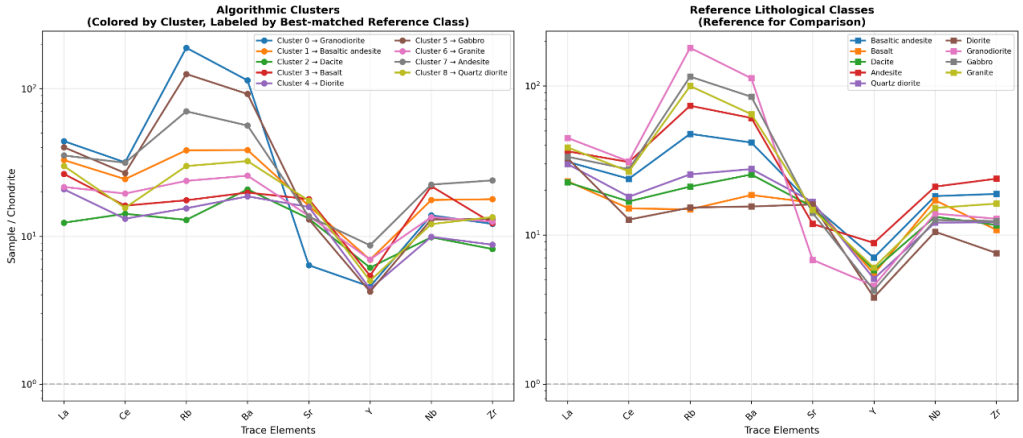

Fig. 11. Normalized trace-element spider diagrams for algorithmic clusters (left) and reference lithological classes (right)

D. Sensitivity and Stability Analysis of Dimensionality Reduction Methods

Given the stochastic nature of nonlinear dimensionality reduction methods such as UMAP and autoencoders, a sensitivity analysis was conducted to evaluate the robustness and consistency of their results. Each method, combined with two normalization approaches (Min-Max and Z-score), was executed over 30 independent runs with different random initializations. Clustering outcomes were assessed using the Adjusted Rand Index (ARI), Purity, and Silhouette Score. The corresponding mean values, standard deviations (SD), variation ranges, and coefficients of variation (CV) are summarized in Table 4.

The results indicate that UMAP-based approaches generally provide more stable outcomes, as evidenced by consistently lower coefficients of variation (CV) across all evaluation metrics compared to the autoencoder. Specifically, UMAP configurations produced more reliable clustering results, with the MinMax combination exhibiting the lowest variability. This demonstrates that UMAP is less sensitive to random initialization under the conditions examined, making it a more dependable choice for practical applications. In contrast, the autoencoder with a seven-dimensional latent representation showed substantially higher variability, particularly in terms of ARI and Silhouette scores. Although high performance was achieved in certain runs, the pronounced sensitivity to initial conditions highlights the need for multiple executions to obtain representative results. This variability emphasizes the importance of repeated experiments when using autoencoders for feature extraction prior to clustering.

Table 4. Sensitivity and stability analysis results

		Mean	SD	Min	Max	CV
MinMax-UMAP(2)	ARI	0.294	0.019	0.2485	0.329	6.45
	Purity	0.6596	0.0162	0.6182	0.688	2.46
	Sil.	0.5261	0.0092	0.4926	0.5557	1.82
MinMax-UMAP(4)	ARI	0.292	0.0105	0.2736	0.317	3.58
	Purity	0.6655	0.0069	0.6531	0.6725	1.04
	Sil.	0.5033	0.0103	0.4741	0.52	2.05
MinMax-AE(7)	ARI	0.2878	0.0370	0.2154	0.3539	12.85
	Purity	0.6553	0.0299	0.5788	0.7035	4.56
	Sil.	0.3668	0.031	0.3144	0.4329	8.46
ZScore-UMAP(2)	ARI	0.2859	0.0206	0.2626	0.3447	7.22
	Purity	0.6589	0.0155	0.6376	0.6977	2.35
	Sil.	0.4988	0.0132	0.4833	0.5512	2.65
ZScore-UMAP(4)	ARI	0.3162	0.022	0.2688	0.3516	6.97
	Purity	0.6828	0.015	0.6512	0.7112	2.2
	Sil.	0.4893	0.009	0.4583	0.5064	1.84
Zscore-AE(7)	ARI	0.2878	0.0335	0.206	0.3568	11.63
	Purity	0.6415	0.0253	0.593	0.6822	3.94
	Sil.	0.3411	0.0234	0.2915	0.4007	6.86

SD= Standard Deviation, CV= Coefficient of Variation

Additionally, the normalization strategy played a significant role in both performance and stability. Z-score normalization combined with UMAP produced the highest average ARI and Purity values. The results presented in Table 4 enable a direct comparison of average clustering performance and methodological stability across all evaluated workflows.

IV. CONCLUSION

This study presents a systematic comparative framework for evaluating dimensionality reduction (DR) techniques in lithological discrimination based on multi-element geochemical data. The results demonstrate that both data normalization strategies and the choice of DR method play critical roles in preserving intrinsic geochemical structures and enhancing clustering performance. Among all evaluated configurations, the combination of Z-score normalization with the nonlinear UMAP method in a four-dimensional embedding consistently achieves superior performance across multiple validation metrics. This workflow effectively captures complex nonlinear geochemical relationships arising from petrogenetic processes and projects them into a discriminative low-dimensional space that facilitates robust cluster separation, outperforming other methods. In contrast, the linear PCA approach provides weaker lithological discrimination, while the autoencoder exhibits lower stability and accuracy for this dataset despite its representational flexibility. Geochemical validation using Ce/La ratios and Harker diagrams confirms that the clusters generated by the Z-score–UMAP–K-means workflow correspond to petrogenetically coherent groups. The method successfully discriminates mafic and felsic end-members while capturing compositional variability in intermediate rock types. Sensitivity analysis over 30 independent runs demonstrates the reproducibility and robustness of the UMAP-based approach, establishing it as a reliable and interpretable pipeline for revealing latent structures in complex geochemical datasets. These practical findings have direct implications for early-stage mineral exploration, lithological mapping, and geochemical anomaly detection.

Although the present analysis focused on a specific igneous rock dataset, the proposed framework can be extended to metamorphic, sedimentary, and mineralized environments. Future research could explore alternative clustering algorithms better suited for non-spherical geochemical distributions, integrate dimensionally reduced features into supervised classification models for predictive lithological mapping, and apply this integrated workflow to multi-scale exploration datasets to enhance exploration targeting efficiency.

REFERENCES

Abbasi, Z., Yang, X., Mohammadoost, H., TaleFazel, E., Hafeez, M., & Shah, A. (2025). Comprehensive geochemical and isotopic constraints on multi-stage magmatism and subsequent Cu–Mo (Au) mineralization in porphyry clusters of Kerman metallogenic belt, Iran: A perspective review. Solid Earth Sciences, 10(2), 100229.

Khashaba, S. M. A., El-Shibiny, N. H., Hassan, S. M., Drueppel, K., & Azer, M. K. (2024). Remote sensing and geochemistry of A-type granites, North Arabian-Nubian Shield: Insights into the origin and evolution of the granitic suites and processes responsible for rare metals enrichment. Ore Geology Reviews, 175, 106391.

Salem, H. S. A., El Fallah, O. A., & El Kammar, M. M. (2024). Hydrochemical study of groundwater in Tazerbo, Libya, using statistical analysis and geochemical modeling. Journal of African Earth Sciences, 218, 105362.

Acosta-Góngora, P., Potter, E. G., Lawley, C. J., Petts, D., & Sparkes, G. (2022). Uraninite chemistry of the Central Mineral Belt, Labrador, Canada: Application of grain-scale unsupervised machine-learning. Journal of Geochemical Exploration, 233, 106910.

Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020, June). Considerably improving clustering algorithms using UMAP dimensionality reduction technique: a comparative study. In International conference on image and signal processing (pp. 317-325). Cham: Springer International Publishing.

Aryafar, A., Shiva, M., & Zaremotlagh, S. (2011). Comparison of rock unit separation and fuzzy logic methods in neutralizing the syngenetic effects in geochemical data, a case study in eastern part of Iran. Journal of Geology Mining Research, 3(1), 7-12.

Badawy, W. M., Dmitriev, A. Y., & Koval, V. Y. (2023). Geochemical ceramic composition dataset using neutron activation and statistical analyses. Data in Brief, 48, 109051.

Bévan, M., Boulvais, P., Hallot, E., Branquet, Y., Gautier, P., Rodriguez Martinez, L., & Audran, B. (2025). Magmatic Differentiation, Magmatic–Hydrothermal Evolution, and Hydrothermal Alteration in Felsic Dikes: Insights from Intrusions in Ultramafic Rocks, Ronda Massif (Spain). The Canadian Journal of Mineralogy and Petrology, 63(4), 325-345.

Canbaz, O., & Karaman, M. (2024). Geochemical characteristics and mapping of Reşadiye (Tokat-Türkiye) bentonite deposits using machine learning and sub-pixel mixture algorithms. Geochemistry, 84(4), 126123.

Cao, G., Zhang, Y., Zhao, H., Cheng, J., Hao, J., Lei, J., ... & Wang, X. (2023). Trace element variations of pyrite in orogenic gold deposits: Constraints from big data and machine learning. Ore Geology Reviews, 157, 105447.

Chen, Y., Chen, B., & Shayilan, A. (2024). Combining categorical boosting and Shapley additive explanations for building an interpretable ensemble classifier for identifying mineralization-related geochemical anomalies. Ore Geology Reviews, 173, 106263.

Doucet, L. S., Tetley, M. G., Li, Z. X., Liu, Y., & Gamaleldien, H. (2022). Geochemical fingerprinting of continental and oceanic basalts: A machine learning approach. Earth-Science Reviews, 233, 104192.

du Bray, E. A., John, D. A., Sherrod, D. R., Evarts, R. C., Conrey, R. M., & Lexa, J. (2006). Geochemical database for volcanic rocks of the western Cascades, Washington, Oregon, and California (No. 155).

Gare, S., Chel, S., Abhinav, T. K., Dhyani, V., Jana, S., & Giri, L. (2022). Mapping of structural arrangement of cells and collective calcium transients: an integrated framework combining live cell imaging using confocal microscopy and UMAP-assisted HDBSCAN-based approach. Integrative Biology, 14(8-12), 184-203.

Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2021). Uniform manifold approximation and projection (UMAP) and its variants: tutorial and survey. arXiv preprint arXiv:2109.02508.

Grunsky, E. (2010). The interpretation of geochemical survey data. Geochemistry-exploration Environment Analysis - GEOCHEM-EXPLOR ENVIRON ANAL 10: 27-74.

Grunsky, E. C., & Caritat, P. D. (2020). State-of-the-art analysis of geochemical data for mineral exploration. Geochemistry: Exploration, Environment, Analysis, 20(2), 217-232.

Al Haj, R., Merheb, M., Halwani, J., & Ouddane, B. (2025). Baseline hydro-geochemical characteristics of groundwater in Abu Ali watershed (Northern Lebanon). Journal of Hydrology: Regional Studies, 57, 102135.

Hansen, T. F., & Aarset, A. (2025). Unsupervised Machine Learning for Data-Driven Rock Mass Classification: Addressing Limitations in Existing Systems Using Drilling Data: TF Hansen, A. Aarset. Rock Mechanics and Rock Engineering, 58(10), 11261-11291.

Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. Complex & Intelligent Systems, 8(3), 2663-2693.

Karthick, K. (2024). Comprehensive overview of optimization techniques in machine learning training. Control Systems and Optimization Letters, 2(1), 23-27.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Sadeghi, M., Casey, P., Carranza, E. J. M., & Lynch, E. P. (2024). Principal components analysis and K-means clustering of till geochemical data: Mapping and targeting of prospective areas for lithium exploration in Västernorrland Region, Sweden. Ore Geology Reviews, 167, 106002.

Stracke, A., Willig, M., Genske, F., Béguelin, P., & Todd, E. (2022). Chemical geodynamics insights from a machine learning approach. Geochemistry, Geophysics, Geosystems, 23(10), e2022GC010606.

Su, Q., Yu, H., Xu, X., Chen, B., Yang, L., Fu, T., ... & 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.

Wang, X., & Chen, Y. (2025). Unsupervised detection of multivariate geochemical anomalies using a high-performance deep autoencoder Gaussian mixture model. Journal of Geochemical Exploration, 271, 107671.

Wang, Y., Yao, H., & Zhao, S. (2016). Auto-encoder based dimensionality reduction. Neurocomputing, 184, 232-242.

Xu, H., Croot, P., & Zhang, C. (2021). Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis. Environment International, 151, 106456.

Xu, Y., & Zuo, R. (2024). Geochemical survey data cube: A useful tool for lithological classification and geochemical anomaly identification. Geochemistry, 84(2), 125959.

Xue, J., Lee, C., Wakeham, S. G., & Armstrong, R. A. (2011). Using principal components analysis (PCA) with cluster analysis to study the organic geochemistry of sinking particles in the ocean. Organic Geochemistry, 42(4), 356-367.

Zaremotlagh, S., Fatemi, S. A., & Azinfar, M. J. (2025). Modeling and Analysis of Joint Systems Using a Combined and Multi-Stage Clustering Approach: A Case Study of the Lucho Granite Mass, Zahedan. Journal of Analytical and Numerical Methods in Mining Engineering 15(44): 49-64.

Zaremotlagh, S., & Hezarkhani, A. (2016). A geochemical modeling to predict the different concentrations of REE and their hidden patterns using several supervised learning methods: Choghart iron deposit, bafq, Iran. Journal of Geochemical Exploration, 165, 35-48.

Zaremotlagh, S., & Hezarkhani, A. (2017). The use of decision tree induction and artificial neural networks for recognizing the geochemical distribution patterns of LREE in the Choghart deposit, Central Iran. Journal of African Earth Sciences, 128, 37-46.

Zhang, Q., Liu, Y., & Fang, H. (2024). Manifold learning-based UMAP method for geochemical anomaly identification. Geochemistry, 84(4), 126157.

Zhao, J., & Chen, S. (2021). Identification of the ore-forming anomaly component by MSVD combined with PCA from element concentrations in fracture zones of the Laochang ore field, Gejiu, SW China. Journal of Earth Science, 32(2), 427-438.